Uma Comparação entre Métodos baseados em Aprendizado de Máquina para inferir número de casos semanais de Dengue

Giovanni E. Zanardo[1], **Éfren L. Souza**[1,2], **Fabíola G. Nakamura**[1], **Eduardo F. Nakamura**[1]

[1] Instituto de Computação – Universidade Federal do Amazonas (UFAM)
[2]Instituto de Engenharia e Geociências – Universidade Federal do Oeste do Pará (UFOPA)

{gio.zanardo, fabiola, nakamura}@icomp.ufam.edu.br, efren.souza@ufopa.edu.br

Abstract. *Arboviruses transmitted by Aedes aegypti and Aedes albopictus are among the leading public health problems, with dengue being the most prominent. Managing dengue epidemics requires advanced preparation; thus, predicting the cases in a specific region can assist in prevention strategies and control the epidemic process. With this in view, this study evaluates the efficiency of classic statistical techniques and machine learning methods in predicting dengue cases from geographic data of San Juan, Puerto Rico. For this, we selected features using the cross-correlation matrix with the total number of weekly dengue cases, which were subsequently filtered by wavelet transformations. The Linear Regression model, using precipitation levels and vegetation filtered by the symmlet wavelet (sym20), showed the best performance on the metrics MAE,* R^2*, MAPE, RMSE, and BIAS.*

Resumo. *As arboviroses transmitidas pelo Aedes aegypti e Aedes albopictus estão entre os principais problemas de saúde pública, sendo a dengue a mais proeminente. O manejo de epidemias de dengue requer preparação avançada; assim, prever os casos em uma região específica pode auxiliar nas estratégias de prevenção e controle do processo epidêmico. Com isso em vista, este estudo avalia a eficácia de técnicas estatísticas clássicas e métodos de aprendizado de máquina na predição de casos de dengue a partir de dados geográficos de San Juan, Porto Rico. Para isso, selecionamos características usando a matriz de correlação cruzada com o número total de casos semanais de dengue, que foram posteriormente filtrados por transformações wavelet. O modelo de Regressão Linear, utilizando níveis de precipitação e vegetação filtrados pela wavelet symmlet (sym20), mostrou o melhor desempenho nas métricas MAE,* R^2*, MAPE, RMSE e BIAS.*

1. Introdução

Arboviroses são doenças transmitidas por arbovírus, sendo estes, vírus transmitidos pela picada de artrópodos hematófagos [da Silva & Angerami 2008]. A temperatura, os ciclos hídricos e umidade são fatores determinantes para a inserção e disseminação de cada arbovirose nos diversos ambientes [Celentano et al. 2008]. Em geral, países tropicais favorecem o desenvolvimento e a proliferação de vetores, aumentando a incidência dessas doenças nessas localidades [Lopes et al. 2014]. Dentre as arboviroses, destaca-se a dengue.

Dengue é causada pelo vírus da dengue, arbovírus da família Flaviviridae, gênero Flavivírus e possui quatro tipos imunológicos: DEN-1, DEN-2, DEN-3 e DEN-4

[Ross 2010]. Em poucos casos, a dengue pode evoluir para uma forma grave, aumentando significativamente o risco de morte. Esse estágio da doença é marcado por sangramento, redução nos níveis de plaquetas no sangue, extravasamento de plasma e queda da pressão arterial para níveis perigosamente baixos. [Ross 2010]. Atualmente, a dengue é a arbovirose mais comum que atinge a humanidade, sendo responsável por cerca de 100 milhões de casos/ano em uma população de risco de 2,5 a 3 bilhões de seres humanos [World Health Organization 2009]. Historicamente, a dengue possui atenção de agências sanitárias, visto que a resolução WHA58.3 da Assembleia Mundial da Saúde de 2005, sobre a revisão do Regulamento Sanitário Internacional (RSI), inclui a dengue como um exemplo de doença que constitui uma emergência de saúde pública de interesse internacional [World Health Organization 2009], com implicações para a segurança sanitária devido à sua capacidade de provocar interrupções e rápida disseminação epidêmica além das fronteiras nacionais.

A disseminação epidêmica é verificada através do aumento significativo na incidência global de dengue nas duas últimas décadas, representando um empecilho para a saúde pública. Entre 2000 e 2019, a Organização Mundial da Saúde (OMS) documentou um aumento de dez vezes nos casos reportados em todo o mundo, de 500.000 para 5,2 milhões [World Health Organization 2023]. O ano de 2019 registrou um pico sem precedentes, com casos reportados em 129 países, seguidos de um ligeiro declínio nos casos entre 2020 e 2022, devido à pandemia de COVID-19 e à menor taxa de notificação. Porém, o segundo semestre de 2023 testemunhou um aumento alarmante de casos em todo o mundo. Este aumento é caracterizado por um crescimento significativo no número, escala e ocorrência simultânea de múltiplos surtos, expandindo-se para regiões anteriormente não afetadas pela dengue, de forma que o total acumulado de casos para o ano de 2023 superou todos os totais anuais anteriores e, em alguns países, a transmissão se estendeu para além das áreas conhecidamente afetadas (América do Sul, México, América Central e países do Caribe). Nas Américas, os casos de dengue aumentaram nas últimas quatro décadas, passando de 1,5 milhão de casos de 1980 a 1989 para 17,5 milhões em 2010-2019. Antes de 2023, o maior número histórico de casos de dengue foi registrado em 2019, com mais de 3,18 milhões de casos, 28.208 casos graves e 1.823 mortes, resultando em uma taxa de letalidade de 0,06 %.

Diferenças climáticas são usadas para explicar a sazonalidade da dengue. Entretanto, essa sazonalidade pode ser influenciada por múltiplos fatores demográficos, como taxas de natalidade, imigração e mobilidade a curto prazo, e ainda pela co-circulação de distintas formas virais, o que dificulta o uso de abordagens estatísticas tradicionais no cenário da predição de surtos [San Martin et al. 2010]. Dessa forma, compreender e predizer surtos de dengue é desafiador, devido à complexa interação entre os determinantes ambientais – fatores que afetam a saúde e o bem-estar humanos – e seus respectivos ambientes [Cabrera et al. 2022]. Essa dinâmica culmina em variações nos padrões de incidência da doença por diferentes regiões geográficas.

