UNIVERSIDADE ESTADUAL DE CAMPINAS
FACULDADE DE ENGENHARIA ELÉTRICA E DE COMPUTAÇÃO

Thiago Ribas Bella

Modelos de Predição de Óbitos por Doenças Cardiovasculares Utilizando Variáveis Ambientais e Estudo Exploratório com Cenários de Mudanças Climáticas

Campinas
2022

Thiago Ribas Bella

Modelos de Predição de Óbitos por Doenças Cardiovasculares Utilizando Variáveis Ambientais e Estudo Exploratório com Cenários de Mudanças Climáticas

Dissertação apresentada à Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Mestre em Engenharia Elétrica na Área de Engenharia de Computação.

Orientadora: Profa. Dra. Paula Dornhofer Paro Costa

Este exemplar corresponde à versão final da dissertação defendida pelo aluno Thiago Ribas Bella, orientado pela Profa. Dra. Paula Dornhofer Paro Costa.

Campinas
2022

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Área de Engenharia e Arquitetura
Rose Meire da Silva - CRB 8/5974

Bella, Thiago Ribas, 1992-
B414m Modelos de predição de óbitos por doenças cardiovasculares utilizando variáveis ambientais e estudo exploratório com cenários de mudanças climáticas / Thiago Ribas Bella. – Campinas, SP : [s.n.], 2022.

Orientador: Paula Dornhofer Paro Costa.
Dissertação (mestrado) – Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.

1. Doenças cardiovasculares. 2. Predição de séries temporais. 3. Mudanças climáticas. I. Costa, Paula Dornhofer Paro, 1978-. II. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Prediction models for cardiovascular disease deaths using environmental variables and an exploratory study based on climate change scenarios
Palavras-chave em inglês:
Cardiovascular deaths
Time series prediction
Climate change
Área de concentração: Engenharia de Computação
Titulação: Mestre em Engenharia Elétrica
Banca examinadora:
Paula Dornhofer Paro Costa [Orientador]
Levy Boccato
André Kazuo Takahata
Data de defesa: 15-08-2022
Programa de Pós-Graduação: Engenharia Elétrica

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0003-1642-7001
- Currículo Lattes do autor: http://lattes.cnpq.br/0613341396317708

Comissão Julgadora – Dissertação de Mestrado

Candidato: Thiago Ribas Bella **RA:** 157414

Data da defesa: 15 de Agosto de 2022

Título da Tese: "Modelos de Predição de Óbitos por Doenças Cardiovasculares Utilizando Variáveis Ambientais e Estudo Exploratório com Cenários de Mudanças Climáticas."

Profa. Dra. Paula Dornhofer Paro Costa (Presidente, FEEC/UNICAMP)
Prof. Dr. Levy Boccato (FEEC/UNICAMP)
Dr. André Kazuo Takahata (ISC/UFABC)

A ata de defesa, com as respectivas assinaturas dos membros da Comissão Julgadora, encontra-se no SIGA (Sistema de Fluxo de Dissertação/Tese) e na Secretaria de Pós-Graduação da Faculdade de Engenharia Elétrica e de Computação.

Agradecimentos

O presente trabalho foi realizado com apoio do CNPq, Conselho Nacional de Desenvolvimento Científico e Tecnológico – Brasil.

Gostaria de agradecer minha orientadora Paula Dornhofer Paro Costa e as professoras e pesquisadoras Eliana Cotta de Faria e Ana Maria Heuminski de Avila pelo apoio e orientação na jornada de ser um pesquisador. Aos colegas do grupo de pesquisa Clima&Saúde pelo suporte técnico e discussões que ajudaram a nortear este trabalho.

Também gostaria de agradecer minha mãe Silvana Maria Correa Pinto por todo apoio e suporte durante meus longos anos de estudo. Por fim, gostaria de agradecer a todos os meus amigos e familiares pelo apoio e pela confiança em mim.

Abstract

Environmental variables play an important role in human health, and climate change can alter these relationships. Cardiovascular diseases (CVDs) are the leading cause of death in the world. In addition to lifestyle, CVDs are also influenced by environmental variables and are subject to climate change. To elaborate strategies for adaptation, mitigation, and prevention of the effects of climate change, tools capable of simulating different future scenarios are necessary, especially in a regionalized way. Thus, the general objective of this work is to generate and compare models of the associations between environmental variables and the number of deaths from CVDs in the city of Campinas, São Paulo. Furthermore, we propose an exploratory study on the impact of different climate change scenarios on the number of deaths from CVDs by 2050. We integrated and curated the databases of deaths from all causes from the Health Department of Campinas. The environmental variables were from the meteorological stations of the Environmental Company of the State of São Paulo, Agronomic Institute of Campinas, Center for Meteorological and Climatic Researches Applied to Agriculture, and Viracopos International Airport. The environmental database includes daily values for temperature, carbon monoxide, particulate matter, relative humidity, and atmospheric pressure. We developed predictive models from the integrated database using two approaches: linear regression with SARIMA errors (LR-SARIMAX) and LSTM recurrent neural network, for daily, weekly and monthly deaths. The grid search technique was adopted to achieve the smallest prediction errors. This technique systematically varies the intrinsic parameters of the models with different combinations of the predictor variables and the number of lags of these variables. Four hundred forty-one models were evaluated using the RMSE and MAPE metrics. The models were compared concerning the data periodicity, model type, variables combination, and the number of lags of environmental variables. The models using monthly data presented prediction errors up to 5 times smaller than the models using data in the other periodicities. Even though the different approaches did not show significant differences in prediction errors, the LR-SARIMAX modeling presented more advantages than the LSTM due to its robustness in relation to data variation and its ability to interpret. The temperature was essential in predicting deaths from CVDs, being present in 12 of the 15 best models. The number of lags of the predictor variables was also relevant in the predictions, presenting values corroborated by the literature. The exploratory study pointed out that with the increase in minimum temperatures due to climate change, a decrease in the number of deaths from CVDs is expected by 2050 for Campinas.

Keywords: Cardiovascular Deaths; Time Series Prediction; Climate Change.

Resumo

As variáveis ambientais desempenham papel importante na saúde humana e as mudanças climáticas podem alterar essas relações. Atualmente, as doenças cardiovasculares (DCVs) são a principal causa de morte no mundo. Além do estilo de vida, as DCVs também sofrem influência de variáveis ambientais e estão sujeitas às mudanças climáticas. Para contribuir com a elaboração de estratégias de adaptação, mitigação e prevenção dos efeitos das mudanças climáticas, se faz necessário ferramentas capazes de simular diferentes cenários futuros, sobretudo de forma regionalizada. Assim, o objetivo geral deste trabalho é gerar e comparar modelos das associações entre variáveis ambientais e o número de mortes por DCVs para a cidade de Campinas, São Paulo. Ainda, propomos um estudo exploratório sobre o impacto de diferentes cenários de mudanças climáticas no número de óbitos por DCVs até 2050. Para isso, integramos e curamos bases de dados de óbitos e de variáveis ambientais. O banco de dados de óbitos por todas as causas foi fornecido pela Secretaria de Saúde de Campinas. O banco de dados das variáveis ambientais foram obtidos das estações meteorológicas da Companhia Ambiental do Estado de São Paulo, do Instituto Agronômico de Campinas, do Centro de Pesquisas Meteorológicas e Climáticas Aplicadas à Agricultura, e do Aeroporto Internacional de Viracopos. A base de dados ambientais inclui valores diários de temperatura, monóxido de carbono, material particulado, umidade relativa e pressão atmosférica. A partir dos dados integrados, desenvolvemos modelos preditivos utilizando duas abordagens distintas: regressão linear com erros SARIMA (LR-SARIMAX) e rede neural recorrente do tipo LSTM, para óbitos diários, semanais e mensais. Para alcançar os menores erro de predição, adotou-se a técnica de *grid search*, que variou sistematicamente os parâmetros intrínsecos dos modelos em conjunto com diferentes combinações das variáveis preditoras e da quantidade de lags dessas variáveis. Foram avaliados 441 modelos através das métricas RMSE e MAPE. Os modelos foram comparados em relação à periodicidade dos dados, ao tipo de modelo, às variáveis utilizadas e a quantidade de lags das variáveis. Os modelos utilizando dados mensais apresentaram erros de predição até 5 vezes menores que os modelos utilizando dados nas outras periodicidades. Apesar das diferentes abordagens não terem apresentado diferenças significativas nos erros de predição, a modelagem a partir do LR-SARIMAX apresentou mais vantagens que o LSTM, devido a sua robustez em relação a variação dos dados e sua capacidade de interpretabilidade. A temperatura desempenhou um papel importante nas predições de óbitos por DCVs, estando presente em 12 dos 15 melhores modelos. A quantidade de lags das variáveis preditoras também foi relevante nas predições, apresentando valores corroborados pela literatura. O estudo exploratório apontou que com o aumento das temperaturas mínimas, devido às mudanças climáticas, espera-se uma diminuição no número de óbitos por DCVs até 2050 para Campinas.

Palavras-chave: Doenças Cardiovasculares; Predição de Séries Temporais; Mudanças climáticas.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

Na sua constituição, a Organização Mundial da Saúde define saúde não como uma mera ausência de doença ou enfermidade, mas como um estado de completo bem-estar físico, mental e social (World Health Organization, 2020). Essa definição abrangente demanda o entendimento de que a saúde das populações depende do adequado funcionamento e da estabilidade dos sistemas físicos e ecológicos da biosfera, frequentemente referenciados como sistemas de suporte à vida (World Health Organization, 2003).

De fato, o papel das variáveis ambientais e das condições climáticas na saúde humana é intuitivo e reconhecido há séculos. As populações humanas sempre foram flageladas, por exemplo, por eventos climáticos extremos como enchentes, ondas de frio, secas e pragas. No entanto, os padrões destes eventos estão mudando, e a quebra da estabilidade dos sistemas naturais da biosfera está impondo novos tipos de ameaças à saúde humana (MCMICHAEL, 1993). As mudanças ambientais globais têm o potencial de afetar a saúde das populações de maneira sem precedentes, com consequências não apenas no presente, mas também no futuro mais distante (LENTON *et al.*, 2019).

Nesse cenário, a definição de estratégias de adaptação, mitigação e prevenção de danos à saúde provocados pelas mudanças ambientais é urgente e desafiadora. Não se trata apenas de modelar as mudanças físicas ambientais e climáticas que estão ocorrendo e prever os seus impactos na saúde, mas também de se estudar, por exemplo, correlações entre clima e saúde ainda desconhecidas enquanto as mudanças acontecem. Ainda, se faz necessário entender uma intrincada teia de relações ambientais e socioeconômicas, capazes de provocar efeitos em cascata ainda imprevisíveis.

Para se estudar e prevenir essa categoria de risco à saúde, ferramentas tradicionais, como a observação empírica e avaliação de risco à saúde baseadas em dados presentes não são suficientes, e devem ser associadas a um novo conjunto de ferramentas, capazes de simular diferentes cenários futuros, e de derivarem modelos preditivos a partir de um grande número de variáveis (MCMICHAEL, 1997). É nesse contexto que o paradigma da ciência guiada por dados - muitas vezes referenciada como Ciência dos Dados, ou *eScience* (no caso de grandes volumes de dados) - se insere (GÓRRIZ *et al.*, 2020), provendo modelos

capazes de anteciparem situações de risco, potencialmente contribuindo para a redução da vulnerabilidade humana às mudanças ambientais globais.

Nessa abordagem, destaca-se, primeiramente, a necessidade de se obter dados curados e integrados de saúde e ambiente (CHAI, 2020). Dados ambientais e de saúde são, essencialmente, volumosos (*big-data*), combinando dados estruturados, semi-estruturados, e não estruturados, coletados de fontes distribuídas, frequentemente armazenados em diferentes formatos, em máquinas que não se comunicam umas com as outras. Dados de saúde, por exemplo, são tipicamente armazenados em servidores de acesso restrito, uma vez que é necessário proteger a privacidade dos indivíduos (NUTLEY; REYNOLDS, 2013), conforme dita a Lei Geral de Proteção dos Dados Pessoais (LGPD) que regula as atividades de tratamento de dados pessoais. A qualidade dos dados também é uma preocupação. Considerando-se ainda os dados de saúde, o processo de transformação da informação em dados digitais ainda é fortemente dependente de intervenção humana, frequentemente resultando em dados imprecisos, incompletos ou inconsistentes. Por outro lado, dados ambientais são tipicamente obtidos de forma automática por uma rede de sensores ou satélites. Neste caso, a qualidade dos dados é fortemente dependente dos equipamentos e técnicas de processamento dos sinais empregadas.

Em segundo lugar, destaca-se a relevância de estudos regionalizados capazes de modelar, de maneira mais precisa, as respostas fenotípicas de uma população ao ambiente, incluindo traços comportamentais e morfológicos (SILVEIRA *et al.*, 2019; SILVEIRA *et al.*, 2021; THOMAS *et al.*, 2018), que influenciam a adaptação humana às diferentes regiões do globo, ao longo do tempo. Isso inclui, por exemplo, o refinamento de modelos teóricos sobre sistemas físicos e biológicos globais para cenários regionais de curto-prazo, particularmente relevante para a geração de alertas antecipados mais precisos.

Finalmente, a geração de modelos integrados de clima e saúde permite a simulação de cenários futuros, como os elaborados pelo Painel Intergovernamental de Mudança Climática (IPCC, do inglês, *Intergovernmental Panel on Climate Change*), abrindo caminhos para a mitigação dos efeitos das mudanças ambientais para a saúde humana, por meio da adoção de medidas antecipadas de adaptação.

Nesse contexto, o presente trabalho visa contribuir com propostas de abordagens e novos resultados associados aos três grandes desafios apontados, tomando como objeto de estudo os impactos de variáveis ambientais em óbitos por doenças cardiovasculares, na cidade de Campinas, no estado de São Paulo.

As doenças cardiovasculares, principalmente a doença isquêmica do coração e o acidente vascular cerebral, são a principal causa de mortalidade global e uma das principais causas para a redução da qualidade de vida no mundo. A prevalência de doenças cardiovasculares no mundo quase dobrou entre os anos de 1990 a 2019 (de cerca de 271 milhões para 523 milhões), e cerca de 18 milhões de pessoas morreram em 2019 (ROTH *et al.*, 2020). No Brasil, as doenças cardiovasculares foram responsáveis por pouco mais de

390 mil mortes em 2018[1]. Além do tabagismo, o sedentarismo, a dieta inadequada e o sobrepeso, os fatores ambientais, tais como a poluição do ar (incluindo a poluição interna às casas por motivo de queima de combustível fóssil para cozer alimentos) e a temperatura fora de patamares considerados ótimos, também caracterizam fatores de risco conhecidos para os óbitos por doenças cardiovasculares (ROTH *et al.*, 2020; MURRAY *et al.*, 2020).

Considerando o desafio da integração dos dados, este trabalho combinou os dados dos óbitos registrados pela Secretaria Municipal de Saúde da cidade de Campinas, filtrados por causas associadas à doenças cardiovasculares, com parâmetros meteorológicos e de poluição, oriundos de estações meteorológicas e de monitoramento da qualidade de ar da cidade de Campinas. Como resultado, além da construção de uma base inédita para a cidade de Campinas, tem-se a consolidação de um método para esse tipo de integração, incluindo o desenvolvimento de rotinas computacionais de pré-processamento e anonimização de dados.

A partir da integração dos dados, este trabalho propôs uma nova abordagem na modelagem das associações existentes entre os valores de diferentes variáveis ambientais e os óbitos por doenças cardiovasculares, visando reproduzir achados da literatura que apontam para um maior risco de óbitos por parâmetros ambientais, tais como valores elevados de material particulado, temperaturas excessivamente baixas ou temperaturas excessivamente altas (ROTH *et al.*, 2020; MURRAY *et al.*, 2020; HAN *et al.*, 2017; GAO *et al.*, 2019; GASPARRINI *et al.*, 2015).

Finalmente, com base nos modelos gerados, realizamos um estudo exploratório do impacto de cenários futuros de mudanças climáticas no número de óbitos por doenças cardiovasculares, com o intuito de implementar uma nova abordagem nas predições de longo prazo.

1.1 Objetivos

O objetivo desse trabalho é o desenvolvimento de um modelo de predição de séries temporais capaz de realizar alertas antecipados para eventos climáticos com potencial de impactar a saúde humana. Ainda, esse trabalho tem o objetivo de oferecer uma nova abordagem para o problema em questão, que será discutida com detalhes no Capítulo 2.

1.2 Perguntas de pesquisa

- **P.1**: É possível utilizar modelos de série temporal para prever o número de óbitos por DCVs com base em dados ambientais?

[1]Dados do Cardiômetro mantido pela Sociedade Brasileira de Cardiologia, acessível pelo endereço de Internet: http://www.cardiometro.com.br/

- **P.2**: É possível realizar previsões de longo prazo utilizando esses modelos?

1.3 Contribuições

Esse trabalho explora modelos de predição de óbitos por doenças cardiovasculares com base em variáveis ambientais. Ainda, faz um estudo de caso a partir de projeções de mudanças climáticas com base em cenários do IPCC. As contribuições desse estudo são:

- **C.1**: Uma nova abordagem na predição de curto e longo prazo de óbitos por doenças cardiovasculares com base em variáveis ambientais;
- **C.2**: Criação de uma base de dados integrada e processada contendo valores de variáveis ambientais e de óbitos por doenças cardiovasculares, para a cidade de Campinas, São Paulo;
- **C.3**: Geração e comparação de modelos capazes de predizer o número de óbitos com base em variáveis ambientais;
- **C.4**: Avaliação do risco futuro de morte por doenças cardiovasculares a partir de projeções de mudança climática.

1.4 Organização

Esse documento está organizado da seguinte forma:

- Capítulo 2 faz uma revisão bibliográfica apresentando conceitos básicos e trabalhos relacionados.
- Capítulo 3 apresenta a metodologia, descrevendo os bancos de dados utilizados e suas transformações, assim como detalhes dos modelos adotados nesse trabalho e a forma de avaliação do desempenho.
- Capítulo 4 expõe os resultados e discute seus significados.
- Capítulo 5 apresenta uma avaliação do risco futuro de óbito por doenças cardiovasculares utilizando nosso modelo juntamente com modelos de projeções do clima sob cenários do IPCC.
- Capítulo 6 apresenta as principais conclusões do trabalho, comentando, também, as limitações da abordagem proposta, e aponta direções para futuras pesquisas no tema.

Capítulo 2

Revisão Bibliográfica

Neste capítulo são apresentados as principais descobertas da literatura sobre o estudo dos impactos de variáveis ambientais no número de óbitos por doenças cardiovasculares em várias regiões do mundo. Ainda, são apresentadas as abordagens de predição de séries temporais exploradas neste trabalho.

2.1 Temperatura e Óbitos por Doenças Cardiovasculares

Da literatura, inúmeros estudos mostram que tanto eventos extremos de frio quanto de calor são prejudiciais à saúde humana. Tipicamente, os trabalhos apresentam uma associação não-linear entre risco relativo de óbito por doença cardiovascular e temperatura de exposição, observando-se aumento do risco a partir de determinados limiares de temperatura, como exemplificado na Figura 2.1, na qual se observa uma elevação do risco relativo para temperaturas médias abaixo de 15 °C ou acima de 25 °C (LIU *et al.*, 2015; HUANG *et al.*, 2017; SILVEIRA *et al.*, 2021).

No entanto, os trabalhos também mostram que os limiares de temperatura variam de acordo com o local ou a região na qual o estudo foi realizado (SILVEIRA *et al.*, 2019; SILVEIRA *et al.*, 2021). Dessa forma, evidencia-se que diferentes populações podem estar mais ou menos adaptadas às variações de temperatura ambiente, e que outros fatores ambientais, tais como níveis de poluição ou acesso a moradias com isolamento térmico, são relevantes para o entendimento do impacto de eventos climáticos extremos na saúde de uma determinada população (HUANG *et al.*, 2017; MOGHADAMNIA *et al.*, 2017; SONG *et al.*, 2017).

Os estudos também mostram que os efeitos dos eventos extremos de frio e calor continuam por dias após o fim da exposição (LIN *et al.*, 2020; SILVEIRA *et al.*, 2019; HUANG *et al.*, 2017). A Figura 2.2 ilustra esse fenômeno, onde exposições a temperaturas elevadas oferecem riscos imediatos e de curta duração (cerca de 5 dias), e nas exposições a

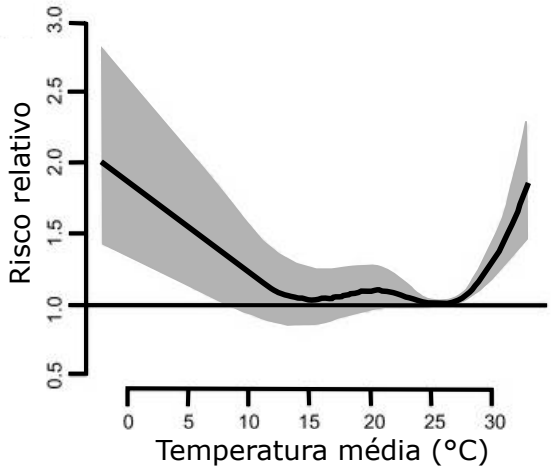


Figura 2.1: Curva de exposição-resposta geral para uma associação genérica entre temperatura e óbitos por doenças cardiovasculares. O sombreado em cinza representa o intervalo de confiança de 95%. Adaptado de (LIU *et al.*, 2015).

baixas temperaturas, os riscos aumentam progressivamente até atingir um pico, por volta do 5º dia, e depois diminuem até desaparecer (LIU *et al.*, 2015). Apesar da Figura 2.2 exibir um aumento do risco relativo 20 dias após a exposição à altas temperaturas, os artigos que abordam a relação entre o calor e as doenças cardiovasculares apontam risco relativo aumentado até no máximo 5 dias após a exposição (LIU *et al.*, 2015).

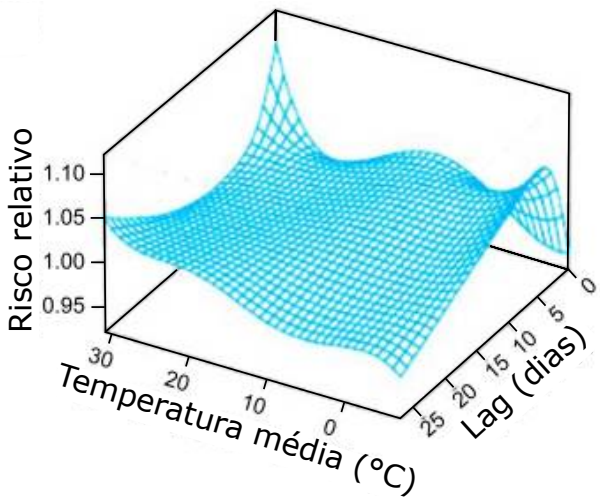


Figura 2.2: Exposição-resposta de óbitos por doenças cardiovasculares por temperatura e lag para uma associação genérica. Adaptado de (LIU *et al.*, 2015).

A temperatura afeta as doenças cardiovasculares de diferentes formas na população. O grupo dos mais susceptíveis, de modo geral, compreende os idosos (>65 anos), pessoas de baixo nível socioeconômico e pessoas com doenças prévias (HUANG *et al.*, 2017; LIU *et al.*, 2015; SONG *et al.*, 2017). Sobre o efeito em doenças cardiovasculares específicas, há trabalhos que mostram que o infarto agudo do miocárdio, por exemplo, é mais afetado, principalmente, em dias quentes (BRAGA *et al.*, 2002), e trabalhos que

apontam a importância dos dias frios nessa causa de óbito (LIU *et al.*, 2015; CHEN *et al.*, 2018).

Destacando a importância dos dias frios, um estudo compreendendo 272 cidades chinesas mostrou que os óbitos por doenças cardiovasculares, tanto de forma geral quanto estratificada por causas específicas, são mais impactados pelo frio moderado do que pelo calor moderado ou por extremos de temperatura. Entretanto, esse fato se deve simplesmente pela maior quantidade de dias sob temperaturas moderadas do que sob temperaturas extremas (CHEN *et al.*, 2018).

