UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Instituto de Ciência e Tecnologia
Câmpus de Sorocaba

KAILANI DE ASSIS TAVERNEZI RODRIGUES

Investigação Causal das Doenças Respiratórias e Doenças Transmitidas por Mosquitos no Brasil por Meio de Aprendizado de Máquina no Contexto das Mudanças Climáticas

Sorocaba
2025

KAILANI DE ASSIS TAVERNEZI RODRIGUES

Investigação Causal das Doenças Respiratórias e Doenças Transmitidas por Mosquitos no Brasil por Meio de Aprendizado de Máquina no Contexto das Mudanças Climáticas

Projeto Final de Curso, apresentado à Universidade Estadual Paulista (UNESP), Instituto de Ciência e Tecnologia, Sorocaba, como parte dos requisitos para obtenção do título de Bacharela em Engenharia de Controle e Automação.

Orientador: Prof. Dr. Leopoldo André Dutra Lusquino Filho

Sorocaba
2025

R696i

Rodrigues, Kailani de Assis Tavernezi

Investigação causal das doenças respiratórias e doenças transmitidas por mosquitos no Brasil por meio do aprendizado de máquina no contexto das mudanças climáticas / Kailani de Assis Tavernezi Rodrigues. -- Sorocaba, 2025

55 p. : il., tabs.

Trabalho de conclusão de curso (Bacharelado - Engenharia de Controle e Automação) - Universidade Estadual Paulista (UNESP), Instituto de Ciência e Tecnologia, Sorocaba

Orientador: Leopoldo André Dutra Lusquino Filho

1. Aprendizado do computador. 2. Causalidade. 3. Mudanças climáticas. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

IMPACTO POTENCIAL DESTA PESQUISA

Este trabalho contribui de forma inovadora para o desenvolvimento de uma metodologia prática e simplificada de inferência causal para dados epidemiológicos a fim de investigar relações causais entre o aumento de casos de doenças e variáveis ambientais e socioeconômicas no contexto das mudanças climáticas, podendo subsidiar o poder público de informações e resultados relevantes para o desenvolvimento de políticas públicas e sanitárias preventivas.

POTENTIAL IMPACT OF THIS RESEARCH

This work contributes in an innovative way to the development of a practical and simplified methodology of causal inference for epidemiological data in order to investigate causal relationships between the increase in cases of diseases and environmental and socioeconomic variables in the context of climate change, which can subsidize the public authorities with information and results relevant to the development of preventive public and health policies.

KAILANI DE ASSIS TAVERNEZI RODRIGUES

INVESTIGAÇÃO CAUSAL DAS DOENÇAS RESPIRATÓRIAS E DOENÇAS TRANSMITIDAS POR MOSQUITOS NO BRASIL POR MEIO DE APRENDIZADO DE MÁQUINA NO CONTEXTO DAS MUDANÇAS CLIMÁTICAS

Projeto Final de Curso apresentado à Universidade Estadual Paulista (UNESP), Instituto de Ciência e Tecnologia, Sorocaba, como parte dos requisitos para obtenção do título de Bacharela em Engenharia de Controle e Automação.

Data de defesa: 22/08/2025

BANCA EXAMINADORA

_____

Prof. Dr. Leopoldo André Dutra Lusquino Filho
UNESP – Instituto de Ciência e Tecnologia – Câmpus de Sorocaba

_____

Profª Dra. Fabiane Mondini
UNESP – Instituto de Ciência e Tecnologia – Câmpus de Sorocaba

_____

Dra. Andressa Maira Sardinha de Souza
UNESP – Instituto de Ciência e Tecnologia – Câmpus de Sorocaba

Dedico este trabalho aos meus pais, Edilaine e Guilherme, e amigos que me apoiaram durante toda a minha trajetória acadêmica.

AGRADECIMENTOS

O presente trabalho foi realizado com apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Brasil. Processo nº 2024/09656-0. Também contou com o apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). Processo: 444734/2023-6.

RESUMO

As mudanças climáticas têm gerado impactos diretos sobre a saúde humana, como o aumento da ocorrência de doenças respiratórias associadas a variações climáticas extremas e à poluição atmosférica, ou com a maior frequência de eventos extremos, como aumento da temperatura e da pluviosidade, impactando diretamente a saúde pública por meio da proliferação de doenças como a dengue. Este trabalho propõe uma abordagem para inferência causal em séries temporais epidemiológicas multivariadas e não lineares, considerando dados de morbidade respiratória e taxas de internação por casos de dengue de municípios brasileiros com autocorrelação espacial positiva entre 1999 e 2021. Foram utilizadas variáveis socioeconômicas, ambientais e demográficas combinadas a diferentes modelos de *machine learning*, incluindo *Random Forest*, ARIMA/ARIMAX e *Long Short-Term Memory* (LSTM), no contexto do algoritmo Aleph, que permite identificar relações causais com base em importância de variáveis e testes estatísticos por permutação. Os resultados indicaram que a *Random Forest* apresentou comportamento mais conservador na seleção de *drivers* causais, enquanto a LSTM foi mais sensível a padrões temporais, selecionando um maior número de variáveis. A rede ARIMA/ARIMAX, por sua vez, mostrou pouca capacidade de inferência causal para os dados utilizados que possuem características adversas. A comparação entre os modelos reforça a necessidade de metodologias robustas e adaptáveis para análise causal em dados epidemiológicos complexos, contribuindo para o monitoramento e prevenção de crises sanitárias associadas ao agravamento das mudanças climáticas.

Palavras-Chave: Inferência Causal; Séries Temporais; Doenças Respiratórias; Dengue.

ABSTRACT

Climate change has generated direct impacts on human health, such as the increased occurrence of respiratory diseases associated with extreme weather variations and air pollution, or the greater frequency of extreme events, such as rising temperatures and rainfall, directly impacting public health through the proliferation of diseases such as dengue. This work proposes an approach for causal inference in multivariate and nonlinear epidemiological time series, considering data on respiratory morbidity and hospitalization rates due to dengue cases from Brazilian municipalities with positive spatial autocorrelation between 1999 and 2021. Socioeconomic, environmental, and demographic variables were combined with different machine learning models, including Random Forest, ARIMA/ARIMAX, and Long Short-Term Memory (LSTM), in the context of the Aleph algorithm, which allows the identification of causal relationships based on variable importance and permutation statistical tests. The results indicated that Random Forest was more conservative in selecting causal drivers, while LSTM was more sensitive to temporal patterns, selecting a larger number of variables. The ARIMA/ARIMAX network, in turn, showed little causal inference capability for the data used, which had adverse characteristics. The comparison between the models reinforces the need for robust and adaptable methodologies for causal analysis in complex epidemiological data, contributing to the monitoring and prevention of health crises associated with worsening climate change.

Keywords: Causal Inference; Time Series; Respiratory Diseases; Dengue.

LISTA DE ILUSTRAÇÕES

Figura 1 Representação esquemática de uma *Random Forest*. 20
Figura 2 Representação de uma célula de memória de uma rede LSTM . . . 22
Figura 3 Diagrama do pipeline implementado. 27
Figura 4 Representação da série temporal sintética. 31
Figura 5 Representação da série temporal sintética reduzida. 32
Figura 6 Plot das variáveis presentes no dataset das Doenças Respiratórias para o município Alta Floresta D'Oeste (RO). 39
Figura 7 Plot das variáveis presentes no dataset da Dengue para o município Alta Floresta D'Oeste (RO). 40
Figura 8 Frequência dos *drivers* selecionados para doenças respiratórias em todos os municípios. 41
Figura 9 Frequência dos *drivers* selecionados para doenças respiratórias em municípios com mais de 500 mil habitantes. 42
Figura 10 Frequência dos *drivers* selecionados para a dengue em todos os municípios. 43
Figura 11 Frequência dos *drivers* selecionados para dengue em municípios com mais de 500 mil habitantes. 44
Figura 12 Frequência dos *drivers* selecionados para doenças respiratórias em municípios com mais de 500 mil habitantes. 45
Figura 13 Frequência dos *drivers* selecionados para dengue em municípios com mais de 500 mil habitantes. 46
Figura 14 Frequência dos drivers selecionados para a dengue em todos os municípios. 47

LISTA DE TABELAS

Tabela 1 – Resultados obtidos para o *dataset* de 3 variáveis sintéticas utilizando o *Random Forest*. 33
Tabela 2 – Resultados obtidos para o *dataset* de 10 variáveis sintéticas utilizando o *Random Forest*. 34
Tabela 3 – Resultados obtidos para o *dataset* de 10 variáveis sintéticas utilizando o LSTM, porém com apenas dois *drivers* máximos. 35
Tabela 4 – Resultados obtidos para o *dataset* de 3 variáveis sintéticas utilizando o LSTM, com 365 pontos. 35
Tabela 5 – Resultados obtidos para o dataset de 3 variáveis sintéticas utilizando o LSTM, com 1825 pontos. 36
Tabela 6 – Resultados obtidos para o dataset de 3 variáveis sintéticas utilizando o LSTM, com 3650 pontos. 36
Tabela 7 – Resultados obtidos para o *dataset* de 3 variáveis sintéticas utilizando o ARIMA/ARIMAX. 36
Tabela 8 – Resultados obtidos para o *dataset* de 10 variáveis sintéticas utilizando o ARIMA/ARIMAX. 37

SUMÁRIO

1 INTRODUÇÃO 13

2 TRABALHOS RELACIONADOS 15
2.1 Análise Causal com Séries Temporais 15
2.2 Análise Causal com Séries Temporais com ML 16
2.3 Aleph 17
2.4 Random Forest 19
2.5 LSTM 21
2.6 ARIMA/ARIMAX 23
2.7 Revisão Bibliográfica de ML em Análise Temporal Epidemiológica 24

3 METODOLOGIA 26
3.1 Tratamento dos Dados 26
3.2 Aleph 26
3.3 Inferência Causal 28

4 EXPERIMENTOS 29
4.1 Setup 29
4.2 Métricas 29
4.3 Dados Sintéticos 30
4.4 Resultados e Discussões - Dados Sintéticos 32
4.4.1 Random Forest 33
4.4.1.1 Experimento 1 33
4.4.1.2 Experimento 2 33
4.4.2 LSTM 34
4.4.2.1 Experimento 1 34
4.4.2.2 Experimento 2 35
4.4.2.3 Experimentos 3 e 4 35
4.4.3 ARIMA/ARIMAX 36
4.4.3.1 Experimento 1 36
4.4.3.2 Experimento 2 37
4.5 Dados Reais - Data SUS 37
4.6 Experimentos com os Dados Reais 38
4.7 Resultados e Discussões - Dados Reais 38
4.7.1 Random Forest 39
4.7.2 LSTM 43

4.7.3 ARIMA/ARIMAX 43
4.7.4 Resultados e Discussões - Comparativo entre os Modelos 44

5 CONCLUSÃO 48

REFERÊNCIAS 50

1 INTRODUÇÃO

As mudanças climáticas têm um impacto evidente e multifacetado no planeta, manifestando-se através do aumento das temperaturas, tempestades mais severas, secas prolongadas e elevação da temperatura dos oceanos, entre outros efeitos (Organização das Nações Unidas, 2025). Observa-se que os efeitos da crise se estendem a impactos diretos à saúde da população, inclusive com o aumento no número de casos de doenças transmitidas por vetores (Souza; Weaver, 2024), ou seja organismos vivos que transmitem doenças de animais para seres humanos, que representam mais de 17% de todas as doenças infecciosas pelo mundo, resultando em 700 mil mortes por ano (Raizada; Mala; Shankar, 2020), sendo algumas delas malária, esquistossomose ("barriga d'água"), tripanossomíase africana ("doença do sono"), leishmaniose, doença de Chagas, febre amarela e dengue, transmitida pelo mosquito Aedes, representando um risco de cerca de 3,9 bilhões de pessoas contraírem a doença por ano, em 129 países, resultando em mais 40 mil mortes por ano, se agravando em países tropicais e subtropicais como o Brasil, sobretudo em populações em maior vulnerabilidade social. Esse aumento é evidenciado pelo Sexto Relatório de Avaliação (AR6) - Relatório de Síntese das Mudanças Climáticas em 2023 (Bo-Tao; Jin, 2021), produzido pelo IPCC, que juntamente indica que, apesar de ser a parte da população que menos contribui para o agravamento da crise climática, as pessoas mais vulneráveis economicamente são as mais afetadas pelas mudanças de temperaturas no planeta (Marotzke; Semmann; Milinski, 2020).

A análise de dados de séries históricas no Brasil mostra a relação entre o aumento de casos de determinadas doenças e eventos climáticos extremos (Ellwanger et al., 2020). Seja pelo aumento do volume de chuvas e consequente desequilíbrio do ecossistema, ou pelas fortes secas que levaram ao processo migratório da população, certas doenças se espalharam em regiões onde antes não havia um número significativo de casos. O aumento constante da temperatura e a ocorrência de eventos climáticos extremos indicam um futuro com maior frequência de epidemias de doenças como dengue, zika e chikungunya no Brasil. Além disso, a Organização Pan-Americana de Saúde (OPAS) (Brasil, 2020) mostra dados que relacionam casos de dengue com o sexo do indivíduo, revelando outras variáveis importantes para a formulação de políticas públicas eficazes na minimização de epidemias. No entanto, encontrar correlações entre diversas variáveis não é suficiente para estabelecer uma relação de causalidade, tal como definida por Granger et al. (Granger, 1969): uma variável X_i causa uma saída Y_i se Y_i pode ser prevista de forma mais eficaz utilizando valores passados de X_i e não apenas os valores de Y_i.

Tendo em vista a quantidade de variáveis existentes no problema exposto, a utilização de modelos computacionais se torna crucial para a análise de séries temporais eficiente, permitindo a determinação das relações de causalidade entre estas variáveis, no entanto, as técnicas tradicionais de inferência causal para séries temporais podem não ser suficientes no contexto deste estudo, devido à complexidade dos sistemas envolvidos (Xu, 2021). A complexidade dos dados reside

em propriedades não-lineares que emergem da interação de seus componentes, e sujeitas a loops de feedback, onde uma variável A é impactada por uma variável B, mas também passa a impactar a si mesma e a variável B. Além disso, esses sistemas são suscetíveis à presença de *black swans* (Batrouni; Bertaux; Nicolle, 2018), eventos inesperados e imprevisíveis que têm grande impacto causal nas variáveis.

Técnicas de *Machine Learning* (ML) para predição de propagação de doenças transmitidas por mosquitos já vem sendo utilizadas, no entanto, apesar da eficiência de tais modelos já ter sido comprovada para a captura de padrões não-lineares em séries históricas epidemiológicas, não existem trabalhos relevantes na literatura sobre o uso da análise causal neste mesmo contexto. Dessa forma, o ML apresenta vantagens sobre os modelos tradicionais para análise de dados, como flexibilidade em trabalhar com dados de alta dimensão e dados irregulares (Cui et al., 2020). Além disso, permite o aprendizado de uma modelagem hierárquica de séries temporais, ou seja, efetivamente as dependências de curto, médio e longo prazo entre variáveis, e compreende relações de causalidade não-lineares (Luo; Peng; Ma, 2020).

Tendo em vista o cenário não-linear da análise das relações espaciais entre fatores socioeconômicos, ambientais e demográficos, usando dados de municípios brasileiros com autocorrelações espaciais positivas, determinadas pelo índice de Moran para morbimortalidade, dentre as diversas técnicas disponíveis para alcançar o objetivo da pesquisa, este trabalho focará no método Aleph (Castro et al., 2023), abordagem essa escolhida devido a sua maior simplicidade de implementação e sua maior explicabilidade, além de ser apropriada para conjuntos de dados com não-linearidades. Considerando a crescente complexidade dos modelos de aprendizado de máquina necessários para lidar com o volume de dados atual, o Aleph propõe o uso de modelos *ensemble*, como *Random Forests*, para avaliar a importância das entradas na previsão e, consequentemente, as relações causais entre variáveis.

Diferentemente do estudo desenvolvido por Castro et al. (Castro et al., 2023), este trabalho inova ao aplicar modelos distintos para análise de causalidade no cenário descrito, uma abordagem inédita até o momento. Inicialmente, foi testado o modelo *Long Short-Term Memory* (LSTM), que se mostrou eficaz e altamente sensível para o processo de inferência causal, mesmo com uma quantidade limitada de dados disponíveis para esse tipo de rede neural (Kandadi; Shankarlingam, 2025), além do custo computacional alto. Em seguida, foi utilizado o modelo *Autoregressive Integrated Moving Average* (ARIMA), que se mostrou inaqueado para o tipo de dados utilizados neste trabalho.

Este trabalho está organizado da seguinte forma: o Capítulo 2 apresenta trabalhos relacionados, contextualizando a análise causal e os modelos de redes neurais abordados. O Capítulo 3 descreve a metodologia da pesquisa, detalhando desde o tratamento de dados aplicado ao conjunto, a arquitetura do Aleph utilizada, à inferência causal feita a partir das predições. O Capítulo 4 detalha e discute os principais resultados obtidos nos experimentos realizados com os diferentes modelos e validação com dados sintéticos e os dados disponibilizados sobre os municípios brasileiros. E por fim, o Capítulo 5 traz as conclusões do estudo.

2 TRABALHOS RELACIONADOS

Devido a complexidade do cenário tratado neste trabalho, é interessante fazer uma revisão bibliográfica criteriosa englobando diversos aspectos relacionados ao tema. A fim de facilitar a compreensão dos temas e destacar de forma específica o que já existe na literatura para embasar este estudo, este capítulo está dividido entre as seções 2.1 e 2.7.