Por consequência, este estudo visa explorar a eficiência preditiva de uma gama de modelos, abrangendo tanto técnicas estatísticas clássicas quanto métodos rasos e profundos de aprendizado de máquina. Entre os modelos estatísticos clássicos avaliados estão o Passeio Aleatório (RW) [Morettin & Toloi 2018], o modelo linear de Holt, a Suavização Exponencial Simples (SES), o modelo de Holt-Winters (ES) e o modelo Au-

torregressivo Integrado de Médias Móveis (ARIMA). Paralelamente, investigamos arquiteturas de aprendizado profundo e modelos rasos regressivos, incluindo Regressão Linear (LR) [James et al. 2023], Support Vector Regression (SVR) [Drucker et al. 1996], Random Forest Regression (RFR) [Breiman 2001], Gradient Boosting Regressor (GBR) [Friedman 2001] e XGBoost [Chen & Guestrin 2016], visando comparar o desempenho individual de cada modelo na predição de número de casos semanais de dengue.

Além da comparação entre modelos, este trabalho verifica estratégias eficientes para a reformatação de conjuntos de dados constituídos por séries temporais [Benidis et al. 2022], adequando-os para aplicação em modelos supervisionados de aprendizado de máquina. A seleção das características mais relevantes e a determinação dos modelos mais eficazes constituem etapas intermediárias para a obtenção de modelos para predição eficiente do número de casos semanais de dengue em San Juan [US National Oceanic and Atmospheric Administration 2017].

As seções do trabalho são estruturadas da seguinte forma. A seção 2 apresenta trabalhos relacionados e suas contribuições. A seção 3 explicita a base de dados trabalhada. A seção 4 define as métricas de avaliação dos resultados. A seção 5 exibe o pré-processamento dos dados. A seção 6 relata os experimentos e detalhes de implementação. A seção 7 analisa os resultados alcançados e os compara com o *baseline* estabelecido. A seção 8 consolida os principais achados, expõe as limitações deste estudo e sugere direções para trabalhos futuros.

2. Trabalhos Relacionados

Panja et al. propuseram um conjunto (*ensemble*) de redes neurais autorregressivas chamado XEWNet (*External Ensemble Wavelet Neural Network*). Essas redes processam séries temporais discretas previamente tratadas pela transformada *wavelet* de Haar. Após o processamento, os valores gerados pelas redes são associados aos níveis de precipitação da região em estudo, resultando na predição final. A precipitação da região demonstrou uma alta causalidade de Granger em relação ao número de casos semanais de dengue, motivando o uso dessa característica como componente auxiliar ou externo do *ensemble*. No total, os autores usaram séries temporais contendo informações geográficas e registros semanais de casos de dengue das regiões de San Juan, Iquitos e Ahmedabad. A preparação dos dados envolveu a aplicação de previsões de curto e longo prazo, com janelas temporais de 26 e 52 semanas, de tal sorte que a janela de 26 semanas apresentou os melhores resultados nas métricas RMSE e MAE, com valores de 7,69 e 5,66, respectivamente, para a cidade de San Juan.

Buczak et al. utilizaram um *ensemble* de três modelos distintos: o Método de Análogos Bidimensionais, que incorpora dados climáticos e o número de casos semanais de dengue; Holt-Winters, com componente aditivo e sazonal, aplicado tanto aos dados brutos quanto aos suavizados pela *wavelet* sym7; e Modelos Históricos Simples. Um conjunto de hiperparâmetros foi testado para cada modelo, e os melhores compuseram o preditor final. O conjunto de dados incluiu informações geográficas e o número de casos semanais de dengue das cidades de San Juan e Iquitos. As métricas utilizadas foram a altura do pico da série temporal, o pico semanal da série e o número de casos em uma temporada específica, estabelecidas conforme as diretrizes de uma competição de predição de dengue. As principais contribuições do estudo são a acurácia na predição

do total de casos semanais e da altura do pico da série temporal para os dados de Iquitos.

Shaikh et al. propuseram a criação de um sistema de aviso prévio para predizer dengue em San Juan e Iquitos, além de prover um sistema inteligente que apresenta possíveis medidas preventivas para os pacientes da região. Um pré-processamento é feito na base de dados para retirar outliers e dados faltantes. Posteriormente, é feita uma seleção de características utilizando o algoritmo heurístico *Neighbour Count-based Dragonfly Electric Fish Optimization* (NC-DEFO). Tais características são levadas ao comitê de classificadores formado pelos modelos ANN, CNN e SVM. Uma vez que a predição para febre de dengue é realizada, o sistema informa possíveis medicamentos para fortalecer o sistema imunológico e prescrições médicas. Os autores empregaram diversas métricas de avaliação, incluindo RMSE, norma L2, norma L1, MAE, norma L-infinito, SMAPE, MASE e MEP, para analisar o desempenho dos modelos. Concluíram que o comitê de classificadores denominado Optimized Ensemble Classifier (OEC) demonstra uma capacidade de predição superior quando comparado com modelos isolados como ANN, CNN, SVM, bem como LSTM-RF, e também com um comitê de classificadores formado por estes três últimos modelos, mas sem a aplicação do NC-DEFO.