Os mecanismos pelos quais a temperatura influencia as DCVs envolvem diversas regulações fisiológicas. Uma delas é a ativação do sistema nervoso simpático e do sistema renina-angiotensina pelo frio, que podem afetar a pressão arterial e levar a hipertensão, que é um fator de risco conhecido para DCVs (LIU *et al.*, 2015). O calor, por sua vez, pode aumentar o fluxo sanguíneo periférico e a sudorese, levando a perda de água do corpo e consequente desidratação. A desidratação por sua vez leva ao aumento da concentração de células no plasma sanguíneo, aumentando assim sua viscosidade. Essas condições desencadeadas pelo calor favorecem o tromboembolismo, assim aumentando o risco de acidentes vasculares (LIU *et al.*, 2015).

Os trabalhos abordados nessa seção evidenciam a variabilidade entre os resultados, possivelmente devido às particularidades tanto do clima quanto das populações estudadas, dessa forma, justificando a realização de estudos regionalizados.

2.2 Poluição e Óbitos por Doenças Cardiovasculares

Quando se trata de poluição, a grande maioria dos artigos aborda majoritariamente o material particulado (RAJAGOPALAN *et al.*, 2018; BOURDREL *et al.*, 2017; DOCKERY, 2009). Esse material é composto por partículas sólidas e líquidas presentes no ar, de fontes naturais ou artificiais, tais como pólen, poeiras, resíduos de combustão, fumaça de cigarro etc. O material particulado é classificado de acordo com o tamanho das partículas, uma vez que diferentes tamanhos possuem diferentes propriedades físico-químicas. O material particulado é geralmente classificado em menores que 2,5 µm (MP2,5) e maiores que 2,5 µm. Não obstante, é comum se utilizar a classificação MP10 (menores que 10 µm) para se referir a um maior intervalo de tamanhos de partículas (QUEIROZ *et al.*, 2007).

Diversos estudos evidenciam os efeitos desse poluente nos óbitos por doenças cardiovasculares, principalmente o MP2,5 e seu efeito a curto prazo (MANNUCCI *et al.*, 2019). Ainda, a quantidade de óbitos por doenças cardiovasculares relacionados ao material particulado (3,3 mi), mundialmente, é superior aos óbitos relacionados a fatores de risco tradicionais, como índice de massa corporal elevado (2,85 mi), níveis elevados de glicose no jejum (2,84 mi) e tabagismo (2,48 mi) (HADLEY *et al.*, 2018).

Informações sobre os mais vulneráveis e susceptíveis ainda são escassas (MANNUCCI *et al.*, 2019; HADLEY *et al.*, 2018), mas já é possível encontrar alguns resultados na literatura. Para a cidade de São Paulo, o risco de óbito por doenças cardiovasculares devido ao material particulado é maior nos grupos de maior nível socioeconômico, mas essa relação pode variar entre regiões (ROMIEU *et al.*, 2012). Ainda, para a cidade de São Paulo, o risco para idosos é o mesmo que para o grupo com todas as idades, enquanto que para a cidade do Rio de Janeiro apenas os idosos são afetados (ROMIEU *et al.*, 2012).

Logo depois de inalado, o material particulado, principalmente de tamanho menor que 2,5 µm, desencadeia a produção de uma série de substâncias inflamatórias no pulmão. Essas substâncias são então liberadas na corrente sanguínea e nas paredes dos vasos, dessa forma, causando uma cascata de reações que levam a ativação sistêmica do processo de coagulação sanguínea favorecendo o processo aterosclerótico (MANNUCCI *et al.*, 2019). Outro mecanismo importante é o estímulo dos nervos sensoriais das vias aéreas, que pode resultar em desequilíbrio do controle autônomo do coração e redução da variação da frequência cardíaca, aumentando assim o risco de arritmia severa e morte súbita (MANNUCCI *et al.*, 2019).

Diferente das associações com a temperatura, em que há mais de um limiar, para a poluição, quanto maior a exposição maior o risco de óbito (LELIEVELD *et al.*, 2019). Assim, a mortalidade mínima por doenças cardiovasculares seria alcançada na ausência de exposição à poluição.

2.3 Previsões

A previsão pode ser entendida como a antecipação do que ainda irá acontecer com base em contexto e dados pré-existentes. Alguns fenômenos podem ser mais previsíveis que outros devido a fatores como: conhecimento sobre o próprio fenômeno e variáveis relacionadas; a quantidade de dados anteriores existentes; e se a própria previsão pode afetar a realidade. Por exemplo, o horário que o sol nasce é altamente previsível devido ao grande conhecimento que se tem sobre o assunto. Em contrapartida, o valor de uma ação na bolsa de valores pode ser complexo de se prever, principalmente devido à grande quantidade de variáveis que podem estar influenciando seu valor e à falta de conhecimento sobre elas (HYNDMAN; ATHANASOPOULOS, 2018).

Uma previsão pode ter caráter qualitativo ou quantitativo a depender da natureza do que se está prevendo. A previsão qualitativa é utilizada quando não há disponibilidade de dados quantitativos sobre o fenômeno ou se os dados quantitativos não são relevantes para a previsão. Em relação as previsões qualitativas, existem abordagens bem definidas para se realizar uma previsão precisa. Nesse caso, se utilizam as previsões de julgamento (*judgemental forecast*, em inglês), que fazem uso de julgamentos intuitivos, opiniões e estimações probabilísticas subjetivas. As previsões quantitativas, por sua vez, são utilizadas

quando há disponibilidade de dados numéricos sobre o passado e se assume que alguns padrões continuarão a se repetir no futuro. A maioria das previsões quantitativas utiliza dados coletados em intervalos regulares no tempo (abordagem de série temporal), ou, então, dados coletados em um único momento (abordagem transversal) (HYNDMAN; ATHANASOPOULOS, 2018).

Em relação aos dados disponíveis, uma previsão pode ainda ser classificada em univariada ou multivariada. As previsões univariadas utilizam apenas dados do presente e do passado da própria série temporal para prever o futuro. As previsões multivariadas utilizam também dados de outras variáveis (preditoras) para prever os valores da variável de interesse (resposta) (HYNDMAN; ATHANASOPOULOS, 2018). A previsão da contagem de óbitos por doenças cardiovasculares por dia utilizando dados de temperatura e poluição é um exemplo de previsão de série temporal multivariada, onde os óbitos representam a variável resposta e a temperatura e poluição representam as variáveis preditoras.

Variáveis preditoras são variáveis que podem estar relacionadas ou podem ajudar a explicar a variável de interesse. Por exemplo, em uma modelagem de demanda horária de eletricidade (DE) poderíamos considerar a temperatura atual, a população, o horário, e o dia da semana. Ainda, inclui-se no modelo um termo de *erro*, que irá permitir variações aleatórias e que, de certo modo, modela as possíveis variáveis relevantes que não foram consideradas. Por possuir variáveis que ajudam a explicar as variações da série de dados, esse modelo é chamado de *modelo explicativo* (HYNDMAN; ATHANASOPOULOS, 2018).

$$\text{DE} = f(\text{temperatura, população, horário, dia da semana, erro}). \tag{2.1}$$

Outra forma de modelar esse problema seria utilizar tanto as variáveis preditoras quanto amostras passadas da própria variável de interesse. Nesse caso tem-se os modelos conhecidos como *modelos de regressão dinâmica* (HYNDMAN; ATHANASOPOULOS, 2018).

$$\text{DE}_{t+1} = f(\text{DE}_t, \text{temperatura, população, horário, dia da semana, erro}), \tag{2.2}$$

onde DE_{t+1} representa, em nosso exemplo, a previsão da demanda de eletricidade na próxima hora e DE_t a demanda de eletricidade no momento presente.

2.3.1 Modelos ARIMA

O modelo autoregressivo integrado de médias móveis (*Autoregressive Integrated Moving Average* - ARIMA, em inglês) é uma das abordagens mais utilizadas na modelagem e previsão de séries temporais. Como o próprio nome diz, uma modelagem ARIMA é composta por um processo autoregressivo (AR), um processo integrativo (I) e um processo de médias móveis (MA) (BOX *et al.*, 2015). No processo autoregressivo se

utiliza a combinação linear dos valores passados da própria série temporal para realizar a previsão. Um modelo autoregressivo de ordem p pode ser descrito como (HYNDMAN; ATHANASOPOULOS, 2018; CHATFIELD, 2000; BOX *et al.*, 2015):

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t, \tag{2.3}$$

onde:

- y_t é o valor da série no tempo t.
- ϕ é o coeficiente do termo autoregressivo.
- ε_t é o termo de erro no tempo t (ruído branco).

Cabe destacar que ε_t segue um processo de ruído branco, ou seja, um processo aleatório de média zero.

O processo de médias móveis faz a previsão a partir de uma combinação linear do valor atual com amostras passadas do erro de regressão. Um processo de médias móveis de ordem ordem q pode ser descrito como (HYNDMAN; ATHANASOPOULOS, 2018; CHATFIELD, 2000):

$$y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}, \tag{2.4}$$

onde:

- y_t é o valor da série no tempo t.
- θ é o coeficiente do termo de médias móveis.
- ε_t é o termo de erro no tempo t (ruído branco).

Esse modelo tem como pressuposto a sua aplicação em séries temporais estacionárias, ou seja, uma série cuja média, variância e função de autocorrelação permanecem constantes ao longo do tempo, conforme exemplificado na Figura 2.3. Entretanto, em dados do mundo real a não-estacionariedade tipicamente se faz presente, sendo necessário realizar a diferenciação dos dados. O processo de diferenciação geralmente transforma uma série não estacionária em estacionária. A operação oposta da diferenciação é a integração, que irá transformar uma série diferenciada na série real novamente. Por esse motivo o modelo ARIMA se trata de um modelo integrado (I) (HYNDMAN; ATHANASOPOULOS, 2018; CHATFIELD, 2000). Assim, um modelo ARIMA(p,d,q) é composto pelos seguintes termos:

- p (AR): ordem do termo autorregressivo.
- d (I): quantidade de diferenciações necessárias.

- q (MA): ordem do termo de médias móveis.

Um modelo ARIMA com todos os termos pode ser descrito como:

$$y'_t = \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t, \tag{2.5}$$

onde:

- y'_t é o valor da série diferenciada no tempo t.
- ϕ é o coeficiente do termo autoregressivo.
- θ é o coeficiente do termo de médias móveis.
- ε_t é o termo de erro no tempo t (ruído branco).

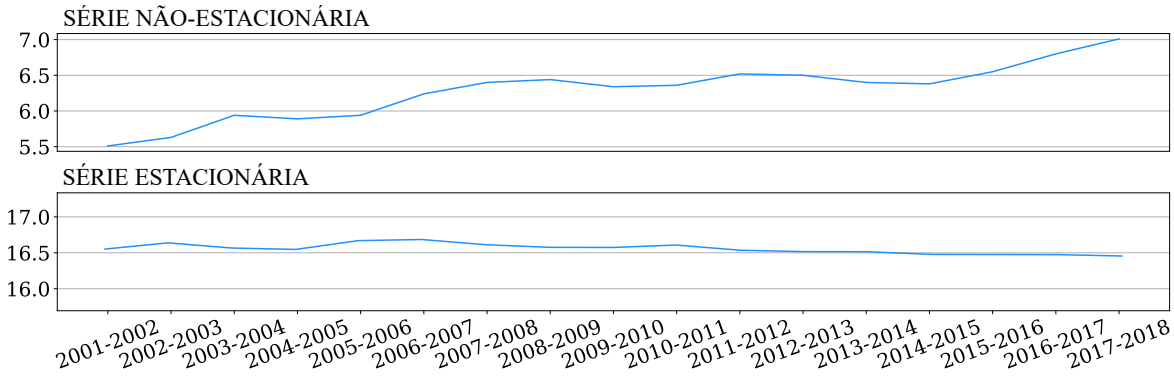


Figura 2.3: Exemplo de estacionariedade em séries temporais genéricas. A figura superior demonstra o comportamento de uma série temporal não-estacionária e a figura inferior uma série estacionária.

Operador Backshift

Ao se trabalhar com valores passados (*lags*) em uma série temporal, o operador *backshift* (B) se torna útil para representar o processo ARIMA de uma maneira compacta. Por exemplo, a representação de um *lag* da variável y_t é igual a (HYNDMAN; ATHANASOPOULOS, 2018):

$$By_t = y_{t-1} \tag{2.6}$$

Do mesmo modo, para representar um valor referente a 12 períodos atrás a notação é igual a:

$$B^{12} y_t = y_{t-12} \tag{2.7}$$

Este operador também é útil para representar a diferenciação da série temporal, conforme mencionado na Seção 2.3.1. Por exemplo, uma série diferenciada duas vezes seria representada como (HYNDMAN; ATHANASOPOULOS, 2018):

$$y''_t = y_t - 2y_{t-1} + y_{t-2} = (1 - 2B + B^2)y_t = (1 - B)^2 y_t, \tag{2.8}$$

onde:

- y''_t é o valor da série duplamente diferenciada no tempo t.
- y_t é o valor da série real no tempo t.
- B é o operador *backshift*.

Modelos SARIMA

O modelo ARIMA ainda pode ser estendido para tratar de dados com sazonalidade (S), sendo, nessa forma, conhecido como modelo SARIMA. Para isso, além dos termos já mencionados no início da seção, incluem-se termos sazonais no modelo. Esses termos são similares aos componentes não-sazonais, i.e., são termos autorregressivos, de diferenciação e de médias móveis, mas referentes a atrasos do período sazonal escolhido (HYNDMAN; ATHANASOPOULOS, 2018). Assim, um modelo SARIMA(pdq)(PQD)s é composto pelos seguintes termos:

- p (AR): ordem do termo autorregressivo.
- d (I): quantidade de diferenciações necessárias.
- q (MA): ordem do termo de médias móveis.
- P (SAR): ordem do termo sazonal autorregressivo.
- D (SI): quantidade de diferenciações sazonais necessárias.
- Q (SMA): ordem do termo sazonal de médias móveis.
- s: número de observações por ano (períodos).

Por exemplo, um modelo SARIMA$(111)(111)_4$ com dados trimestrais pode ser representado como:

$$(1 - \phi_1 B)\ (1 - \Phi_1 B^4)(1 - B)(1 - B^4)y_t = (1 + \theta_1 B)\ (1 + \Theta_1 B^4)\varepsilon_t. \tag{2.9}$$

onde:

- y_t é o valor da série no tempo t.

- ϕ é o coeficiente do termo autoregressivo.
- θ é o coeficiente do termo de médias móveis.
- Φ é o coeficiente do termo sazonal autoregressivo.
- Θ é o coeficiente do termo sazonal de médias móveis.
- B é o operador *backshift*.
- ε_t é o termo de erro no tempo t (ruído branco).

Regressão Linear com erros SARIMA

As abordagens descritas até agora são frequentemente usadas para modelar séries temporais. Contudo, elas tem originalmente um caráter univariado e, consequentemente, não oferecem possibilidade de utilização de múltiplas variáveis preditoras em sua modelagem. Uma vez que esse trabalho deseja modelar a contagem de óbitos em conjunto com variáveis do clima, modelos multivariados são fundamentais. A regressão linear com erros SARIMA (LR-SARIMAX, onde 'X' se refere às variáveis eXógenas preditoras) é uma abordagem que pode oferecer esses recursos por se tratar de um *modelo de regressão dinâmica* (HYNDMAN; ATHANASOPOULOS, 2018).

Uma regressão linear convencional encontraria problemas na modelagem de séries temporais. Dados de séries temporais comumente são autocorrelacionados e não-estacionários, podendo levar a regressões espúrias, ou seja, uma relação estatística sem relação de causa-efeito (HYNDMAN; ATHANASOPOULOS, 2018).

Na abordagem a partir do LR-SARIMAX as variáveis preditoras são ajustadas conforme uma regressão linear comum, enquanto os erros da regressão linear (η_t), que dessa vez podem estar correlacionados, são ajustados em um processo SARIMA, conforme mostram as equações a seguir:

$$y_t = \beta_t x_t + n_t \tag{2.10}$$

$$\phi_p(B)\Phi_P(B^s)\Delta^d\Delta_s^D n_t = \theta_q(B)\Theta_Q(B^s)\epsilon_t, \tag{2.11}$$

onde:

- y_t é o valor da variável resposta no tempo t.
- β_t é o coeficiente da variável preditora x_t.
- x_t é o valor da variável preditora no tempo t.
- n_t é o termo de erro da regressão linear no tempo t que se assume seguir um processo SARIMA.

- $\phi_p(B)$ são os termos autorregressivos não-sazonais.
- $\Phi_P(B^s)$ são os termos autorregressivos sazonais.
- Δ^d e Δ_s^D é a série diferenciada d vezes não-sazonalmente, e D vezes sazonalmente, respectivamente.
- $\theta_q(B)$ são os termos de médias móveis não-sazonais.
- $\Theta_Q(B^s)$ são os termos de médias móveis sazonais.
- ϵ_t é o termo de erro do modelo SARIMA no tempo t (ruído branco).

2.3.2 Redes Neurais Artificiais

As redes neurais artificiais são modelos computacionais inspirados no funcionamento dos neurônios humanos. A arquitetura mais simples de uma rede neural artificial é conhecida como *perceptron* e está ilustrada na Figura 2.4. Primeiramente, esse modelo computa a soma ponderada dos dados de entrada $\mathbf{X}$ (X1, X2 e X3) através dos pesos sinápticos $\mathbf{W}$ (w1, w2 e w3). Por fim, aplica-se uma função de ativação $\varphi(\cdot)$, como por exemplo uma função degrau (Heaviside), que irá definir o valor de saída dos dados (GÉRON, 2019). A saída do modelo $h_w(\mathbf{X})$ pode ser expressa matematicamente da seguinte forma:

$$h_w(\mathbf{X}) = \varphi(\mathbf{X}^T\mathbf{W}) \tag{2.12}$$

Uma extensão emblemática e de grande importância é a rede perceptron de múltiplas camadas (*multilayer perceptron* - MLP, em inglês), na qual várias camadas formadas por neurônios do tipo perceptron (Figura 2.4) são colocadas em sequência e processam a informação em uma única direção, desde a entrada até a saída, conforme ilustrado na Figura 2.5. No exemplo da Figura 2.5, também pode se notar que os neurônios recebem um sinal de polarização (*Bias*) em conjunto com os dados de entrada, que também serão ajustados e considerados para gerar o valor de saída (GÉRON, 2019; BROWNLEE, 2016). A grande vantagem do uso das redes neurais é sua capacidade de aproximação universal (HORNIK *et al.*, 1990). Assim, uma rede neural tem a capacidade de mapear funções que não seriam possíveis através de um modelo ARIMA, por exemplo. Uma vez que não se conhece a relação exata entre as variáveis de clima e os óbitos na cidade de Campinas, esse modelo se torna de grande utilidade.

Treinamento das redes neurais

Por treinamento entende-se o processo de ajuste dos parâmetros internos de uma rede neural, i.e., ajuste dos valores dos pesos sinápticos $\mathbf{W}$ e termos de *bias* (Figuras 2.4 e 2.5). O treinamento supervisionado é uma das diversas formas de se treinar uma rede.

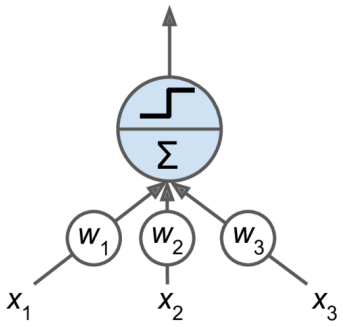


Figura 2.4: Arquitetura de um *perceptron*. Adaptado de (GÉRON, 2019).

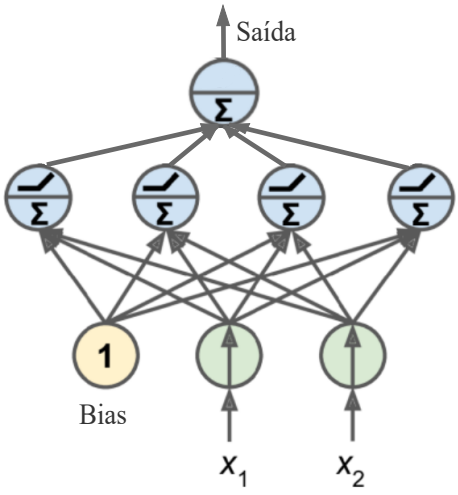


Figura 2.5: Arquitetura de um *perceptron* multicamadas para função de regressão. Adaptado de (GÉRON, 2019).

Esse tipo de treinamento ocorre quando a rede neural aprende a partir de exemplos rotulados, ou seja, quando a rede gera um resultado e o compara com o resultado real. Primeiramente, os dados de entrada são multiplicados pelos pesos sinápticos iniciais, que passam por uma função de ativação somada ao termo *bias* e geram um resultado de saída (previsão). Essa previsão é comparada com o resultado real através de uma função de perda, que calcula quanto a resposta do modelo diferiu dos dados reais. Então, busca-se minimizar a função de perda com respeito aos parâmetros livres da rede, o que constitui um problema de otimização não-linear irrestrita. O processo de treinamento, então, é realizado com o auxílio de algoritmos iterativos baseados no gradiente, como o gradiente descendente estocástico (*Stochastic Gradient Descent* - SGD, em inglês) e o Adam, os quais exploram a noção de retropropagação do erro (error backpropagation) para determinar o gradiente em relação aos parâmetros de camadas internas da arquitetura. (GÉRON, 2019; BROWNLEE, 2016; KINGMA; BA, 2015).

Além dos parâmetros que serão definidos ao fim do processo de treinamento, as redes neurais ainda possuem parâmetros a serem definidos antes do treinamento e que exercem influência no processo de aprendizagem (hiperparâmetros). Dentre esses hiperparâmetros podemos citar a quantidade de camadas ocultas da rede, o tipo de função de ativação e a quantidade de neurônios em cada camada (GÉRON, 2019). Os detalhes da configuração das redes utilizadas nesse trabalho se encontram no Capítulo 3.

Rede Neural LSTM

As arquiteturas de redes neurais descritas acima se tratam de redes conhecidas como FNNs (*feedforward neural networks*, em inglês), onde a informação é processada em uma única direção (Figura 2.5) e são capazes de processar informações de dados anteriores apenas quando essas são utilizadas como dados de entrada. Entretanto, existe outra arquitetura que é muito utilizada na previsão de séries temporais, sendo conhecidas como RNNs (*Recurrent Neural Networks*, em inglês). A diferença fundamental desse tipo de arquitetura para as *feedforward* está na sua capacidade de memória devido à presença de realimentações. Uma possibilidade simples para obter uma RNN consiste em permitir que o sinal de saída de um neurônio também seja utilizado como sinal de entrada do mesmo neurônio, conforme demonstrado na Figura 2.6 A. Figura 2.6 B ilustra o funcionamento desse neurônio no tempo, onde o valor de saída y é utilizado em conjunto com os dados de entrada x (GÉRON, 2019).