2.1 ANÁLISE CAUSAL COM SÉRIES TEMPORAIS

A definição clássica de causalidade é: X causa Y (representado por X $\rightarrow$ Y) se, e somente se, uma intervenção ou mudança feita em X provoca um efeito em Y. Nessa relação, X é chamado de variável causadora (ou *driver*), e Y é o alvo (*target*) ou o efeito (Castro et al., 2023). Ou seja, se mudar X leva a uma mudança observável em Y, podemos dizer que X tem um efeito causal sobre Y. Já correlação é definida como uma medida estatística que expressa o grau e a direção da relação linear entre duas variáveis. Quando duas variáveis variam juntas de forma consistente diz-se que elas estão correlacionadas. A correlação pode ser positiva, quando ambas aumentam ou diminuem juntas, negativa, quando uma variável aumenta a outra diminui ou nula, que é quando não há relação linear clara entre as variáveis (Moore; McCabe; Craig, 2009). Portanto, duas variáveis podem ser correlatas mas não causais, uma vez que correlação diz apenas sobre associações, mas não sobre direção e influência causal entre as variáveis.

Uma das primeiras abordagens a determinar causalidade é proposta por Granger et al. (Granger, 1969), que é baseada na simples ideia de que a causa, ao mesmo tempo que precede o efeito, ajuda a prevê-lo (Seth; Barrett; Barnett, 2015). A causalidade de Granger é um teste estatístico usado para verificar se uma série temporal X ajuda a prever uma série temporal Y, baseado na análise da hipótese de os valores passados de X, além dos dos próprios valores passados de Y, ajudarem a prever os valores de Y, caso isso seja verdade, diz-se que X "causa" Y no sentido de Granger (Granger, 1969).

Formalmente, esse tipo de relação pode ser representado por um modelo autorregressivo bivariado, que considera a influência das defasagens de ambas as variáveis e, eventualmente, também de seus valores contemporâneos. A seguir, apresentam-se as equações 1 e 2 que descrevem esse sistema:

$$X_t + b_0 Y_t = \sum_{j=1}^{m} a_j X_{t-j} + \sum_{j=1}^{m} b_j Y_{t-j} + \varepsilon_t' \tag{1}$$

$$Y_t + c_0 X_t = \sum_{j=1}^{m} c_j X_{t-j} + \sum_{j=1}^{m} d_j Y_{t-j} + \varepsilon_t''. \tag{2}$$

A causalidade de Granger é amplamente aplicada em diversas áreas. Na economia, por

exemplo, é usada para investigar se as taxas de juros influenciam a inflação, como discutido por Stock & Watson. (Stock; Watson, 2001). A ideia é verificar se variações passadas em juros ajudam a prever mudanças futuras na inflação. Na neurociência, Seth et al. (Seth; Barrett; Barnett, 2015) aplicam a técnica para identificar relações direcionais entre regiões cerebrais com base em registros temporais de EEG ou fMRI. Aqui, o cuidado central é garantir a estacionariedade dos sinais e evitar interpretações errôneas em presença de variáveis ocultas. Em ciências ambientais, estudos como Liang et al. (Liang, 2015) exploram causalidade entre temperatura oceânica e padrões de precipitação.

Entretanto, a causalidade de Granger possui limitações. Primeiramente, é suposto linearidade, ou seja, as relações entre as variáveis são descritas por combinações lineares dos valores passados, assumindo que a influência de X sobre Y pode ser representada por uma soma ponderada dos valores anteriores das variáveis, sem interações ou efeitos não-lineares (Barnett; Barrett; Seth, 2009). A causalidade de Granger requer séries estacionárias, ou seja, possui propriedades estatísticas como média, variância e autocorrelação que não mudam com o tempo (Hamilton, 1994). Além disso, não distingue entre correlação indireta (via uma terceira variável) e causalidade direta, que seria quando duas variáveis parecem fortemente correlacionadas mas não necessariamente tenham uma relação causal entre elas, na verdade existe uma terceira variável que influencia ambas as variáveis (Pearl, 2003). Para que o conceito seja mais compreensível, considere o exemplo de que no verão, tanto os casos de afogamento quanto as vendas de sorvete aumentam, em uma análise simplória, poderia se dizer que o consumo de sorvete causa afogamentos, entretanto, existe uma terceira variável que é o aumento da temperatura durante o verão, que justifica muito melhor o aumento de afogamentos, pois as pessoas tendem a entrar no mar com mais frequência para se refrescar, e também tendem a comprar mais sorvetes com temperaturas mais altas, logo, casos de afogamentos e vendas de sorvete são correlacionadas, e ambas causadas pelo aumento da temperatura. Por fim, é sensível ao número de lags, ou seja, valores passados da série temporal, e à seleção incorreta do modelo autorregressivo linear (como VAR).

Tendo em vista tais limitações, em contextos mais complexos como para dados ambientais, os métodos tradicionais de análise causal se tornam insuficientes para capturar dinâmicas não-lineares, interações de alta ordem e relações temporais mais sutis. Nesse cenário, técnicas baseadas em *Machine Learning* (ML) e *Deep Learning* (DL) vem sendo mais utilizadas para modelar tais padrões complexos, oferecendo alternativas mais robustas para a inferência causal nestes contextos, devido a capacidade de tais redes de lidar com séries temporais de alta dimensionalidade, não lineares e não estacionárias.

2.2 ANÁLISE CAUSAL COM SÉRIES TEMPORAIS COM ML

A aplicação de *Machine Learning* para a inferência causal em séries temporais surgiu como resposta às limitações dos métodos tradicionais. A princípio, as abordagens focavam em modelos

de aprendizado supervisionado, como árvores de decisão e *Random Forests*, que já possibilitam a análise de dados de alta dimensionalidade e não-linearidades de forma mais efetiva. Runge et al. (Runge et al., 2019), propuseram uma estrutura de inferência causal baseada em decomposição de informações e aprendizado de grafos causais, combinando testes estatísticos com seleção de variáveis por meio de algoritmos de aprendizado.

Com o avanço dos estudos na área de redes neurais, surgiram métodos baseados em arquiteturas profundas. O *Neural Granger Causality* (Tank et al., 2018) foi um marco importante, pois reformulou o teste de Granger utilizando Redes Neurais Recorrentes e *Multi Layer Perceptron*, permitindo a captura de dependências não-lineares entre variáveis. Tal ideia evoluiu ao se utilizar *Attention Mechanisms*, redes *Transformers* e Autoencoders Variacionais para inferência causal, como em Nauta et al. (Nauta; Bucur; Seifert, 2019) que propôs o *Temporal Causal Discovery Framework* (TCDF) que utiliza redes neurais convolucionais com mecanismos de atenção para identificar causas temporais e seus respectivos atrasos, aprendendo grafos causais diretamente a partir dos dados.

Redes neurais profundas foram utilizadas por Ahmad et al. (Ahmad; Shadaydeh; Denzler, 2021) combinadas à técnica de *knockoff* para inferência causal em séries temporais não-lineares, onde os *knockoffs* geram contrafactuais para realizar os testes de hipótese nas relações causais, o que vai além dos limites da causalidade de Granger tradicional. Mais recentemente também abordaram a inferência causal em sistemas de séries temporais não-lineares multivariadas, como em dados climáticos e dados neurológicos, através da metodologia de inferência baseada em grupos através do algoritmo DeepAR, entretanto, assumindo estacionariedade, suficiência causal e invariância no modelo (Ahmad; Shadaydeh; Denzler, 2024).

Por outro lado, Faruque et al. (Faruque et al., 2024) desenvolve uma técnica de deep learning chamada *Time-Series Causal Neural Network* (TS-CausalNN) para descobrir relações causais atuais e passadas simultaneamente, utilizando uma arquitetura que combina blocos convolucionais com camadas causais personalizadas, com a vantagem de que essa abordagem lida naturalmente com dados não-lineares e não estacionários.

Apesar dos trabalhos apresentados que utilizam abordagens de *machine* e *deep learning*, tais abordagens ainda são pouco exploradas para a inferência causal em séries temporais multivariadas, não-lineares e não estacionárias. Nesse sentido, um estudo foi desenvolvido para a elaboração de um algoritmo que tem como objetivo realizar a descoberta causal, estabelecendo as relações causais incluindo *lags* de tempo (Castro et al., 2023).

2.3 ALEPH

O trabalho elaborado por Castro et al. (Castro et al., 2023) foca na inferência causal de séries temporais levando em consideração a descoberta causal com defasagens temporais, que diz sobre o atraso de tempo entre a causa e o efeito. A ideia de inferir causalidade é baseada em poder identificar quais preditores, dentro de um grupo de variáveis, o algoritmo considera mais

importantes na predição de uma variável alvo (*target*), realizando testes estatísticos para avaliar a significância dessas predições e poder estabelecer relações causais baseadas nesses resultados.

O pipeline base dessa metodologia, que foi nomeada de Aleph, considera as seguintes etapas:

1) Obter um *baseline*, que é o ponto de partida para a avaliação da métrica, a partir da predição da *target* utilizando somente os valores presentes da própria *target* como preditora.

2) Utilizar as demais variáveis (*features*) como preditoras da *target*, sendo consideradas como potenciais *drivers*, ou seja, *features* que possivelmente tem uma relação causal com a *target*.

O algoritmo completo deste método é baseado em 11 passos:

I) Dentre as variáveis do dataset, escolhe-se a *target*;

II) É criado um vetor com os valores passados da *target*;

III) Os valores do vetor criado em II são utilizados para treinar o modelo ML e prever o valor atual da *target*;

IV) Com a predição realizada em III, é definido o *baseline* inicial a partir da métrica utilizada na predição;

V) Um conjunto de *drivers* potenciais é escolhido;

VI) Se o conjunto definido em V não é vazio, uma variável destas é escolhida;

VII) Um vetor é criado usando os valores passados da variável escolhida em VI;

VIII) Utilizando o vetor de valores passados da *target* criado em II, os valores dos vetores criados na etapa VII para os valores passados dos *drivers* selecionados na etapa X, e o vetor criado em VII para a variável dos *drivers* potenciais em análise neste momento, o modelo ML é treinado e o valor atual da *target* é previsto.

IX) A métrica da predição é feita em VIII é comparada ao *baseline* atual, se esse valor é melhor do que o *baseline*, e o teste estatístico de significância é aprovado, avança-se para a etapa X, caso não, pula-se para a etapa XI.

X) O *driver* testado é adicionado à lista de *drivers* selecionados.

XI) É considerado que essa variável não é um *driver* para a *target* e ele então é ignorado.

Os passos V a XI são realizados até que todas as variáveis no conjunto de potenciais *drivers* tenham sido testadas.

Ao comparar os métodos de inferência causal com séries temporais baseados em *machine learning* descritos na literatura recente com o método proposto por Castro et al. (Castro et al.,

2023), observa-se que há diferenças significativas tanto na complexidade quanto na abordagem metodológica. As técnicas baseadas em redes neurais profundas, como o Neural Granger Causality (Tank et al., 2018), o TCDF (Nauta; Bucur; Seifert, 2019) e o TS-CausalNN (Faruque et al., 2024), são projetadas para capturar dependências temporais complexas e não-lineares, muitas vezes combinando mecanismos como atenção, convoluções e geração de contrafactuais. Essas abordagens oferecem alta flexibilidade e desempenho em cenários de grande volume de dados e elevada complexidade estrutural. Contudo, apresentam limitações associadas à baixa interpretabilidade, alto custo computacional e maior exigência de dados para treinamento eficaz.

Por outro lado, o método Aleph adota uma estratégia mais simples e interpretável, baseada na utilização de modelos de *ensemble*, como *Random Forest* e CatBoost, em conjunto com métricas de importância de variáveis e testes estatísticos via permutação. Essa abordagem permite capturar relações causais não-lineares de forma direta, especialmente por meio da comparação entre modelos que utilizam apenas o histórico da variável alvo e modelos que incorporam o histórico de potenciais variáveis causadoras. Apesar de não modelar explicitamente contrafactuais, nem capturar representações latentes como as redes profundas, o método se destaca por ser mais robusto em cenários de menor volume de dados, além de oferecer maior transparência e interpretabilidade dos resultados. Dessa forma, representa uma alternativa metodológica viável, particularmente em contextos aplicados que exigem simplicidade, clareza na análise e eficiência computacional.

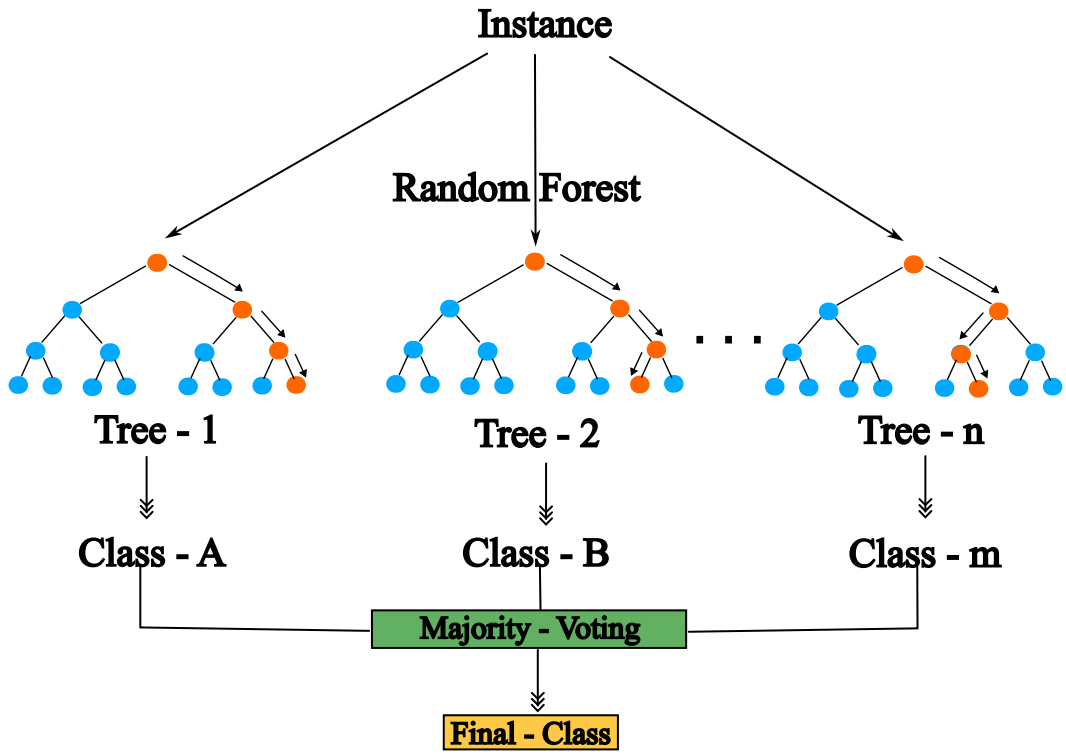
Para a determinação de potenciais *drivers*, qualquer mecanismo de regressão serviria, entretanto, para a descoberta causal proposta se faz necessário a utilização de algoritmos de *machine learning* que quantifique a importância de cada *feature* na predição da *target*, como *Random Forest*, que é um algoritmo *ensemble* que fornecem a informação sobre a importância da variável. Tendo em vista isso, é cabível explorar diferentes modelos de redes neurais para a realização da tarefa proposta pelo Aleph, como *Random Forest*, *Long Short-Term Memory* (LSTM) e *AutoRegressive Integrated Moving Average* (ARIMA).

2.4 RANDOM FOREST

Os modelos *Random Forest* são algoritmos de aprendizado de máquina supervisionados, classificado como *ensemble learning*, baseados em árvores de decisão, que são modelos usados para classificação e regressão, com um funcionamento baseado em uma estrutura hierárquica semelhante a um árvore, onde os dados são divididos repetidamente com base em condições simples de decisão, representadas por nós internos da árvores, e cada ramo desses nós representa o resultado da condição, tendo o nó folha do ramo como a previsão final (Mitchell, 1999). A Figura 1 traz um diagrama representativo de uma *Random Forest*.

A ideia central das *Random Forests* é construir um grande número de árvores de decisão durante o treinamento e, em seguida, agregá-las (*ensemble*) para produzir a previsão final. Quando se trata de uma tarefa de regressão, é utilizada a média das saídas das árvores, quando

Figura 1 – Representação esquemática de uma *Random Forest*.



Fonte: Criação própria
Legenda: Os círculos representam os nós das árvores de decisão, e os traços que os ligam os ramos. o nó folha é representado pelo último círculo laranja.

se trata de uma tarefa de classificação, é utilizado um sistema de votação majoritária (*majority vote*).

Random Forests tem se destacado na análise e previsão de séries temporais multivariadas, sua capacidade de capturar relações não lineares e sua robustez frente a ruídos e colineraridades entre variáveis o tornam adequado para diversas áreas de conhecimento. Em ciências ambientais, por exemplo, vêm sendo utilizados para modelar processos complexos, como no estudo de Zeng et al. (Li et al., 2020) onde um modelo *Random Forest* foi empregado em conjunto com técnicas de permutação para inferência causal entre umidade do solo e precipitação, demonstrando a utilidade do modelo em contextos climáticos complexos. No contexto epidemiológico, são aplicados na modelagem de surtos de doenças infectocontagiosas e na compreensão de fatores de riscos associados, como para prever a incidência de diarreia infecciosa em uma província chinesa (Fang et al., 2020), incorporando variáveis climáticas, defasagens temporais e sazonais, ou para estimar frações atribuíveis à população em um estudo caso-controle de infecções por *Salmonella enterica* (Gu et al., 2015), onde o modelo se mostrou eficaz em sistematizar exposições hierarquizadas e complexas, fornecendo estimativas robustas de risco populacional. Bontempi et al. (Bontempi; Taieb; Borgne, 2012) também discutem estratégias de aprendizado de máquina para previsão

de séries temporais, destacando o papel dos métodos baseados em árvores como alternativas promissoras aos modelos lineares tradicionais, especialmente em contextos onde o conhecimento prévio sobre a estrutura dos dados é limitado.