3. Base de Dados

A base de dados [US National Oceanic and Atmospheric Administration 2017] é formada da quantidade de casos de dengue reportados semanalmente na cidade de San Juan (Porto Rico) no período de 1990 a 2010. A base possui formato tabular, composta por 936 instâncias, constituída de 24 atributos, em que a maioria dos atributos representa uma série temporal. De forma explícita, os atributos são provenientes de quatro fontes distintas: GHCNd (Global Historical Climatology Network daily) – banco de dados integrado de resumos climáticos diários de estações de superfície terrestres ao redor do mundo; PERSIANN – algoritmo de recuperação de precipitação baseado em satélite que fornece informações de precipitação quase em tempo real usando redes neurais; NOAA NCPE (National Centers for Environmental Prediction); e NDVI (Índice de Vegetação de Diferença Normalizada) da NOAA. Abaixo listamos os atributos usados por fonte:

- GHCNd: cidade em questão, semana de registro, ano de registro, temperatura máxima, temperatura mínima, temperatura média, total de precipitação em milímetros e faixa de temperatura diurna;
- PERSIANN: precipitação total em milímetros;
- NOAA NCPE: precipitação total em milímetros, temperatura média do ponto de orvalho, temperatura média do ar, umidade relativa média, umidade específica média, precipitação em quilograma por metro quadrado, temperatura máxima do ar, temperatura mínima do ar, temperatura, média do ar e faixa de temperatura média do período diurno;
- NDVI: quantidade de pixels a sudeste do centróide da cidade, quantidade de pixels a sudoeste do centróide da cidade, quantidade de pixels a nordeste do centróide da cidade e quantidade de pixels a noroeste do centróide da cidade.

4. Métricas

Para avaliar os modelos, seis métricas foram utilizadas: Erro Absoluto Médio (MAE), Percentual de Erro Médio Absoluto (MAPE), a Raiz Quadrada do Erro Quadrático Médio

(RMSE), Viés Médio (BIAS) e R^2 (coeficiente de determinação), dadas, respectivamente, por

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (1), \qquad \text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad (2),$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (3), \qquad \text{BIAS} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i) \quad (4),$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (5),$$

onde consideramos y_i como os valores reais, $\hat{y}_i$ como os valores preditos e n como o número total de observações.

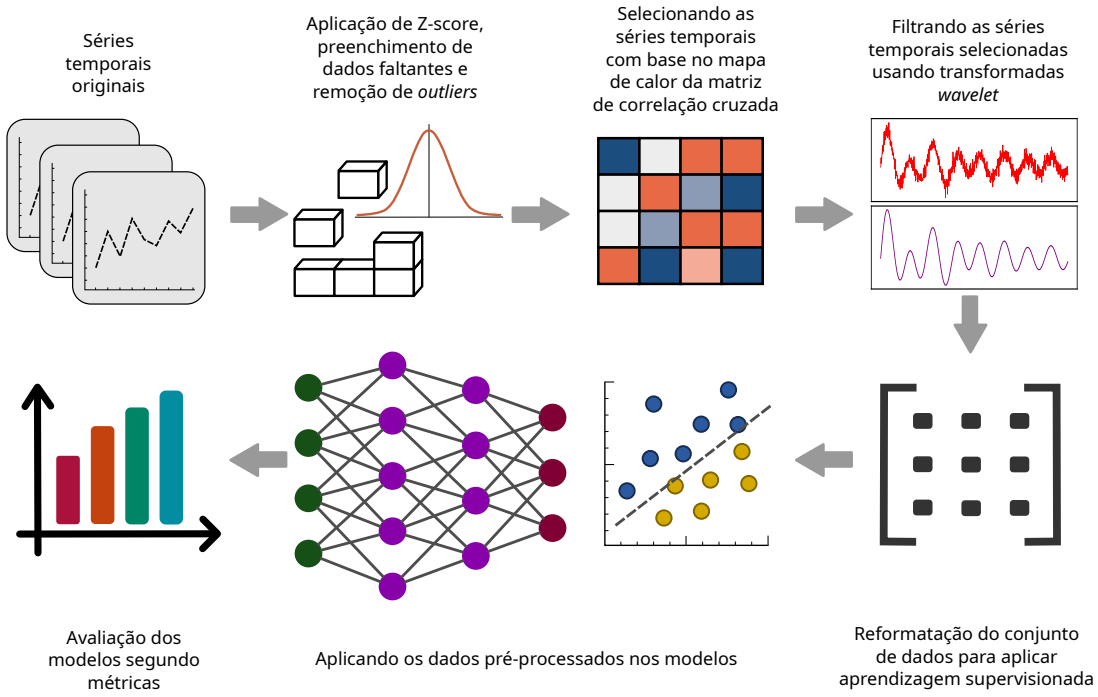

Figura 1. Panorama geral da experimentação.

5. Pré-processamento

Primeiro, os dados ausentes são preenchidos por meio de interpolação polinomial. Em seguida, apenas a coluna de casos totais de dengue é selecionada para a aplicação do Z-score [Kreyszig 2010]. Logo, cada instância dessa coluna é subtraída da média dos

valores da coluna e dividida pelo desvio padrão. Após este processo, as instâncias com Z-score maior que -3 e menor que 3 são selecionadas. Esta abordagem identifica valores aberrantes (*outliers*) ao excluir pontos que estão a mais de três desvios padrão da média.

Uma série temporal é caracterizada por sua tendência, sazonalidade e resíduo [Morettin & Toloi 2018]. Algoritmos estatísticos clássicos como ARIMA, exigem estacionariedade da série por hipótese. Portanto, aplica-se o Teste Aumentado de Dickey-Fuller [Fuller 1976] para verificação da estacionariedade das séries.

Com este pré-processamento, alguns modelos são testados e verificamos, após a visualização gráfica da série temporal do total de casos e dos valores preditos, que certos pontos não são captados pelos modelos, por mais que um ajuste fino fosse realizado. Assim, ainda no processo de retirada de *outliers*, há a extração manual destes pontos aberrantes na base de dados de San Juan. Com este tratamento inicial feito, passamos para formatação dos dados nos modelos testados.

Para os modelos estatísticos clássicos abordados, como Holt-Winters ou Suavização Exponencial, modelo linear de Holt, Suavização Exponencial Simples, ARIMA e Passeio Aleatório, tomamos apenas a coluna de casos totais dos conjuntos de dados, porque esses modelos realizam predição apenas em contexto univariado [Morettin & Toloi 2018].