De fato, a capacidade de memória das RNNs representa uma vantagem em relação às FNNs. No entanto, as RNNs não conseguem fazer uso de dados mais antigos da série (memória de longo prazo), devido ao desvanecimento de gradiente durante o processo de treinamento, que impede o ajuste dos pesos dos neurônios. Nesse sentido, as células LSTM (*Long Short-Term Memory*, em inglês), ilustrada na Figura 2.7, apresentam superioridade em relação a uma RNN padrão por serem capazes de também detectar relações de longo

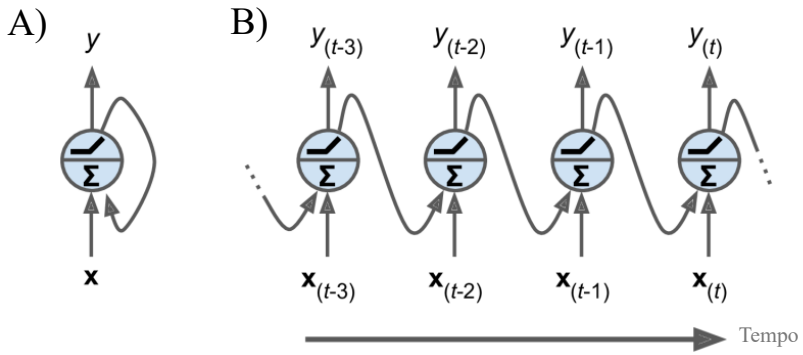


Figura 2.6: Arquitetura básica de uma rede neural recorrente (A) e seu desdobramento no tempo (B). Adaptado de (GÉRON, 2019).

prazo nos dados (GÉRON, 2019; BROWNLEE, 2017). Em um modelo de previsão de óbitos baseado em variáveis do clima essa característica se torna interessante, pois a rede seria capaz de detectar tanto as relações de curto prazo entre as variáveis, conforme descrito na Seção 2.1, quanto possíveis efeitos sazonais de longo prazo. As equações 2.13 sumarizam como os dados são computados em uma célula LSTM.

O processamento de dados em uma célula LSTM pode ser segmentado para melhor compreensão (Figura 2.7 e Equação 2.13). O vetor de estados $c(t)$ processa quais dados serão armazenados, descartados, ou reaproveitados na saída $h(t)$. A passagem de informação no vetor de estados é controlada por três portas: porta de esquecimento, porta de entrada e porta de saída. A porta de esquecimento $f(t)$ determina quais informações serão esquecidas a partir dos dados de saída da célula anterior e dos novos dados de entrada(Figura 2.7 e Equação 2.13). No próximo passo, a porta de entrada, composta por $g(t)$ e $i(t)$, determina quais informações novas serão adicionadas ao vetor de estados $c(t)$. Por fim, a porta de saída $o(t)$ processa a saída $y(t)$ da rede com base nos dados de entrada $x(t)$, na saída anterior $h(t-1)$ e no vetor de estados $c(t)$.

$$
\begin{aligned}
\mathbf{c}(\mathrm{t}) &= \mathbf{f}(\mathrm{t}) \otimes \mathbf{c}_{(t-1)} + \mathbf{i}(\mathrm{t}) \otimes \mathbf{g}(\mathrm{t}) \\
\mathbf{h}(\mathrm{t}) &= \mathbf{y}(\mathrm{t}) = \mathbf{o}(\mathrm{t}) \otimes \tanh(\mathbf{c}_{(t)}) \\
\mathbf{f}(\mathrm{t}) &= \sigma(\mathbf{W}_{xf}^T \mathbf{x}_{(t)} + \mathbf{W}_{hf}^T \mathbf{h}_{(t-1)} + \mathbf{b}_i) \\
\mathbf{g}(\mathrm{t}) &= \tanh(\mathbf{W}_{xg}^T \mathbf{x}_{(t)} + \mathbf{W}_{hg}^T \mathbf{h}_{(t-1)} + \mathbf{b}_g) \\
\mathbf{i}(\mathrm{t}) &= \sigma(\mathbf{W}_{xi}^T \mathbf{x}_{(t)} + \mathbf{W}_{hi}^T \mathbf{h}_{(t-1)} + \mathbf{b}_i) \\
\mathbf{o}(\mathrm{t}) &= \sigma(\mathbf{W}_{xo}^T \mathbf{x}_{(t)} + \mathbf{W}_{ho}^T \mathbf{h}_{(t-1)} + \mathbf{b}_o)
\end{aligned}
\tag{2.13}
$$

onde:

- $\mathbf{W}_{xi}$,$\mathbf{W}_{xf}$,$\mathbf{W}_{xo}$,$\mathbf{W}_{xg}$ são os pesos das matrizes associados aos dados de entrada.
- $\mathbf{W}_{hi}$,$\mathbf{W}_{hf}$,$\mathbf{W}_{ho}$,$\mathbf{W}_{hg}$ são os pesos das matrizes associados à saída da célula no momento anterior h_{t-1}.

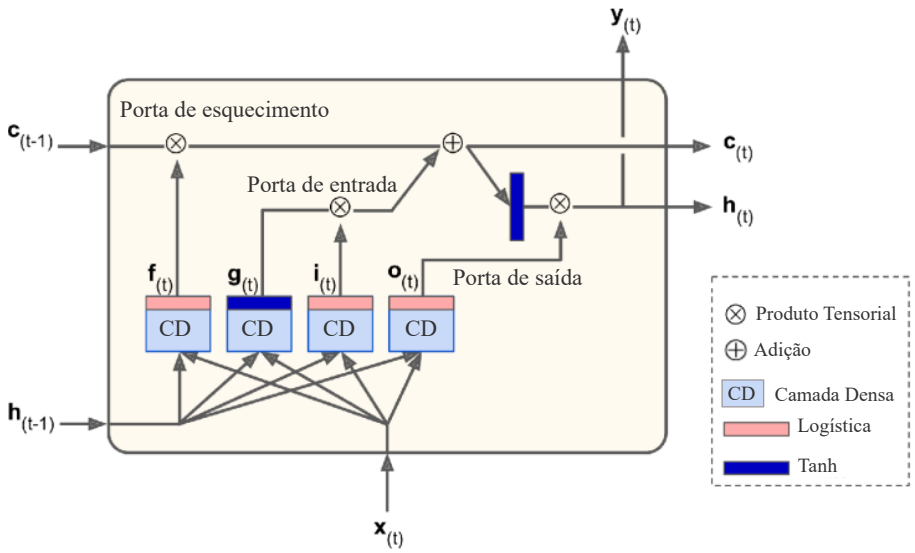


Figura 2.7: Arquitetura de uma célula LSTM. Adaptado de (GÉRON, 2019).

- $\mathbf{b}_i$,$\mathbf{b}_f$,$\mathbf{b}_o$,$\mathbf{b}_g$ são os sinal de polarização (*bias*) associados a cada uma das quatro camadas densas.

2.3.3 Avaliação do Modelo

Uma vez que os modelos foram ajustados se faz necessário avaliar seu desempenho, que é realizado através da medição de métricas de erro. Uma prática comumente utilizada é a divisão dos dados disponíveis em conjuntos de treino e teste, tipicamente na proporção de 80% e 20%, respectivamente. Dessa forma, ajusta-se o modelo com os dados de treino e avalia-se o quanto os modelos erram em dados inéditos, i.e., no conjunto de teste. A avaliação com dados de teste é fundamental, pois uma situação indesejada que pode ocorrer é o modelo apresentar erros baixos quando avaliado com dados de treino, mas apresentar erros altos com dados de teste. Nesse caso, diz-se que o modelo apresenta sobreajuste (GÉRON, 2019; HYNDMAN; ATHANASOPOULOS, 2018).

A depender da tarefa que o modelo realizará (regressão ou classificação), existem métricas específicas e que são amplamente utilizadas. A previsão de séries temporais na contagem de óbitos a partir de variáveis contínuas, como temperatura e poluição, trata-se de uma tarefa de regressão. Nesse caso, existem métricas de erro dependentes da escala dos dados, ou seja, que não podem ser comparadas com diferentes conjuntos de dados, e métricas que não dependem de escala, podendo ser utilizadas para comparar diferentes modelos que utilizam dados com diferentes unidades de medida (HYNDMAN; ATHANASOPOULOS, 2018).

Dentre as métricas dependentes de escala mais populares estão a raiz do erro quadrático médio (*Root Mean Squared Error* - RMSE, em inglês) e o erro médio absoluto (*Mean Absolute Error* - MAE, em inglês). A principal diferença entre essas métricas

está na forma com que elas penalizam os erros, sendo que para o RMSE, erros maiores serão mais penalizados, enquanto para o MAE, a penalidade é linear. Ambas métricas apresentam valores na mesma unidade de medida da variável resposta, facilitando sua interpretação (HYNDMAN; ATHANASOPOULOS, 2018). Essas métricas podem ser matematicamente expressas da seguinte forma:

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}, \tag{2.14}$$

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|, \tag{2.15}$$

onde n é o tamanho amostral, y_j o valor real e $\hat{y}_j$ o valor previsto.

Já a métrica independente de escala mais utilizada é o erro percentual absoluto médio (*Mean Absolute Percentage Error* - MAPE, em inglês). Essa métrica avalia a porcentagem em que o modelo errou em relação ao valor real (HYNDMAN; ATHANASOPOULOS, 2018), e pode ser matematicamente expressa como:

$$MAPE = \frac{1}{n}\sum_{j=1}^{n}\left|\frac{(y_j - \hat{y}_j)}{y_j}\right|, \tag{2.16}$$

onde n é o tamanho amostral, y_j o valor real e $\hat{y}_j$ o valor previsto.

Neste trabalho optou-se pelo RMSE como principal métrica de erro, uma vez que ela penaliza erros muito grandes, tanto para mais quanto para menos. Ainda, para comparar modelos com dados em diferentes escalas, como é o caso de contagem de óbitos por dia, semana e mês, optou-se pela métrica MAPE.

No contexto desse trabalho serão utilizados a regressão linear com erros SARIMA (LR-SARIMAX) e a rede neural recorrente do tipo LSTM. A LR-SARIMAX foi escolhida pelo fato de ser uma abordagem capaz de lidar com regressão de série temporal multivariável, como é o caso da modelagem da quantidade de óbitos por DCVs e de valores de variáveis ambientais, e por considerar explicitamente aspectos temporais das variáveis como a sazonalidade, através dos termos sazonais do modelo. A LSTM foi escolhida por ser capaz de modelar possíveis relações não lineares entre as variáveis e por sua capacidade de memória, que pode ser útil na modelagem de variáveis que continuam exercendo efeito por dias após a exposição, como é o caso da temperatura.

2.4 Trabalhos Relacionados

Assim como apresentado no Capítulo 1, o objetivo desse trabalho é o desenvolvimento de um modelo de predição de séries temporais capaz de realizar alertas antecipados para eventos climáticos com potencial de impactar a saúde humana (óbitos por DCVs).

No limite do nosso conhecimento, esse é o primeiro trabalho que aborda a predição de óbitos por DCVs a partir de uma regressão linear com erros SARIMA e de uma rede neural recorrente LSTM.

Conforme apresentado na Tabela 2.1, onde estão listados os trabalhos relacionados, é possível observar que este trabalho oferece uma nova abordagem para o problema. O trabalho de maior similaridade, que de fato realiza uma tarefa de regressão na predição[1] de óbitos, o faz em duas etapas, e apenas para longo prazo, utilizando somente a temperatura como variável preditora (SILVEIRA *et al.*, 2021). Silveira *et al.* (2021) primeiramente estimam a relação entre a variável temperatura e os óbitos por DCVs através de um Modelo Não-linear de Tempo Tardio (*Distributed Lag Non-linear Model - DLNM*, em inglês), e depois, extrapolam essa relação na projeção[2] de óbitos para cenários de mudanças climáticas a partir de uma função Log-linear. Outros autores também utilizam o DLNM para avaliar a relação entre as variáveis temperatura e óbitos por DCVs. Porém, nenhum dos trabalhos realiza predições de curto prazo (SILVEIRA *et al.*, 2021; SILVEIRA *et al.*, 2019; HUANG *et al.*, 2017; SON *et al.*, 2016).

Wang *et al.* (2020) preenchem essa lacuna das predições de curto prazo, inclusive utilizando rede neural LSTM e outras variáveis preditoras além da temperatura. Todavia, os autores trabalham apenas com a contagem de pacientes ambulatoriais, i.e., pacientes atendidos e liberados no mesmo dia.

Alguns trabalhos bem interessantes, mas distantes deste trabalho por abordarem a tarefa de classificação, avaliam se os pacientes desenvolverão DCVs com base em diversas variáveis preditoras, tais como idade, sexo e nível de colesterol (GHOSH *et al.*, 2021; MOHAN *et al.*, 2019). Entretanto, esses trabalhos realizam a predição de forma individualizada para cada paciente, além de se tratar de uma abordagem completamente diferente, fugindo do escopo das séries temporais.

Por fim, existem diversos autores que utilizam abordagens semelhantes às propostas neste trabalho, evidenciando o potencial de tais abordagens, mas consideram outros tipos de dados, ainda se tratando de óbitos, como por exemplo óbitos por COVID-19 (KUMAR *et al.*, 2020), por acidente de trânsito (WAI *et al.*, 2019) e taxas de mortalidade por DCVs e infantil (BANERJEE; HUTH, 2020; PURWANTO *et al.*, 2010).

2.5 Considerações Finais

Esse capítulo apresentou os conceitos básicos e os trabalhos relacionados no qual este trabalho se baseia. No começo do capítulo foi apresentada a relação entre temperatura, poluição e óbitos por doenças cardiovasculares. Os estudos mostram a relação temporal entre as variáveis e a variação dos efeitos entre diferentes regiões, destacando a importância

[1] *predição*: previsão baseada em variáveis com valores reais
[2] *projeção*: previsão baseada em cenários variáveis com valores simulados

Tabela 2.1: Trabalhos relacionados listados por autor, abordagem e propósito. Para fins de comparação, este trabalho está apresentado na primeira linha e destacado em negrito. Os trabalhos relacionados estão ordenados por ordem de relevância e similaridade a partir da segunda linha.

AUTORES	ABORDAGEM	PROPÓSITO
Este trabalho	**- LR-SARIMAX** **- LSTM**	**- Predição de óbitos por DCVs** **- Projeção de óbitos por DCVs sob cenários de mudanças climáticas**
Silveira *et al.* (2021)	- Modelo Não-linear de Tempo Tardio - Extrapolação Log-linear	- Estimação da relação entre temperatura e óbitos por DCVs - Projeção de óbitos por DCVs sob cenários de mudanças climáticas
Wang *et al.* (2020)	- LSTM	- Predição do número de pacientes ambulatoriais por DCVs
Silveira *et al.* (2019)	- Modelo Não-linear de Tempo Tardio	- Estimação da relação entre temperatura e óbitos por DCVs
Huang *et al.* (2017)	- Modelo Não-linear de Tempo Tardio	- Estimação da relação entre temperatura e óbitos por DCVs
Son *et al.* (2016)	- Modelo Linear Generalizado	- Estimação da relação entre temperatura e óbitos por DCVs
Ghosh *et al.* (2021)	- Árvores de Decisão - Gradient Boosting	- Predição de DCV para cada indivíduo (classificador)
Mohan *et al.* (2019)	- Máquina de Vetores de Suporte - Árvore de Decisão - Naive Bayes - Modelo Linear Generalizado	- Predição de DCV para cada indivíduo (classificador)
Kumar *et al.* (2020)	- ARIMA - LSTM	- Previsão de óbitos por COVID-19
Wai *et al.* (2019)	- Modelo Linear Generalizado - ARIMA	- Previsão de óbitos por acidente de trânsito
Banerjee e Huth (2020)	- ARIMA	- Projeção de taxa de DCVs
Purwanto *et al.* (2010)	- ARIMA - Perceptron Multicamadas	- Predição da taxa de mortalidade infantil

de estudos locais. Ainda, o capítulo apresentou a definição de série temporal, caracterizando a abordagem desse trabalho na modelagem da quantidade de óbitos por DCVs em conjunto com valores de variáveis ambientais. O capítulo também apresentou as ferramentas mais utilizadas na modelagem de séries temporais e descreveu os conceitos básicos de cada ferramenta, assim como descreveu a metodologia para avaliar tais modelos. Por fim, foram apresentados os trabalhos relacionados, ressaltando a sua contribuição, no sentido de propor uma nova abordagem na predição de séries temporais de óbitos por DCVs a partir de variáveis ambientais.

Capítulo 3

Metodologia

Este capítulo está dividido em cinco seções. A Seção 3.1 apresenta a visão geral da metodologia adotada nesse trabalho. A Seção 3.2 faz um detalhamento sobre os bancos de dados utilizados e descreve seu pré-processamento para composição do banco de dados integrado. A Seção 3.3 descreve o processo de seleção e pré-processamento das variáveis utilizadas nos modelos de predição. A Seção 3.4 apresenta as configurações dos modelos utilizados e a forma como foram avaliados. Por fim, a Seção 3.5 traz as considerações finais do capítulo.

3.1 Visão Geral da Metodologia

A metodologia desse trabalho pode ser sumarizada pela sequência de etapas descritas a seguir:

- **Etapa 1 - Captação, Análise Exploratória e Transformação dos Dados:** A base de dados de óbitos foi obtida por meio de parceria com a Secretaria Municipal de Saúde de Campinas (SMSC) e seu uso para fins de pesquisa foi autorizado pelo Comitê de Ética em Pesquisa da Unicamp (CAAE 95503318.4.0000.5404), uma vez que contém dados sensíveis, de acordo com a Lei Geral de Proteção de Dados Pessoais (LGPD). Os bancos de dados de variáveis meteorológicas foram obtidos por meio de parceria com o Centro de Pesquisas Meteorológicas e Climáticas Aplicadas à Agricultura da Unicamp (CEPAGRI), aeroporto de Viracopos (VCP) e Instituto Agronômico de Campinas (IAC). Adicionalmente, dados de poluição atmosférica foram obtidos por meio do sistema informatizado de monitoramento de ar[1] da Companhia Ambiental do Estado de São Paulo (CETESB). O processo de captação incluiu a documentação das bases de dados por meio de reuniões técnicas com especialistas, que foram fundamentais para o entendimento das variáveis presentes nos bancos de dados. Concluído o processo de documentação das bases, realizou-se

[1] <https://cetesb.sp.gov.br/ar/qualar/>

o processo de análise exploratória e transformação dos dados, como descrito na Seção 3.2.

- **Etapa 2 - Integração de Bases de Dados de Clima e Saúde:** Nesta etapa foi criada uma base de dados integrada com parâmetros ambientais e de óbitos por todas as causas. A integração foi realizada tomando-se como referência inicial os seguintes bancos de dados e variáveis:

  – A caracterização do clima foi realizada por dados diários de umidade relativa, temperaturas médias, máximas e mínimas das estações meteorológicas do Instituto Agronômico de Campinas (período de 1890 a 2018), do aeroporto de Viracopos (período de 1998 a 2018) e do Centro de Pesquisas Meteorológicas e Climáticas Aplicadas à Agricultura (período de 1997 a 2018).

  – Informações diárias sobre poluição atmosférica foram obtidas pelo sistema de informação QUALAR, disponibilizado pela CETESB, no período de 2001 a 2019, para três estações de monitoramento em Campinas, identificadas como Centro, Taquaral e Vila União. Os índices utilizados são concentrações diárias de monóxido de carbono (CO) e material particulado (MP10).

  – Número de óbitos diários por todas as causas no município de Campinas no período de 2000 a 2019. Os óbitos estão categorizados por causa básica. A causa básica de um óbito é definida como a doença ou lesão que iniciou a cadeia de acontecimentos patológicos que conduziram diretamente à morte, ou as circunstâncias do acidente ou violência que produziram a lesão fatal (Ministério da Saúde : FUNASA, 2001).

- **Etapa 3 - Geração e Comparação de Modelos:** Construção de modelos de regressão de série temporal capazes de predizer o número de óbitos DCVs. Nesta etapa, foram construídos dois modelos: um modelo de regressão linear com erros SARIMA (LR-SARIMAX) e um modelo de redes neurais recorrentes do tipo LSTM (*Long Short-Term Memory*, em inglês) (GOODFELLOW *et al.*, 2016). Para os dados de entrada, foram utilizadas as variáveis ambientais, tanto para o dia atual quanto para L dias anteriores, uma vez que após a exposição, a temperatura ainda pode continuar impactando a saúde, conforme apresentado na Seção 2.1 do Capítulo 2. O desempenho dos modelos foi avaliado através da raiz do erro quadrático médio (ou *Root Mean Square Error*, RMSE, em inglês) e do erro percentual absoluto médio (ou *Mean Absolute Percentage Error*, MAPE, em inglês).

- **Etapa 4 - Estudo Exploratório de Risco Futuro:** Estudo do risco futuro para óbitos causados por doenças cardiovasculares baseado em dados gerados pelo modelo climático regional Eta, considerando um cenário de altas emissões de gases

estufa (mais pessimista, RCP 8,5) e um cenário de emissões intermediárias (RCP 4,5) (CHOU *et al.*, 2014a). Esse etapa será discutida no Capítulo 5.

3.2 Captação, Análise Exploratória e Transformação dos Dados

A cidade de Campinas está localizada no Sudeste do Brasil no bioma Mata Atlântica com clima subtropical úmido de inverno seco (Cwa), segundo a classificação de Köppen-Geiger (BECK *et al.*, 2018). A cidade tem 794,5 km^2 e uma população estimada de 1.223.237 habitantes, a maioria entre 15 e 45 anos. A captação dos dados ambientais e de saúde consistiu na reunião de planilhas oriundas de diversas instituições da cidade. Os dados de óbitos foram obtidos da Secretaria Municipal de Saúde de Campinas; os dados meteorológicos foram obtidos do aeroporto de Viracopos, do CEPAGRI e do IAC; por fim, os dados de poluição foram obtidos de diversas estações de monitoramento da CETESB. A Figura 3.1 ilustra a localização das estações meteorológicas.

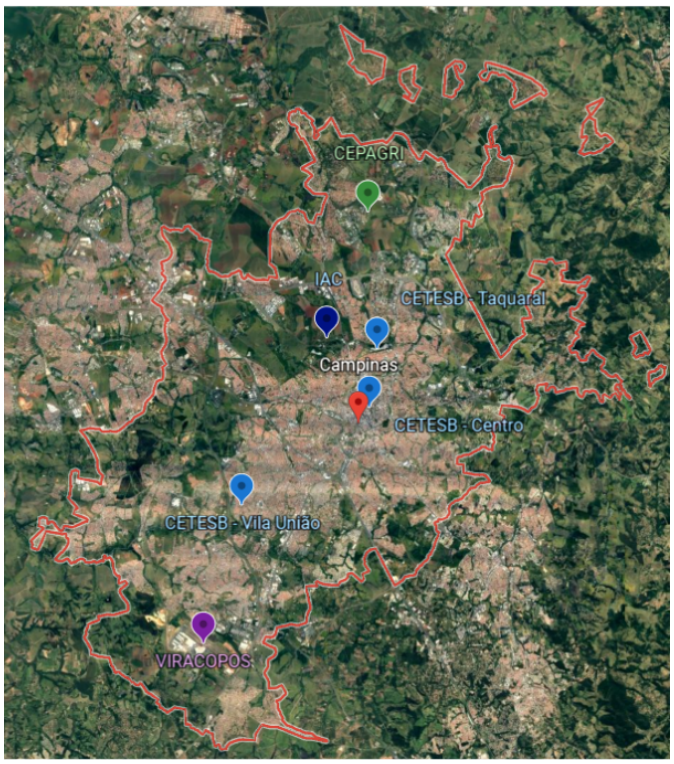


Figura 3.1: Localização das estações meteorológicas. O ponto em verde indica a estação do CEPAGRI; o ponto em azul escuro indica a estação do IAC; os pontos em azul claro indicam as estações da CETESB (de cima para baixo: Taquaral, Centro e Vila União); o ponto em roxo indica a estação do aeroporto de Viracopos; e o ponto em vermelho, ao centro, indica o centro da cidade de Campinas. As linhas em vermelho delimitam os limites da cidade.

Uma vez que os dados provêm de fontes distintas e registram diferentes informa-

ções ou grandezas, se faz necessário um pré-processamento antes de sua integração em um banco de dados unificado. Além disso, bancos de dados do mundo real frequentemente apresentam problemas, como ausência de valores em determinados campos, informações divergentes ou duplicadas, ou até mesmo informações desnecessárias criadas pelo sistema de gerenciamento dos dados.