Além disso, o modelo *Random Forest* tem sido explorado em abordagens de inferência causal, particularmente por meio da técnica de *permutation importance*, que permite estimar o quanto a aleatorização de uma variável reduz a capacidade preditiva do modelo, fornecendo uma medida indireta de sua influência causal. Essa abordagem tem ganhado relevância em contextos onde há interesse em identificar variáveis com forte contribuição preditiva, mesmo em cenários altamente não lineares.

2.5 LSTM

Redes LSTM são uma classe de Redes Neurais Recorrentes (RNNs), projetadas para superar dificuldades que as RNNs tradicionais têm ao aprender dependências de longo prazo, chamado de desvanecimento ou explosão do gradiente (Bengio; Simard; Frasconi, 1994). Apresentada por Hochreiter & Schmidhuber (Hochreiter; Schmidhuber, 1997), redes LSTM utilizam arquiteturas baseadas em portas que controlam o fluxo de dados dentro da célula de memória, permitindo que a rede armazene e recupere informações relevantes por longos intervalos de tempo. A Figura 2 ilustra as portas dentro de um célula de memória, com os inputs, output e funções de ativação.

A célula LSTM é composta por três portas principais e dois estados, porta de esquecimento (f_t), que decide quais informações do estado anterior da célula devem ser esquecidas, estado candidato da célula ($\tilde{c}_t$), que gera novos conteúdos candidatos que podem ser adicionados à memória da célula se a porta de entrada (i_t) decidir que as informações novas devem ser armazenadas na célula de memória, estado da célula (c_t) que é a memória a longo prazo da rede, que vai sendo atualizada ao longo do tempo com base no que é esquecido e no que é aprendido e a porta de saída (o_t), que decide quais partes do estado da célula atual vão ser expostas como saída da célula naquele instante (h_t), que serve tanto de *output* atual da rede quanto como entrada para o próximo passo.

As equações 3 a 8 e descrevem a dinâmica de uma célula LSTM no instante de tempo *t* são:

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \tag{3}$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \tag{4}$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \tag{5}$$

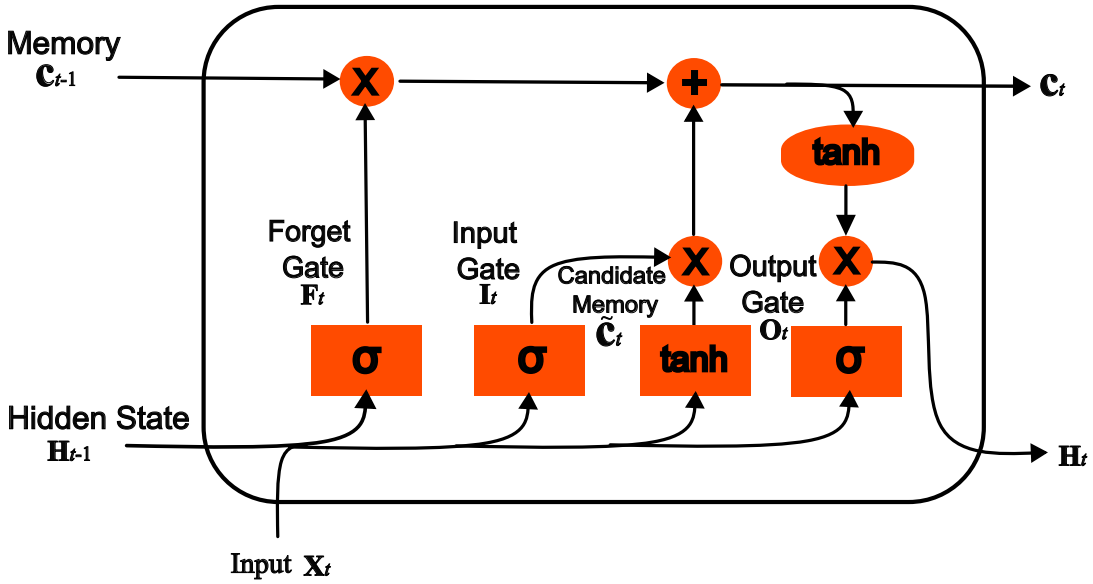
$$\tilde{c}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c) \tag{6}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{7}$$

$$h_t = o_t \odot \tanh(c_t) \tag{8}$$

Onde, x_t é a entrada no tempo *t*, h_{t-1} é o estado oculto anterior, σ é a função sigmóide

Figura 2 – Representação de uma célula de memória de uma rede LSTM



Fonte: Criação própria
Legenda: σ representa a função de ativação sigmóide, *tanh* a função de ativação tangente hiperbólica e os círculos representam os operadores.

logística, $\odot$ denota a multiplicação elemento a elemento (produto de Hadamard) e W, U e b são os pesos e vieses aprendidos durante o treinamento.

Devido à capacidade de capturar dependências de longo prazo e lidar com dados sequenciais de forma eficiente, redes LSTM vem sendo aplicadas em diversos domínios como finanças, energia e clima, sendo utilizadas para prever preços de ações e volatilidade do mercado (Fischer; Krauss, 2018), modelar o consumo e demanda de energética (Marino; Amarasinghe; Manic, 2016) e prever variações meteorológicas e padrões sazonais utilizando redes convolucionais (Shi et al., 2015). Além disso, LSTMs também vem sendo aplicadas em tarefas de inferência causal em séries temporais, sobretudo em cenários não lineares e multivariados, como proposto por Tank et al. (Tank et al., 2018) que combina a estrutura de redes neurais com o critério de Granger para inferir relações causais ao treinar uma rede LSTM com penalidades específicas que promovem a seleção de variáveis causais relevantes.

Apesar das vantagens apresentadas, o uso de redes LSTM em contextos como o do presente trabalho, pode apresentar limitações práticas importantes. A principal delas está relacionada ao alto custo computacional associado ao treinamento dessas redes, especialmente quando múltiplas camadas e unidades ocultas são utilizadas (Greff et al., 2016). Além disso, o desempenho da

LSTM é altamente dependente de uma quantidade adequada de dados para capturar padrões temporais com robustez. Em cenários com séries temporais curtas, lacunas nos dados ou janelas de observação reduzidas, como é comum em municípios com baixa densidade populacional ou registros incompletos, a rede pode sofrer com *overfitting* ou incapacidade de aprender padrões significativos (Karim et al., 2017; A.; Mahmood, 2019), prejudicando tanto a acurácia quanto a interpretação causal dos resultados. Assim, embora as LSTM sejam poderosas, seu uso exige cuidados quando aplicadas a domínios com recursos limitados, como é o caso de dados epidemiológicos esparsos.

2.6 ARIMA/ARIMAX

O modelo ARIMA é um modelo autoregressivo, ou seja, baseado a ideia de que o valor atual de uma série pode ser explicado pelo seus p valores passados, onde p determina o número de passos no passado são necessários para prever o valor atual (Hyndman; Athanasopoulos, 2021), que é composto por três elementos fundamentais, representados pelos parâmetros (p, d, q), que resultam em um modelo que é utilizado para modelar séries temporais que se tornam estacionárias após d diferenciações.

A primeira componente é a componente Autoregressiva, representada pelo parâmetro p, que refere-se a parte do modelo que explica que o valor atual da série através de uma combinação linear de seus próprios valores passados. Isso pode ser expresso matematicamente pela equação 9:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t \tag{9}$$

Onde ϕ são os coeficientes de regressão e ε_t é termo de erro aleatório.

A componente Integrada, representada pelo parâmetro d, representa o número de diferenciações necessárias para tornar a série estacionária, ou seja, estabilizar a média e remover tendências. A diferenciação de primeira ordem pode ser representada pela equação 10:

$$y'_t = y_t - y_{t-1} \tag{10}$$

Por fim, a componente de Média Móvel, representado pelo parâmetro q, relaciona o valor atual da série com uma média móvel dos erros passados, representado pela equação 11:

$$y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} \tag{11}$$

Onde ϕ são os coeficientes da média móvel e μ é a média da série.

O modelo ARIMA é comumente utilizado para modelar séries temporais univariadas, sobretudo quando se pretende considerar o contexto histórico da série na inferência de valores futuros. De acordo com Box et al. (Box Gwilym M. Jenkins; Ljung, 2015), é uma das abordagens mais consolidadas da estatística aplicada a séries temporais. É um modelo especialmente útil pois pelo fato de a estacionariedade da série ser essencial no uso do ARIMA, porque os métodos de

inferência utilizados na estimação do modelo partem do pressuposto de que o processo gerador dos dados é estável (Wei, 2006), o próprio modelo pode garantir a estacionariedade da série através das d diferenciações sucessivas aplicadas.

Na economia, o ARIMA é frequentemente empregado para prever indicadores macroeconômicos, como feito por Box et al. (Box et al., 1976) que demonstraram o uso do ARIMA para modelar séries econômicas, e seu método se tornou base para previsões em bancos centrais e instituições financeiras. Na demografia, o modelo é bastante utilizado para projetar populações e taxas de natalidade e mortalidade, como aplicado por Hyndman & Booth (Hynsman; Booth, 2008) para projetar séries demográficas em combinação com métodos de decomposição. Na engenharia, o ARIMA é aplicado para controle de processos, manutenção preditiva e previsão de consumo energético. Um exemplo de uso do modelo é para previsão de cargas elétricas em sistemas de energia como apresentado por Hippert et al. (Hippert; Pedreira; Souza, 2001), que avaliaram a eficácia do modelo e de redes neurais em tal aplicação.

Uma extensão importante do modelo ARIMA é o ARIMAX (*AutoRegressive Integrated Moving Average with eXogenous variables*), que, enquanto o modelo tradicional depende apenas da própria série temporal para as previsões, o ARIMAX permite incluir informações externas que podem influenciar diretamente no comportamento da série analisada, o que permite identificar o impacto que outras variáveis com dinâmica própria podem ter na variável de interesse. Matematicamente pode ser representado pela equação 12:

$$y_t = \beta_0 + \sum_{i=1}^{p} \phi_i \, y_{t-i} + \sum_{j=1}^{k} \beta_j \, x_{j,t} + \varepsilon_t + \sum_{l=1}^{q} \theta_l \, \varepsilon_{t-l} \tag{12}$$

Onde X_t representa a(s) variável(eis) exógena(s), β_i são os coeficientes associados a essas variáveis, e os demais termos seguem a notação do modelo ARIMA tradicional. A incorporação de variáveis exógenas pode melhorar substancialmente a acurácia das previsões quando há uma relação causal clara e mensurável entre as variáveis (Box et al., 1976).

2.7 REVISÃO BIBLIOGRÁFICA DE ML EM ANÁLISE TEMPORAL EPIDEMIOLÓGICA

Métodos baseados em *machine learning* estão sendo utilizados também no contexto epidemiológico. Técnicas como *Random Forest*, *Support Vector Machine* (SVM), redes neurais profundas como LSTM e XGBoost têm sido amplamente empregadas na previsão da disseminação de doenças infecciosas, análise de risco e detecção precoce de surtos. Tais modelos se mostram particularmente úteis na integração de múltiplas fontes de dados heterogêneos, como registros clínicos, mobilidade populacional e dados ambientais, e na captura de dinâmicas não lineares entre variáveis.

No contexto da pandemia de COVID-19, Yang et al. (Yang et al., 2020) utilizam um modelo SEIR (*Susceptible-Exposed-Infectious-Recovered*), que é um modelo compartimental clássico na epidemiologia, modificado para considerar efeitos das intervenções de saúde pública, como

a quarentena, que serve de base determinística para representar a dinâmica da transmissão da doença, aplicado em conjunto com o modelo LSTM para prever o número de novas infecções da doença ao longo do tempo na China. Ainda no recorte geográfico chinês, redes LSTM foram comparadas a modelos SARIMA, SARIMA-LSTM e uma versão híbrida com análise de componentes singulares (SSA), em um estudo na província de Shanxi, para determinar o melhor desempenho na previsão de casos de influenza (Zhao et al., 2023).

Outras comparações entre modelos foram realizadas com o intuito de determinar a forma mais eficaz de prever surtos de doenças. Um exemplo é o trabalho de Kane et al. (Kane et al., 2014), que comparou modelos ARIMA e *Random Forest* para prever surtos de influenza aviária H5N1 no Egito, resultando que a *Random Forest* obteve melhores resultados no erro médio quadrático e na capacidade preditiva. Também notou-se a superioridade das *Random Forest* com os modelos ARIMA e ARIMAX na previsão da incidência de doença diarreica infecciosa na província de Jiangsu, China (Zhou; Sun; Zheng, 2020). Neste trabalho, incorporou-se variáveis meteorológicas defasadas e dados de casos anteriores ao modelo *Random Forest*, que apresentou um erro percentual médio (MAPE) superior ao apresentado pelos modelos baseados em ARIMA/ARIMAX.

Entretanto, apesar de já existirem inúmeros trabalhos na literatura que utilizam *machine learning* para previsão de surtos, mortalidade ou disseminação de doenças, estudos que utilizam de tais técnicas para inferência causal no contexto epidemiológico, a fim de não só poder estimar surtos de com base nos casos da doença mas sim considerar outros fatores que podem ter relações causais com tais surtos, por exemplo, ainda são emergentes e desafiadores, tendendo a exigir a combinação de técnicas estatísticas clássicas e métodos modernos.

3 METODOLOGIA

A execução dos experimentos deste trabalho depende de uma série de etapas de trabalho, o primeiro passo é o tratamento dos dados, que precisam ser trabalhados para que estejam balanceados e no formato correto para que a inferência causal utilizando *machine learning* seja possível, os detalhes deste tratamento estão na seção 3.1. Depois de tratados, os dados passam pelo algoritmo do Aleph que faz as predições e com os resultados da métrica realiza a inferência causal, os detalhes deste algoritmo estão na seção 3.2 com o diagrama da Figura 3 ilustrando o pipeline do Aleph. A interpretação da inferência causal resultante do algoritmo está descrita na seção 3.3.

3.1 TRATAMENTO DOS DADOS

Para que os resultados da inferência causal sejam garantidos como corretos, é necessário que os dados utilizados estejam dentro de padrões. Um ponto muito importante é a granularidade dos dados, ou seja, a periodicidade deles, que deve ser suficiente para subsidiar eficientemente o modelo que fará a inferência causal. Tendo-se uma periodicidade pequena, o dataset é limitante, portanto, se faz necessário que sejam aplicadas técnicas de interpolação como Jittering (Flores; Tito-Chura; Apaza-Alanoca, 2021) e Spline Cúbica (McKinley; Levine, 2014) para aumentar a granularidade, ou seja, digamos que para o período de um mês a quantidade de dados disponíveis seja pequena, portanto, é necessário interpolar dados sintéticos para que nesse período hajam mais dados disponíveis. Além disso, a distribuição de dados deve ser balanceada para que o modelo preditivo não tire conclusões errôneas baseadas em um conjunto específico de dados em um intervalo de tempo. E por fim, o Aleph trabalha somente com dados numéricos, então toda informação não pertinente ou não numérica deve ser retirada antes do treinamento do modelo, como informações de data e códigos de identificação de municípios.

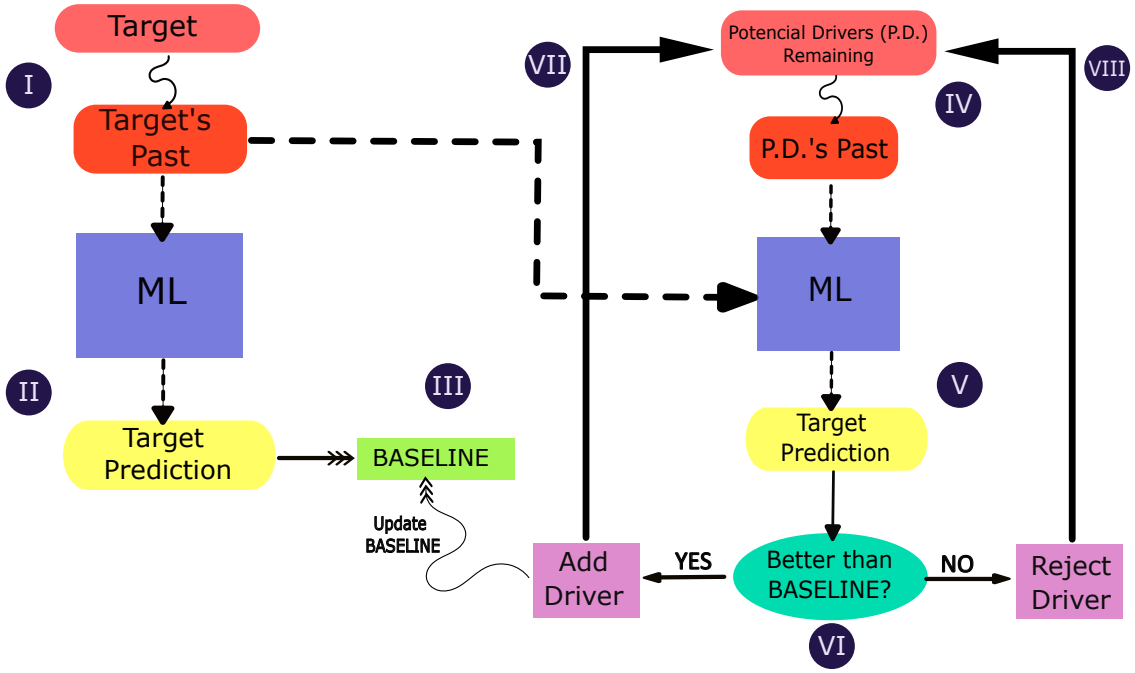
3.2 ALEPH

Uma vez que os dados tenham sido tratados, é possível aplicá-los à arquitetura do Aleph. Em resumo, essa arquitetura propõe que, uma vez definido o *baseline* da métrica escolhida, através da predição univariada, que considera somente a *target* e seus valores passados para prever a *target*, outras variáveis são escolhidas dentro de um grupo de possíveis *drivers* e essas são testadas uma a uma para verificar se a adição delas na predição da *target* apresentou melhora na métrica com relação ao *baseline* ou não. Caso essa nova variável tenha melhorado a métrica, ela é submetida à um teste estatístico a fim de verificar se a hipótese nula (h_0) é rejeitada, e então é adicionada como *driver* da *target*.