Os modelos rasos e profundos de aprendizado de máquina necessitam de reformatação da base de dados para aplicação de aprendizado supervisionado. A literatura aborda três estratégias principais: a predição *one-step*, *multi-step* e a *multi-output* [Benidis et al. 2022]. Dentro destas estratégias, duas abordagens são testadas, sendo estas, a abordagem inocente (univariada) e a não inocente (multivariada). De um lado, a inocente consiste na construção de janelas temporais baseadas no uso exclusivo da coluna de casos totais semanais de dengue. Por outro lado, a não inocente usa mais de uma coluna para formar a janela temporal, na qual cada elemento da janela representa um vetor de dimensão correspondente ao número de características escolhidas.

De maneira intuitiva, para a formatação supervisionada desejada, necessitamos formar duas matrizes denotadas por X (matriz formada pelas janelas temporais) e Y (a matriz com os valores alvos, formada por valores da coluna de casos totais e/ou mais colunas). De maneira precisa, seja $T > 0$ o tamanho da janela, $c > 0$ o número de colunas (características selecionadas) e $n > 0$ o número de instâncias do conjunto de dados. No caso *one-step*, $X \in \mathbb{R}^{(n-T)\times T\times c}$ é a matriz de janelas temporais, onde temos $n - T$ janelas, em que cada janela tem tamanho T e cada elemento da janela tem tamanho c (c -upla). A matriz $Y \in \mathbb{R}^{(n-T)\times 1}$ é a coluna alvo, formada apenas pelos casos totais. Para o caso *multi-step*, aproveitamos o treinamento do modelo feito de maneira *one-step*. Inicialmente, define-se o vetor $J \in \mathbb{R}^{T\times c}$ como a primeira janela dos dados de teste. Realiza-se a predição do valor $p \in \mathbb{R}^c$ com base em J. Em seguida, cada elemento de J com índice i é deslocado para o índice $i - 1$, em que o primeiro elemento de J ocupará agora a última posição de J. Ao término, substitua o elemento da última posição do vetor J com o valor p e repita o processo até que chegue na quantidade de amostras de teste previamente definida. Para o *multi-output*, temos $X \in \mathbb{R}^{(n-T_x-T_y)\times T_x\times c_1}$ e $Y \in \mathbb{R}^{(n-T_x-T_y)\times T_y\times c_2}$, em que T_x e T_y não são necessariamente iguais, c_1 são as características utilizadas como base na predição e c_2 são as características dos valores preditos.

Em uma palavra, usamos T_x c_1-uplas para predizer T_y c_2-uplas. Portanto, repare que para *one-step* e *multi-step*, o caso inocente toma $c = 1$ e o não inocente toma $c > 1$. Para o caso *multi-output*, se $c_1 = c_2 = 1$, tem-se o caso inocente. Se $c_1 > 1$ e $c_2 > 1$, temos o caso não inocente.

No contexto não inocente, faz-se uma engenharia de características para selecionar aquelas que geram maior poder preditivo, ou seja, escolhem-se as características que captam a tendência geral dos dados e que possuem melhor desempenho em métricas. Com efeito, inicialmente observamos a matriz de correlação cruzada [Derrick & Thomas 2004] (Figura 2) entre o número de casos semanais de dengue e as características representadas pelos índices das linhas da Figura 2: temperatura mínima (0), total de precipitação em milímetros da mensuração feita pelo NOAA (1), reanálise da precipitação (2), quantidades de pixels relativos à vegetação (3 - 6), reanálise de temperatura máxima (7), reanálise de temperatura mínima (8), reanálise da variação de temperatura diurna (9), análise da temperatura diurna (10), temperatura máxima da semana (11), total de precipitação em milímetros do satélite PERSIANN (12), reanálise da temperatura do ar (13), reanálise do ponto do orvalho (14), reanálise do nível de precipitação em quilograma por metro quadrado (15), reanálise do percentual de umidade relativa (16), reanálise do percentual de umidade específica (17) e temperatura média (18).

Diante desta análise, realizamos diversas combinações, onde uma delas consiste na escolha de características (linhas do mapa de calor) com maior correlação positiva (linhas de índices 2, 3, 4, 5, 8, 9, 10, 12, 13) e *lag* temporal de tamanho 17. Porém, a melhor combinação de características obtida neste trabalho consiste nos índices de 0 a 6 (linhas de 0 a 6 do mapa de calor) mais a característica que representa o número de casos totais e *lag* temporal de tamanho 22. Portanto, constatamos que o nível de precipitação e a vegetação do ambiente são fatores mais relevantes para a predição do número de casos semanais, o que ressalta, de forma empírica, os resultados obtidos por Santos et al. (2019) e Panja et al. (2023).

Após selecionar características relevantes de nossos conjuntos de dados, aplicamos a transformada *wavelet* a cada série temporal selecionada. Este método [Strang & Nguyen 1996] divide o sinal em componentes de alta frequência, chamados detalhes, e baixa frequência, chamados aproximação. O processo de decomposição em níveis é um hiperparâmetro na análise por *wavelets*. Em cada nível de decomposição, o componente de aproximação do nível anterior é novamente dividido em uma nova aproximação e detalhes. Os coeficientes de detalhes de cada nível representam diferentes bandas de frequência. Por exemplo, no primeiro nível, os detalhes capturam as frequências mais altas do sinal original, enquanto no segundo nível, eles representam as frequências mais altas dentro da metade inferior do espectro de frequência do sinal original, e assim por diante. O número máximo de níveis de decomposição depende tanto do comprimento do sinal quanto do tipo de *wavelet* utilizado, e a escolha adequada do número de níveis é essencial para evitar decomposições em que os coeficientes de aproximação ou detalhe tornam-se insignificantes ou não informativos. Em geral, níveis mais altos permitem uma análise mais fina dos detalhes do sinal, enquanto níveis mais baixos são melhores para identificar tendências gerais. Ao longo das experimentações, testamos todas as *wavelets* discretas disponíveis na biblioteca PyWavelets [Lee et al. 2019] e deixamos o nível de decomposição igual a um. Por fim, com as características desejadas