3.2.1 Óbitos por Doenças Cardiovasculares

O banco de dados de óbitos registra pessoas que faleceram em Campinas e eram residentes do município, que faleceram no município e não eram residentes, ou que eram residentes, mas faleceram em outro município. Originalmente, o banco de dados registra 83 campos de informação de um total de 160174 óbitos ocorridos entre 2001 e 2018. As variáveis descrevem, dentre outras informações, a identificação do indivíduo e do bairro de sua residência, data, hora e local do ocorrido, a causa básica e a circunstância da morte. No pré-processamento dos dados foram removidos óbitos infantis e fetais, duplicatas, óbitos sem data de nascimento ou com sexo ignorado. Após pré-processamento, que também resultou em sua anonimização, restaram as variáveis de idade, sexo e cor do indivíduo, data do óbito e sua causa básica.

3.2.2 Dados Meteorológicos

O banco de dados meteorológicos do CEPAGRI é baseado em registros a cada 10 minutos. Esse banco contém informações de direção e velocidade do vento, radiação solar, fluxo de calor no solo, pluviosidade, pressão e umidade relativa do ar, e temperaturas do ar e do solo. Para o presente trabalho foram selecionadas as variáveis de pressão, umidade relativa e temperatura do ar. A partir dos registros selecionados foi construída uma nova base com os valores diários de temperatura máxima e mínima, pressão atmosférica média e umidades relativas mínima, média e máxima.

O banco de dados do IAC contém registros diários de pluviosidade, e temperaturas máximas e mínimas do ar. No contexto desse trabalho descartou-se os dados de pluviosidade.

O banco de dados da estação meteorológica do aeroporto de Viracopos também foi fornecido com dados diários. Este banco contém registros de temperaturas máxima e mínima, umidades relativas máxima e mínima, pressão atmosférica média, direção e velocidade do vento, precipitação, e dados sobre nuvens. Neste trabalho foram selecionadas apenas as variáveis de temperaturas máxima e mínima, umidades relativas máxima e mínima, além da pressão atmosférica média.

3.2.3 Dados de Poluição

O banco de dados da CETESB é proveniente das estações de monitoramento de qualidade do ar localizadas nos bairros Taquaral, Vila União e Centro de Campinas. As estações de monitoramento reportam dados horários de monóxido de carbono (CO), material particulado de 10 µm (MP10) e menor ou igual a 2,5 µm (MP2,5), óxidos de nitrogênio (NO, NO_2 e NOx), ozônio (O_3), direção e velocidade do vento, temperatura, pressão e umidade relativa do ar, e radiação solar e ultravioleta. Entretanto, as únicas variáveis coletadas de forma consistente e durante o mesmo período do banco de dados de óbitos (2001-2018) foram CO e MP10, provenientes da estação do Centro de Campinas. Os dados horários de CO e MP10 foram transformados em dados diários médio, máximo e mínimo.

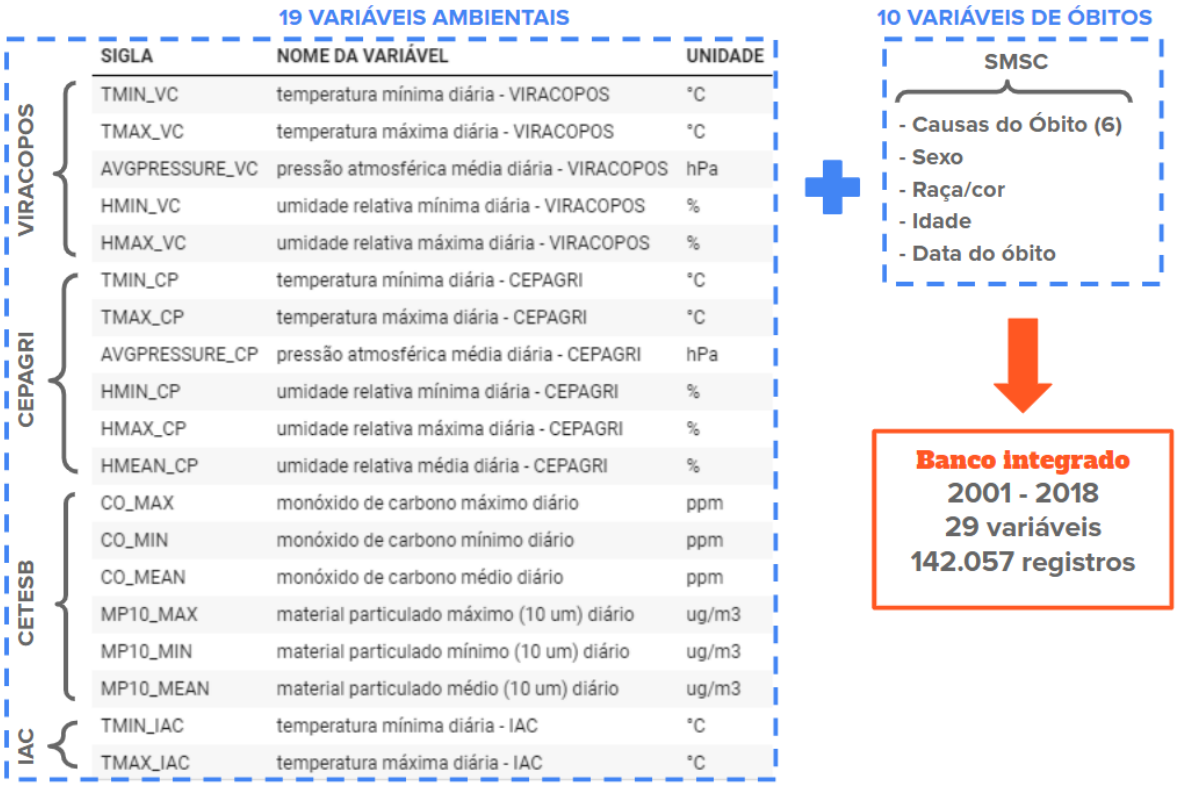
3.2.4 Integração de Bases de Clima e Saúde

Os dados de saúde e ambiente foram integrados em uma única base através das datas em comum, representando uma das grandes contribuições desse trabalho (BELLA *et al.*, 2022). No limite do nosso conhecimento, essa é a primeira base de dados pública que integra dados de parâmetros ambientais a óbitos por todas as causas. A disponibilidade de dados da CETESB foi responsável pelo limite inferior do banco de dados integrado, contendo informações somente a partir de 2001. Em contrapartida, os dados das estações meteorológicas foram responsáveis pelo limite superior, contendo informações até 2018. O banco de dados integrado resultou em 142057 observações e 19 variáveis ambientais preditoras, totalizando 29 variáveis (Figura 3.2 e Tabela 3.1) e é de domínio público com livre acesso para uso (BELLA *et al.*, 2022).

3.3 Seleção e Pré-processamento das Variáveis

Foi realizado um recorte no banco de dados integrado para compor os dados que foram utilizados nos modelos. Selecionou-se os dados referentes aos óbitos que tiveram como causa básica doenças cardiovasculares, identificados em I00 até I99 pelo código do sistema de Classificação Internacional de Doenças (CID-10). Após seleção, os registros de óbitos foram transformados em número de óbitos por dia.

Nos dados de poluição, optou-se por descartar valores máximos e mínimos diários e manter apenas os valores médios, pois esses valores estão mais próximos dos valores que a população está exposta na maior parte do tempo. Em relação às temperaturas, foram adotados as variáveis da estação meteorológica do IAC devido à reconhecida precisão de seus sensores, conforme informado por especialista da área. Os dados de pressão atmosférica média foram provenientes do aeroporto de Viracopos, uma vez que esses sensores apresentaram problemas no período de interesse para na estação do CEPAGRI.



	SIGLA	NOME DA VARIÁVEL	UNIDADE
VIRACOPOS	TMIN_VC	temperatura mínima diária - VIRACOPOS	°C
	TMAX_VC	temperatura máxima diária - VIRACOPOS	°C
	AVGPRESSURE_VC	pressão atmosférica média diária - VIRACOPOS	hPa
	HMIN_VC	umidade relativa mínima diária - VIRACOPOS	%
	HMAX_VC	umidade relativa máxima diária - VIRACOPOS	%
CEPAGRI	TMIN_CP	temperatura mínima diária - CEPAGRI	°C
	TMAX_CP	temperatura máxima diária - CEPAGRI	°C
	AVGPRESSURE_CP	pressão atmosférica média diária - CEPAGRI	hPa
	HMIN_CP	umidade relativa mínima diária - CEPAGRI	%
	HMAX_CP	umidade relativa máxima diária - CEPAGRI	%
	HMEAN_CP	umidade relativa média diária - CEPAGRI	%
CETESB	CO_MAX	monóxido de carbono máximo diário	ppm
	CO_MIN	monóxido de carbono mínimo diário	ppm
	CO_MEAN	monóxido de carbono médio diário	ppm
	MP10_MAX	material particulado máximo (10 um) diário	ug/m3
	MP10_MIN	material particulado mínimo (10 um) diário	ug/m3
	MP10_MEAN	material particulado médio (10 um) diário	ug/m3
IAC	TMIN_IAC	temperatura mínima diária - IAC	°C
	TMAX_IAC	temperatura máxima diária - IAC	°C

Figura 3.2: Representação da composição do banco de dados integrado.

Para umidade, utilizou-se valores médios fornecidos pelo CEPAGRI, pois é a única métrica de umidade projetada pelo modelo climático regional Eta.

As séries históricas também apresentaram dados faltantes, que foram substituídos pelas médias totais de suas respectivas séries de dados (considerando todo o conjunto de dados), uma vez que os modelos não lidam com valores ausentes. A variável CO médio apresentou ausência de 328 medidas, o MP10 médio apresentou ausência de 242 medidas, a umidade relativa média apresentou 55, e a temperatura máxima e a pressão atmosférica apresentaram apenas 1 medida ausente. A temperatura mínima não apresentou medidas ausentes.

Para lidar com o problema da multicolinearidade nos modelos, descartou-se variáveis preditoras correlacionadas entre si que apresentaram fator de inflação de variância (Variance Inflation Factor - VIF, em inglês) $\geq$10 (O'BRIEN, 2007). O VIF é uma métrica utilizada para detectar multicolinearidade entre variáveis preditoras e pode ser expresso matematicamente como:

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2}, \tag{3.1}$$

onde: VIF_i é o fator de inflação de variância da i-ésima variável preditora e R_i^2 é o coeficiente de determinação entre a i-ésima variável preditora e as demais.

Entretanto, procurou-se manter as variáveis temperatura e poluição, devido à

sua importância para o problema (Capítulo 2, Seções 2.1 e 2.2), descartando assim suas contrapartes correlacionadas. Após descarte, restaram as variáveis temperatura mínima (T_MIN), monóxido de carbono médio (CO) e material particulado (10 µm) médio (MP10). As estatísticas descritivas dessas variáveis se encontram no Capítulo 4.

Os dados foram divididos em conjuntos de variáveis resposta (óbitos) e variáveis preditoras (ambientais), com 80% dos dados para treino e 20% para teste (validação), sendo a porção de treino referente ao período de 2001/Jan a 2015/Jun, e a porção de teste de 2015/Jul a 2018/Dez. O desempenho dos modelos é avaliado através do conjunto de dados de teste. Os dados são divididos entre conjunto de treino e teste. Os modelos têm acesso apenas ao conjunto de treino, que é utilizado para ajustá-los durante o processo de treinamento, enquanto que o conjunto de teste (dados reais) é comparado com as predições dos modelos.

Para cada tipo de modelo, utilizou-se as variáveis nas agregações diária, semanal e mensal, apresentando 6574, 939, e 216 observações, respectivamente. A agregação semanal e mensal foi realizada através da soma dos valores diários das variáveis.

3.4 Geração e Avaliação de Modelos

Este estudo utilizou regressão linear com erros SARIMA e rede neural recorrente do tipo LSTM para modelar a relação entre as variáveis ambientais e os óbitos por DCVs. Cada uma dessas abordagens requer configurações intrínsecas definidas a priori. Com o intuito de alcançar os menores erros de predição, adotou-se a técnica de busca em grade (*grid search*, em inglês), que consiste em variar sistematicamente essas configurações intrínsecas em conjunto com diferentes combinações das variáveis preditoras a fim de se determinar a configuração de melhor desempenho segundo critérios previamente estabelecidos.

3.4.1 Regressão Linear com Erros SARIMA

Foi realizado um *grid search* variando a combinação das variáveis preditoras e a quantidade de unidades de atrasos (*lags*) dessas variáveis para cada agregação de dados. A quantidade lags das variáveis preditoras variou de 1 a 15 para a agregação diária, e de 1 a 3 para as agregações semanal e mensal. Para cada lag adicionado, a quantidade de lags já presente no modelo foi mantida. Por exemplo, para determinado modelo que utiliza 3 lags, a previsão no tempo $t + 1$ utilizará os dados de entrada referente às variáveis ambientais no tempo $t - 2$ em conjunto com as mesmas variáveis no tempo $t - 1$ e t.

Foram ajustados diversos modelos com diferentes parâmetros para cada combinação de variáveis e seus lags, conforme ilustrado na Figura 3.3. Os principais parâmetros a serem definidos em um modelo LR-SARIMAX são a quantidade de diferenciações da série temporal e a quantidade de termos autoregressivos e de médias móveis, conforme mencio-

nado no Capítulo 2, Seção 2.3.1 (HYNDMAN; ATHANASOPOULOS, 2018). Dentre os modelos ajustados através do estimador de máxima verossimilhança com otimizador de Broyden-Fletcher-Goldfarb-Shanno (BFGS), foi escolhido o que apresentou menor Critério de Informação de Akaike (*Akaike Information Criterion*- AIC, em inglês) (HYNDMAN; KHANDAKAR, 2008).

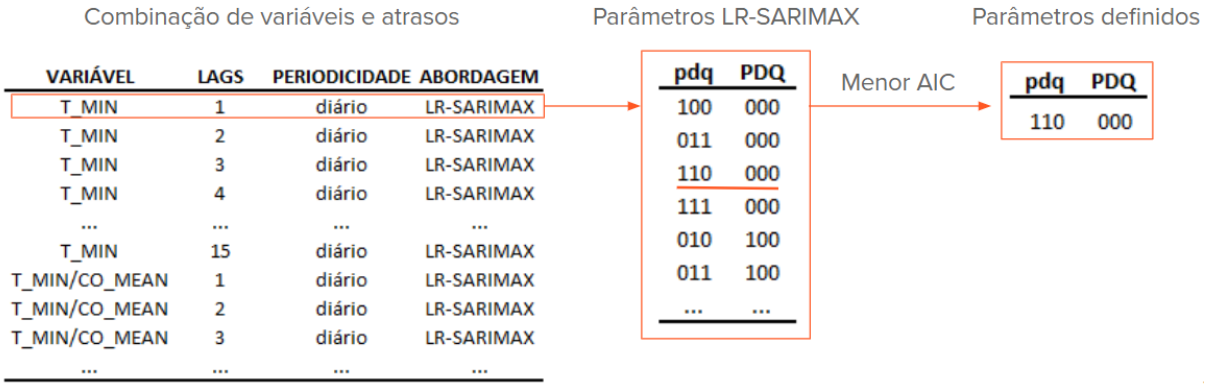


Figura 3.3: Representação do processo de escolha dos parâmetros dos modelos LR-SARIMAX.

3.4.2 Rede Neural LSTM

Para os modelos LSTM, foi adotado o mesmo processo de *grid search* usada no LR-SARIMAX, mas com parâmetros intrínsecos pré-definidos e para duas variações da rede neural. Os principais parâmetros a serem definidos em um modelo LSTM são o otimizador (VANI; RAO, 2019), a função de perda (WANG *et al.*, 2022), o tamanho do *batch* e o número de épocas (BROWNLEE, 2017) necessárias para treinar o modelo. Foi utilizado o erro quadrático médio como função de perda e otimizador do tipo Adam. Foi utilizado um tamanho de *batch* de 300 amostras para agregação de dados diária e semanal, e um limite de 300 épocas. Para frequência mensal, esses parâmetros foram definidos como 5 e 600, respectivamente. Além disso, utilizou-se um método de parada antecipada para evitar sobreajuste (YING, 2019).

Implementou-se o *grid search* em uma rede neural LSTM superficial e uma profunda para cada agregação de dados, referidos daqui para frente como LSTM-shallow e LSTM-deep, respectivamente. Redes LSTM-shallow, também conhecidas como *Vanilla*-LSTM, possuem uma arquitetura simples conforme definida no artigo original que a descreveu pela primeira vez (HOCHREITER; SCHMIDHUBER, 1997; BROWNLEE, 2017). Nesse trabalho, a rede LSTM-shallow foi composta por uma camada de entrada, uma única camada oculta do tipo LSTM com três neurônios e ativação tangente hiperbólica, e uma camada densa (totalmente conectada) de saída (BROWNLEE, 2017). As redes LSTM-deep, também conhecidas como *Stacked*-LSTM, seguem uma arquitetura semelhante,

mas com mais camadas ocultas (BROWNLEE, 2017). Uma rede multicamadas pode ser mais eficiente na aproximação de funções do que uma rede superficial, mas com um custo computacional mais alto (BROWNLEE, 2017; LECUN *et al.*, 2015). Neste trabalho, a rede LSTM-deep foi definida com cinco camadas ocultas de 16, 32, 64, 32 e 16 neurônios, respectivamente (Figura 3.4).

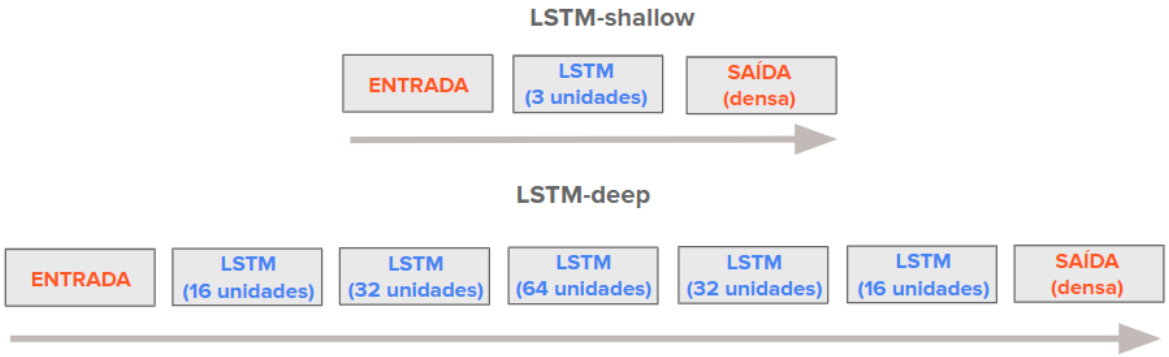


Figura 3.4: Arquiteturas das redes neurais LSTM-shallow e LSTM-deep. A seta em cinza indica o sentido dos dados.

3.4.3 Avaliação dos Modelos

Com base no conjunto de dados de teste, foram avaliadas as predições dos modelos usando as métricas RMSE e MAPE, conforme apresentadas no Capítulo 2, Seção 2.3.3. A métrica RMSE representa o quanto o modelo errou em média na previsão de óbitos na mesma unidade de medida da variável resposta. A métrica MAPE é a diferença média, em porcentagem, entre o valor previsto e o observado. Essas métricas foram utilizadas devido à facilidade de interpretação e amplo uso em aprendizado de máquina.

Como as agregações dos dados estão em diferentes escalas (diário, semanal e mensal), a comparação entre esses grupos pela métrica RMSE levaria a conclusões equivocadas, tornando a métrica MAPE uma melhor escolha neste caso. Portanto, o MAPE foi utilizado para comparar modelos entre diferentes agregações e o RMSE para comparar modelos de mesma agregação.

3.5 Considerações Finais

Nesse capítulo foram apresentados os dados utilizados nesse trabalho, assim como os tratamentos aplicados antes de comporem o banco de dados integrado de ambiente e saúde. Cabe destacar que uma das contribuições desse trabalho foi a consolidação do banco de dados integrado, que foi tornado público e poderá servir para futuros trabalhos e fins educacionais. Também foram apresentadas a seleção e as transformações necessárias

realizadas nos dados antes de serem utilizados como dado de entrada dos modelos. Por fim, foram descritas as configurações dos modelos e a forma com que os dados foram utilizados, assim como o método de avaliação dos desempenhos.

Os scripts python utilizados nas seções 3.3, 3.4 e 3.4.3 e os bancos de dados estão públicos para uso (BELLA; COSTA, 2022; BELLA *et al.*, 2022).

Tabela 3.1: Variáveis do banco de dados integrado. As variáveis ambientais selecionadas para alimentar os modelos estão em negrito.

SIGLA	VARIÁVEL	FONTE	UNIDADE
DATE	data do óbito	SMSC	-
HORAOBITO	hora do óbito	SMSC	horas
IDADE	idade	SMSC	anos
SEXO	sexo	SMSC	-
RACACOR	cor	SMSC	-
LINHAA	causa terminal do óbito	SMSC	CID10
LINHAB	causa precedente do óbito	SMSC	CID10
LINHAC	causa precedente do óbito	SMSC	CID10
LINHAD	causa precedente do óbito	SMSC	CID10
LINHAII	outras causas relacionadas	SMSC	CID10
CAUSABAS	causa básica (desencadeante) do óbito	SMSC	CID10
CO_MAX	monóxido de carbono máximo diário	CETESB	ppm
CO_MIN	monóxido de carbono mínimo diário	CETESB	ppm
CO_MEAN	**monóxido de carbono médio diário**	**CETESB**	**ppm**
MP10_MAX	material particulado 10 µm máximo diário	CETESB	ug/m^3
MP10_MIN	material particulado 10 µm mínimo diário	CETESB	ug/m^3
MP10_MEAN	**material particulado 10 µm médio diário**	**CETESB**	**ug/m^3**
TMIN_IAC	**temperatura mínima diária**	**IAC**	**°C**
TMAX_IAC	temperatura máxima diária	IAC	°C
TMIN_VC	temperatura mínima diária	VIRACOPOS	°C
TMAX_VC	temperatura máxima diária	VIRACOPOS	°C
AVGPRESSURE_VC	pressão atmosférica média diária	VIRACOPOS	hPa
HMIN_VC	umidade mínima diária	VIRACOPOS	%
HMAX_VC	umidade máxima diária	VIRACOPOS	%
TMIN_CP	temperatura mínima diária	CEPAGRI	°C
TMAX_CP	temperatura máxima diária	CEPAGRI	°C
AVGPRESSURE_CP	pressão atmosférica média diária	CEPAGRI	hPa
HMIN_CP	umidade mínima diária	CEPAGRI	%
HMAX_CP	umidade máxima diária	CEPAGRI	%

Capítulo 4

Resultados

O presente capítulo apresenta e discute os resultados da aplicação da metodologia descrita no Capítulo 3. O capítulo está dividido em quatro seções. A Seção 4.1 apresenta uma análise descritiva com o intuito de contextualizar e compreender as relações entre as variáveis utilizadas nesse trabalho. A Seção 4.2 apresenta os resultados da otimização dos parâmetros, assim como as predições de menor erro em cada agregação de dados. A avaliação da robustez comparando as diferentes abordagens está descrita na Seção 4.3. Por fim, a Seção 4.4 traz as considerações finais do capítulo.

4.1 Análise Descritiva das Variáveis

As variáveis no período de 2001 a 2018 utilizadas nesse trabalho são: contagem de óbitos por doenças cardiovasculares (DCVs), temperatura mínima, monóxido de carbono médio e material particulado médio diários (Figura 4.1). Na Tabela 4.1, essas variáveis estão sumarizadas com os dados agregados semanalmente e mensalmente. A Figura 4.1 apresenta o comportamento geral das variáveis ao longo do tempo. O teste de análise de tendência de Mann-Kendall (BARI *et al.*, 2016; MANN, 1945) mostra que a variável óbito exibe tendência de aumento, enquanto as variáveis monóxido de carbono médio e material particulado médio apresentam tendências de diminuição ($p < 5\%$). A Figura 4.2 mostra a média móvel diária dessas variáveis agrupadas em períodos de dois anos e corrobora as tendências encontradas pelo teste de Mann-Kendall. Destaca-se ainda que o aumento de óbitos por doenças cardiovasculares no período (14,11%) foi superior ao aumento populacional de Campinas (10,55%), evidenciando um aumento efetivo no número de óbitos por DCVs na cidade (IBGE, 2012).