Esse teste estatístico consiste em embaralhar essa variável aleatóriamente dentro do *dataset* e realizar novamente o treinamento, fazendo isso repetidamente, em um número de iterações pré estabelecido, e avaliar se ao fim desses repetidos embaralhamentos houve piora na previsão da *target* através de um teste probabilístico de significância. Caso isso seja verdade, então a hipótese nula é rejeitada e a melhora da métrica não foi uma coincidência, confirmando a hipótese alternativa que diz que a variável tem informação causal real, e essa variável é adicionada à lista de *drivers*, e o *baseline* é atualizado para o valor da métrica gerado por esse *driver*, e uma nova variável é testada nesse mesmo processo, até que todas tenham sido testadas e a lista final de *drivers* selecionados tenha sido gerada.

O diagrama da Figura 3 ilustra essa sequência de execuções:

Figura 3 – Diagrama do pipeline implementado.



Fonte: Criação própria

Legenda: I) Selecionar a variável *target* e os seus valores passados. II) Utilizar os valores da *target* para treinar o modelo ML e prever a *target*. III) Estabelecer o *baseline*. IV) Das variáveis restantes, selecionar um potencial *driver* e seus valores passados. V) Utilizar os valores passados da *target* e os valores do potencial *driver* para treinar e prever a *target*. VI) Avaliar se a inclusão do potencial *driver* melhorou a métrica na predição. VII) Caso sim, houve melhora, a variável é adicionada à lista de *drivers* selecionados, e o *baseline* é atualizado, e os passos IV, V e VI são feitos novamente até que o conjunto esteja vazio. VIII) Caso não, não houve melhora, a variável é rejeitada como *driver* e os passos IV, V e VI são feitos novamente até que o conjunto esteja vazio.

Como dito na seção 2.3, o Aleph comporta diferentes modelos de *machine learning*, desde

que quantifique a importância de cada variável na predição, também é necessário que o modelo lide com séries temporais multivariadas, não estacionárias e não-lineares. A arquitetura utilizada não difere o modelo na sua capacidade de realizar a inferência causal, eventualmente um modelo pode se mostrar superior a outro em um determinado conjunto de dados, entretanto, o *pipeline* do Aleph garante a inferência causal de toda forma, o que permite que diferentes topologias, como sejam exploradas para que se obtenham dados comparativos.

3.3 INFERÊNCIA CAUSAL

Para o contexto deste trabalho, o Aleph foi estruturado de forma a fornecer como resultado os valores de *baseline* obtidos com as iterações dos potenciais *drivers* e os nomes das variáveis que foram selecionadas como *drivers*, dessa forma, se torna fácil compreender quais as relações causais estabelecidas através das predições. Além disso, um documento de saída foi gerado para o acompanhamento do treinamento, registrando todos os municípios que foram treinados, identificando a *target* e quais variáveis foram consideradas nas predições, os valores de métrica geradas para cada potencial *driver* testado e o valor do *baseline* atual, também foi gerado um documento que registra a *target*, todos os *drivers* selecionados para cada município, o *baseline* atual (que corresponde ao valor da métrica gerado por esse *driver*) e o resultado do teste estatístico de hipótese.

4 EXPERIMENTOS

4.1 SETUP

A aluna utilizou o próprio notebook para realizar os experimentos, com os seguintes recursos disponíveis:

- **Processador:** 11th Gen Intel® Core™ i7-1165G7 @ 2.80GHz
- **Sistema:** Intel® 3501 x64-based PC
- **Placa de vídeo:** Intel® Iris® Xe Graphics
- **Memória RAM:** 8GB
- **Armazenamento:** 256GB ADATA IM2P33F3A SSD NVMe (PCIe)

4.2 MÉTRICAS

Dentre inúmeras métricas disponíveis para avaliação das predições realizadas pelo modelo ML, a métrica escolhida para os experimentos realizados foi a *Symmetric Mean Absolute Percentage Error* (SMAPE), que busca quantificar o erro percentual entre os valores previstos e os valores reais de forma simétrica, tratando os erros de superestimação e subestimação de forma equilibrada.

Matematicamente o SMAPE é definida pela equação 1:

$$\text{SMAPE} = \frac{100\%}{n} \sum_{t=1}^{n} \frac{|y_t - \hat{y}_t|}{\frac{|y_t| + |\hat{y}_t|}{2}} \tag{1}$$

sendo, y_t o valor real da saída (*target*) no instante *t*, $\hat{y}_t$ o valor previsto da saída no instante *t* e n o número total de observações, ou seja, a janela de tempo em que se pretende ter previsões da saída.

O SMAPE calcula o erro absoluto entre o valor real e o valor previsto e o normaliza pela média dos valores absolutos real e previsto, fazendo com que essa métrica seja menos sensível a escalas e mais equilibrado quando os valores são muito baixos ou muito altos. O resultado da equação 1 é um valor percentual simétrico, que permite visualizar se o erro percentual foi positivo ou negativo, o que torna muito mais fácil a interpretação dos resultados quando comparado a métricas não relativas como MAE ou MSE.

Para determinados experimentos, também foram calculados os valores de *Recall* (também conhecido como sensibilidade) e Especificidade, com o objetivo de mensurar a capacidade do modelo de identificar corretamente os casos positivos e negativos, pois em alguns experimentos o modelo demonstrou algumas falhas nas identificações dos *drivers*, sendo nesse cenário, um caso

positivo um *driver* identificado que de fato faz parte da equação da *target* (como por exemplo, na equação 5 , da seção 4.3, onde a variável X3 de fato está na equação e caso fosse identificada como *driver* estaria correta) e um caso negativo um *driver* não identificado que de fato não está na equação da *target* (como por exemplo, na equação 5, da seção 4.3, a variável X2 não está, portanto, se ela não foi identificada como *driver* estaria correta). As equações 2 e 3 demonstram como essas métricas são calculadas.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{Especificidade} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{3}$$

Um verdadeiro positivo seria a identificação correta de um *driver* que está na equação da *target*, um falso positivo seria a identificação errônea de um *driver* que não está na equação da *target*, um verdadeiro negativo seria a não identificação de um *driver* que não está na equação da *target* e um falso negativo seria a não identificação de um *driver* que não está na equação da *target*. O valor resultante dessas métricas é um valor de 0 a 1, sendo 1 para o *recall* e para a especificidade uma taxa de acerto de 100%.

4.3 DADOS SINTÉTICOS

A fim de testar a topologia com dados conhecidos, optou-se inicialmente realizar experimentos utilizando dados sintéticos gerados a partir das equações, apresentadas em *Time series causal relationships discovery through feature importance and ensemble models* (Castro et al., 2023), e com a aplicação de sementes, que garantem que mesmo utilizando funções randômicas para a geração dos dados, os dados serão sempre os mesmos em todos os experimentos. As equações estão apresentadas abaixo:

$$X_t^0 = 0.3X_{t-1}^0 + 0.4X_{t-2}^0 + \eta^0 \tag{4}$$

$$\begin{aligned} X_t^1 = {} & 0.4X_{t-1}^1 + 0.5X_{t-4}^1 + 0.3(X_{t-2}^1)^2 + 0.4X_{t-5}^0 \\ & + 0.3(X_{t-9}^0)^2 + 0.3(X_{t-1}^1)^2 + 0.4X_{t-7}^4 + 0.4X_{t-3}^6 + 0.6X_{t-8}^6 + \eta^1 \end{aligned} \tag{5}$$

$$X_t^2 = 0.4X_{t-1}^2 + 0.2X_{t-2}^2 + \eta^2 \tag{6}$$

$$X_t^3 = 0.1X_{t-3}^3 + 0.3X_{t-2}^3 + 0.4X_{t-6}^2 + \eta^3 \tag{7}$$

$$X_t^4 = 0.3X_{t-4}^4 + 0.3X_{t-1}^4 + \eta^4 \tag{8}$$

$$X_t^5 = 0.5X_{t-5}^5 + 0.2X_{t-5}^5 + \eta^5 \tag{9}$$

$$X_t^6 = 0.4X_{t-6}^6 + 0.4X_{t-9}^6 + \eta^6 \tag{10}$$

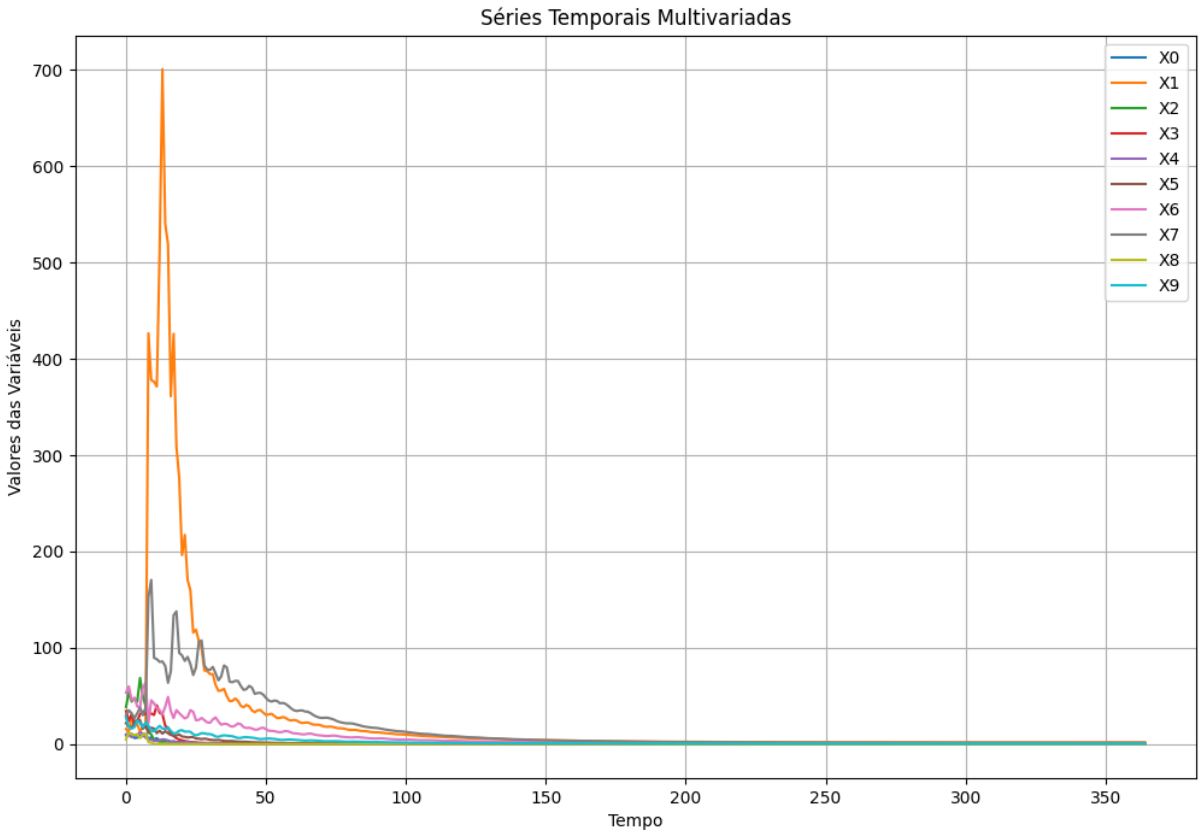
$$\begin{aligned} X_t^7 = {} & 0.5X_{t-9}^7 + 0.8X_{t-1}^7 + 0.7X_{t-6}^7 + 0.4X_{t-5}^7 \\ & + 0.6X_{t-8}^5 + 0.9X_{t-1}^8 + 0.7(X_{t-2}^8)^2 + 0.7X_{t-3}^9 + 0.8X_{t-4}^9 + \eta^7 \end{aligned} \tag{11}$$

$$X_t^8 = 0.1X_{t-1}^8 + 0.1X_{t-2}^8 + \eta^8 \tag{12}$$

$$X_t^9 = 0.3X_{t-7}^9 + 0.5X_{t-2}^8 + \eta^9 \tag{13}$$

A Figura 4 mostra o comportamento dos dados sintéticos gerados a partir das equações 4 a 13.

Figura 4 – Representação da série temporal sintética.



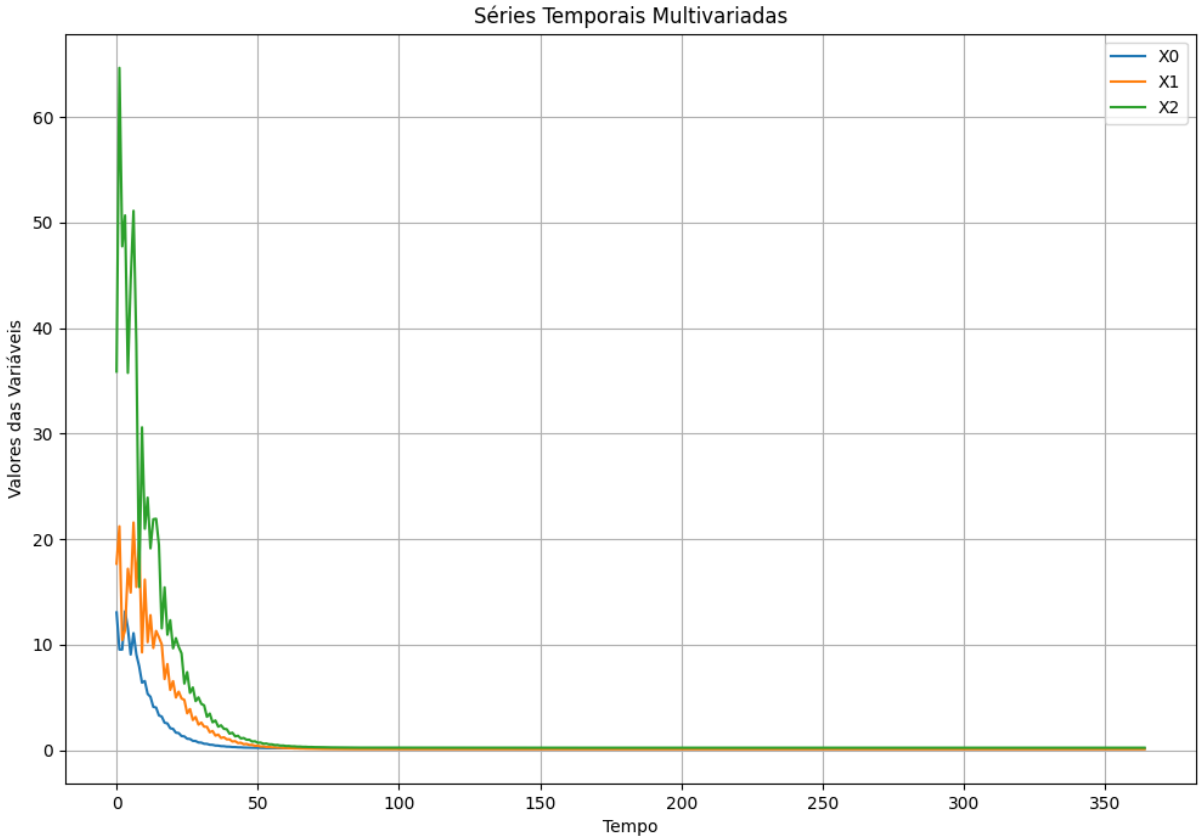
Fonte: Criação própria
Legenda: Gráfico gerado a partir dos dados sintéticos para as séries temporais multivariadas ao longo do período de um ano (365 pontos).

A princípio foram utilizadas todas as 10 equações apresentadas, analisando as predições caso

a caso como se cada variável deste *dataset* fosse a *target*, mas o custo computacional se tornou excessivo para certos modelos e decidiu-se por testar apenas com 3 variáveis a fim de otimizar os experimentos.

No caso em que utilizou-se apenas 3 equações, foram utilizadas as equações 5, 6 e 7, que tem suas séries representadas na Figura 5.

Figura 5 – Representação da série temporal sintética reduzida.



Fonte: Criação própria
Legenda: Gráfico gerado a partir das equações 5, 6 e 7 ao longo do período de um ano (365 pontos).

Para ambos os casos, os resultados estão detalhados na seção 4.4.

4.4 RESULTADOS E DISCUSSÕES - DADOS SINTÉTICOS

Os experimentos foram realizados utilizando 3 modelos de redes neurais: *Random Forest*, LSTM e ARIMA/ARIMAX. Para melhor compreensão dos resultados, eles serão descritos em subseções separadas pelo modelo treinado.

É válido pontuar que neste momento, o *pipeline* de treinamento estava sendo da seguinte forma: o modelo era treinado para cada *target*, fazendo iterações nas previsões adicionando a cada iteração uma nova variável dentro do potenciais *drivers* e fazendo uma nova previsão, depois de concluída todas as previsões para todas as *targets*, foi feito o embaralhamento de

Tabela 1 – Resultados obtidos para o *dataset* de 3 variáveis sintéticas utilizando o *Random Forest*.

Variável que está sendo embaralhada		X0	X1	X2	X0	X1	X2	X0	X1	X2		X0	X1	X2
Target	*Drivers*	*Drivers* *PE			*Recall*			Especificidade			SMAPE	SMAPE *PE		
X0	X0	X0	X0	X0	1,000	1,000	1,000	1,000	1,000	1,000	41,32	191,27	41,32	41,32
X1	X0 e X1	X1	X1	X1	0,500	0,500	0,500	1,000	1,000	1,000	41,32	43,09	190,14	43,09
X2	X0 e X2	X2	X2	X2	0,333	0,333	0,333	0,000	0,000	0,000	41,32	43,62	43,62	186,95

*PE = Pós-Embaralhamento

cada variável e feita novamente todas as previsões, depois, foram feito os cálculos de *recall* e especificidade.