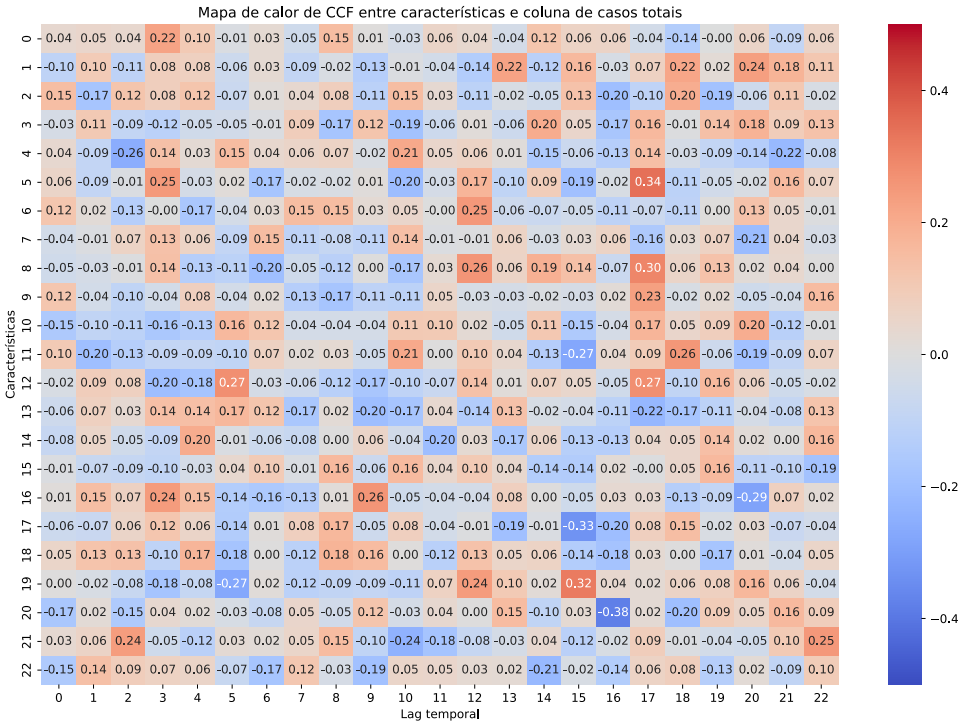

Figura 2. Mapa de Calor da Correlação Cruzada entre as características e a coluna de casos totais.

e transformadas pela *wavelet* escolhida em dois contextos distintos, construímos as matrizes X e Y definidas acima. Posteriormente, dividimos estes dados em treino/teste na proporção 80/20 e passamos para a fase de experimentação dos modelos e comparação de resultados obtidos.

6. Experimentação

Os modelos estatísticos foram implementados com a biblioteca statsmodels [Seabold & Perktold 2010]. Os modelos de aprendizado de máquina foram criados usando scikit-learn, com ajuste apenas no SVR para usar *kernel* linear. Os modelos profundos foram desenvolvidos em tensorflow versão 2.15.0, explorando três arquiteturas: LSTNet [Lai et al. 2017]; combinação de CNN e LSTM; e uma versão modificada desta última com componente autorregressivo, similar à LSTNet. As configurações incluíram função de ativação ReLu, função de perda MAE, e otimizador Adam com taxa de aprendizado de 0,01. Utilizou-se a camada de *dropout* visando prevenir *overfitting*. Em cada rede, usamos parada antecipada baseada na MAE com paciência de 10 épocas, e o treinamento foi definido para 150 épocas, mas geralmente encerrado antes pelo critério de parada. As camadas convolutivas contaram com 16 filtros de tamanho dois e as recorrentes com 32 unidades. Testamos a presença e ausência de camadas de *pooling*, inspirados pela arquitetura da LSTNet. A melhor janela para predição *one-step* foi de $T = 22$ e para *multi-output*, utilizamos $T_x = 22$ e $Ty = 11$.

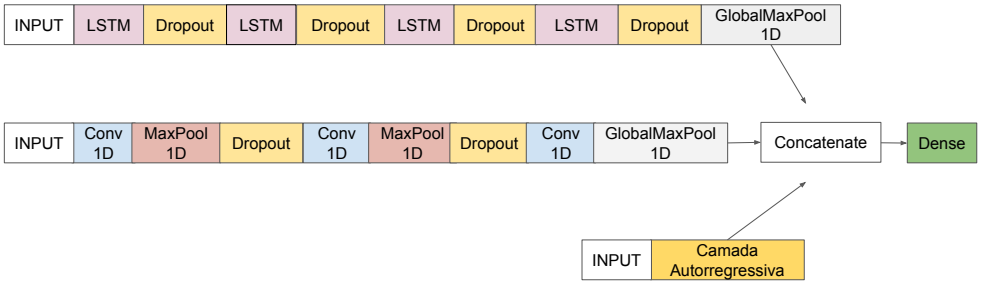

Figura 3. Modelo CNN-LSTM-LR com uso de *pooling*.