A temperatura mínima, por sua vez, foi a única variável que não exibiu clara tendência de aumento ou diminuição ($p = 0,31$). Entretanto, é possível observar ciclos de sazonalidade da temperatura mínima, associados a dias mais quentes entre dezembro e março (verão) e dias mais frios entre junho e setembro (inverno), como apresentado na Figura 4.3. Os óbitos por DCVs também apresentam sazonalidade, com picos ocorrendo

nos períodos mais frios, como ilustrado na Figura 4.3. Também é possível perceber que na maioria dos dias de verão (faixas vermelho claro) ocorrem menos de dez óbitos, enquanto no inverno (faixas azul ciano) esse limiar é ultrapassado com maior frequência.

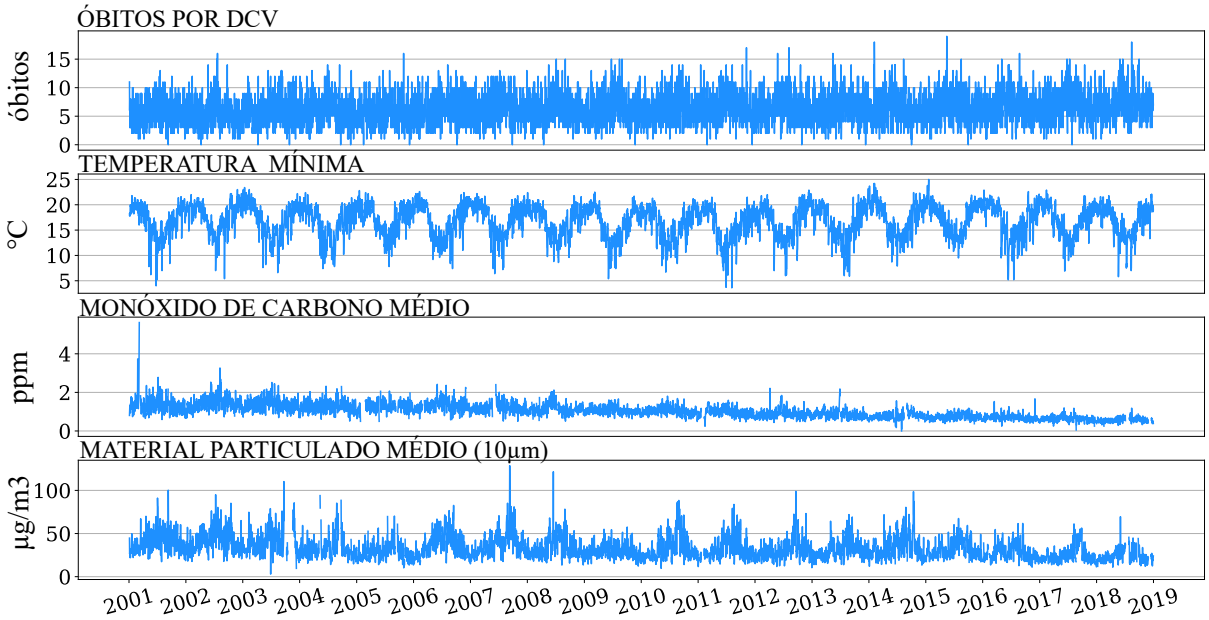


Figura 4.1: Série temporal diária das variáveis no período entre 2001 a 2018.

Tabela 4.1: Estatística descritiva dos dados agrupados por dia, semana e mês no período de 2001 a 2018 para todas as variáveis utilizadas. ÓBITOS: óbitos por doenças cardiovasculares, T_MIN: Temperatura mínima, CO: Monóxido de carbono médio e MP10: Material particulado (10 µm) médio. Os dados semanais e mensais foram agregados por soma a partir dos dados diários.

		ÓBITOS	T_MIN(°C)	CO(ppm)	MP10(µg/m^3)
Diário	média	6,3	16,6	1,1	33,3
	desvio padrão	2,7	3,3	0,4	12,7
	mínimo	0,0	3,6	0,0	3,0
	máximo	19,0	25,0	5,6	128,7
Semanal	média	44,9	116,4	7,0	224,0
	desvio padrão	8,8	20,9	2,5	77,4
	mínimo	11,0	18,2	0,0	0,0
	máximo	81,0	157,9	20,2	560,1
Mensal	média	191,4	506,4	30,6	974,9
	desvio padrão	26,4	81,0	9,6	276,4
	mínimo	139,0	301,0	4,5	173,0
	máximo	286,0	649,0	54,5	1805,7

4.2 Modelos de Predição com Parâmetros Otimizados

Nesta seção são sumarizados e discutidos os resultados obtidos aplicando-se a técnica de otimização de parâmetros via *grid search*. Em outras palavras, os modelos com melhor desempenho foram encontrados a partir da combinação de diferentes agregações

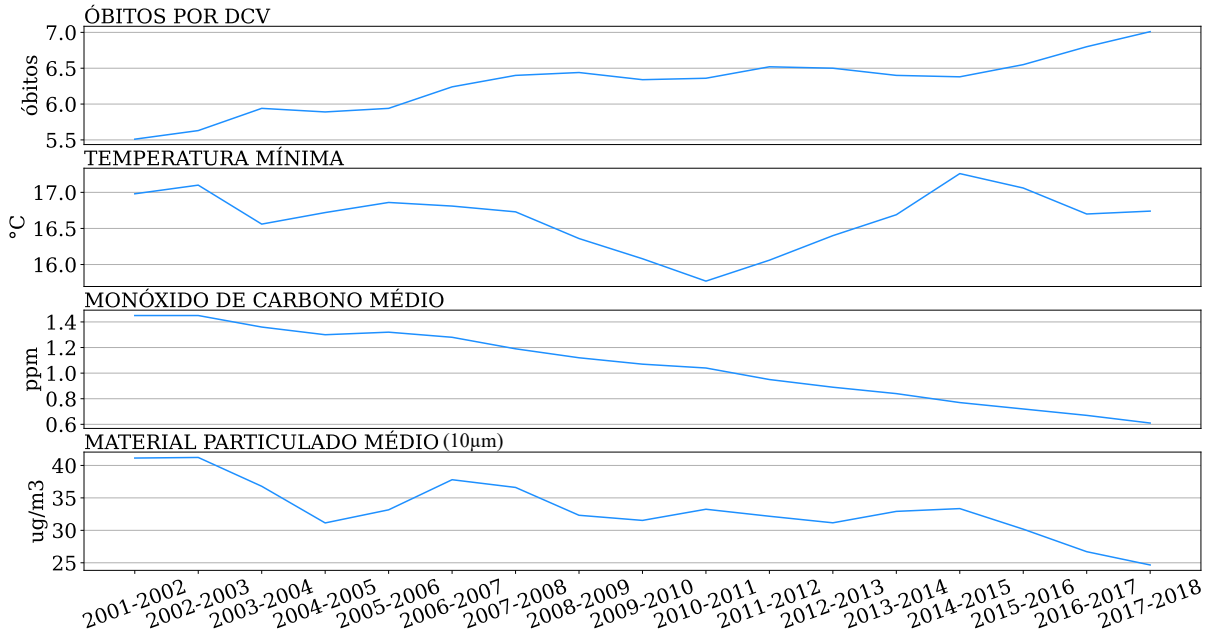


Figura 4.2: Média móvel diária das variáveis a cada dois anos entre 2001 e 2018.

de dados (diário, semanal e mensal), conjuntos de variáveis, parâmetros associados aos atrasos considerados (*lags*) e abordagens de modelagem de série temporal. Ao todo foram avaliados 441 modelos através das métricas RMSE e MAPE. A Tabela 4.2 apresenta os 15 modelos que obtiveram melhor desempenho de predição considerando o conjunto de dados de teste (validação), segundo a métrica RMSE. As seções seguintes discutem os principais achados derivados dessa tabela.

4.2.1 Erro de Predição e Periodicidade da Contagem de Óbitos

A Tabela 4.2 evidencia que os modelos que apresentaram menor erro de predição são os modelos com periodicidade mensal da contagem de óbitos. É possível observar a consistência desse resultado quando avaliamos os erros MAPE estratificados por agregação, que resultou em média e desvio padrão de 8,4±1,4, 14,5±1,1 e 43,2±1,6 (óbitos) para os dados mensais, semanais e diários, respectivamente. A Figura 4.4 complementa essa informação, mostrando as predições de menor erro em cada agregação de dados (diária, semanal ou mensal).

Ainda, é possível observar que os melhores modelos em cada periodicidade de dados apresentaram desempenho superior a um preditor *naive*, i.e., um preditor que prevê apenas a média histórica de óbitos a cada ponto da série temporal. As métricas RMSE e MAPE dos melhores modelos em cada periodicidade de dados, e sua respectiva comparação com um preditor *naive*, estão apresentados na Tabela 4.3.

Entende-se que a modelagem de predição mensal seja favorável para o desenvolvimento de sistemas de alerta antecipado que possibilitem a adequação da infraestrutura de atendimento e a adequada conscientização da população, tanto pelo magnitude do

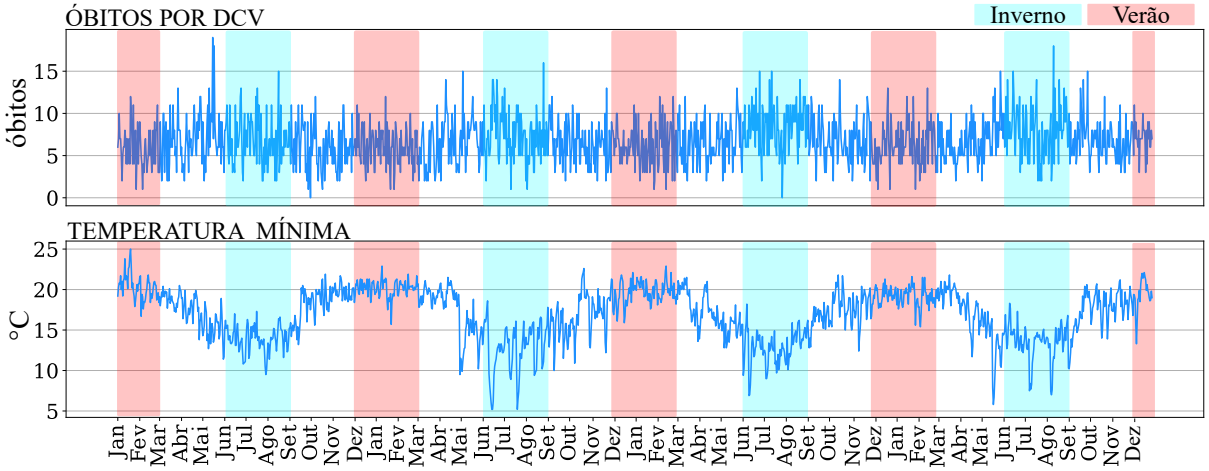


Figura 4.3: Recorte dos óbitos por DCVs e temperatura mínima diários no período de janeiro de 2015 a dezembro de 2018. Observe que ambas as variáveis apresentam variações sazonais e flutuam em direções opostas.

erro apresentada pelo modelo mensal, quanto pela suficiência de tempo na tomada de decisão (UNISDR, 2006).

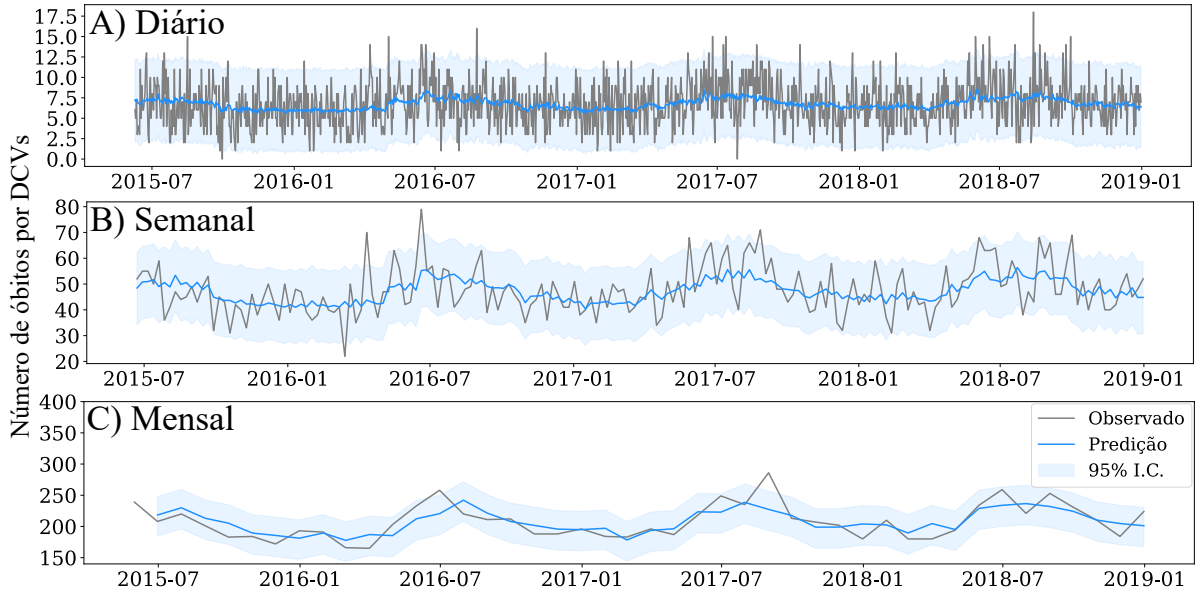


Figura 4.4: Predição dos dados de teste dos modelos de menor erro em cada agregação de dados. A faixa em azul claro representa o intervalo de confiança de 95%. É possível observar que o modelo com dados mensais (C) é o que melhor acompanha a variação dos dados observados (em cinza).

4.2.2 Importância das Variáveis de Entrada na Modelagem

A temperatura mínima se mostrou como uma variável de entrada relevante para as predições. Essa variável estava presente em 12 dos 15 melhores modelos, sendo que três desses modelos utilizam apenas a temperatura mínima como variável de entrada, variando apenas a quantidade de atrasos das variáveis preditoras (Tabela 4.2). Essa relevância

Tabela 4.2: Melhores 15 modelos ordenados de forma crescente com base no RMSE. T_MIN: Temperatura mínima, CO: Monóxido de carbono médio e MP10: Material particulado (10 µm) médio. Os modelos de menor erro em cada abordagem estão destacados em negrito.

MODELO	AGREGAÇÃO	VARIÁVEIS	ATRASOS	RMSE
LR-SARIMAX	**mensal**	**T_MIN/CO/MP10**	**1**	**17,6**
LR-SARIMAX	mensal	T_MIN	2	18,8
LR-SARIMAX	mensal	T_MIN/MP10	1	19,0
LR-SARIMAX	mensal	T_MIN	1	19,2
LR-SARIMAX	mensal	T_MIN/CO	1	19,4
LR-SARIMAX	mensal	MP10	2	19,6
LR-SARIMAX	mensal	T_MIN/CO/MP10	2	19,7
LR-SARIMAX	mensal	T_MIN/CO	3	19,8
LR-SARIMAX	mensal	T_MIN	3	19,9
LR-SARIMAX	mensal	T_MIN/MP10	3	19,9
LR-SARIMAX	mensal	T_MIN/CO	2	20,3
LSTM-deep	**mensal**	**T_MIN/CO/MP10**	**3**	**20,4**
LR-SARIMAX	mensal	MP10/ CO	3	20,4
LSTM-shallow	mensal	T_MIN/CO/MP10	3	20,7
LR-SARIMAX	mensal	MP10	1	20,7

Tabela 4.3: Métricas de erro de um preditor *naive* em comparação com os melhores modelos em cada periodicidade de dados.

	Diário		Semanal		Mensal	
	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
Naive	2,7	43,4	9,3	14,4	25,8	9,7
LR-SARIMAX	2,6	43,1	7,8	13,4	17,4	6,4

também ficou evidente nos modelos que utilizaram apenas uma variável por vez. Nesses modelos, o uso da temperatura mínima como variável de entrada retornou erros menores que os modelos que utilizaram somente monóxido de carbono ou material particulado (Figura 4.5). A importância da variável temperatura na predição de óbitos por DCVs observada neste trabalho corrobora os achados de diversos estudos que avaliam o impacto das variáveis ambientais nas mortes por doenças cardiovasculares, conforme apresentado no Capítulo 2.

4.2.3 Atrasos das Variáveis de Entrada

O número de unidades de atraso (*lags*) das variáveis preditoras utilizadas como dados de entrada que resultou em menores erros variou entre os modelos testados. Nos dados diários, o LR-SARIMAX necessitou de pelo menos cinco *lags* para alcançar o mesmo

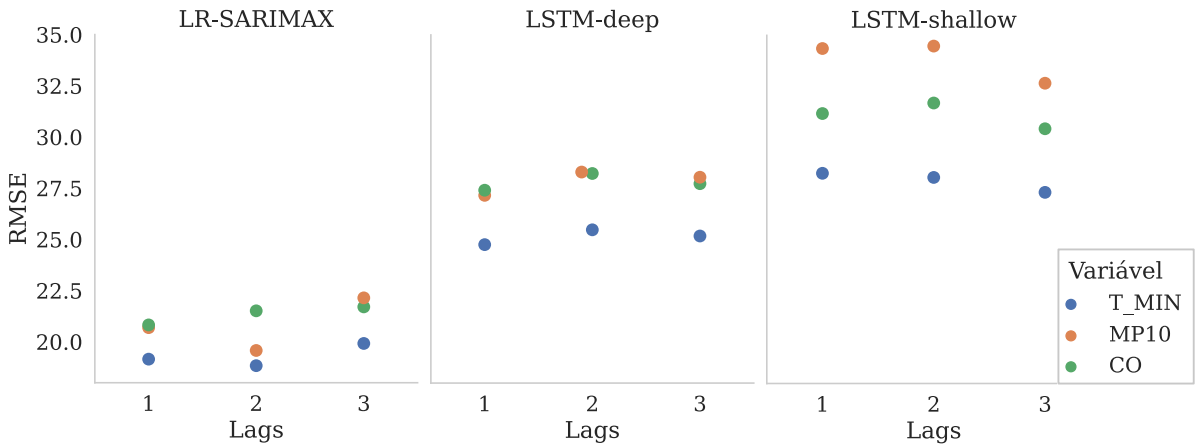


Figura 4.5: Raiz do erro quadrático médio (RMSE) em modelos com dados agregados por mês que utilizaram apenas uma variável de entrada durante o *grid search*. T_MIN: Temperatura mínima, CO: Monóxido de carbono médio e MP10: Material particulado (10 µm) médio.

nível de erro que os modelos LSTM, como exemplificado na Figura 4.6, na qual estão representados os melhores modelos por abordagem (legenda) utilizando dados diários, e mostrando a raiz do erro quadrático médio (eixo y) para cada quantidade de atrasos (eixo x) das variáveis preditoras (legenda). Nos dados semanais, no geral, a maior quantidade de atrasos, i.e., três semanas, resultou em menores erros. Para os dados mensais, a quantidade de atrasos que resultou em menores erros variou entre os modelos. Não obstante, os 5 melhores modelos utilizaram, principalmente, apenas um atraso da variável preditora (Tabela 4.2).

Quando comparamos os modelos de menor erro em cada agregação de dados, observamos que as quantidades de atraso das variáveis preditoras foram similares. Nos dados diários, os cinco melhores modelos utilizaram em média 13,4 atrasos das variáveis preditoras, enquanto os modelos com dados semanais utilizaram em média 2,6 semanas (∼18 dias) de atrasos. Os modelos com dados mensais, no geral, utilizaram um único mês de atraso (Tabela 4.2), que é a quantidade mínima considerada. Essa faixa de valores das quantidade de atraso é coerente com resultados da literatura, que reportam temperaturas baixas exercendo efeitos que duram cerca de 20 dias (SILVEIRA *et al.*, 2019; ANDERSON; BELL, 2009; LIN *et al.*, 2020). Ao utilizar essa quantidade de atrasos, os efeitos das altas temperaturas também são considerados, pois exercem efeito por aproximadamente 5 dias (SILVEIRA *et al.*, 2019; HUANG *et al.*, 2017; LIN *et al.*, 2020).

Para os modelos LSTM, a quantidade de atrasos não foi um parâmetro relevante (Figura 4.6), especialmente para os dados diários, evidenciando a capacidade de memória desse tipo de abordagem (BROWNLEE, 2017).

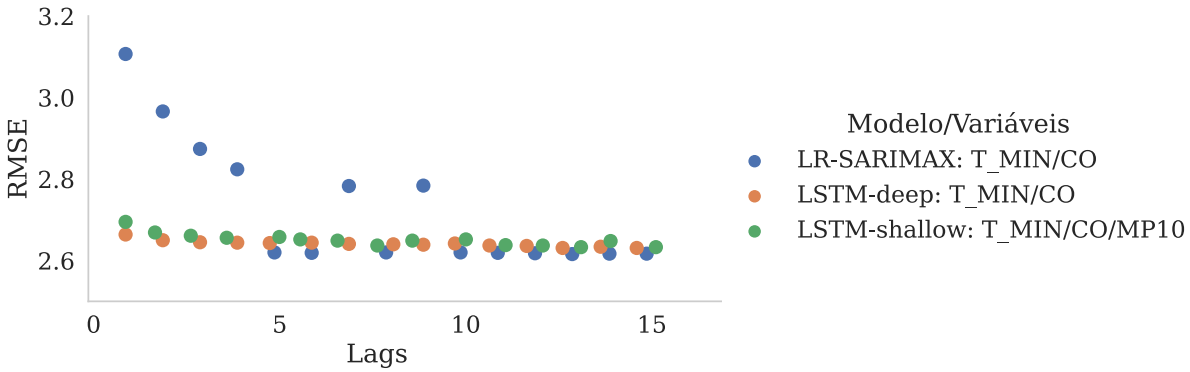


Figura 4.6: Raiz do erro quadrático médio (RMSE) nos modelos de menor erro, para cada abordagem, para dados diários. T_MIN: Temperatura mínima, CO: Monóxido de carbono médio e MP10: Material particulado (10 µm) médio.

4.2.4 Desempenho das Abordagens LR-SARIMAX e LSTM

De uma forma geral, os modelos que utilizaram o LR-SARIMAX apresentaram melhor desempenho preditivo que os modelos baseados na abordagem LSTM. Considerando-se dados diários, as diferentes abordagens apresentaram erros similares. Entretanto, como mencionado na Seção 4.2.3, os modelos LR-SARIMAX necessitaram de pelo menos cinco atrasos das variáveis preditoras para alcançar os mesmos erros que os modelos LSTM (Figura 4.6). Nos dados semanais e mensais, os modelos que utilizaram o LR-SARIMAX apresentaram, em geral, os melhores desempenhos, seguido pelas variações LSTM-deep e LSTM-shallow. Considerando todos os modelos implementados, a abordagem utilizando regressão linear (LR-SARIMAX) foi a que apresentou os melhores desempenhos, como evidenciado pela Tabela 4.2.

4.2.5 Análise das Predições

Os melhores modelos diários, semanais e mensais foram capazes de capturar as variações sazonais com um intervalo de confiança que não ultrapassou os valores observados (Figura 4.4). Entretanto, os modelos com dados diário e semanal (A e B) não foram capazes de prever os picos e os vales da série temporal. O melhor modelo encontrado no *grid search* foi o LR-SARIMAX utilizando todas as variáveis com dados mensais com apenas um atraso (Figura 4.4-C), ou seja, variáveis preditoras com dados do mês atual para prever óbitos por doenças cardiovasculares no mês seguinte. Esse modelo apresentou erro RMSE de 17,6 óbitos e erros MAPE de 6,4%.