4.4.1 Random Forest

A primeira rede a ser testada foi a *Random Forest*, a fim de testar o algoritmo de Castro et al. (Castro et al., 2023) de forma mais fidedigna e comprovar as conclusões obtidas neste trabalho.

4.4.1.1 Experimento 1

Inicialmente foram utilizados apenas o *dataset* de dados sintéticos baseado em 3 equações, considerando o período de 1 ano (365 pontos). A Tabela 1 resume os principais resultados obtidos para este caso.

Nesse experimento, observa-se que os resultados de SMAPE são razoáveis (representam um erro na predição 41,32% acima dos valores reais), e a repetição deste resultado em todos os casos se dá pela natureza determinística do modelo para sementes, os valores de *recall* e especificidade para o caso onde a *target* foi X0 demonstram que a taxa de acerto do modelo foi de 100%, assim como para a especificidade para X1 como *target*, entretanto o *recall* em X1 e ambas em X2 demonstram uma taxa de erro considerável. Os valores de SMAPE pós-embaralhamento são coerentes uma vez que quando a variável está embaralhada o modelo faz previsões mais distantes ao considerar esta variável.

4.4.1.2 Experimento 2

Em seguida foram realizados experimentos considerando o *dataset* completo (10 variáveis), também para o período de 1 ano. Os resultados estão descritos na Tabela 2.

Neste caso, os valores de SMAPE também foram satisfatórios (42,27% na maior parte dos experimentos), exceto nas duas últimas *targets* pois os valores são excessivamente baixo, indicando a possibilidade de ter havido algum erro, porém como nesta etapa o objetivo eram apenas de testes iniciais sobre o modelo, optou-se por não investigar esses resultados mais profundamente por uma questão de otimização do tempo de trabalho. Além disso, os resultados de *recall* neste experimento se mostraram mais variáveis, tendo situações onde o modelo obteve 100% de acerto e outras em que as taxas de erro foram mais altas, indicando possíveis falhas. Já a especificidade obteve bons resultados em todos os casos, com uma taxa ligeiramente menor

Tabela 2 – Resultados obtidos para o *dataset* de 10 variáveis sintéticas utilizando o *Random Forest*.

Variável que está sendo embaralhada		X0	X0	X0		X0
Target	Drivers	Drivers *PE	Recall	Especificidade	SMAPE	SMAPE *PE
X0	X0	X0 e X8	1,000	0,889	42,27	199,01
X1	X0 e X1	X1	0,200	1,000	42,27	96,24
X2	X0 e X2	X2	0,500	1,000	42,27	45,90
X3	X0 e X3	X3	0,500	1,000	42,27	49,30
X4	X0 e X4	X4	1,000	1,000	42,27	48,92
X5	X0 e X5	X5	1,000	1,000	42,27	108,01
X6	X0 e X6	X6	1,000	1,000	42,27	97,38
X7	X0 e X7	X7	0,167	1,000	42,27	97,29
X8	X8	X8	1,000	1,000	0,00	0,00
X9	X8 e X9	X9	1,000	1,000	0,00	97,57

*PE = Pós-Embaralhamento

de acerto para a *target* X0. Com o embaralhamento das variáveis novamente observa-se uma coerência no valor mais distante para o caso da *target* X0, uma vez que essa foi a variável embaralhada (é válido pontuar que nesse experimento somente a variável X0 foi embaralhada para otimizar o tempo de estudo e reduzir o custo computacional e a escolha desta variável foi randômica), mas também nota-se que os resultados de SMAPE variaram muito mais, obtendo valores altos em alguns casos e valores extremamente baixos em outros, o que também indica um possível caso de atenção sobre o modelo.

4.4.2 LSTM

Em seguida, foi testada a rede LSTM, pois foi a rede inicialmente proposta para as hipóteses pensadas na idealização deste trabalho.

4.4.2.1 Experimento 1

Devido aos custos computacionais naturalmente envolvidos com esse modelo, em razão da complexidade da sua estrutura, foram feitos mais testes, iniciando a análise considerando apenas a mais um *driver* além da *target* para as previsões, com o *dataset* de 10 variáveis e 365 pontos (período de 1 ano). A Tabela 3 destaca os resultados obtidos neste primeiro teste.

Esses resultados se mostram semelhantes aos da *Random Forest* no que diz respeito à *recall* e especificidade, porém, os valores de SMAPE apresentam um erro maior na predição, indicando a dificuldade percebida ao longo dos experimentos em usar a LSTM em uma quantidade relativamente baixa de dados, como no *dataset* utilizado neste trabalho.

Tabela 3 – Resultados obtidos para o *dataset* de 10 variáveis sintéticas utilizando o LSTM, porém com apenas dois *drivers* máximos.

Variável que está sendo embaralhada		X1	X1	X1		X1
Target	*Drivers*	*Drivers* *PE	*Recall*	Especificidade	SMAPE	SMAPE *PE
X0	X0 e X9	X0 e X8	1,000	0,889	199,97	199,98
X1	X1 e X8	X1 e X6	0,400	1,000	53,93	169,98
X2	X2 e X5	X1 e X2	0,500	0,875	199,80	198,93
X3	X0 e X3	X3 e X7	0,500	0,875	199,66	197,65
X4	X4 e X6	X1 e X4	1,000	0,889	199,63	199,90
X5	X5	X0 e X5	1,000	0,889	199,97	199,80
X6	X5 e X6	X2 e X6	1,000	0,889	179,43	182,05
X7	X4 e X7	X7 e X8	0,333	1,000	135,89	166,06
X8	X2 e X8	X8	1,000	1,000	107,70	61,34
X9	X2 e X9	X5 e X9	1,000	0,889	187,03	143,91

*PE = Pós-Embaralhamento

Tabela 4 – Resultados obtidos para o *dataset* de 3 variáveis sintéticas utilizando o LSTM, com 365 pontos.

Target	*Drivers*	SMAPE
X0	X0	199,70
X1	X0, X1 e X2	199,94
X2	X2	199,86

4.4.2.2 Experimento 2

A fim de analisar o máximo de combinações possíveis de *drivers* nos demais experimentos optou-se por utilizar o *dataset* de apenas 3 equações, considerando 365 pontos (período de 1 ano), levando em conta o custo computacional envolvido com a LSTM. Além disso, a partir dos resultados obtidos para a primeira parte do experimento, tendo em vista já a percepção da possibilidade de a LSTM não ser adequada para a quantidade de pontos e dados disponíveis, a fim de otimizar o tempo de trabalho, optou-se por não realizar o embaralhamento das variáveis para fazer novas previsões. Os resultados deste experimento estão descritos na Tabela 4.

Os resultados deste experimento reforçaram a hipótese de que os dados não seriam suficientes para esse modelo, mas ainda sim, tentou-se mais algumas alternativas para oferecer mais dados de treinamento nos experimentos seguintes.

4.4.2.3 Experimentos 3 e 4

Os experimentos 3 e 4 testaram o aumento no número de pontos para o treinamento do modelo, novamente sem o embaralhamento para otimizar o tempo de trabalho e diminuir o tempo computacional. No experimento 3 foi utilizado o *dataset* de 3 variáveis, porém, com 1825 pontos, que correspondem ao período de 5 anos de dados. Já o experimento 4 utilizou o mesmo *dataset*, mas com 3650 pontos, equivalente ao período de 10 anos. As Tabelas 5 e 6 mostram os resultados.

Tabela 5 – Resultados obtidos para o dataset de 3 variáveis sintéticas utilizando o LSTM, com 1825 pontos.

Target	*Drivers*	SMAPE
X0	X0 e X2	199,65
X1	X0, X1 e X2	199,98
X2	X0 e X2	199,78

Tabela 6 – Resultados obtidos para o dataset de 3 variáveis sintéticas utilizando o LSTM, com 3650 pontos.

Target	*Drivers*	SMAPE
X0	X0, X1 e X2	199,85
X1	X0 e X1	199,86
X2	X2	199,90

Tabela 7 – Resultados obtidos para o *dataset* de 3 variáveis sintéticas utilizando o ARIMA/ARIMAX.

Variável que está sendo embaralhada		X0	X1	X2	X0	X1	X2	X0	X1	X2		X0	X1	X2
Target	*Drivers*	*Drivers* *PE			*Recall*			Especificidade			SMAPE	SMAPE *PE		
X0	X0	X0	X0	X0	1,000	1,000	1,000	1,000	1,000	1,000	0,00	193,64	0,00	0,00
X1	X0 e X1	X1	X1	X1	0,500	0,500	0,500	1,000	1,000	1,000	0,00	0,00	183,54	0,00
X2	X0 e X2	X2	X2	X2	0,500	0,500	0,500	1,000	1,000	1,000	0,00	0,00	0,00	193,93

*PE = Pós-Embaralhamento

E mesmo com o aumento da quantidade de dados, os resultados para o SMAPE não foram satisfatórios, confirmando que a rede LSTM poderia não ser adequada para este trabalho.

Também foi feita uma tentativa de executar um experimento utilizando as 10 variáveis disponíveis, mas com todas as combinações possíveis de *drivers*, mas devido a falta de poder computacional após muitas horas de treinamento o sistema teve seus recursos esgotados e não foi possível concluir o experimento.

Ainda sim, posteriormente a rede foi testada utilizando os dados reais com o código do algoritmo otimizado para redução do tempo e custos computacionais e os resultados serão apresentados na seção 4.5.2.

4.4.3 ARIMA/ARIMAX

Por fim, testou-se a rede ARIMA/ARIMAX, tendo em vista os resultados não satisfatórios obtidos nos testes da rede LSTM.

4.4.3.1 Experimento 1

O primeiro experimento realizado considerou apenas 3 equações em um período de 1 ano. Os resultados estão apresentados na Tabela 7.

Os resultados deste experimento se assemelham aos obtidos com a rede *Random Forest*. Observa-se que os valores de SMAPE indicam que não houveram erros na predição, o que

Tabela 8 – Resultados obtidos para o *dataset* de 10 variáveis sintéticas utilizando o ARIMA/ARIMAX.

Variável que está sendo embaralhada		X0	X0	X0		X0
Target	*Drivers*	*Drivers* *PE	*Recall*	Especificidade	SMAPE	SMAPE *PE
X0	X0	X0	1,000	1,000	0,00	199,09
X1	X0 e X1	X1	0,200	1,000	0,00	94,91
X2	X0 e X2	X2	0,500	1,000	0,00	0,00
X3	X0 e X3	X3	0,500	1,000	0,00	0,00
X4	X0 e X4	X4	1,000	1,000	0,00	0,00
X5	X0 e X5	X5	1,000	1,000	0,00	42,37
X6	X0 e X6	X6	1,000	1,000	0,00	95,06
X7	X0 e X7	X7	0,167	1,000	0,00	93,79
X8	X8	X8	1,000	1,000	0,00	0,00
X9	X8 e X9	X5 e X9	1,000	1,000	0,00	94,92

*PE = Pós-Embaralhamento

chama a atenção para o treinamento com os demais dados. Novamente, os valores resultantes pós-embaralhamento são coerentes com a variável que está sendo embaralhada.

4.4.3.2 Experimento 2

O último experimento utilizando dados sintéticos considerou as 10 equações em um período de 1 ano. Os resultados estão apresentados na Tabela 8.

Os resultados desse experimento mantiveram um padrão semelhante ao que ocorreu no experimento 1 com a rede ARIMA/ARIMAX, apresentou uma melhoria na especificidade tanto em relação ao experimento com 3 variáveis quanto ao experimento com 10 variáveis com a *Random Forest*, o *recall* não apresentou mudanças significativas quanto aos demais experimentos, e os resultados pós-embaralhamento apresentaram algumas mudanças com relação ao padrão observado no experimento 1, indicando alguns possíveis erros na inferência causal.

Finalizados os experimentos com os dados sintéticos, foram analisados os resultados e percebeu-se a necessidade de otimização do código para reduzir o tempo de treinamento dos modelos e custos computacionais, para então aplicar os dados reais ao Aleph. Os resultados dessa implementação serão apresentados na seção 4.7.

4.5 DADOS REAIS - DATA SUS

Após a realização dos testes com os dados sintéticos, dois *datasets* formados a partir de dados públicos foram aplicados aos algoritmos. Um dos *datasets* utilizado nestes experimentos foi formado a partir de informações sobre óbitos e internações ocasionadas por doenças do aparelho respiratório (CID-10: J00-J99) no Brasil, entre 1999 e 2021, estratificados por sexo, faixa etária, escolaridade e raça/cor. Além disso, foram calculadas taxas de morbimortalidade ajustadas por população, a fim de reduzir distorções em municípios de pequeno porte. Dados socioeconômicos,

como PIB (Produto Interno Bruto) e população municipal, foram obtidos do IBGE, considerando a evolução das malhas territoriais ao longo do tempo. Também foram incorporadas variáveis ambientais, como o índice de urbanização, calculado com dados do MapBiomas e imagens de satélite, e dados de emissões atmosféricas (CO_2, CH_4 e N_2O) provenientes da base EDGAR, interpolados para escala municipal. O outro *dataset* foi formado a partir de dados de estatísticas vitais de doenças 10ª revisão da Classificação Estatística Internacional de Doenças e Problemas Relacionados à Saúde (CID-10), estabelecida pela Organização Mundial da Saúde (OMS), sendo coletados manualmente dados de morbidade e mortalidade para a dengue (CID-10: A90) entre os anos de 1999 e 2021, para todo o território nacional, além da inclusão de variáveis socioeconômicas e ambientais tal qual feito no primeiro *dataset*. As Figuras 6 e 7 mostram o comportamento dos dados para as doenças do aparelho respiratório e para a dengue, respectivamente, para o município de Alta Floresta D'Oeste (RO), escolhido aleatoriamente para amostragem.

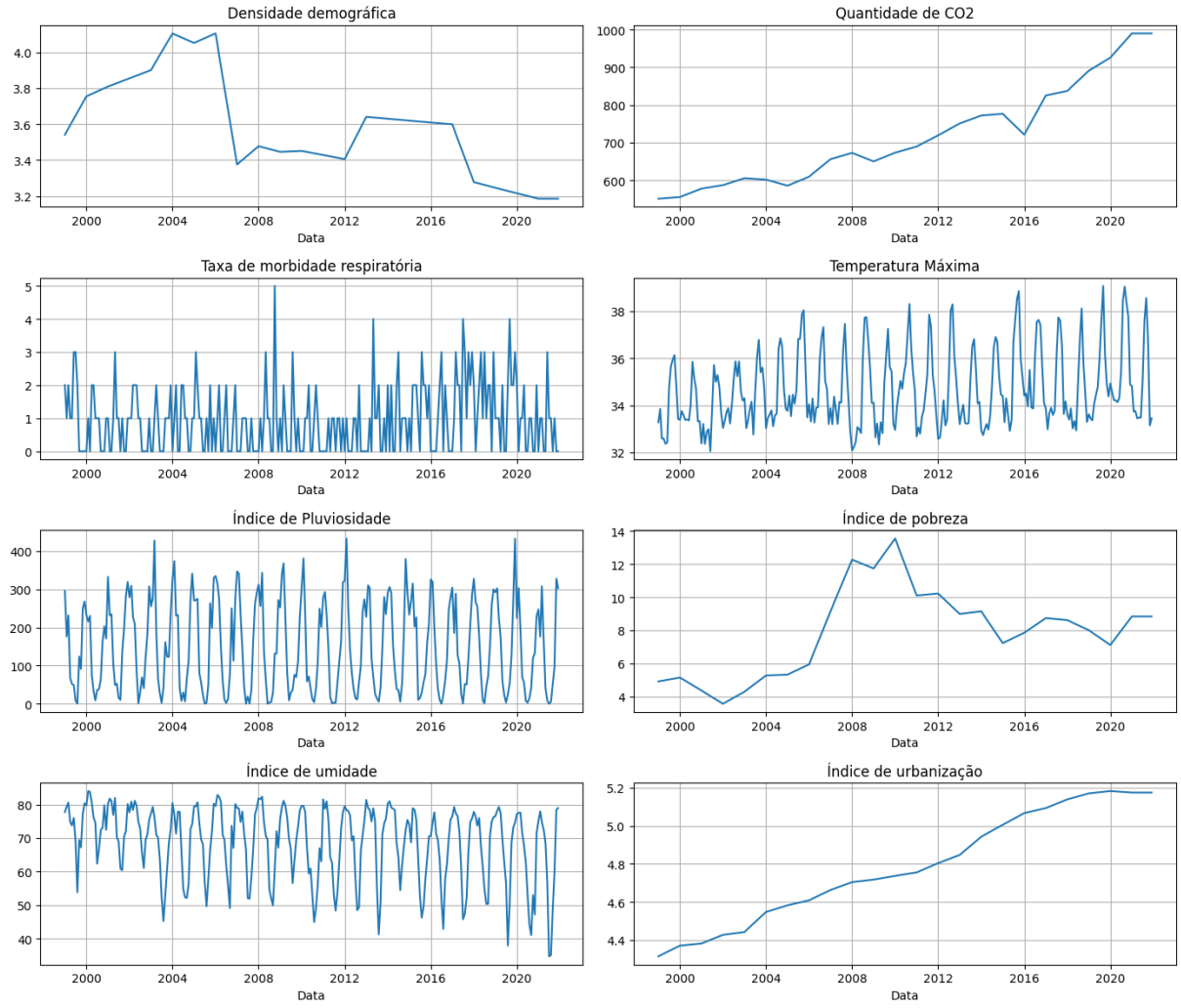
4.6 EXPERIMENTOS COM OS DADOS REAIS

Para estes experimentos, o *pipeline* utilizado é exatamente o apresentado na Figura 3. A mudança foi feita pois percebeu-se a necessidade de otimização do código para redução do tempo e custo computacional, além de uma melhor apresentação do código para facilitação da reprodutibilidade e generalização. Para os dados reais, também foram utilizadas as redes *Random Forest*, ARIMA/ARIMAX e LSTM. Para as redes *Random Forest* e ARIMA/ARIMAX foram feitos 2 tipos experimentos, um considerando todas as cidades brasileiras que possuem autocorrelação espacial positiva para morbimortalidade, tanto para a dengue, quanto para as doenças do aparelho respiratório, e outro considerando apenas cidades, que estão dentro destes conjuntos, que possuem mais do que 500 mil habitantes, também para ambos os dados. Já a rede LSTM foi aplicada apenas para municípios com mais de 500 mil habitantes, pois a rede possui alta complexidade e a quantidade de dados é muito alta, portanto, devido ao poder computacional disponível, aplicar a totalidade dos dados se tornou inviável para esta rede, mas como este experimento foi realizado para as 3 redes empregadas, foi possível estabelecer um comparativo dos resultados obtidos em cada uma delas. Os resultados desses experimentos estão apresentados na seção 4.7.