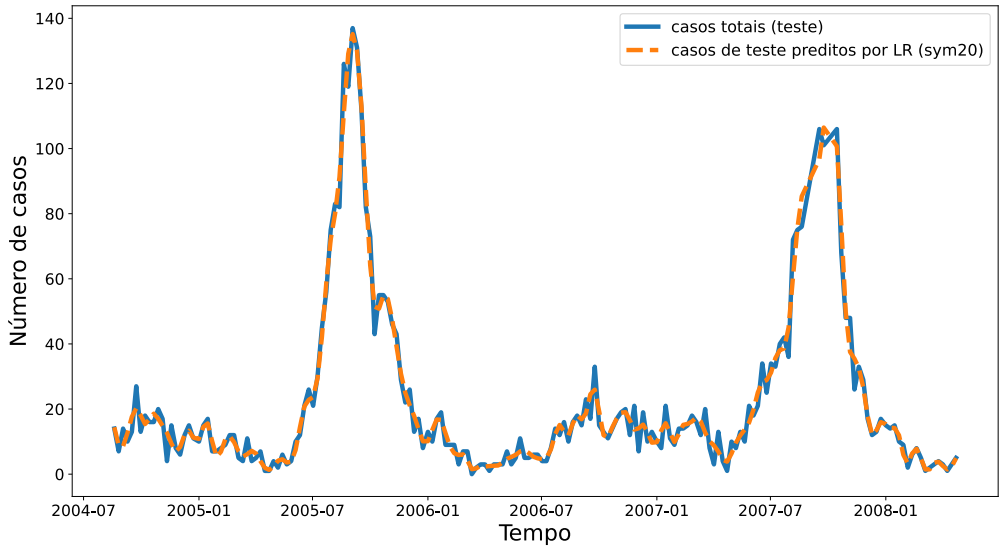

Figura 4. Comparação entre predições de LR e casos totais de dengue não filtrados.

7. Resultados

Na Tabela 1, visando obter consistência nos resultados, comparamos os resultados preditos com a coluna de casos totais filtrada pela *wavelet* em questão. Entretanto, como o objetivo principal consiste em comparar os resultados preditos com a realidade apresentada, temos na Tabela 2 os resultados preditos comparados com a coluna de casos totais não filtrada dos melhores modelos obtidos na Tabela 1. Além disso, a Figura 4 mostra o ajuste dos resultados preditos para com a coluna de casos totais não filtrada.

[2]Na Tabela 1, o símbolo * indica que as camadas de convolução são seguidas de *pooling*. Já ** indica que as camadas de convolução não são seguidas de camada de *pooling*. Note a piora sem o uso de *pooling*.

		TREINO					TESTE				
ANÁLISE	**MODELO**	MAE	R^2	MAPE	RMSE	BIAS	MAE	R^2	MAPE	RMSE	BIAS
INOCENTE (UNIVARIADA)	SES	8,24	0,934	2,41e+13	14,13	-0,012	3,13e+14	-0,141	17,86	33,38	-11,75
	HOLT	8,31	0,934	4,13e+13	14,17	0,959	2,03e+15	-8,159	80,43	94,56	75,52
	ES	9,46	0,924	2,88e+13	15,17	-0,138	5,32e+15	-95,08	271,04	306,27	0,0
	RW	8,25	0,934	2,41e+13	14,14	-0,012	1,69e+14	0,868	6,94	11,37	0,043
	ARIMA	8,21	0,939	6,02e+13	13,58	-0,038	8,70e+14	-0,127	27,53	33,17	0,0
	LR one-step	8,28	0,942	5,58e+13	13,37	8,95e-15	2,41e+14	0,862	7,37	11,59	0,755
	SVR one-step	8,03	0,938	4,80e+13	13,86	-1,13	1,94e+14	0,874	6,97	11,09	-0,328
	GBR one-step	4,48	0,989	7,49e+13	5,81	-4,10e-16	1,38e+14	0,867	6,67	11,39	1,084
	RFR one-step	3,19	0,988	2,08e+13	6,03	0,045	1,63e+14	0,861	7,04	11,64	1,786
	XGB one-step	0,113	0,99999	1,70e+12	0,164	0,000236	2,27e+14	0,829	7,75	12,92	0,639
	LR multi-step	8,28	0,942	5,58e+13	13,37	8,95e-15	7,55e+14	-0,028	24,49	31,68	5,49
	SVR multi-step	8,03	0,938	4,80e+13	13,86	-1,13	1,87e+14	-0,298	19,17	35,59	-17,08
	GBR multi-step	4,48	0,989	7,49e+13	5,81	-4,10e-16	4,16e+14	-0,021	17,99	31,56	-9,37
	RFR multi-step	3,19	0,988	2,08e+13	6,03	0,045	1,87e+14	-0,298	19,19	35,59	-17,14
	XGB multi-step	0,113	0,99999	1,70e+12	0,164	0,000236	5,96e+14	-0,284	26,44	35,41	6,56
	LR multi-output	18,05	0,57	2,74e+14	36,87	4,61e-16	3,92e+14	0,34	17,19	25,29	5,12
	RFR multi-output	5,59	0,95	3,95e+13	11,92	0,03	2,41e+14	-0,78	21,67	41,46	11,00
	XGB multi-output	0,13	1,00	1,93e+12	0,19	-0.00017	0,30	0,74	9,67	22,44	0,73
	GBR multi-output	39,52	-0,79	8,11e+14	74,92	0,00	9,44e+14	-2,25	32,43	56,06	9,60
	SVR multi-output	28,54	-0,08	4,42e+14	58,09	-13,05	5,88e+14	-0,19	22,05	33,98	-1,46
NÃO INOCENTE (MULTIVARIADA)	RFR (rbio2.4)	3,24	0,955	0,126	6,00	-0,049	3,76	0,946	0,204	6,18	-0,650
	XGB (rbio2.4)	2,29	0,988	0,126	3,05	0,000	3,71	0,937	0,204	6,64	-0,318
	GB (rbio2.4)	1,95	0,992	0,111	2,53	-0,000	3,84	0,940	0,216	6,50	-0,408
	LR (rbio2.4)	2,16	0,986	0,174	3,37	-0,000	3,09	0,975	0,534	4,22	-0,297
	SVR (rbio2.4)	1,85	0,977	0,124	4,24	0,189	2,29	0,979	0,247	3,86	-0,092
	CNN-LSTM (rbio2.4)*	3,53	0,956	0,213	5,71	-684,93	4,18	0,945	0,290	6,27	-391,85
	CNN-LSTM (rbio2.4)**	5,60	0,889	1,083	8,76	-103,92	15,38	0,250	0,543	18,81	-1909,09
	LSTNet (rbio2.4)	19,20	-10,80	1,61	29,20	12420,50	14,80	-9,35	1,40	25,27	1850,00
	CNN-LSTM-LR (rbio2.4)	1,89	0,98	0,12	3,22	1,15	3,14	0,97	0,51	4,23	-89,67
	RFR (coif8)	3,12	0,960	0,124	5,63	-0,034	3,74	0,949	0,206	5,98	-0,705
	XGB (coif8)	1,72	0,994	0,102	2,25	0,0009	3,93	0,933	0,204	6,84	-0,583
	GB (coif8)	1,79	0,993	0,103	2,33	-0,0000013	3,84	0,941	0,206	6,47	-0,695
	LR (coif8)	0,28	0,9998	0,025	0,39	0,000006	0,40	0,9996	0,076	0,51	-0,032
	SVR (coif8)	0,81	0,997	0,052	1,62	0,097	0,99	0,997	0,105	1,48	-0,039
	CNN-LSTM (coif8)*	4,91	0,936	0,387	6,72	2158,42	6,76	0,876	0,642	8,48	20,16
	LSTNet (coif8)	18,20	-80,02	1,02	29,50	9641,68	14,42	-68,40	0,80	25,85	1083,45
	CNN-LSTM-LR (coif8)	1,45	0,99	0,090	2,20	-49,01	5,20	0,94	2,33	6,68	-7,86
	RFR (db11)	3,016	0,964	0,121	5,363	0,013	3,652	0,954	0,202	5,697	-0,776
	XGB (db11)	2,471	0,987	0,131	3,234	0,0023	3,663	0,947	0,203	6,100	-0,694
	GBR (db11)	1,697	0,994	0,098	2,196	4,55e-16	3,792	0,951	0,213	5,881	-0,535
	LR (db11)	0,513	0,999	0,044	0,736	-2,30e-12	0,765	0,999	0,175	0,966	-0,070
	SVR (db11)	1,039	0,995	0,058	1,957	0,095	1,252	0,995	0,108	1,917	0,0026
	CNN-LSTM-LR (db11)	0,26	0,99	0,024	0,38	12,70	0,40	0,99	0,083	0,51	0,19
	RFR (sym20)	3,24	0,962	0,128	5,46	-0,020	3,89	0,946	0,211	6,20	-0,837
	XGB (sym20)	1,80	0,993	0,106	2,32	0,001	4,01	0,934	0,212	6,83	-0,822
	GB (sym20)	1,69	0,994	0,102	2,15	0,0	4,20	0,924	0.218	7,32	-0,458
	LR (sym20)	0,22	0,999	0,017	0,30	0,0	0,29	0,999	0,032	0,38	-0,014
	SVR (sym20)	0,80	0,997	0,053	1,46	0,094	1,04	0,997	0,084	1,52	-0,058
	LSTNet (sym20)	17,90	-94,91	0,92	28,85	8410,00	14,94	-77,93	0,75	26,00	731,16
	CNN-LSTM-LR (sym20)	0,21	0,99	0,016	0,30	31,43	0,30	0,99	0,034	0,38	7,00