A Figura 4.7 mostra o quanto o modelo errou ponto a ponto nas previsões utilizando dados mensais e reflete os valores utilizados para calcular o erro RMSE. É possível observar que a maior parte das predições nos dados de teste (22 pontos) previu mais óbitos que os valores observados, em contraste às 15 predições subestimadas. Visando a aplicação do modelo como um sistema de alerta antecipado, as predições em que o

modelo subestimou o número de óbitos são consideradas indesejáveis. Uma modelagem mais adequada dos picos e vales da série temporal caracteriza um importante aspecto a ser tratado em trabalhos futuros.

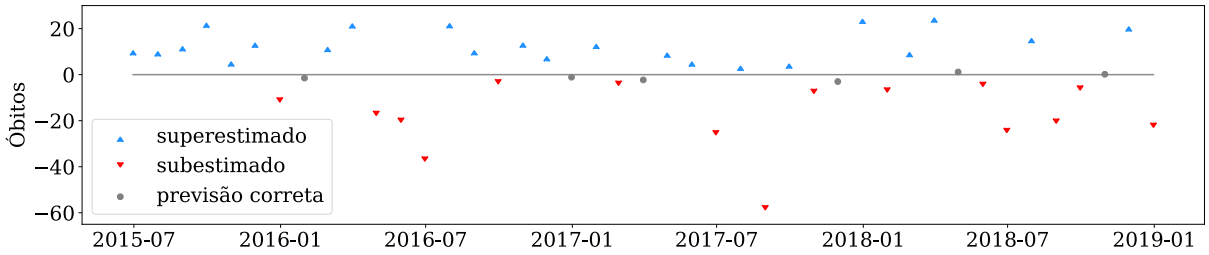


Figura 4.7: Diferença entre os valores previstos e observados. A linha em cinza faz referência ao ponto onde a previsão é isenta de erro. Os pontos em cinza são as predições isentas de erro. Os triângulos em azul são as predições superestimadas. Os triângulos invertidos em vermelho são as predições subestimadas.

4.2.6 Sistema de Alerta Antecipado

Um sistema de alerta antecipado efetivo é composto por quatro elementos principais: conhecimento do risco, monitoramento e serviço de alerta, disseminação e comunicação, e capacidade de resposta (UNISDR, 2006). A Figura 4.8 apresenta um fluxograma de uma possível proposta de um sistema de alerta antecipado associada ao presente trabalho.

O *conhecimento do risco* envolve a coleta e análise dos dados, considerando a relação entre as variáveis em questão, conforme apresentado neste capítulo e nos Capítulos 2 e 3. O *monitoramento e serviço de alerta* se trata do núcleo de um sistema de alerta antecipado e envolve o monitoramento das variáveis relacionadas ao risco e a geração de predições precisas, como é o caso do modelo LR-SARIMAX mensal baseado nas variáveis de temperatura e poluição.

A *disseminação e comunicação*, por sua vez, refere-se a etapa de levar a predição, transformada em uma mensagem clara e simples, àqueles que estão sob risco e são os maiores interessados. Para alcançar esses público de forma efetiva, essa etapa deve ser feita através de diversos canais de comunicação como mensagens SMS, transmissão em telejornais locais e conscientização nas unidades básicas de saúde (UBS) (UNISDR, 2006).

Por fim, a *capacidade de resposta* diz respeito ao nível de preparo dos envolvidos para lidar com um evento de ameaça a saúde a partir do alerta. Nessa etapa, os programas de conscientização para a população e o treinamento dos profissionais de saúde desempenham um papel importante (UNISDR, 2006).

Na Figura 4.8, este trabalho não se preocupou em desenvolver sistemas de captação de dados e armazenamento contínuos e apenas utilizou informações de séries históricas pré-existentes. A partir dos dados destas séries históricas, o foco do presente trabalho foi a exploração de diferentes modelagens de predição e, dentre elas, a identificação

da mais adequada para a implementação do núcleo do serviço de alerta. Finalmente, o trabalho também não aborda as etapas de disseminação e comunicação do sistema de alerta.

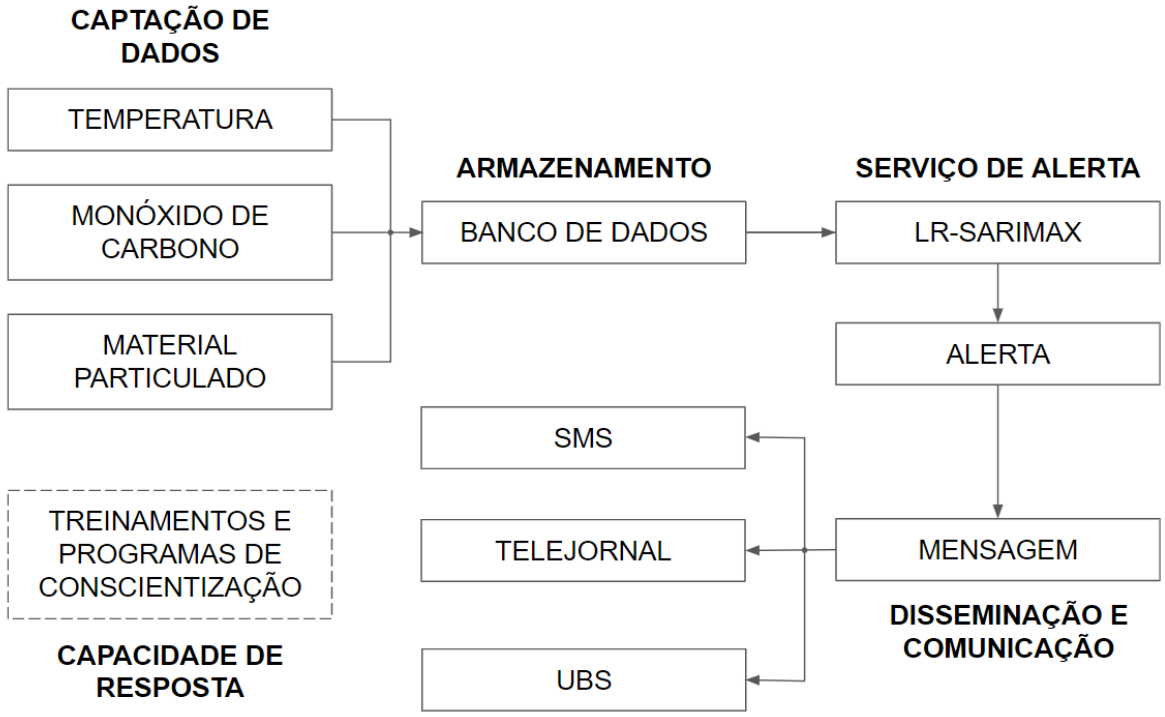


Figura 4.8: Fluxograma dos elementos de um sistema de alerta antecipado.

4.3 Validação Cruzada

Para avaliar a robustez dos resultados da predição e comparar os desempenhos do LR-SARIMAX e LSTM, foi realizada uma validação cruzada de série temporal no melhor modelo em cada abordagem (destacados em negrito na Tabela 4.2). Como a validação cruzada de série temporal não pode ser aleatorizada, pois alteraria a estrutura temporal dos dados, utilizou-se a técnica de janela deslizante em todas as variáveis, onde o tamanho amostral foi reduzido (de 216 para 171 amostras) com o intuito de criar um espaço para a movimentação da janela. A partir da primeira janela (observações 0 a 171), avançou-se 5 passos temporais por vez, resultando em um total de 10 amostras, conforme exemplificado na Figura 4.9. A divisão dos dados seguiu a mesma proporção adotada no *grid search*, sendo 80% para treino e 20% para teste. Para cada um dos 10 momentos da janela deslizante, em cada um dos modelos, obteve-se o RMSE para os dados de teste.

Na Figura 4.10 os pontos pretos representam a distribuição dos erros RMSE obtidos para cada uma das 10 janelas de teste, enquanto os boxplots que os envolvem destacam os quartis relevantes. O gráfico permite observar que o modelo LR-SARIMAX apresenta menor variabilidade dos erros. Esse modelo apresentou um RMSE médio de 18,1

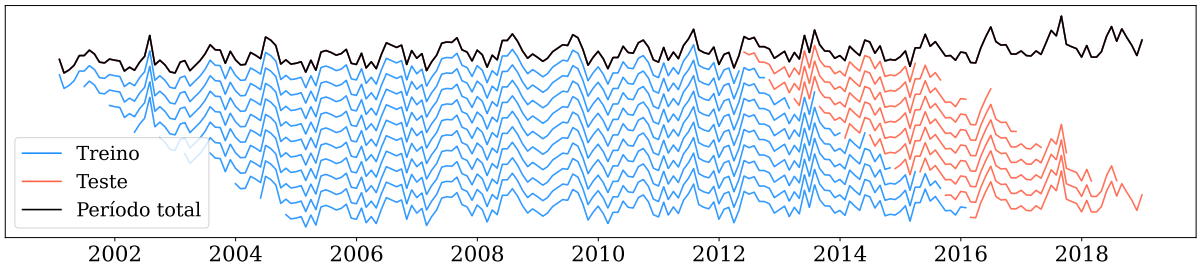


Figura 4.9: Divisão dos dados aplicado na validação cruzada. A série temporal completa (2001-2018) que serviu como base está representada na parte superior em preto. As porções de treino da janela deslizante estão representadas em azul e as porções de teste em vermelho. Cada uma das 10 linhas abaixo da série completa (em preto) representa um momento na janela deslizante em que os modelos foram treinados e testados. A janela deslizante foi adotada em todas as variáveis utilizadas.

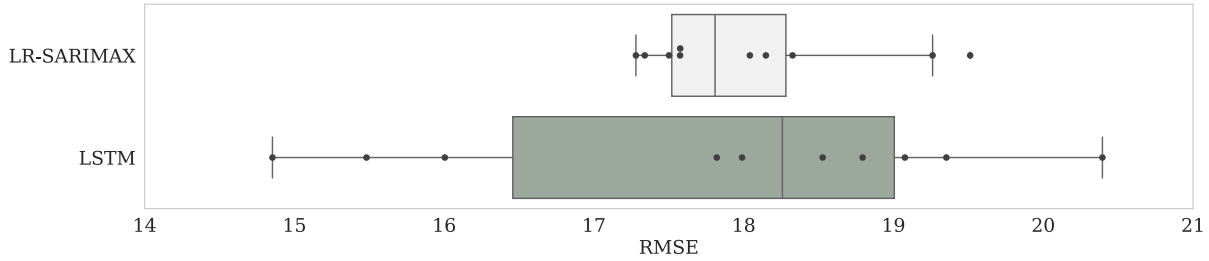


Figura 4.10: Boxplots representando os erros nos dados de teste nos dois melhores modelos em cada abordagem. O boxplot em cinza claro representa o diagrama do LR-SARIMAX e em cinza escuro do LSTM. Os pontos em preto representam os erros nos dados de teste para cada um dos 10 momentos da janela deslizante.

e desvio padrão de 0,8 na validação cruzada. O modelo LSTM, por sua vez, apresentou um erro médio de 17,8, mas com mais que o dobro de variação, exibindo desvio padrão de 1,8. Apesar da diferença na variação dos erros, não houve diferença estatisticamente significativa entre as duas abordagens de acordo com o teste de hipótese U de Mann-Whitney ($p > 5\%$).

4.4 Considerações Finais

Esse capítulo apresentou e discutiu os resultados das análises realizadas nos dados e da otimização dos modelos. A variável temperatura apresentou uma sazonalidade bem marcada, mas foi a única variável que não apresentou tendência, enquanto as outras variáveis apresentaram tendência de aumento (óbitos) ou diminuição (poluição). A variável temperatura também se mostrou aquela de maior relevância para as predições. Os óbitos também apresentaram sazonalidade, com os picos ocorrendo nos dias mais frios. Os melhores modelos exibiram similaridade no número de *lags* das variáveis preditoras, apresentando entre 13 a 18 dias, conforme encontrado na literatura.

No contexto das análises realizadas, o modelo LR-SARIMAX se destacou pelos menores erros de predição e maior estabilidade (evidenciada pela menor variabilidade

dos erros na comparação com a abordagem LSTM). Um argumento adicional a favor da abordagem LR-SARIMAX baseia-se na sua interpretabilidade intrínseca, que possibilita estudar os resultados do modelo a partir dos valores de seus coeficientes. No entanto, no contexto dos dados utilizados neste trabalho, os coeficientes das variáveis de poluição (monóxido de carbono e material particulado) resultaram em correlação fraca com a variável resposta (regressores com $p > 0{,}05$). Apesar deste resultado não inviabilizar o uso do modelo obtido como modelo de predição, não é adequado realizar análises de mecanismos de causalidade utilizando-se o modelo obtido.

Finalmente, destaca-se que a agregação mensal de dados resultou em menores erros em comparação com as outras agregações. O fato de o melhor modelo realizar predições com um mês de antecedência é interessante para a implementação de sistemas de alerta para tomadores de decisões traçarem estratégias de adaptação (CONLON *et al.*, 2011). Especificamente para a região estudada, é possível utilizar as predições em conjunto com o Programa Saúde da Família, que integra ações de saúde pública com cuidados e tratamentos, mas foca em práticas preventivas, como visitas domiciliares e educação sanitária (PERES *et al.*, 2006). Nesse sentido, seria possível alertar a população, especialmente os mais vulneráveis, quando as predições apontarem picos de mortalidade, oferecendo também o suporte a essas famílias.

Capítulo 5

Estudo Exploratório do Impacto das Mudanças Climáticas nos Óbitos por Doenças Cardiovasculares em Campinas

No Capítulo 4, foram avaliados os desempenhos de modelos de predição de contagem de óbitos (diária, semanal ou mensal) por DCVs, na cidade de Campinas, em função de variáveis ambientais. Em particular, mostrou-se que a modelagem LR-SARIMAX, alimentada por dados mensais, resultou nos menores erros de predição (vide Seção 4.2, Tabela 4.2).

Por meio de um estudo exploratório, o presente capítulo busca evidenciar como modelos de predição são ferramentas importantes para a melhor compreensão dos impactos das mudanças climáticas na saúde humana, contribuindo para o desenvolvimento de estratégias de adaptação. A própria definição de adaptação, no contexto das mudanças climáticas, já engloba a ideia do uso de modelos de previsão na tomada de ações. O IPCC (*Intergovernmental Panel on Climate Change*, em inglês) define adaptação como o processo de ajuste da população à mudanças atuais ou esperadas do ambiente, no sentido de se preparar para possíveis efeitos deletérios ou, então, se aproveitar de situações benéficas (IPCC, 2022). Ainda, os modelos de previsão podem ser usados como ferramentas para auxiliar tomadores de decisão e as partes interessadas a escolherem os melhores caminhos a seguir (World Health Organization, 2003).

Outro aspecto relevante desse estudo se deve ao fato de a predição ser de alta resolução, sendo personalizada para a cidade de Campinas. Até 2050, os riscos associados às mudanças climáticas estão mais relacionados à vulnerabilidade e exposição da população do que nas diferenças entre os cenários de mudanças do clima (IPCC, 2022). Isso reforça ainda mais a necessidade de estudos regionalizados, uma vez que diferentes regiões podem

responder de forma diferente às mudanças do clima (IPCC, 2022).

O estudo proposto avalia o impacto de diferentes cenários de mudanças climáticas na contagem de óbitos por doenças cardiovasculares em Campinas, até 2050. Para isso, utilizamos como dados de entrada os valores de temperatura mínima gerados a partir de modelos regionais de projeção climática futura, ajustados para diferentes cenários de mudanças climáticas padronizados pelo IPCC. Tais dados de temperaturas mínimas são então submetidos ao modelo de predição LR-SARIMAX mensal correspondente, que fornece uma projeção futura da contagem de óbitos mensais na cidade de Campinas, até 2050, em diferentes cenários de mudanças climáticas.

O estudo conduzido ilustra como as abordagens de modelagem adotadas neste trabalho podem ser aplicadas para monitorar e avaliar o impacto dos diferentes cenários das mudanças climáticas na saúde, caracterizando uma importante ferramenta de tomada de decisão antecipada por gestores da área da saúde.

Este capítulo está dividido em cinco seções. A Seção 5.1 faz uma introdução sobre os modelos climáticos e os cenários de mudanças climáticas. A Seção 5.2 descreve a metodologia adotada nesse estudo exploratório. A Seção 5.3 apresenta uma análise descritiva dos dados de temperaturas observadas e previstas sob cenários de mudanças climáticas. A Seção 5.4 apresenta as projeções de óbitos por DCVs até 2050. Por fim, a Seção 5.5 traz as considerações finais deste estudo.

Cabe aqui definir alguns termos que são amplamente utilizados no capítulo:

- **Dados observados**: Dados reais mensurados, sejam eles de temperatura ou contagem de óbitos;
- **Previsão**: Antecipação de determinado fenômeno, podendo ser uma predição ou uma projeção.
- **Predição**: Previsão baseada em dados observados;
- **Projeção**: Previsão baseada em valores simulados, como, por exemplo, a *projeção dos óbitos até 2050*, onde realiza-se uma previsão com base em temperaturas esperadas sob cenários de mudanças climáticas.

5.1 Modelos de Projeção do Clima

A principal abordagem de projeção do clima sob cenários de mudanças climáticas é realizada através de modelos climáticos globais (*Global Climate Models*, GCM, em inglês) (CHOU *et al.*, 2014b).

Os GCMs são modelos de simulação que realizam a projeção do clima a partir de cálculos que consideram os princípios físicos gerais da dinâmica dos fluidos e termodinâmica e descrevem as interações globais entre atmosfera, oceano e superfície terrestre (JUNIOR

et al., 2016). Esses modelos são robustos, no sentido de considerarem diversas variáveis e processos do sistema terrestres. De uma forma geral, um GCM é constituído por componentes dinâmicos, como, por exemplo, o efeito da rotação da terra e o transporte vertical da umidade do ar; componentes físicos, como a formação de nuvens e a produção de chuva; e variáveis que descrevem o estado da atmosfera, como temperatura, umidade e pressão (JUNIOR *et al.*, 2016).

As simulações dos GCMs demandam grande poder computacional e uma estratégia típica para a resolução de equações complexas, as quais consideram inúmeras variáveis, é a diminuição da resolução espacial. Nessa abordagem, grandes áreas de 100 a 300 quilômetros quadrados são tratadas de maneira uniforme, desconsiderando-se eventuais diferenças microclimáticas daquela região (JUNIOR *et al.*, 2016; CHOU *et al.*, 2014c; CHOU *et al.*, 2014b).

Uma abordagem de maior interesse para escalas menores são os modelos climáticos regionais (*Regional Climate Models*, RCM, em inglês). Esses modelos baseiam-se nos modelos globais para realizar a redução de escala para uma resolução de dezenas de quilômetros. Os RCMs oferecem resolução menor a partir de dados da região de interesse em conjunto com informações dos GCMs, podendo estar aninhados a um ou mais modelos globais (JUNIOR *et al.*, 2016). Além disso, os RCMs podem ser forçados também por cenários de mudanças climáticas como os Caminhos Representativos de Concentração.

Os Caminhos Representativos de Concentração (*Representative Concentration Pathways*, RCPs, em inglês), caracterizam diferentes projeções de concentração de gases de efeito estufa, descritos em relatórios emitidos pelo IPCC. Esses cenários são identificados de acordo com sua respectiva forçante radiativa, ou seja, o balanço entre a radiação solar absorvida e radiada de volta para o espaço. Quanto maior a forçante radiativa, maior o desbalanço no sentido de aquecimento global.

O quinto relatório do IPCC trata de quatro cenários RCPs: um cenário otimista (RCP2,6), onde o desbalanço é fixado em 2,6 W/m^2 com subsequente declínio após o ano 2100; dois cenários intermediários (RCPs 4,5 e 6,0), nos quais o desbalanço se estabiliza em 4,5 W/m^2 ou 6,0 W/m^2 após 2100; e um cenário de altas emissões (RCP8.5), em que a forçante ultrapassa 8,5 W/m^2 após 2100 e ainda continua a aumentar por alguns anos (PACHAURI *et al.*, 2014). Em termos de temperatura, o cenário RCP 4,5 prevê um aumento da temperatura média de superfície global entre 1,1°C a 2,6°C, e o cenário mais pessimista, RCP 8,5, prevê um aumento entre 2,6°C e 4,8°C até o fim do século XXI, conforme apresentado na Figura 5.1 (PACHAURI *et al.*, 2014). Esses aumentos utilizam como referência o período de 1986 a 2005.

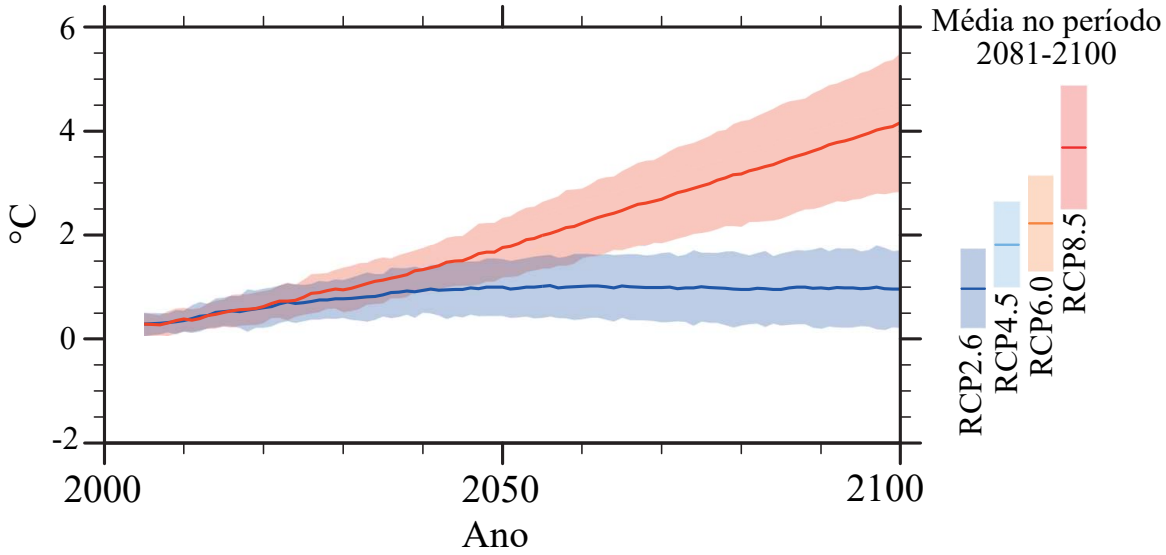


Figura 5.1: Mudança de temperatura média global da superfície do planeta até 2100. A mudança é referente ao período entre 1986 a 2005. No centro é destacada a mudança de temperatura anualmente para os cenários RCP2,6 e RCP8,5. No canto direito são mostradas as médias de mudança de temperatura no período entre 2081 a 2100 para todos os cenários. Adaptado de (PACHAURI *et al.*, 2014)

5.2 Desenho do Estudo

O presente estudo visou avaliar como as projeções do clima sob diferentes cenários de mudanças climáticas afetam a contagem de óbitos por DCVs na cidade de Campinas. Considerando-se que o modelo climático não faz projeções de dados de poluição, utilizou-se o modelo mensal LR-SARIMAX baseado em dados de entrada de temperatura mínima (segunda linha da Tabela 4.2). Dados de projeção de temperatura mínima até o ano de 2050 foram submetidos ao modelo, considerando-se dois cenários de mudanças climáticas: RCP4,5 (cenário intermediário) e RCP8,5 (cenário de altas emissões), conforme definições apresentadas na Seção 5.1.

Os dados de projeções de temperatura utilizados nesse estudo foram obtidos a partir do modelo climático regional Eta (MESINGER *et al.*, 2012), gerados pelo Centro de Previsão de Tempo e Estudos Climáticos (CPTEC) e disponibilizados na Plataforma de Projeções de Mudança do Clima para a América do Sul Regionalizadas pelo Modelo Eta (PROJETA) (CHOU *et al.*, 2014c; CHOU *et al.*, 2014b), que oferece projeções do clima a partir de 2006.