4.7 RESULTADOS E DISCUSSÕES - DADOS REAIS

Nesta seção os resultados serão apresentados para cada um dos experimentos descritos na seção 4.6, e essa descrição será feita em subseções que separam os experimentos pela rede utilizada. Além disso, para os dados reais foram gerados histogramas que representam a frequência de seleção de *drivers* em cada situação, permitindo que sejam analisadas tanto a consistência da inferência causal entre as redes, quanto a coerência dos *drivers* selecionados de uma perspectiva sanitária e socioambiental.

Figura 6 – Plot das variáveis presentes no dataset das Doenças Respiratórias para o município Alta Floresta D'Oeste (RO).



Fonte: Criação própria

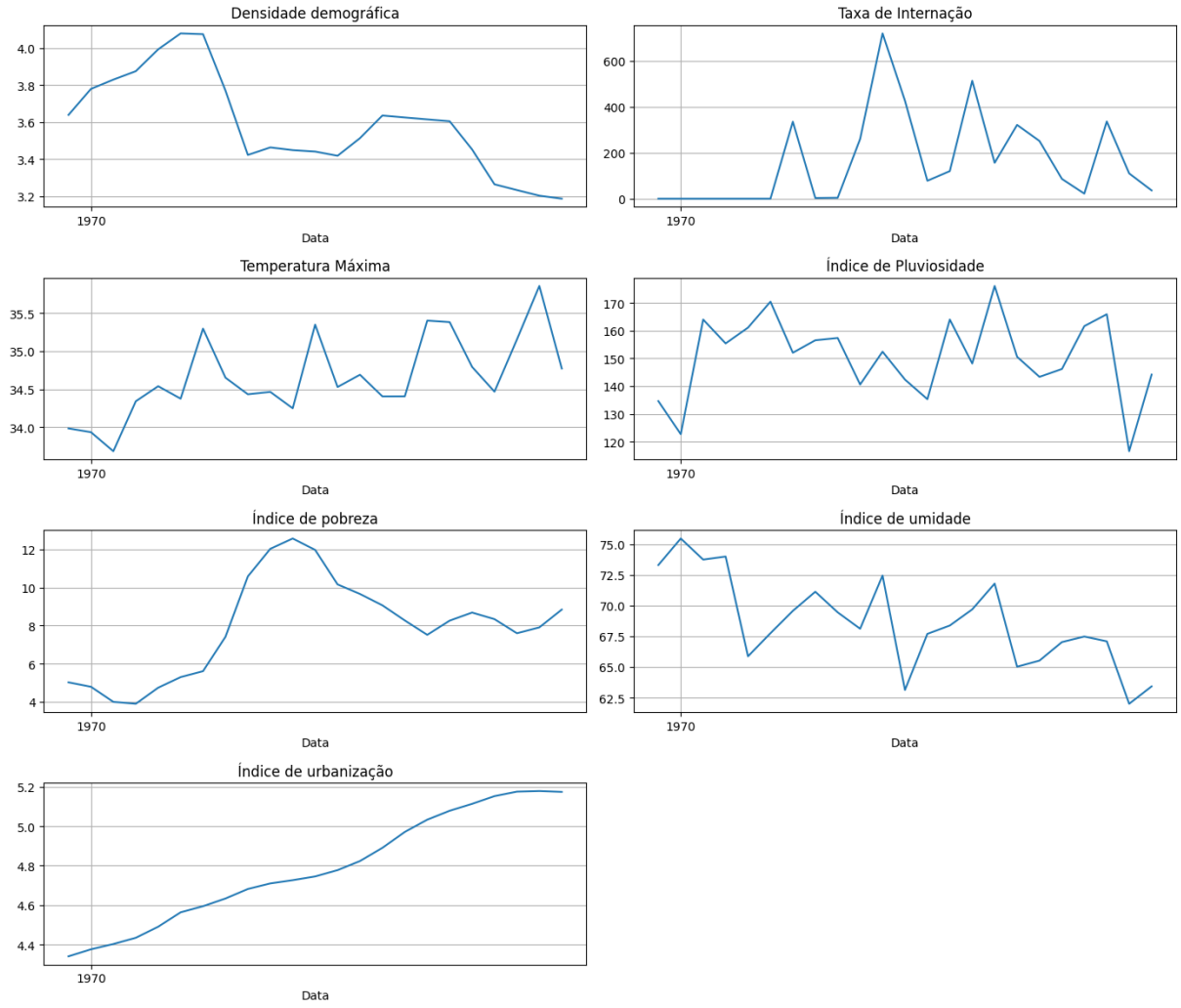
Legenda: Representação visual das séries temporais utilizadas no *dataset* para os experimentos sobre Doenças Respiratórias. As variáveis representadas são: densidade demográfica, quantidade de CO_2, taxa de mortalidade respiratória (sendo essa a variável alvo do experimento), temperatura máxima, índice de pluviosidade, índice de pobreza, índice de umidade e índice de urbanização.

4.7.1 Random Forest

Conforme dito anteriormente, para essa rede foram realizados dois experimentos em cada *dataset*, um que considera todas as cidades disponíveis no conjunto de dados, e um que considera apenas as cidades com mais de 500 mil habitantes. Os histogramas de cada experimento para os dados das doenças respiratórias estão apresentados nas Figuras 8 e 9 e para os dados da dengue nas Figuras 10 e 11.

A Figura 8 mostra que a rede *Random Forest* selecionou como principais *drivers* a quantidade de CO_2 e o índice de pobreza seguido da densidade demográfica e temperatura máxima. Do ponto de vista sanitário, a emissão de CO_2, apesar de não ser gás mais poluente, as altas concentrações dessa substância costumam estar relacionadas com outro poluentes mais agressivos, e a exposição

Figura 7 – Plot das variáveis presentes no dataset da Dengue para o município Alta Floresta D'Oeste (RO).

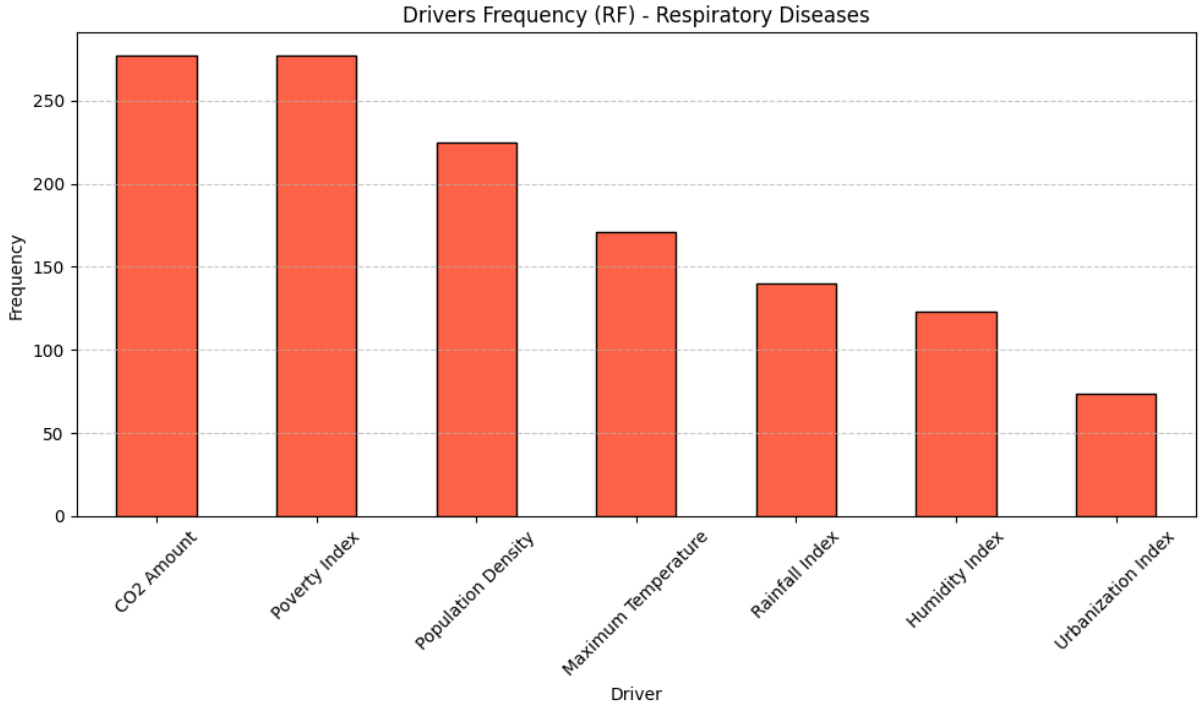


Fonte: Criação própria

Legenda: Representação visual das séries temporais utilizadas no *dataset* nos experimentos sobre a Dengue. As variáveis representadas são: densidade demográfica, taxa de internação por dengue (sendo essa a variável alvo do experimento), temperatura máxima, índice de pluviosidade, índice de pobreza, índice de umidade e índice de urbanização.

prolongada à poluição atmosférica está associada com doenças respiratórias, tal qual como asma, bronquite e pode aumentar a mortalidade em razão causas respiratórias (Schraufnagel et al., 2019). Por outro lado, o índice de pobreza não parece estar tão diretamente relacionado ao aumento de casos de doenças respiratórias, mas sim evidencia uma parcela da população que tem menos acesso a serviços de saúde, podendo agravar esses quadros quando eles vem a acontecer, e tendo menos acesso a programas de vacinação, e também menos acesso a educação, dificultando a conscientização sobre cuidados necessários para evitar tais doenças. Já a densidade demográfica é um fator muito relevante, pois doenças respiratórias virais são transmitidas entre seres humanos, como a influenza e a COVID-19 por exemplo, portanto, quanto maior a densidade demográfica, maior a probabilidade de transmissão da doença (Neiderud, 2015). Teixeira et al. (Gonçalves;

Figura 8 – Frequência dos *drivers* selecionados para doenças respiratórias em todos os municípios.

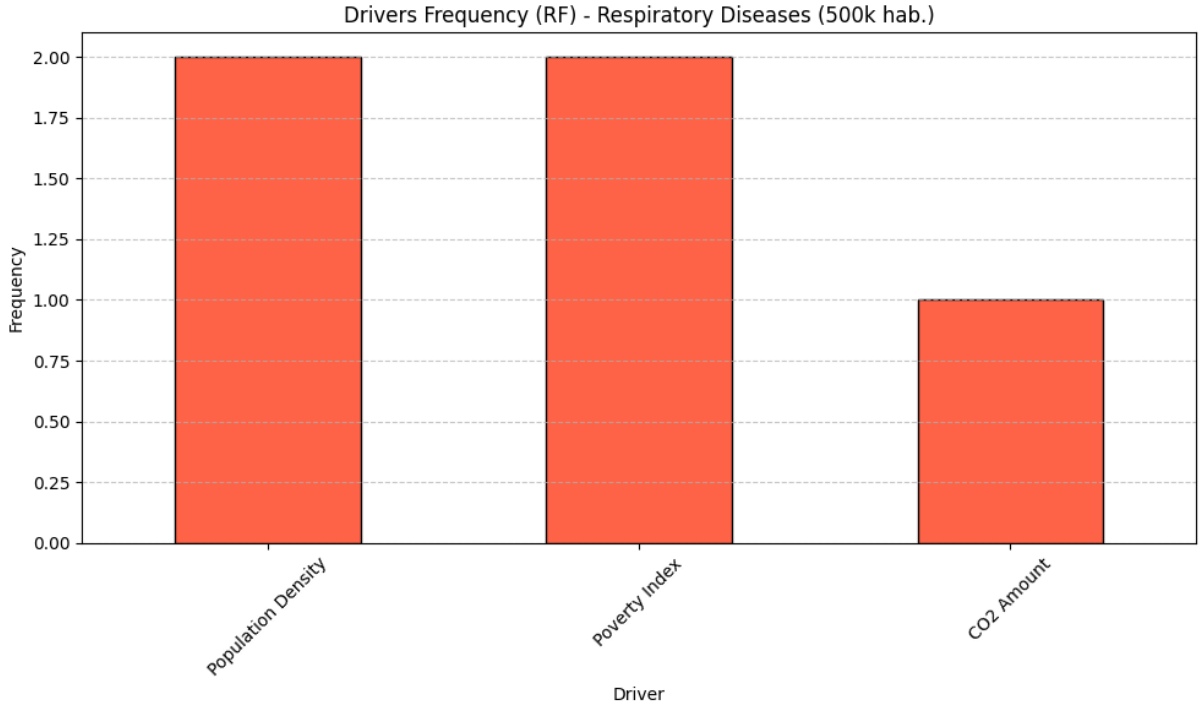


Fonte: Criação própria
Legenda: Histograma feito para representar a frequência em que os *drivers* foram selecionados ao longo do treinamento, utilizando a *Random Forest*, para as doenças respiratórias considerando todos os municípios.

Coelho, 2010) mostra um outro ponto de vista sobre a relação do aumento de temperatura e o aumento de casos de doenças respiratórias, indicando que com o aumento da diferença de temperatura entre os meses de abril e maio nos últimos anos, a entrada na estação do inverno está se tornando mais abrupta, com isso, surge um maior problema de termo-regulação em indivíduos adaptados ao clima mais ameno de abril, e pessoas com tal problema se tornam mais sensíveis, gerando doenças respiratórias e cardiovasculares, sendo a população infantil e geriátrica mais suscetíveis. A Figura 9 segue o mesmo padrão de identificação de *drivers* que o apresentado, entretanto em uma frequência bem mais baixa, sobretudo quando comparado com às demais redes que terão seus resultados apresentados nas próximas subseções. A análise sobre as causas dessa identificação de menor frequência será feita ao final desta seção quando todos os resultados tiverem sido apresentados e uma análise comparativa das redes possa ser feita.

Com relação a dengue, a Figura 10 destaca como principais *drivers* o índice de pobreza, a temperatura máxima, índice de pluviosidade e densidade demográfica. Assim como analisado para as doenças respiratórias, o índice de pobreza não impacta diretamente o número de casos, mas sim evidencia as consequências da desigualdade social. Da mesma forma, pessoas em maior vulnerabilidade social têm menos acesso à saúde e educação, mas para a dengue existe um agravante até mais relevante, a falta de saneamento básico, principalmente as deficiências quanto

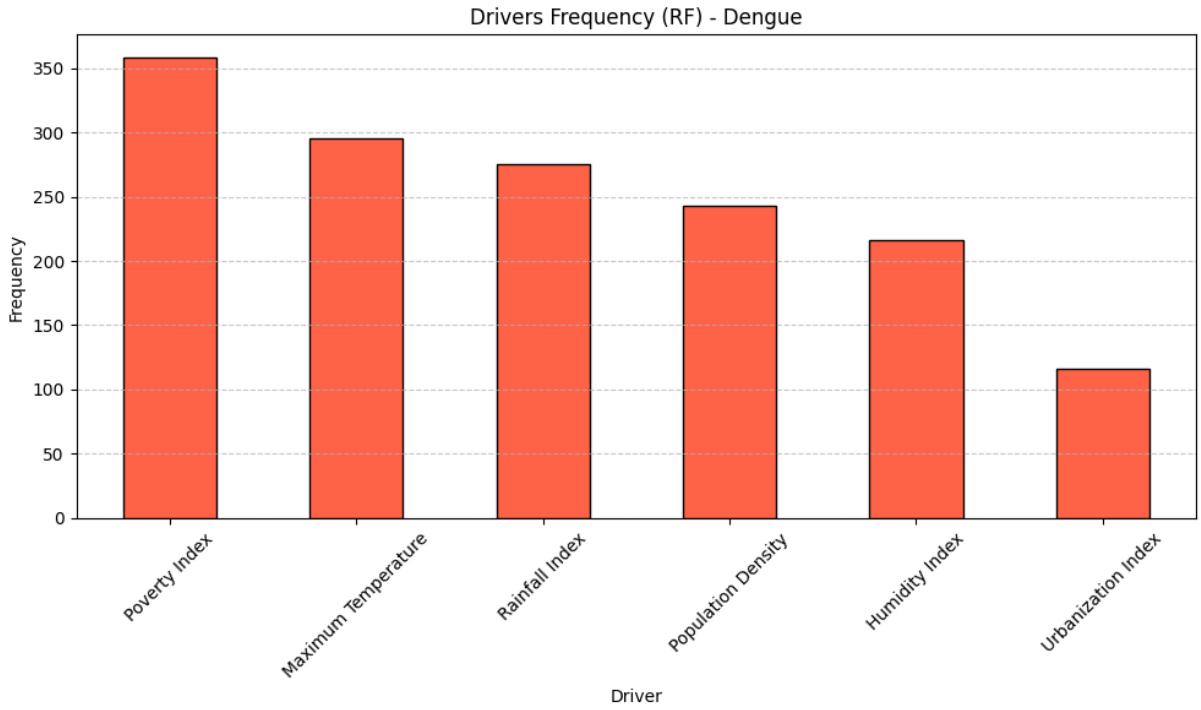
Figura 9 – Frequência dos *drivers* selecionados para doenças respiratórias em municípios com mais de 500 mil habitantes.