Tabela 1. Resultados obtidos com os modelos.

	TREINO					TESTE				
MODELO	MAE	R^2	MAPE	RMSE	BIAS	MAE	R^2	MAPE	RMSE	BIAS
LR (sym20)	3,15	0,97	0,18	4,87	-0,003	2,72	0,98	0,20	3,80	0,05
CNN-LSTM-LR (sym20)	3,15	0,97	0,18	4,87	-1,39	2,79	0,98	0,20	3,80	-9,64

Tabela 2. Comparação com a coluna original de casos totais.

Em comparação ao *baseline* [Shaikh et al. 2023], que usa algoritmos genéticos na fase de pré-processamento e um *ensemble* de modelos (ANN, SVR e CNN) para predição, os modelos da Tabela 2 obtiveram uma superioridade de 1,90 em RMSE. Além disso, o modelo LR (sym20) apresenta resultados melhores do que os de *Panja et al.* em termos de MAE e RMSE, com reduções de 2,94 e 3,89, respectivamente.

Uma limitação significativa deste estudo é a falta de generalização do modelo mais eficaz, uma vez que tanto o treinamento quanto os testes se limitaram a dados de San Juan. Ademais, o reduzido número de instâncias pode ter beneficiado modelos rasos em detrimento de modelos profundos. Adicionalmente, o grande número de características gerado no momento da estratégia de janelamento pode ter prejudicado o desempenho de modelos como o SVR e as árvores de regressão, devido à alta dimensionalidade, resultando na "maldição da dimensionalidade".

8. Conclusões

Neste estudo, investigamos métodos para predizer o número semanal de casos de dengue em San Juan, Porto Rico, usando três abordagens: modelagem estatística tradicional, aprendizado de máquina raso e aprendizado de máquina profundo. Realizamos análises univariadas e multivariadas para estes modelos. O Passeio Aleatório (RW) destacou-se entre os modelos estatísticos (SES, Holt, ES, ARIMA, RW), enquanto o ARIMA, apesar de convergir, teve predições imprecisas. Nos modelos rasos (RFR, SVR, XGB, GB, LR) e profundos (CNN-LSTM, CNN-LSTM-LR, LSTNet), optamos por estratégias univariadas e multivariadas, priorizando predições *one-step* após baixo desempenho em abordagens *multi-step* e *multi-output*. A análise da Tabela 1, mostra que o modelo de Regressão Linear (LR) com pré-processamento via *wavelet* sym20 foi o mais eficaz entre os rasos, e o CNN-LSTM-LR se destacou entre os profundos, utilizando *max pooling*. Ao comparar os resultados com dados não filtrados, observamos na Tabela 2 que o modelo LR (sym20) se destacou em termos de MAE, apresentando um BIAS de 0,05. Por outro lado, o CNN-LSTM-LR mostrou subnotificação de casos, com um BIAS de -9,64. A investigação indicou que o CNN-LSTM-LR, em grande parte, apenas replicava o desempenho de seu componente autorregressivo, alcançando resultados similares ao do modelo LR (sym20).