A plataforma PROJETA [1] disponibiliza dados de projeção climática regionalizados para o Brasil. Após o preenchimento de uma requisição online, os dados são disponibilizados por email em formatos CSV, XML, JSON ou GEOJSON. Na requisição dos dados é necessário informar o cenário climático, a frequência dos dados (anual, mensal, diária), a localização desejada para as projeções (município ou coordenadas), as variáveis climáticas

[1] <https://projeta.cptec.inpe.br>

de interesse, o formato do arquivo contendo as projeções e o período de início e fim da série temporal.

O modelo regional utilizado depende das condições iniciais fornecidas pelo modelo climático global HadGEM2-ES e dos cenários de concentração dos gases de efeito estufa RCP4,5 e RCP8,5 (CHOU *et al.*, 2014c; CHOU *et al.*, 2014b). O modelo climático regional oferece resolução de 5 km centrados em uma coordenada específica, conforme informado na requisição dos dados. Para esse estudo foram utilizadas as coordenadas em graus da estação meteorológica do IAC onde os dados observados foram coletados (latitude: -22.871843, longitude: -47.077706).

Por fim, comparou-se os valores observados com as previsões, tanto de temperaturas quanto de óbitos. Para conseguir comparar os óbitos observados com as predições e as projeções sob cenários RCP, utilizou-se o período de 2016 a 2018, que é referente aos dados de teste do modelo de predição, conforme apresentado no Capítulo 3, Seção 3.3. Dessa forma, os grupos a serem comparados para as temperaturas e óbitos são:

Temperaturas:

- Temperaturas observadas (2016-2018)
- Temperaturas sob RCPs (2016-2018)
- Temperaturas sob RCPs (2016-2050)

Óbitos por DCVs:

- Óbitos observados (2016-2018)
- Óbitos preditos (2016-2018)
- Óbitos projetados sob RCPs (2016-2018)
- Óbitos projetados sob RCPs (2016-2050)

5.3 Projeções de Temperatura

As projeções de temperatura para ambos cenários de mudanças climáticas apresentaram valores superiores aos observados, conforme mostrado na Figura 5.2. As temperaturas de fato observadas no período entre 2016 e 2018 foram em média três graus mais baixas que as previstas no mesmo período sob os cenários RCP4,5 e RCP8,5, indicando que Campinas ainda não chegou nesses cenários. As médias de temperatura sob os cenários de mudanças climáticas não apresentaram diferenças significativas entre elas. Entretanto, para o período de 2016 a 2050 (Tabela 5.1) a diferença entre os cenários se torna clara, principalmente nos dados agrupados por mês (Figura 5.3), evidenciando que cenários de

maior emissão resultam em maiores temperaturas. Na Figura 5.3 é possível observar que as temperaturas sob cenário RCP8,5 alcançam valores mais altos que o cenário RCP4,5.

Ainda, é possível perceber que apesar do aumento médio na temperatura sob os cenários de mudanças climáticas, os extremos inferiores continuam apresentando valores tão baixos quanto os valores observados, principalmente no cenário de maior forçante radiativa (RCP8,5), conforme destacado nas setas na Figura 5.2. Esses valores indicam que o aumento da intensidade de eventos de frio extremo previstos numa escala macro para a América do Sul podem também afetar a região de Campinas (KODRA *et al.*, 2011).

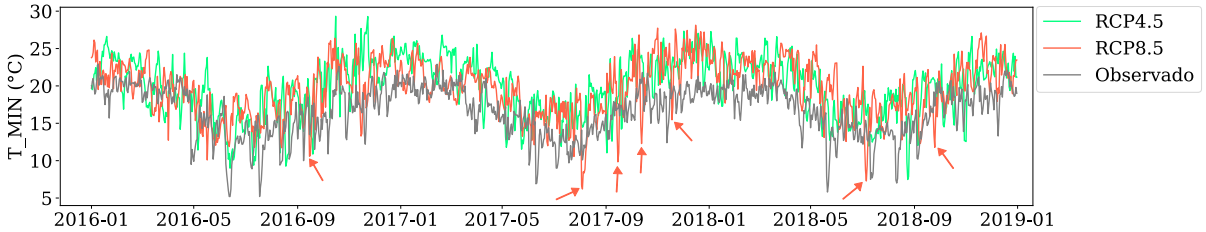


Figura 5.2: Temperatura mínima diária observada e projetada sob cenários de mudanças climáticas entre 2016 e 2018. As setas em vermelho indicam extremos de temperatura sob o cenário RCP8,5. As temperaturas sob os cenários RCP4,5 e 8,5 (linhas em verde e vermelho, respectivamente) apresentaram valores maiores que a temperatura observada (em cinza) para o mesmo período.

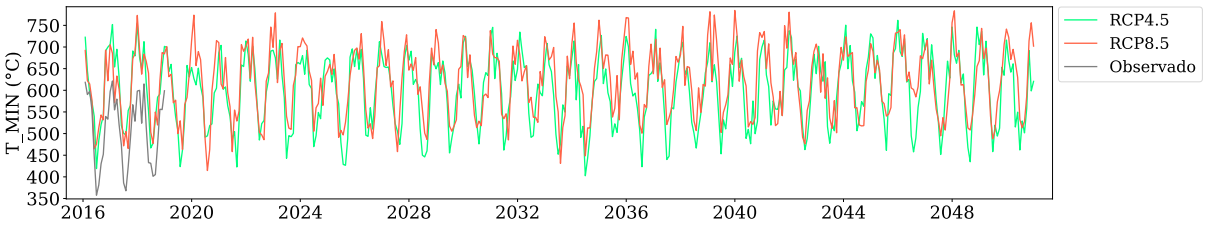


Figura 5.3: Temperatura mínima observada (2016-2018) e projetada sob cenários de mudanças climáticas (2016-2050) agrupadas por mês. As temperaturas sob o cenário RCP4,5 (linha verde) apresentaram valores menores que sob o cenário RCP8,5 (linha vermelha).

Tabela 5.1: Estatísticas descritivas da temperatura mínima diária observada e projetada sob cenários de mudanças climáticas. As estatísticas dos valores observados correspondem ao período entre 2016 e 2018. As estatísticas dos valores projetados sob os cenários RCP4,5 e 8,5 estão apresentados tanto no período de 2016 a 2018 quanto de 2016 a 2050. D.P.: Desvio padrão; mín.: valor mínimo da série; máx.: valor máximo da série.

		2016-2018		2016-2050	
	OBSERVADO	RCP4,5	RCP8,5	RCP4,5	RCP8,5
média	16,7	19,9	19,7	19,7	20,4
D.P.	3,3	3,9	3,6	3,6	3,7
mín.	5,2	7,5	6,2	3,2	4,3
máx.	22,9	29,3	28,1	29,6	30,1

5.4 Previsões de Contagens de Óbitos por DCVs em Campinas

A Figura 5.4 apresenta os valores observados e as previsões da contagem de óbitos por mês no período entre 2016 e 2018. Em cinza, a figura mostra os óbitos observados, em azul, os óbitos preditos, e em verde e vermelho, os óbitos projetados com base nos cenários de mudanças climáticas. Comparando os óbitos observados (cinza) e preditos (azul) com as projeções (verde e vermelho), observamos que se esperaria em média cerca de 20 óbitos a menos sob os cenários de mudanças climáticas (Figura 5.4 e Tabela 5.2). Ainda, o teste de Mann-Kendall sugere ausência de tendência nos óbitos sob os cenários RCP ($p < 5\%$), tanto entre 2016 a 2018 quanto entre 2016 a 2050, em contraste com a tendência de aumento encontrado nos óbitos observados no período de 2001 a 2018, como mencionado no Capítulo 4.

Não houve diferença na contagem de óbitos médio entre os dois cenários RCP no período entre 2016 e 2018 (Mann-Whitney U test: $p > 5\%$), conforme exibido na Figura 5.4 e Tabela 5.2. Contudo, no período entre 2016 e 2050 (Figura 5.5 e Tabela 5.2), observamos que o cenário RCP4,5 apresenta em média três óbitos a mais por mês que o cenário RCP8,5 (Mann-Whitney U test: $p < 5\%$).

Como se pode observar, as projeções indicam uma diminuição na contagem de óbitos por DCVs. Ainda, indicam que o cenário mais quente (RCP8,5) representaria menos óbitos por DCVs que o cenário intermediário (RCP4,5).

Esses resultados são corroborados por um estudo similar que utilizou a abordagem de Modelo Não-linear de Tempo Tardio e Extrapolação Log-linear, conforme mencionado na Tabela 2.1 do Capítulo 2. Esse estudo projetou o excesso de mortalidade[2] por DCVs relacionadas à temperatura sob cenários de mudanças climáticas (RCP4,5 e RCP8,5) e apontou que na maior parte do Brasil os óbitos aumentariam. Entretanto, para algumas cidades do sul e sudeste os óbitos diminuiriam, pelo menos até 2050, como exemplificado na Figura 5.6A, que apresenta o excesso de mortalidade por DCVs para os cenários RCP4,5 e RCP8,5 na cidade de São Paulo (SILVEIRA *et al.*, 2021).

A explicação para esse fenômeno é que a diminuição na contagem de óbitos relacionados ao frio seria maior que o aumento na contagem de óbitos relacionados ao calor, resultando em uma diminuição nos óbitos como um todo (SILVEIRA *et al.*, 2021). Na prática, o deslocamento da faixa de temperaturas no sentido de aumento resulta em menor risco relativo de óbito (Figura 5.6B), mas só até certo ponto, uma vez que temperaturas muito elevadas também representam risco de saúde (SILVEIRA *et al.*, 2021). Assim, existe uma faixa de temperatura considerada ideal, onde o risco de óbito seria mínimo.

Outro ponto que corrobora esses achados é a correlação entre temperatura e

[2]Excesso de mortalidade: Aumento percentual da contagem de mortes em relação ao esperado/observado.

óbitos por DCVs encontrada na análise descritiva (Seção 4.1) do Capítulo 4. A correlação mostra que a contagem de óbitos por DCVs nas baixas temperaturas (inverno) é maior que nas altas temperaturas (verão). Dessa forma, com o aumento das temperaturas devido às mudanças climáticas, se espera que a contagem de óbitos diminua, pelo menos até certo ponto (SILVEIRA *et al.*, 2021).

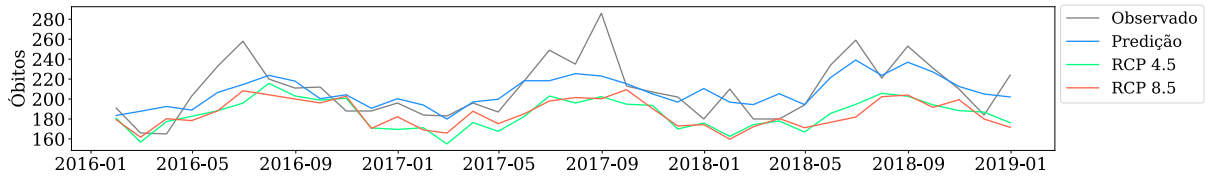


Figura 5.4: Óbitos por DCVs no período de 2016 a 2018. Em cinza, são os óbitos observados, em azul, os óbitos preditos, e em verde e vermelho são os óbitos projetados sob os cenários RCP4,5 e RCP8,5, respectivamente.

Tabela 5.2: Estatísticas descritivas dos óbitos acumulados por mês. Na tabela são apresentados os óbitos observado, predito baseado em temperaturas observadas e projetados sob cenários de mudanças climáticas. As estatísticas dos valores observados e preditos baseado em temperaturas observadas correspondem ao período entre 2016 e 2018. As estatísticas dos valores previstos sob os cenários RCP4,5 e 8,5 são apresentados tanto no período de 2016 a 2018 quanto de 2016 a 2050. D.P.: Desvio padrão; mín.: valor mínimo da série; máx.: valor máximo da série.

			2016-2018		2016-2050	
	OBSERVADO	PREDITO	RCP4,5	RCP8,5	RCP4,5	RCP8,5
média	209,8	207,1	184,6	185,4	185,4	182,8
D.P.	28,2	14,8	15,0	14,1	12,1	11,7
mín.	165	179,9	155,0	159,7	155,0	157,3
máx.	286	239,1	215,7	209,4	215,7	212,7

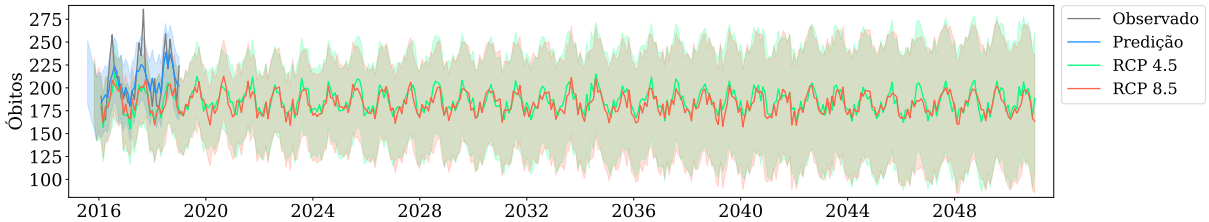


Figura 5.5: Óbitos por DCVs no período de 2016 a 2050. Em cinza, são os óbitos observados, em azul, os óbitos preditos com base nas temperaturas observadas, em verde e vermelho são os óbitos preditos com base nas temperaturas projetadas sob os cenários RCP4,5 e RCP8,5, respectivamente. As áreas sombreadas representam o intervalo de confiança de 95%.

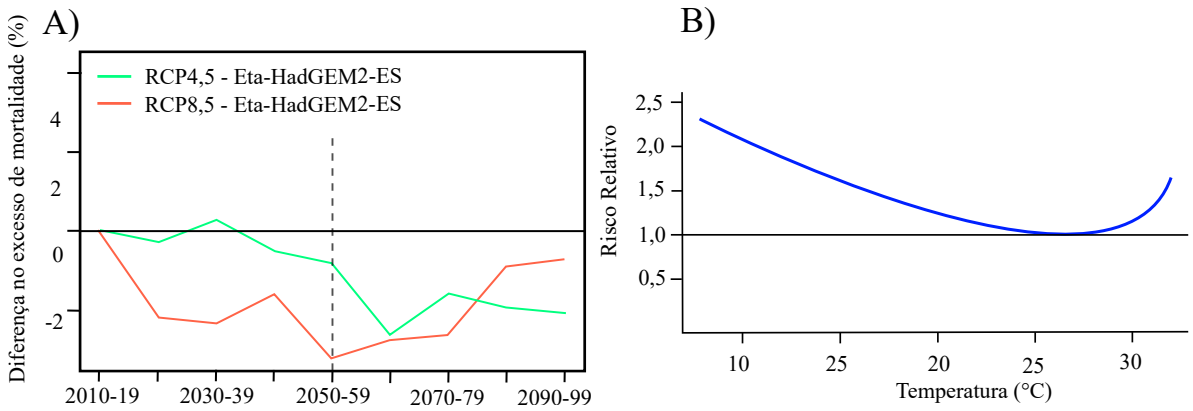


Figura 5.6: A) Projeção da diferença no excesso de mortalidade por DCVs no período de 2010 a 2100 sob cenários de mudanças climáticas na cidade de São Paulo. A linha vertical tracejada indica o ano de 2050. B) Risco relativo de óbitos por DCVs por temperatura média na cidade de São Paulo. Em verde, está o excesso de mortalidade sob cenário RCP4,5 e em vermelho o excesso de mortalidade sob cenário RCP8,5. Adaptado de (SILVEIRA *et al.*, 2021)

5.5 Considerações Finais

As projeções do clima para Campinas, obtidas através do modelo climático regional Eta-HadGEM2-ES, apontam aumento nas temperaturas mínimas para os cenários RCP4,5 e RCP8,5 até 2050. Contudo, a comparação entre as temperaturas observadas e projetadas (2016 a 2018) mostra que a cidade ainda não alcançou o patamar previsto pelos cenários de mudanças climáticas.

Um dos melhores modelos de predição de óbitos desenvolvido nesse estudo projeta que o aumento da temperatura, devido às mudanças climáticas, resultará na diminuição da contagem de óbitos por DCVs até 2050. Esse resultado é corroborado por Silveira *et al.* (2021), que realizaram um estudo de maior resolução abordando 21 cidades brasileiras. De acordo com as projeções de Silveira *et al.* (2021), diversas cidades do sul e sudeste terão a contagem de óbitos por DCVs reduzida devido ao aumento da temperatura.

Apesar desse estudo ter sido corroborado por outro trabalho, ele apresenta limitações. As projeções da contagem de óbitos até 2050 foram baseadas em um modelo treinado com dados de temperatura entre 2001 e 2015. Sendo assim, a projeção assume que a relação entre as variáveis permaneça constante para ser extrapolável até 2050. Outra limitação, entretanto inerente à tarefa de projeção de longo prazo, é o nível de incerteza das projeções, que aumenta conforme aumentamos o horizonte temporal da análise.

Capítulo 6

Conclusão

O objetivo desse trabalho foi o desenvolvimento de um modelo de predição de séries temporais capaz de realizar alertas antecipados para eventos climáticos com potencial de impactar a saúde humana, em específico, os óbitos por doenças cardiovasculares (DCVs). No limite do nosso conhecimento, esse é o primeiro trabalho que aborda a predição de óbitos por DCVs a partir da regressão linear com erros SARIMA (LR-SARIMAX) e de rede neural recorrente LSTM. As perguntas de pesquisa respondidas nesse trabalho foram:

- **P.1**: É possível utilizar modelos de série temporal para prever o número de óbitos por DCVs com base em dados ambientais?
- **P.2**: É possível realizar previsões de longo prazo utilizando esses modelos?

6.1 Predição de Curto Prazo

Como contribuição, esse trabalhou criou uma base de dados integrada e processada contendo valores de variáveis ambientais e de óbitos por todas as causas, para a cidade de Campinas, São Paulo, a qual representa a primeira base de dados integrada pública de clima e saúde, até o nosso conhecimento. Essa base de dados foi disponibilizada publicamente e poderá ser usada para fins educacionais e de pesquisa (BELLA *et al.*, 2022).

A partir do banco de dados integrado, foram gerados modelos preditivos de série temporal da contagem de óbitos por DCVs com base em variáveis ambientais. Tanto a abordagem utilizada quanto os modelos treinados representam contribuições deste trabalho, uma vez que se trata do primeiro estudo a realizar predições de curto prazo para contagem de óbitos por DCVs utilizando variáveis ambientais, conforme discutido no Capítulo 2, Seção 2.4. Ainda, esses modelos podem servir de base para sistemas de alerta antecipado, visando o bem estar da população e o auxílio a tomadores de decisão. Dessa forma, esse trabalho responde à primeira pergunta de pesquisa, mostrando a possibilidade de prever

óbitos por DCVs com base em variáveis ambientais utilizando a abordagem de séries temporais.

Com relação às abordagens de previsão, esse estudo utilizou a técnica de grid search em modelos LR-SARIMAX e em redes neurais LSTM para alcançar os menores erros de predição. Essa técnica consistiu na variação sistemática dos parâmetros intrínsecos dos modelos em conjunto com diferentes combinações das variáveis preditoras e da quantidade de atrasos dessas variáveis. Ao todo, foram avaliados 441 modelos através das métricas RMSE e MAPE. Os modelos foram comparados em relação às variáveis utilizadas, à quantidade de lags, à periodicidade dos dados e ao tipo de modelo.

Analisando os modelos de menor erro, o trabalho evidenciou que a temperatura desempenha um papel importante nas predições de óbitos. A quantidade de atrasos das variáveis preditoras também foi relevante nas predições, corroborando achados anteriores da literatura que mostram que as variáveis ambientais continuam exercendo efeito após o fim da exposição. Apesar das diferentes abordagens não terem apresentado diferenças significativas em relação ao erro de predição, para as especificidades desse trabalho, a modelagem a partir do LR-SARIMAX se mostrou mais robusta em relação à variação dos dados que o LSTM. Também foi possível concluir que os modelos utilizando dados na periodicidade mensal apresentaram erros de predição significativamente menores que os modelos utilizando as periodicidades diária ou semanal.

6.2 Predição de Longo Prazo

Além das predições de curto prazo, esse trabalho realizou um estudo exploratório sobre o impacto das mudanças climáticas no número de óbitos para a região, com base em projeções do clima para Campinas até a metade do século. Esse estudo deixa como contribuições uma nova abordagem para predições de longo prazo e projeções do cenário futuro de óbitos por DCVs para Campinas. Ainda, responde à segunda pergunta de pesquisa, evidenciando a possibilidade de realização de previsões de longo prazo com base em cenários de mudanças climáticas.

O estudo exploratório apontou que com o aumento das temperaturas mínimas, devido às mudanças climáticas, espera-se uma diminuição da contagem de óbitos por DCVs até 2050 para Campinas. Esse resultado é corroborado por Silveira *et al.* (2021), que mostram que embora isso não seja verdadeiro para a maioria das regiões brasileiras, esse efeito pode ocorrer para algumas cidades do sul e sudeste. Isso pode acontecer pelo fato da diminuição no número de óbitos relacionados ao frio ser maior que o aumento no número de óbitos relacionados ao calor.

6.3 Limitações e Trabalhos Futuros

Visando a aplicação do modelo como um sistema de alerta antecipado, predições subestimadas do número de óbitos são indesejadas, uma vez que estariam subestimando uma potencial ameaça à vida. Os modelos que utilizaram dados com menor periodicidade apresentaram os menores erros, mas, ainda apresentaram predições subestimadas. Dessa forma, a busca por modelagens mais sensíveis à predição de eventos extremos, que resultam em picos da série temporal, é de fundamental importância em trabalhos futuros que tenham como objetivo aplicar o modelo em sistemas de alerta.

Uma importante limitação do estudo exploratório realizado se deve às incertezas intrínsecas da previsão de longo prazo. Entretanto, essa é uma limitação comum a estudos que realizam esse tipo de previsão. Até a metade do século, outras variáveis podem influenciar o número de óbitos, diminuindo o desempenho do modelo. Ainda, a mesma variável pode apresentar relação futura diferente daquela em que o modelo foi treinado. Mesmo modelos com excelentes desempenhos podem apresentar problemas futuros, pois uma vez que são implementados, vidas serão salvas e diminuirão as correlações entre as variáveis ambientais e os óbitos por DCVs. Assim, esse estudo exploratório tem como pressuposto que a relação entre as variáveis não se modificará significativamente até 2050.

Referências

ANDERSON, B. G.; BELL, M. L. Weather-related mortality: how heat, cold, and heat waves affect mortality in the United States. *Epidemiology (Cambridge, Mass.)*, v. 20, n. 2, p. 205–213, mar. 2009. ISSN 1531-5487.

BANERJEE, S.; HUTH, J. K. Time-series study of cardiovascular rates in India: A systematic analysis between 1990 and 2017. *Indian Heart Journal*, v. 72, n. 3, p. 194–196, maio 2020. ISSN 0019-4832. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S0019483220301206>.

BARI, S. H.; RAHMAN, M. T. U.; HOQUE, M. A.; HUSSAIN, M. M. Analysis of seasonal and annual rainfall trends in the northern region of Bangladesh. *Atmospheric Research*, v. 176-177, p. 148–158, jul. 2016. ISSN 0169-8095. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S0169809516300254>.

BECK, H. E.; ZIMMERMANN, N. E.; MCVICAR, T. R.; VERGOPOLAN, N.; BERG, A.; WOOD, E. F. Present and future Köppen-Geiger climate classification maps at 1-km resolution. *Scientific Data*, v. 5, n. 1, p. 180214, out. 2018. ISSN 2052-4463. Available from Internet: <https://www.nature.com/articles/sdata2018214>.