Fonte: Criação própria
Legenda: Histograma feito para representar a frequência em que os *drivers* foram selecionados ao longo do treinamento, utilizando a *Random Forest*, para as doenças respiratórias considerando os municípios com mais de 500 mil habitantes em 2021.

ao abastecimento de água à coleta de lixo, que contribuem para a proliferação de mosquitos Aedes que transmitem a dengue (Tauil, 2010). Além disso, o aumento da temperatura e do índice de pluviosidade contribuem para aumentar o número de criadouros do mosquito da dengue (Meira et al., 2021), facilitando assim o desenvolvimento do vetor. Já a densidade demográfica maior, favorece a contaminação, pois o mosquito passa a ter maior facilidade em se alimentar de sangue humano e transmitir o vírus, devido à maior disponibilidade de hospedeiros, além disso, a densidade demográfica também está relacionada com o índice de urbanização, que apesar de aparecer em menor relevância, também se encontra no histograma da Figura 10, e cidades mais urbanizadas tendem a ter um menor escoamento de água, favorecendo a criação de mosquitos transmissores da dengue. A Figura 11 demonstra um padrão semelhante, porém, nota-se que neste caso o índice de pobreza aparece com bem menos relevância do que na Figura 10. Uma possível explicação para isso é que cidades mais populosas, como as utilizadas nesse experimento, tendem a ter melhores serviços de saneamento básico. Segundo a 15ª edição do Ranking de Saneamento, feito em 2023 pelo instituto Trata Brasil (Ranking…, 2023), os 100 maiores municípios do país apresentaram cobertura média de abastecimento de água, por exemplo, de cerca de 93,9%, enquanto a média nacional foi de aproximadamente 83,1%. A melhor prestação de serviços de saneamento básico pode estar relacionada a uma maior arrecadação tributária

Figura 10 – Frequência dos *drivers* selecionados para a dengue em todos os municípios.



Fonte: Criação própria
Legenda: Histograma feito para representar a frequência em que os *drivers* foram selecionados ao longo do treinamento, utilizando a *Random Forest*, para a dengue considerando todos os municípios.

em municípios mais populosos, além de maiores pressões políticas, entre outros fatores sócio econômicos.

4.7.2 LSTM

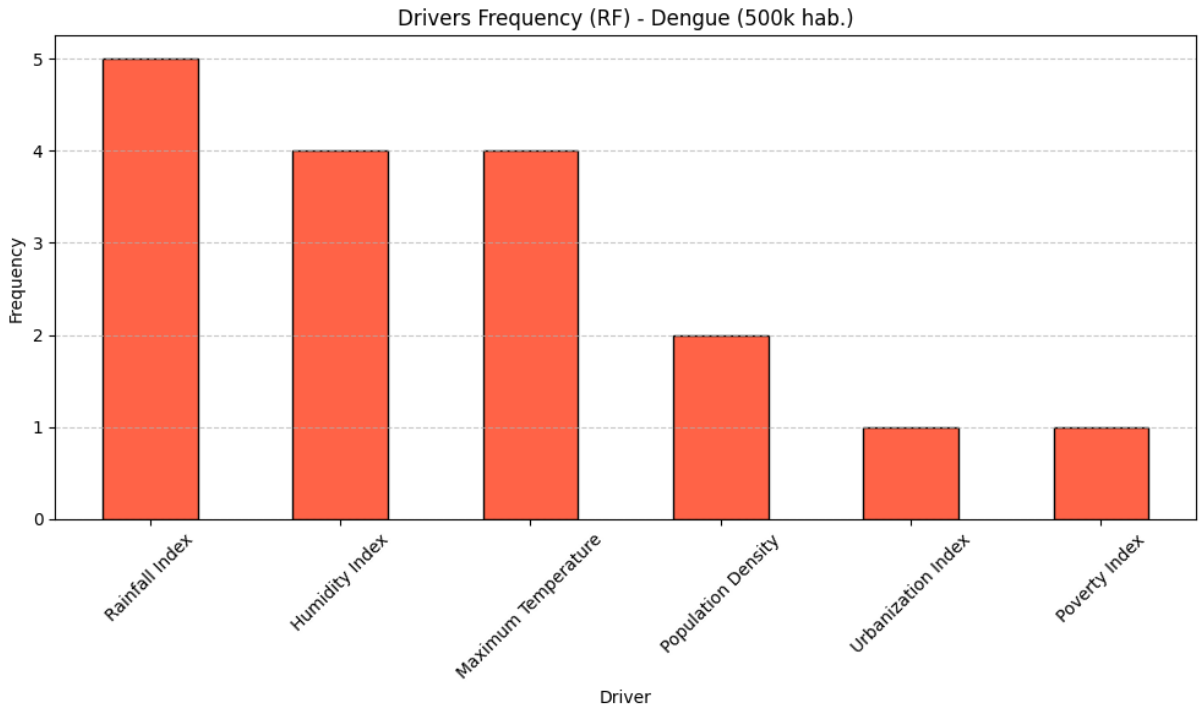
Já neste caso, foram feitos apenas um experimento para cada conjunto de dados, o que considera apenas os municípios com mais de 500 mil habitantes dentro dos conjuntos. O histograma do experimento para os dados das doenças respiratórias está apresentado na Figura 12 e para os dados da dengue na Figura 13.

Do ponto de vista sanitário, observa-se que a LSTM obteve uma boa performance na seleção de *drivers*, mantendo a coerência discutida na subseção 4.7.1 para as doenças respiratórias, e sendo ainda melhor para a dengue, selecionando majoritariamente os *drivers* que possuem uma relação mais direta, vale o destaque para a maior relevância do índice de urbanização para a dengue (Figura 13), que está relacionado com a menor capacidade de escoamento das cidades, aumentando a proliferação dos mosquitos, tal como discutido anteriormente.

4.7.3 ARIMA/ARIMAX

Por fim, assim como para a *Random Forest*, foram realizados 2 experimentos para cada *dataset*, considerando todos os municípios e apenas os que tem mais de 500 mil habitan-

Figura 11 – Frequência dos *drivers* selecionados para dengue em municípios com mais de 500 mil habitantes.



Fonte: Criação própria
Legenda: Histograma feito para representar a frequência em que os *drivers* foram selecionados ao longo do treinamento, utilizando a *Random Forest*, para a dengue considerando os municípios com mais de 500 mil habitantes em 2021.

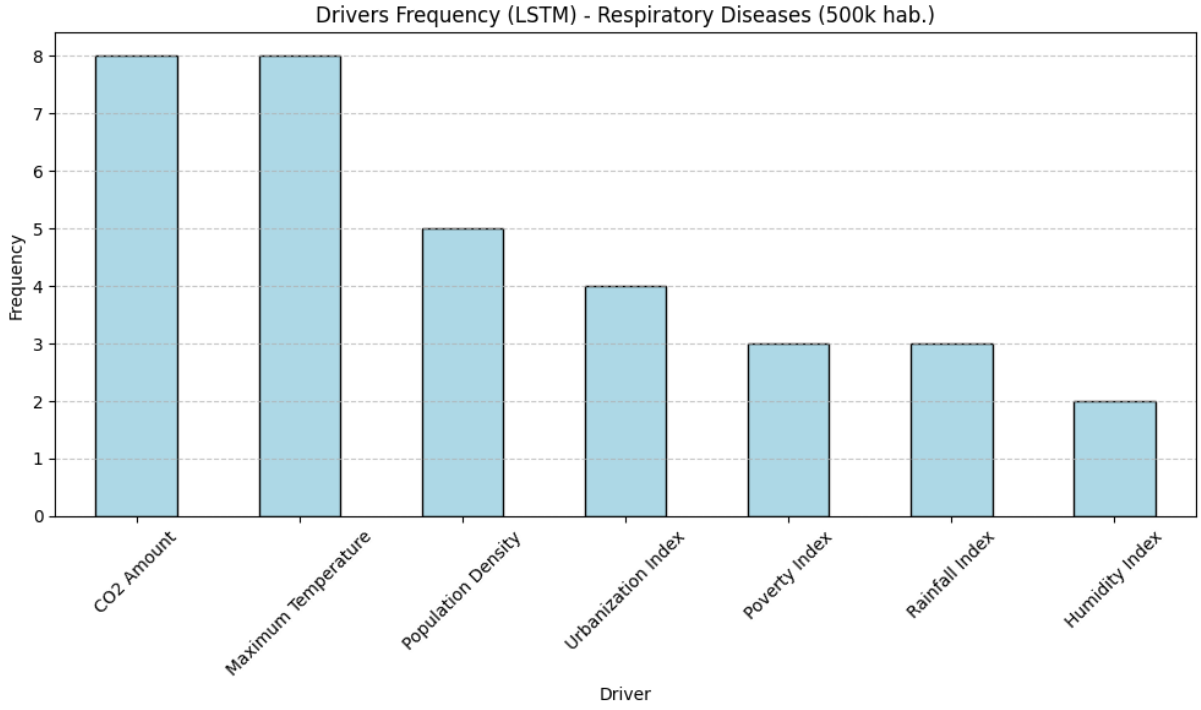
tes. O histograma da Figura 14 ilustra a seleção de *drivers* para os experimentos com a rede ARIMA/ARIMAX.

A Figura 14 mostra que apenas um dos experimentos teve algum *driver* selecionado, e nesse caso somente um para todo o *dataset* de municípios e variáveis. Apesar de diferir consideravelmente dos resultados da inferência causal obtidos com as demais redes, o resultado de que nenhuma variável exógena trouxe melhorias estatisticamente significativas para a predição da *target*, ou seja, nenhum *driver* foi selecionado, pode ser explicado pela forte autocorrelação de algumas variáveis combinado ao tamanho relativamente pequeno dos *datasets* que fez com que a abordagem autorregressiva do ARIMA, que tende a preferir utilizar apenas a variável endógena (*target*) para prever os seus valores futuros, sem a necessidade de inclusão de outras variáveis na predição (Box et al., 1976), fazendo com que esse modelo não seja o mais adequado para dar suporte ao algoritmo do Aleph.

4.7.4 Resultados e Discussões - Comparativo entre os Modelos

De forma comparativa, observa-se que a *Random Forest* apresenta um bom desempenho na seleção de *drivers* com base na importância preditiva, mostrando coerência em diversos casos. No entanto, a LSTM demonstrou uma sensibilidade ainda maior para inferência causal,

Figura 12 – Frequência dos *drivers* selecionados para doenças respiratórias em municípios com mais de 500 mil habitantes.

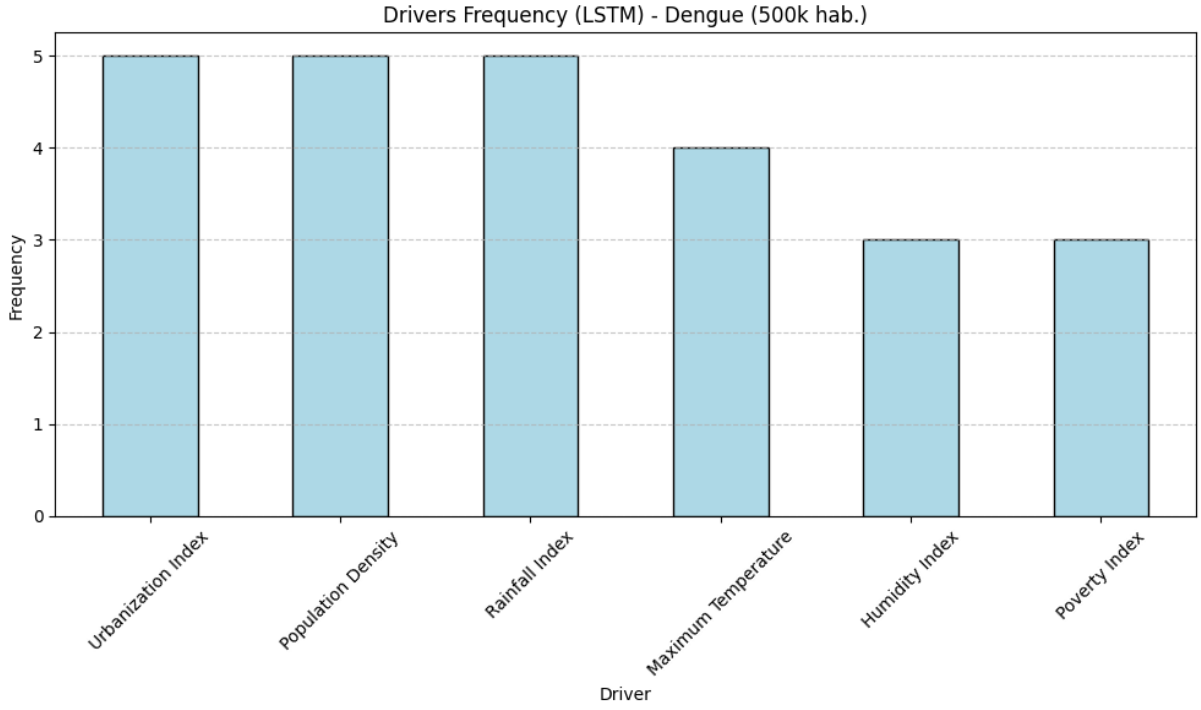


Fonte: Criação própria
Legenda: Histograma feito para representar a frequência em que os *drivers* foram selecionados ao longo do treinamento, utilizando a LSTM, para as doenças respiratórias considerando os municípios com mais de 500 mil habitantes em 2021.

selecionando mais *drivers* e, em alguns casos, como nas análises relacionadas à dengue, de forma mais consistente. Esse comportamento pode ser atribuído à capacidade da LSTM de modelar dependências temporais de longo prazo e de capturar dinâmicas não lineares complexas (Hochreiter; Schmidhuber, 1997), características inerentes aos fenômenos epidemiológicos. Enquanto a *Random Forest* considera apenas relações instantâneas ou fixas, a LSTM utiliza sua memória interna para compreender padrões sequenciais e causais ao longo do tempo, o que a torna mais sensível a variações temporais sutis. Entretanto, a desvantagem da LSTM neste trabalho foi o seu alto custo computacional, que não permitiu que todos os experimentos fossem realizados, limitando o objeto de estudo da pesquisa.

Adicionalmente, um fenômeno conhecido como *masking effect* pode explicar por que a *Random Forest* tende a selecionar menos variáveis em certos contextos. Quando múltiplas variáveis estão altamente correlacionadas entre si, como pode ocorrer em dados ambientais e socioeconômicos, a importância atribuída a uma variável pode ser "mascarada" pela presença de outra com alta correlação, mesmo que ambas sejam relevantes (Strobl et al., 2008), e isso se dá pois a rede se baseia na redução do erro de previsão ao construir *splits*, que são as divisões que ocorrem em cada nó das árvores de decisão. Isso compromete a detecção de drivers causais redundantes, reduzindo a sensibilidade do modelo. Esse efeito é particularmente problemático

Figura 13 – Frequência dos *drivers* selecionados para dengue em municípios com mais de 500 mil habitantes.

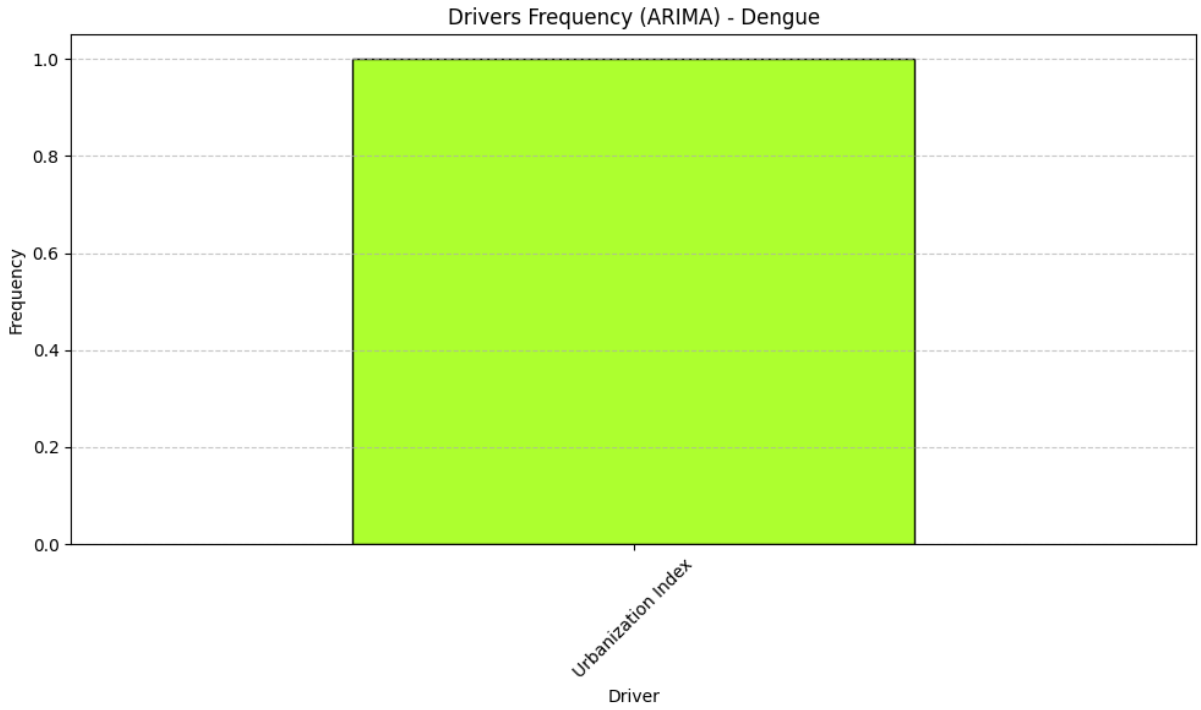


Fonte: Criação própria
Legenda: Histograma feito para representar a frequência em que os *drivers* foram selecionados ao longo do treinamento, utilizando a LSTM, para a dengue considerando os municípios com mais de 500 mil habitantes em 2021.

em contextos de autocorrelação entre variáveis explicativas, fazendo com que a *Random Forest* negligencie variáveis com influência real sobre a variável alvo.