Para trabalhos futuros, propomos explorar pré-processamentos como PCA adaptado para séries temporais e avaliar arquiteturas de codificador-decodificador com atenção autorregressiva, como Transformers [Vaswani et al. 2023], visando melhorar a predição de dengue. Além disso, expandiremos a aplicação dos modelos em mais *datasets* para generalizar os resultados obtidos.

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES-PROEX) - Código de Financiamento 001. Este trabalho foi parcialmente financiado pela Fundação de Amparo à Pesquisa do Estado do Amazonas – FAPEAM – por meio do projeto POSGRAD 2024/2025. O trabalho também conta com o apoio do Centro de Inovação em Inteligência Artificial para a Saúde (CIIA-Saúde), Proc. 2020/09866-4 - FAPESP/MCTIC/CGI. Esta pesquisa, realizada no âmbito do Projeto Samsung-UFAM de Ensino e Pesquisa (SUPER), de acordo com o Artigo 39 do Decreto n°10.521/2020, foi financiada pela Samsung Eletrônica da Amazônia Ltda, nos termos da Lei Federal n°8.387/1991, através do convênio 001/2020 firmado com a UFAM e FAEPI, Brasil.

Referências

Benidis, K., Rangapuram, S. S., Flunkert, V., Wang, Y., Maddix, D., Turkmen, C., Gasthaus, J., Bohlke-Schneider, M., Salinas, D., Stella, L., Aubet, F.-X., Callot, L., & Januschowski, T. (2022). Deep learning for time series forecasting: Tutorial and literature survey. *ACM Computing Surveys*, 55(6):1–36.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Buczak, A. L., Baugher, B., Moniz, L. J., Bagley, T., Babin, S. M., & Guven, E. (2018). Ensemble method for dengue prediction. *PLOS ONE*, 13(1):e0189988.

Cabrera, M., Leake, J., Naranjo-Torres, J., Valero, N., Cabrera, J. C., & Rodríguez-Morales, A. J. (2022). Dengue prediction in latin america using machine learning and the one health perspective: A literature review. *Tropical Medicine and Infectious Disease*, 7(10).

Celentano, D. D., Sifakis, F., Go, V., & Davis, W. (2008). Changing sexual mores and disease transmission. In *The Social Ecology of Infectious Diseases*, pages 50–76. Elsevier.

Chen, T. & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16. ACM.

da Silva, L. J. & Angerami, R. N. (2008). *Viroses emergentes no Brasil*. Editora Fiocruz.

Derrick, T. & Thomas, J. (2004). *Time-Series Analysis: The Cross-Correlation Function*, pages 189–205. Human Kinetics Publishers, Champaign, Illinois. Posted with permission.

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., & Vapnik, V. (1996). Support vector regression machines. In *Proceedings of the 9th International Conference on Neural Information Processing Systems*, page 155–161, Cambridge, MA, USA. MIT Press.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5).

Fuller, W. A. (1976). *Introduction to statistical time series*. Probability & Mathematical Statistics S. John Wiley & Sons, Nashville, TN.

Guo, P., Liu, T., & Zhang, Q. e. a. (2017). Developing a dengue forecast model using machine learning: a case study in china. *PLoS Negl. Trop. Dis.*, 11(10).

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An introduction to statistical learning*. Springer International Publishing, Cham, Switzerland, 1 edition.

Kreyszig, E. (2010). *Advanced Engineering Mathematics 10E*. John Wiley & Sons, Chichester, England.

Lai, G., Chang, W., Yang, Y., & Liu, H. (2017). Modeling long- and short-term temporal patterns with deep neural networks. *CoRR*, abs/1703.07015.

Lee, G. R., Gommers, R., Waselewski, F., Wohlfahrt, K., & Leary, A. (2019). Pywavelets: A python package for wavelet analysis. *Journal of Open Source Software*, 4(36):1237.

Lopes, N., Nozawa, C., & Linhares, R. E. C. (2014). Características gerais e epidemiologia dos arbovírus emergentes no brasil. *Revista Pan-Amazônica de Saúde*, 5(3).

Morettin, P. A. & Toloi, C. M. (2018). *Análise de séries temporais*. Blucher.

Panja, M., Chakraborty, T., Nadim, S. S., Ghosh, I., Kumar, U., & Liu, N. (2023). An ensemble neural network approach to forecast dengue outbreak based on climatic condition. *Chaos, Solitons & Fractals*, 167:113124.

Ross, T. M. (2010). Dengue virus. *Clinics in Laboratory Medicine*, 30(1):149–160.

San Martin, J., Solorzano, J., & Guzman, M. e. a. (2010). The epidemiology of dengue in the americas over the last three decades: a worrisome reality. *Am. J. Trop. Med. Hyg.*, 82(1):128–135.

Santos, C. A. G., Guerra-Gomes, I. C., Gois, B. M., Peixoto, R. F., Keesen, T. S. L., & da Silva, R. M. (2019). Correlation of dengue incidence and rainfall occurrence using wavelet transform for joão pessoa city. *Science of The Total Environment*, 647:794–805.

Seabold, S. & Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.

Shaikh, M. S. G., SureshKumar, D. B., & Narang, D. (2023). Development of optimized ensemble classifier for dengue fever prediction and recommendation system. *Biomedical Signal Processing and Control*, 85:104809.

Strang, G. & Nguyen, T. (1996). *Wavelets and filter banks*. Wellesley-Cambridge Press, Wellesley, MA, 2 edition.

US National Oceanic and Atmospheric Administration (2017). Dengue forecasting project website. Acessado em 21 de fevereiro de 2024.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). Attention is all you need.

World Health Organization (2009). *Dengue: Guidelines for diagnosis, treatment, prevention and control*. World Health Organization, Genève, Switzerland.

World Health Organization (2023). Dengue - global situation. [Online; accessed 12-29-2023].