BELLA, T. R.; COSTA, P. D. P. *Prediction Models for Cardiovascular Disease Deaths*. 2022. Available from Internet: <https://github.com/climate-and-health-datasci-Unicamp/cvd-deaths-from-env>.

BELLA, T. R.; LAZARI, J. P. d.; OLIVEIRA, D. S. d.; COELHO, R. P.; COROZOLLA, W.; BEZERRA, L. M.; AVILA, A. M. H. d.; COSTA, P. D. P.; FARIA, E. C. de. *Human death records and environmental parameters of the city of Campinas, state of São Paulo, Brazil, 2001-2018*. Repositório de Dados de Pesquisa da Unicamp, 2022. Available from Internet: <https://doi.org/10.25824/redu/UPCDSK>.

BOURDREL, T.; BIND, M.-A.; BÉJOT, Y.; MOREL, O.; ARGACHA, J.-F. Cardiovascular effects of air pollution. *Archives of Cardiovascular Diseases*, v. 110, n. 11, p. 634–642, nov. 2017. ISSN 1875-2136. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S1875213617301304>.

BOX, G.; JENKINS, G.; REINSEL, G.; LJUNG, G. *Time Series Analysis: Forecasting and Control*. Wiley, 2015. (Wiley Series in Probability and Statistics). ISBN 9781118674925. Available from Internet: <https://books.google.com.br/books?id=rNt5CgAAQBAJ>.

BRAGA, A. L. F.; ZANOBETTI, A.; SCHWARTZ, J. The effect of weather on respiratory and cardiovascular deaths in 12 U.S. cities. *Environmental Health Perspectives*, v. 110, n. 9, p. 859–863, set. 2002. ISSN 0091-6765. Available from Internet: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1240983/>.

BROWNLEE, J. *Deep Learning With Python: Develop Deep Learning Models on Theano and TensorFlow Using Keras*. [S.l.]: Machine Learning Mastery, 2016.

BROWNLEE, J. *Long Short-term Memory Networks with Python: Develop Sequence Prediction Models with Deep Learning*. Jason Brownlee, 2017. Available from Internet: <https://books.google.com.br/books?id=ONpdswEACAAJ>.

CHAI, C. P. The importance of data cleaning: Three visualization examples. *CHANCE*, Taylor Francis, v. 33, n. 1, p. 4–9, 2020. Available from Internet: <https://doi.org/10.1080/09332480.2020.1726112>.

CHATFIELD, C. *Time-Series Forecasting*. New York: Chapman and Hall/CRC, 2000. ISBN 978-0-429-12635-2.

CHEN, R.; YIN, P.; WANG, L.; LIU, C.; NIU, Y.; WANG, W.; JIANG, Y.; LIU, Y.; LIU, J.; QI, J.; YOU, J.; KAN, H.; ZHOU, M. Association between ambient temperature and mortality risk and burden: time series study in 272 main Chinese cities. *The BMJ*, v. 363, p. k4306, out. 2018. ISSN 0959-8138. Available from Internet: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6207921/>.

CHOU, S. C.; LYRA, A.; MOURÃO, C.; DERECZYNSKI, C.; PILOTTO, I.; GOMES, J.; BUSTAMANTE, J.; TAVARES, P.; SILVA, A.; RODRIGUES, D. *et al.* Assessment of climate change over south america under rcp 4.5 and 8.5 downscaling scenarios. *American Journal of Climate Change*, Scientific Research Publishing, v. 3, n. 05, p. 512, 2014.

CHOU, S. C.; LYRA, A.; MOURÃO, C.; DERECZYNSKI, C.; PILOTTO, I.; GOMES, J.; BUSTAMANTE, J.; TAVARES, P.; SILVA, A.; RODRIGUES, D.; CAMPOS, D.; CHAGAS, D.; SUEIRO, G.; SIQUEIRA, G.; MARENGO, J. Assessment of Climate Change over South America under RCP 4.5 and 8.5 Downscaling Scenarios. *American Journal of Climate Change*, v. 3, n. 5, p. 512–527, dez. 2014. Number: 5 Publisher: Scientific Research Publishing. Available from Internet: <http://www.scirp.org/Journal/Paperabs.aspx?paperid=52887>.

CHOU, S. C.; LYRA, A.; MOURÃO, C.; DERECZYNSKI, C.; PILOTTO, I.; GOMES, J.; BUSTAMANTE, J.; TAVARES, P.; SILVA, A.; RODRIGUES, D.; CAMPOS, D.; CHAGAS, D.; SUEIRO, G.; SIQUEIRA, G.; NOBRE, P.; MARENGO, J. Evaluation of the Eta Simulations Nested in Three Global Climate Models. *American Journal of Climate Change*, v. 3, n. 5, p. 438–454, dez. 2014. Number: 5 Publisher: Scientific Research Publishing. Available from Internet: <http://www.scirp.org/Journal/Paperabs.aspx?paperid=52877>.

CONLON, K. C.; RAJKOVICH, N. B.; WHITE-NEWSOME, J. L.; LARSEN, L.; O'NEILL, M. S. Preventing cold-related morbidity and mortality in a changing climate. *Maturitas*, v. 69, n. 3, p. 197–202, jul. 2011. ISSN 0378-5122. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S0378512211001162>.

DOCKERY, D. W. Health effects of particulate air pollution. *Annals of Epidemiology*, v. 19, n. 4, p. 257–263, abr. 2009. ISSN 1873-2585.

GAO, J.; YU, F.; XU, Z.; DUAN, J.; CHENG, Q.; BAI, L.; ZHANG, Y.; WEI, Q.; YI, W.; PAN, R. *et al.* The association between cold spells and admissions of ischemic stroke

in hefei, china: Modified by gender and age. *Science of The Total Environment*, Elsevier, v. 669, p. 140–147, 2019.

GASPARRINI, A.; GUO, Y.; HASHIZUME, M.; LAVIGNE, E.; ZANOBETTI, A.; SCHWARTZ, J.; TOBIAS, A.; TONG, S.; ROCKLÖV, J.; FORSBERG, B. *et al.* Mortality risk attributable to high and low ambient temperature: a multicountry observational study. *The Lancet*, Elsevier, v. 386, n. 9991, p. 369–375, 2015.

GÉRON, A. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. [S.l.]: "O'Reilly Media, Inc.", 2019. Google-Books-ID: HnetDwAAQBAJ. ISBN 978-1-4920-3259-5.

GHOSH, P.; AZAM, S.; JONKMAN, M.; KARIM, A.; SHAMRAT, F. M. J. M.; IGNATIOUS, E.; SHULTANA, S.; BEERAVOLU, A. R.; BOER, F. D. Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms With Relief and LASSO Feature Selection Techniques. *IEEE Access*, v. 9, p. 19304–19326, 2021. ISSN 2169-3536. Conference Name: IEEE Access.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A.; BENGIO, Y. *Deep learning*. [S.l.]: MIT press Cambridge, 2016. v. 1.

GÓRRIZ, J. M.; RAMÍREZ, J.; ORTÍZ, A.; MARTÍNEZ-MURCIA, F. J.; SEGOVIA, F.; SUCKLING, J.; LEMING, M.; ZHANG, Y.-D.; ÁLVAREZ-SÁNCHEZ, J. R.; BOLOGNA, G.; BONOMINI, P.; CASADO, F. E.; CHARTE, D.; CHARTE, F.; CONTRERAS, R.; CUESTA-INFANTE, A.; DURO, R. J.; FERNÁNDEZ-CABALLERO, A.; FERNÁNDEZ-JOVER, E.; GÓMEZ-VILDA, P.; GRAÑA, M.; HERRERA, F.; IGLESIAS, R.; LEKOVA, A.; LOPE, J. de; LÓPEZ-RUBIO, E.; MARTÍNEZ-TOMÁS, R.; MOLINA-CABELLO, M. A.; MONTEMAYOR, A. S.; NOVAIS, P.; PALACIOS-ALONSO, D.; PANTRIGO, J. J.; PAYNE, B. R.; LÓPEZ, F. de la P.; PINNINGHOFF, M. A.; RINCÓN, M.; SANTOS, J.; THURNHOFER-HEMSI, K.; TSANAS, A.; VARELA, R.; FERRÁNDEZ, J. M. Artificial intelligence within the interplay between natural and artificial computation: Advances in data science, trends and applications. *Neurocomputing*, v. 410, p. 237–270, out. 2020. ISSN 0925-2312. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S0925231220309292>.

HADLEY, M. B.; BAUMGARTNER, J.; VEDANTHAN, R. Developing a Clinical Approach to Air Pollution and Cardiovascular Health. *Circulation*, v. 137, n. 7, p. 725–742, fev. 2018. ISSN 0009-7322, 1524-4539. Available from Internet: <https://www.ahajournals.org/doi/10.1161/CIRCULATIONAHA.117.030377>.

HAN, J.; LIU, S.; ZHANG, J.; ZHOU, L.; FANG, Q.; ZHANG, J.; ZHANG, Y. The impact of temperature extremes on mortality: a time-series study in jinan, china. *Bmj Open*, British Medical Journal Publishing Group, v. 7, n. 4, p. e014741, 2017.

HOCHREITER, S.; SCHMIDHUBER, J. Long Short-Term Memory. *Neural Computation*, v. 9, n. 8, p. 1735–1780, nov. 1997. ISSN 0899-7667. Conference Name: Neural Computation.

HORNIK, K.; STINCHCOMBE, M.; WHITE, H. Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural Networks*, v. 3, n. 5, p. 551–560, jan. 1990. ISSN 0893-6080. Available from Internet: <https://www.sciencedirect.com/science/article/pii/0893608090900056>.

HUANG, J.; TAN, J.; YU, W. Temperature and Cardiovascular Mortality Associations in Four Southern Chinese Cities: A Time-Series Study Using a Distributed Lag Non-Linear Model. *Sustainability*, v. 9, n. 3, p. 321, fev. 2017. ISSN 2071-1050. Available from Internet: <http://www.mdpi.com/2071-1050/9/3/321>.

HYNDMAN, R.; ATHANASOPOULOS, G. *Forecasting: Principles and Practice*. 2nd. ed. Australia: OTexts, 2018.

HYNDMAN, R. J.; KHANDAKAR, Y. Automatic time series forecasting: The forecast package for r. *Journal of Statistical Software, Articles*, v. 27, n. 3, p. 1–22, 2008. ISSN 1548-7660. Available from Internet: <https://www.jstatsoft.org/v027/i03>.

IBGE. Censo brasileiro de 2010. 2012. Rio de Janeiro. Available from Internet: <https://censo2010.ibge.gov.br/resultados.html>.

IPCC. Summary for policymakers. In: _____. *Climate Change 2022: Impacts, Adaptation, and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge, UK: Cambridge University Press, 2022.

Modelos climáticos: Uma revisão da modelagem numérica. In: JUNIOR, J. Z.; FURTADO, A. T.; PFEIFFER, C. C. (Ed.). *Planejamento da produção de cana-de-açúcar no contexto das mudanças climáticas globais*. SciELO - Editora da Unicamp, 2016. p. 115–130. ISBN 978-85-268-1499-8. Available from Internet: <http://www.jstor.org/stable/10.7476/9788526814998.12>.

KINGMA, D. P.; BA, J. Adam: A method for stochastic optimization. In: BENGIO, Y.; LECUN, Y. (Ed.). *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. [s.n.], 2015. Available from Internet: <http://arxiv.org/abs/1412.6980>.

KODRA, E.; STEINHAEUSER, K.; GANGULY, A. R. Persisting cold extremes under 21st-century warming scenarios. *Geophysical Research Letters*, v. 38, n. 8, 2011. ISSN 1944-8007. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2011GL047103. Available from Internet: <https://onlinelibrary.wiley.com/doi/abs/10.1029/2011GL047103>.

KUMAR, M.; GUPTA, S.; KUMAR, K.; SACHDEVA, M. SPREADING OF COVID-19 IN INDIA, ITALY, JAPAN, SPAIN, UK, US: A Prediction Using ARIMA and LSTM Model. *Digital Government: Research and Practice*, v. 1, n. 4, p. 1–9, dez. 2020. ISSN 2691-199X, 2639-0175. Available from Internet: <https://dl.acm.org/doi/10.1145/3411760>.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. *Nature*, v. 521, n. 7553, p. 436–444, maio 2015. ISSN 1476-4687. Available from Internet: <https://doi.org/10.1038/nature14539>.

LELIEVELD, J.; KLINGMüLLER, K.; POZZER, A.; PöSCHL, U.; FNAIS, M.; DAIBER, A.; MüNZEL, T. Cardiovascular disease burden from ambient air pollution in Europe reassessed using novel hazard ratio functions. *European Heart Journal*, v. 40, n. 20, p. 1590–1596, maio 2019. ISSN 0195-668X, 1522-9645. Available from Internet: <https://academic.oup.com/eurheartj/article/40/20/1590/5372326>.

LENTON, T. M.; ROCKSTRÖM, J.; GAFFNEY, O.; RAHMSTORF, S.; RICHARDSON, K.; STEFFEN, W.; SCHELLNHUBER, H. J. *Climate tipping points—too risky to bet against*. [S.l.]: Nature Publishing Group, 2019.

LIN, Y.-K.; SUNG, F.-C.; HONDA, Y.; CHEN, Y.-J.; WANG, Y.-C. Comparative assessments of mortality from and morbidity of circulatory diseases in association with extreme temperatures. *Science of The Total Environment*, v. 723, p. 138012, jun. 2020. ISSN 0048-9697. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S0048969720315254>.

LIU, C.; YAVAR, Z.; SUN, Q. Cardiovascular response to thermoregulatory challenges. *American Journal of Physiology - Heart and Circulatory Physiology*, v. 309, n. 11, p. H1793–H1812, dez. 2015. ISSN 0363-6135. Available from Internet: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4698386/>.

MANN, H. B. Nonparametric Tests Against Trend. *Econometrica*, v. 13, n. 3, p. 245–259, 1945. ISSN 0012-9682. Publisher: [Wiley, Econometric Society]. Available from Internet: <https://www.jstor.org/stable/1907187>.

MANNUCCI, P. M.; HARARI, S.; FRANCHINI, M. Novel evidence for a greater burden of ambient air pollution on cardiovascular disease. *Haematologica*, v. 104, n. 12, p. 2349–2357, dez. 2019. ISSN 1592-8721.

MCMICHAEL, A. *Planetary overload : global environmental change and the health of the human species*. [S.l.]: Cambridge U.P., 1993. ISBN 0-521-45759-9.

MCMICHAEL, A. Integrated assessment of potential health impact of global environmental change: prospects and limitations. *Environmental Modeling & Assessment*, Springer, v. 2, n. 3, p. 129–137, 1997.

MESINGER, F.; CHOU, S. C.; GOMES, J. L.; JOVIC, D.; BASTOS, P.; BUSTAMANTE, J. F.; LAZIC, L.; LYRA, A. A.; MORELLI, S.; RISTIC, I.; VELJOVIC, K. An upgraded version of the Eta model. *Meteorology and Atmospheric Physics*, v. 116, n. 3, p. 63–79, maio 2012. ISSN 1436-5065. Available from Internet: <https://doi.org/10.1007/s00703-012-0182-z>.

Ministério da Saúde : FUNASA. *Manual de instruções para o preenchimento da declaração de óbito*. Brasília/DF: Ascom, 2001.

MOGHADAMNIA, M. T.; ARDALAN, A.; MESDAGHINIA, A.; KESHTKAR, A.; NADDAFI, K.; YEKANINEJAD, M. S. Ambient temperature and cardiovascular mortality: a systematic review and meta-analysis. *PeerJ*, v. 5, p. e3574, 2017. Publisher: PeerJ Inc.

MOHAN, S.; THIRUMALAI, C.; SRIVASTAVA, G. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. *IEEE Access*, v. 7, p. 81542–81554, 2019. ISSN 2169-3536. Conference Name: IEEE Access.

MURRAY, C. J.; ARAVKIN, A. Y.; ZHENG, P.; ABBAFATI, C.; ABBAS, K. M.; ABBASI-KANGEVARI, M.; ABD-ALLAH, F.; ABDELALIM, A.; ABDOLLAHI, M.; ABDOLLAHPOUR, I. *et al.* Global burden of 87 risk factors in 204 countries and territories, 1990–2019: a systematic analysis for the global burden of disease study 2019. *The Lancet*, Elsevier, v. 396, n. 10258, p. 1223–1249, 2020.

NUTLEY, T.; REYNOLDS, H. Improving the use of health data for health system strengthening. *Global Health Action*, Taylor Francis, v. 6, n. 1, p. 20001, 2013. PMID: 28140939. Available from Internet: <https://doi.org/10.3402/gha.v6i0.20001>.

O'BRIEN, R. M. A Caution Regarding Rules of Thumb for Variance Inflation Factors. *Quality & Quantity*, v. 41, n. 5, p. 673–690, out. 2007. ISSN 1573-7845. Available from Internet: <https://doi.org/10.1007/s11135-006-9018-6>.

PACHAURI, R. K.; ALLEN, M. R.; BARROS, V. R.; BROOME, J.; CRAMER, W.; CHRIST, R.; CHURCH, J. A.; CLARKE, L.; DAHE, Q.; DASGUPTA, P. *et al. Climate change 2014: synthesis report. Contribution of Working Groups I, II and III to the fifth assessment report of the Intergovernmental Panel on Climate Change*. [S.l.]: Ipcc, 2014.

PERES, E. M.; ANDRADE, A. M.; POZ, M. R. D.; GRANDE, N. R. The practice of physicians and nurses in the Brazilian Family Health Programme – evidences of change in the delivery health care model. *Human Resources for Health*, v. 4, n. 1, p. 25, nov. 2006. ISSN 1478-4491. Available from Internet: <https://doi.org/10.1186/1478-4491-4-25>.

PURWANTO, D.; ESWARAN, C.; LOGESWARAN, R. A Comparison of ARIMA, Neural Network and Linear Regression Models for the Prediction of Infant Mortality Rate. In: *2010 Fourth Asia International Conference on Mathematical/Analytical Modelling and Computer Simulation*. Kota Kinabalu, Malaysia: IEEE, 2010. p. 34–39. ISBN 978-1-4244-7196-6. Available from Internet: <http://ieeexplore.ieee.org/document/5489680/>.

QUEIROZ, P. G. M.; JACOMINO, V. M. F.; MENEZES, M. Â. d. B. C. Composição elementar do material particulado presente no aerossol atmosférico do município de Sete Lagoas, Minas Gerais. *Química Nova*, v. 30, p. 1233–1239, out. 2007. ISSN 0100-4042, 1678-7064. Publisher: Sociedade Brasileira de Química. Available from Internet: <http://www.scielo.br/j/qn/a/NkWDdqrL44kQDbyHjrxFd8p/?lang=pt>.

RAJAGOPALAN, S.; AL-KINDI, S. G.; BROOK, R. D. Air Pollution and Cardiovascular Disease: JACC State-of-the-Art Review. *Journal of the American College of Cardiology*, v. 72, n. 17, p. 2054–2070, out. 2018. ISSN 0735-1097. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S0735109718383554>.

ROMIEU, I.; GOUVEIA, N.; CIFUENTES, L.; LEON, A. D.; JUNGER, W.; VERA, J.; STRAPPA, V.; MAGALI, H.-D.; MIRANDA-SOBERANIS, V.; ROJAS-BRACHO, L.; CARBAJAL-ARROYO, L.; TZINTZUN-CERVANTES, G. Multicity study of air pollution and mortality in Latin America (the ESCALA Study). *Research report (Health Effects Institute)*, v. 171, p. 5–86, out. 2012.

ROTH, G. A.; MENSAH, G. A.; JOHNSON, C. O.; ADDOLORATO, G.; AMMIRATI, E.; BADDOUR, L. M.; BARENGO, N. C.; BEATON, A. Z.; BENJAMIN, E. J.; BENZIGER, C. P. *et al.* Global burden of cardiovascular diseases and risk factors, 1990–2019: Update from the gbd 2019 study. *Journal of the American College of Cardiology*, Elsevier, 2020.

SILVEIRA, I. H.; CORTES, T. R.; OLIVEIRA, B. F. A. de; JUNGER, W. L. Projections of excess cardiovascular mortality related to temperature under different climate change scenarios and regionalized climate model simulations in Brazilian cities. *Environmental Research*, v. 197, p. 110995, jun. 2021. ISSN 1096-0953.

SILVEIRA, I. H.; OLIVEIRA, B. F. A.; CORTES, T. R.; JUNGER, W. L. The effect of ambient temperature on cardiovascular mortality in 27 Brazilian cities. *Science of The Total Environment*, v. 691, p. 996–1004, nov. 2019. ISSN 0048-9697. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S0048969719330669>.

SON, J.-Y.; GOUVEIA, N.; BRAVO, M. A.; FREITAS, C. U. de; BELL, M. L. The impact of temperature on mortality in a subtropical city: effects of cold, heat, and heat waves in São Paulo, Brazil. *International Journal of Biometeorology*, v. 60, n. 1, p. 113–121, jan. 2016. ISSN 1432-1254. Available from Internet: <https://doi.org/10.1007/s00484-015-1009-7>.

SONG, X.; WANG, S.; HU, Y.; YUE, M.; ZHANG, T.; LIU, Y.; TIAN, J.; SHANG, K. Impact of ambient temperature on morbidity and mortality: an overview of reviews. *Science of the Total Environment*, v. 586, p. 241–254, 2017. Publisher: Elsevier.

THOMAS, H.; DIAMOND, J.; VIECO, A.; CHAUDHURI, S.; SHINNAR, E.; CROMER, S.; PEREL, P.; MENSAH, G. A.; NARULA, J.; JOHNSON, C. O.; ROTH, G. A.; MORAN, A. E. Global Atlas of Cardiovascular Disease 2000-2016: The Path to Prevention and Control. *Global Heart*, v. 13, n. 3, p. 143–163, set. 2018. ISSN 2211-8179.

UNISDR. *Developing early warning systems, a checklist: third international conference on early warning (EWC III), 27-29 March 2006, Bonn, Germany*. [s.n.], 2006. Available from Internet: <https://www.undrr.org/publication/developing-early-warning-systems-checklist-third-international-conference-early-warning>.

VANI, S.; RAO, T. V. M. An experimental approach towards the performance assessment of various optimizers on convolutional neural network. In: *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*. [S.l.: s.n.], 2019. p. 331–336.

WAI, A. H. C.; SENG, S. Y.; FEI, J. L. W. Fatality Involving Road Accidents in Malaysia: A Comparison between Three Statistical Models. In: *Proceedings of the 2019 2nd International Conference on Mathematics and Statistics - ICoMS'19*. Prague, Czech Republic: ACM Press, 2019. p. 101–105. ISBN 978-1-4503-7168-1. Available from Internet: <http://dl.acm.org/citation.cfm?doid=3343485.3343494>.

WANG, C.; QI, Y.; ZHU, G. Deep learning for predicting the occurrence of cardiopulmonary diseases in Nanjing, China. *Chemosphere*, v. 257, p. 127176, out. 2020. ISSN 0045-6535. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S0045653520313692>.

WANG, Q.; MA, Y.; ZHAO, K.; TIAN, Y. A Comprehensive Survey of Loss Functions in Machine Learning. *Annals of Data Science*, v. 9, n. 2, p. 187–212, abr. 2022. ISSN 2198-5812. Available from Internet: <https://doi.org/10.1007/s40745-020-00253-5>.

World Health Organization. *Climate change and human health : risks and responses /ceditors : A. J. McMichael ... [et al.]*. [S.l.]: World Health Organization, 2003. Prepared jointly by the World Health Organization, the World Meteorological Organization and the United Nations Environment Programme p.

World Health Organization. Publications. *Basic documents*. 49th ed. ed. [S.l.]: World Health Organization, 2020. 238 p. p.

YING, X. An overview of overfitting and its solutions. *Journal of Physics: Conference Series*, IOP Publishing, v. 1168, p. 022022, feb 2019. Available from Internet: <https://doi.org/10.1088/1742-6596/1168/2/022022>.