Além disso, observou-se que a rede ARIMA/ARIMAX entre as três redes estudadas é a menos adequada, tendo em vista a sua característica autorregressiva que pode prejudicar a inferência causal em contextos de forte autocorrelação de variáveis como o contexto epidemiológico. Essa observação se dá por algumas características construtivas das redes ARIMA. A principal delas é que mesmo utilizando variáveis exógenas (modelo ARIMAX), o modelo ARIMA se baseia principalmente nos valores passados da *target* para a predição, e se existe uma forte autocorrelação das variáveis e a *target* já explica bem a si mesma, o modelo atribui menor importância estatística às variáveis exógenas, fazendo com que o modelo as ignore ou subestime, causando o efeito *masking effect* também (Validated, 2016). Outro ponto sobre a inclusão de variáveis é a consideração de linearidade e estacionariedade feita pelo modelo, que utiliza dos seus parâmetros internos para a garantia dessas condições, fazendo com que o impacto das relações causais entre as variáveis seja diminuído e a essas relações sejam descartadas, dificultando a extração de causas dinâmicas reais (Wang et al., 2010). Além disso, o modelo ARIMA/ARIMAX não possui uma estrutura natural para a geração de contrafactuais, processo necessário para processos de inferência causal do tipo A/B, que é baseado na ideia de existirem

Figura 14 – Frequência dos drivers selecionados para a dengue em todos os municípios.



Fonte: Criação própria
Legenda: Histograma feito para representar a frequência em que os *drivers* foram selecionados ao longo do treinamento, utilizando a ARIMA/ARIMAX, para a dengue considerando todos os municípios.

duas possibilidades, uma onde o *driver* testado existe e outra onde ele não existe, e mesmo com a utilização de mecanismos para a geração dos contrafactuais como o uso do embaralhamento, que simula a possibilidade de ausência do *driver* para o comparativo, devido a presunção de linearidade e estacionariedade, esse processo se torna limitado à efeitos lineares, imediatos e invariantes no tempo, o que não reflete a realidade dos dados (Menchetti; Cipollini; Mealli, 2021). Por outro lado, o modelo *Random Forest* e LSTM conseguem lidar melhor com a questão dos contrafactuais, uma vez que a estrutura deles naturalmente suporta a predição com dados não lineares e não estacionários sem a necessidade de linearização.

5 CONCLUSÃO

O presente trabalho está inserido no contexto da análise causal de séries temporais aplicadas à epidemiologia, com especial atenção aos efeitos das mudanças climáticas e variáveis socio-ambientais na morbidade respiratória e de dengue no Brasil. Diante da crescente preocupação com os impactos ambientais sobre a saúde pública, torna-se essencial entender como fatores como poluição atmosférica, temperatura, umidade e condições socioeconômicas influenciam a incidência de doenças na população. Entretanto, a complexidade inerente a esses sistemas, muitas vezes marcados por não linearidades e interações multivariadas, demanda o uso de abordagens mais robustas do que os testes estatísticos tradicionais.

Neste trabalho, desenvolveu-se e aplicou-se o sistema Aleph, um *framework* para inferência causal em séries temporais que permite a combinação de diferentes preditores baseados em técnicas de *machine* e *deep learning*, como, ARIMA, *Random Forest* e LSTM, juntamente com testes estatísticos para a realização da tarefa de forma simples e robusta. A metodologia foi validada com dados sintéticos e aplicada a um extenso conjunto de dados reais, composto por séries temporais socioeconômicas e ambientais mensais de municípios brasileiros entre 1999 e 2021, com foco em morbidade respiratória e dengue, baseado na métrica SMAPE para avaliação de desempenho e uma abordagem baseada em embaralhamento (*permutation test*) para inferência. Com isso, foi possível comparar os diferentes modelos quanto à sua capacidade de detectar relações causais consistentes a fim de fazer uma análise sobre os pontos positivos e negativos da aplicação de cada modelo para o contexto epidemiológico.

A realização deste trabalho permitiu a obtenção de resultados relevantes para a área de estudo. Os experimentos relacionados às doenças respiratórias indicaram como principais variáveis que possuem efeito causal a quantidade de CO_2, altamente relacionada aos índices de poluição, temperatura máxima, que possui relação com a queda de imunidade da população com quedas mais abruptas de temperatura que se tornaram mais significativas nos últimos anos e densidade demográfica que é diretamente relacionada à maior facilidade de transmissão de doenças respiratórias virais. Já para os experimentos relacionados a dengue, as principais variáveis causais foram o índice de urbanização, diretamente relacionado ao escoamento de águas pluviais nas cidades, densidade demográfica, que também se relacionada com a facilitação da transmissão da doença pelo contato vetor-humano, o índice de pluviosidade e a temperatura máxima, que são fatores muito relevantes para a proliferação do mosquito. Do ponto de vista computacional, os experimentos permitiram uma análise comparativa da aplicação causal de três modelos consolidados de *machine learning* baseados em redes neurais, permitindo a compressão sobre a eficiência da aplicação de cada modelo no contexto epidemiológico, devido às suas características construtivas e complexidade estrutural. A aplicação do algoritmo Aleph utilizando diferentes redes neurais se mostrou academicamente valiosa, permitindo consolidar o *framework* puramente *data-driven*, ou seja, que realiza a tomada de decisão baseada em dados reais, em

contextos diferentes do aplicado por Castro et al. (Castro et al., 2023), que pela primeira vez foi aplicado em uma utilização prática do modelo para a inferência causal em sistemas com alto grau de não linearidade como o domínio epidemiológico. Também foi possível comprovar a sua capacidade de generalização, se mostrando igualmente eficiente para o uso com a *Random Forest* e com a LSTM. Como fruto deste trabalho, conclui-se que as redes LSTM são altamente adequadas para a inferência causal de séries temporais epidemiológicas, como as estudadas aqui, mostrando que a sua sensibilidade para captura de padrões temporais complexos se faz bastante útil em contextos de variáveis fortemente correlacionadas, com única desvantagem a necessidade de maiores recursos para que experimentos mais completos sejam feitos tendo em vista a complexidade computacional do modelo. Além disso, comprovou-se que a rede ARIMA/ARIMAX se mostra menos adequada para experimentos como esse.

Por fim, é notável o potencial da utilização das redes LSTM para essa tarefa, mas se fazem necessários mais experimentos com *datasets* mais completos e diversificados, utilizando recursos mais robustos que suportem a complexidade imposta pela rede. Também se mostra válido aprofundar mais os estudos relacionados à utilização das redes ARIMA/ARIMAX junto com o algoritmo Aleph não somente para dados epidemiológicos, mas para outros cenários a fim de encontrar a melhor aplicação desta rede no processo de inferência causal.

REFERÊNCIAS

A., S.; MAHMOOD, A. Review of deep learning algorithms and architectures. *IEEE Access*, v. 7, p. 53040–53065, 2019.

AHMAD, W.; SHADAYDEH, M.; DENZLER, J. Causal inference in non-linear time-series using deep networks and knockoff counterfactuals. **2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)**, IEEE, 2021.

AHMAD, W.; SHADAYDEH, M.; DENZLER, J. Deep learning-based group causal inference in multivariate time-series. **Cornell University**, 2024.

BARNETT, L.; BARRETT, A. B.; SETH, A. K. Granger causality and transfer entropy are equivalent for gaussian variables. **Physical Review Letters**, American Physical Society, v. 103, p. 238701, 2009.

BATROUNI, M.; BERTAUX, A.; NICOLLE, C. Scenario analysis, from bigdata to black swan. **Computer Science Review**, Elsevier, v. 28, p. 131–139, 2018.

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

BO-TAO, Z.; JIN, Q. Changes of weather and climate extremes in the ipcc ar6. **Advances in Climate Change Research**, v. 17, n. 6, p. 713, 2021.

BONTEMPI, G.; TAIEB, S. B.; BORGNE, Y.-A. L. **Machine Learning Strategies for Time Series Forecasting**. [S.l.]: Second European Summer School, 2012.

BOX, G. E. P. et al. Analysis: Forecasting and control. **San francisco**, 1976.

BOX GWILYM M. JENKINS, G. C. R. G. E. P.; LJUNG, G. M. **Time Series Analysis: Forecasting and Control**. [S.l.]: Wiley, 2015.

BRASIL, O. Organização pan-americana de saúde. **Folha informativa–COVID-19 (doença causada pelo novo coronavírus). Disponível em: paho. org/bra/index. php**, 2020.

CASTRO, M. et al. Time series causal relationships discovery through feature importance and ensemble models. **Scientific Reports**, Nature Publishing Group UK London, v. 13, n. 1, p. 11402, 2023.

CUI, P. et al. Causal inference meets machine learning. In: **Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining**. [S.l.: s.n.], 2020. p. 3527–3528.

ELLWANGER, J. H. et al. Beyond diversity loss and climate change: Impacts of amazon deforestation on infectious diseases and public health. **Anais da Academia Brasileira de Ciências**, SciELO Brasil, v. 92, 2020.

FANG, X. et al. Forecasting incidence of infectious diarrhea using random forest in jiangsu province, china. **BMC Infectious Diseases**, 2020.

FARUQUE, O. et al. Ts-causalnn: Learning temporal causal relations from non-linear non-stationary time series data. **Cornell University**, 2024.

FISCHER, T.; KRAUSS, C. Deep learning with long short-term memory networks for financial market predictions. **European Journal of Operational Research**, v. 270, n. 2, p. 654–669, 2018.

FLORES, A.; TITO-CHURA, H.; APAZA-ALANOCA, H. Data augmentation for short-term time series prediction with deep learning. 2021.

GONçALVES, F. L. T.; COELHO, M. S. Z. S. Variação da morbidade de doenças respiratórias em função da variação da temperatura entre os meses de abril e maio em são paulo. **Ciência e Natura**, v. 32, n. 1, p. 103–117, 2010.

GRANGER, C. W. J. Investigating causal relations by econometric models and cross-spectral methods. **Econometrica**, Econometric Society, v. 37, n. 3, p. 424–438, 1969.

GREFF, K. et al. Lstm: A search space odyssey. **IEEE transactions on neural networks and learning systems**, IEEE, v. 28, n. 10, p. 2222–2232, 2016.

GU, W. et al. Use of random forest to estimate population attributable fractions from a case-control study of salmonella enterica serotype enteritidis infections. **Epidemiology & Infection**, Cambridge University Press, v. 143, n. 13, p. 2786–2794, 2015.

HAMILTON, J. D. **Time Series Analysis**. [S.l.]: Princeton University Press, 1994.

HIPPERT, H. S.; PEDREIRA, C. E.; SOUZA, R. C. Neural networks for short-term load forecasting: a review and evaluation. **IEEE Transactions on Power Systems**, v. 16, n. 1, p. 44–55, 2001.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. **Neural Computation**, v. 9, n. 8, p. 1735–1780, 1997.

HYNDMAN, R. J.; ATHANASOPOULOS, G. **Forecasting: Principles and Practice**. [S.l.]: Monash university, 2021.

HYNSMAN, R. J.; BOOTH, H. Stochastic population forecasts using functional data models for mortality, fertility and migration. **Internationa Journal of Forecasting**, v. 24, n. 3, p. 323–342, 2008.

KANDADI, T.; SHANKARLINGAM, G. Drawbacks of lstm algorithm: A case study. 2025.

KANE, M. J. et al. Comparison of arima and random forest time series models for prediction of avian influenza h5n1 outbreaks. **BMC Bioinformatics**, v. 15, p. 276, 2014. Disponível em: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-15-276.

KARIM, F. et al. Lstm fully convolutional networks for time series classification. **IEEE access**, IEEE, v. 6, p. 1662–1669, 2017.

LI, L. et al. A causal inference model based on random forests to identify the effect of soil moisture on precipitation. **American Metereological Society**, 2020.

LIANG, X. S. Normalizing the causality between time series. **Physical Review E**, American Physical Society, v. 92, p. 022126(6), 2015.

LUO, Y.; PENG, J.; MA, J. When causal inference meets deep learning. **Nature Machine Intelligence**, Nature Publishing Group UK London, v. 2, n. 8, p. 426–427, 2020.

MARINO, D. L.; AMARASINGHE, K.; MANIC, M. Building energy load forecasting using deep neural networks. In: IEEE. **IECON 2016-42nd annual conference of the IEEE industrial electronics society**. [S.l.], 2016. p. 7046–7051.

MAROTZKE, J.; SEMMANN, D.; MILINSKI, M. The economic interaction between climate change mitigation, climate migration and poverty. **Nature Climate Change**, Nature Publishing Group UK London, v. 10, n. 6, p. 518–525, 2020.

MCKINLEY, S.; LEVINE, M. **Cubic spline interpolation**. 2014. Acesso em: 27 jun. 2025. Disponível em: https://www.rajgunesh.com/resources/downloads/numerical/cubicsplineinterpol.pdf.

MEIRA, M. C. R. et al. Influência do clima na ocorrência de dengue em um município brasileiro de tríplice fronteira. **Cogitare Enfermagem**, v. 26, 2021.

MENCHETTI, F.; CIPOLLINI, F.; MEALLI, F. Estimating the causal effect of an intervention in a time series setting: the c-arima approach. **Econometrics**, Cornell University, 2021.

MITCHELL, T. M. **Machine Learning**. [S.l.]: McGraw-Hill, 1999.

MOORE, D. S.; MCCABE, G. P.; CRAIG, B. A. **Introduction to the Practice of Statistics**. [S.l.]: WH Freeman New York, 2009. v. 4.

NAUTA, M.; BUCUR, D.; SEIFERT, C. Causal discovery with attention-based convolutional neural networks. **Machine Learning and Knowledge Extraction**, v. 1, n. 1, p. 312–340, 2019.

NEIDERUD, C. J. How urbanization affects the epidemiology of emerging infectious diseases. **Infection Ecology & Epidemiology**, v. 5, n. 1, 2015.

Organização das Nações Unidas. **Causas e Efeitos das Mudanças Climáticas**. 2025. Página web da ONU. Acessado em 13 de maio de 2025. Disponível em: https://www.un.org/pt/climatechange/science/causes-effects-climate-change.

PEARL, J. **Causality: Models, Reasoning, and Inference**. [S.l.]: Cambridge University Press, 2003. v. 19(4).

RAIZADA, S.; MALA, S.; SHANKAR, A. Vector borne disease outbreak prediction by machine learning. In: IEEE. **2020 International conference on smart technologies in computing, electrical and electronics (ICSTCEE)**. [S.l.], 2020. p. 213–218.

RANKING do Saneamento 2023. Trata Brasil, 2023. Disponível em: https://tratabrasil.org.br/ranking-do-saneamento-2023/. Acesso em: 30 jul. 2025.

RUNGE, J. et al. Inferring causation from time series in earth system sciences. **Nature Communications**, 2019.

SCHRAUFNAGEL, D. E. et al. Air pollution and noncommunicable diseases: A review by the forum of international respiratory societies' environmental committee, part 2: Air pollution and organ systems. **Chest**, v. 155, n. 2, p. 417–426, 2019.

SETH, A. K.; BARRETT, A. B.; BARNETT, L. Granger causality analysis in neuroscience and neuroimaging. **JNeurosci**, Society for Neuroscience, v. 35(8), p. 3293–3297, 2015.

SHI, X. et al. Convolutional lstm network: A machine learning approach for precipitation nowcasting. **Advances in neural information processing systems**, v. 28, 2015.

SOUZA, W. M. de; WEAVER, S. C. Effects of climate change and human activities on vector-borne diseases. **Nature Reviews Microbiology**, Nature Publishing Group UK London, p. 1–16, 2024.

STOCK, J. H.; WATSON, M. W. Vector autoregressions. **Journal of Economic Perspectives**, American Economic Association, v. 15(4), p. 101–115, 2001.

STROBL, C. et al. Conditional variable importance for random forests. **BMC bioinformatics**, Springer, v. 9, n. 1, p. 307, 2008.

TANK, A. et al. Neural granger causality for nonlinear time series. **IEEE Transactions on Pattern Analysis and Machine Interlligence**, IEEE, v. 44(8), 2018.

TAUIL, P. L. Urbanização e ecologia da dengue. **Ciência e Natura**, v. 32, n. 1, p. 103–117, 2010.

VALIDATED, C. **ARIMA forecast insensitive to inclusion of exogenous variables**. 2016. Disponível em: https://stats.stackexchange.com/questions/200739/arima-forecast-insensitive-to-inclusion-of-exogenous-variables. Acesso em: 30 jul. 2025.

WANG, H. et al. A methodoly for calculating the contribuition of exogenous variables to arimax predictions. v. 32, n. 1, p. 103–117, 2010.

WEI, W. W. S. Univariate and multivariate methods. **TIME SERIES ANALYSIS**, Addison-Wesley Publishing Company, 2006.

XU, J. G. e B. Complex systems, emergence, and multiscale analysis: A tutorial and brief survey. **Applied Sciences**, MDPI, v. 11, n. 12, p. 5736, 2021.

YANG, Z. et al. Modified seir and ai prediction of the epidemics trend of covid-19 in china under public health interventions. **Journal of Thoracic Disease**, v. 12, n. 3, p. 165–174, 2020.

ZHAO, Z. et al. Study on the prediction effect of a combined model of sarima and lstm based on ssa for influenza in shanxi province, china. **BMC Infectious Diseases**, Springer, v. 23, n. 1, p. 71, 2023.

ZHOU, F.; SUN, H.; ZHENG, S. Forecasting incidence of infectious diarrhea using random forest in jiangsu province, china. **BMC Infectious Diseases**, v. 20, p. 222, 2020. Disponível em: https://bmcinfectdis.biomedcentral.com/articles/10.1186/s12879-020-4930-2.