Revista Brasileira de Computação Aplicada, Novembro, 2022

DOI: 10.5335/rbca.v14i3.13483
Vol. 14, Nº 3, pp. 37–50
Homepage: seer.upf.br/index.php/rbca/index

ARTIGO ORIGINAL

Aplicação de séries temporais na estimação do número de casos de dengue em Cascavel-PR

Application of time series to estimate the number of dengue cases in Cascavel-PR

Angelo José Orssatto[1], Claudia Brandelero Rizzi[1], Rogério Luis Rizzi[1], André Luiz Brun[1]

[1]Universidade Estadual do Oeste do Paraná - UNIOESTE

angelojorssatto@gmail.com,claudia.rizzi@unioeste.br,rogerio.rizzi@unioeste.br,andre.brun@unioeste.br

Recebido: 30/04/2022. **Revisado:** 10/10/2022. **Aceito:** 30/10/2022.

Resumo

A dengue é uma das doenças mais importantes atualmente, ameaçando aproximadamente metade da população mundial, principalmente as regiões próximas aos trópicos, com climas tropicais e subtropicais, em áreas urbanas e suburbanas. O número de casos e mortes decorrentes da doença tem aumentado consideravelmente, implicando em maiores impactos sociais e econômicos. Uma boa estratégia para o combate ao vetor e ao tratamento apropriado aos contaminados é a preparação adequada dos órgãos responsáveis. Assim, estratégias que possam predizer a ocorrência de surtos da dengue desempenham papel primordial. Neste estudo fez-se a avaliação de diversas estratégias na estimação do número de casos positivos de dengue para o município de Cascavel, Paraná. Avaliou-se os métodos de Média Móvel, Suavização Exponencial, ARIMA, SVM, MLP, Florestas Aleatórias e Redes Neurais Recorrentes. As informações utilizadas no estudo foram o número de casos positivos no período de 2007 a 2014, o mês de ocorrência, a temperatura média, o índice pluviométrico e a umidade relativa do ar. O melhor desempenho foi alcançado pelo MLP com valor de 3,163 para a raiz do erro quadrático médio (RMSE). Em seguida, os melhores desempenhos foram obtidos pelos algoritmos de Floresta Aleatório e Média Móvel, com RMSE de 3,264 e 4,123, respectivamente.

Palavras-Chave: Aedes; Aprendizagem de Máquina; Modelos de Regressão; Surtos de Dengue.

Abstract

Dengue fever is one of the most important diseases today, threatening about half of the world's population, especially in tropical regions with tropical and subtropical climates, in urban and suburban areas. The number of disease cases and deaths has increased significantly, leading to greater social and economic impacts. A good strategy to control the vector and properly treat those infected is the adequate preparation of the relevant agencies. Therefore, strategies that can predict the occurrence of dengue outbreaks play a key role. In this study, several strategies were evaluated to estimate the number of positive dengue cases for the municipality of Cascavel, Paraná. The methods evaluated were moving average, exponential smoothing, ARIMA, SVM, MLP, random forests, and recurrent neural networks. The information used in the study was the number of positive cases from 2007 to 2014, mounth of occurance, average temperature, rainfall index, and relative humidity. The MLP achieved the best performance with a value of 3.163 for the root mean square error (RMSE). The Random Forest and moving average algorithms also performed the best, with an RMSE of 3.264 and 4.123, respectively.

Keywords: Aedes; Machine Learning; Regression Models; Dengue Outbreaks.

1 Introdução

A dengue é uma doença infecciosa causada por um vírus da família *Flaviridae* transmitida principalmente pela picada do mosquito *Aedes aegypti*. Os diferentes sorotipos (DENV-1 à DENV-4) podem causar enfermidades graves e mortais. Os principais sintomas são febre alta, dores de cabeça, dores no corpo e nas articulações, fraqueza, dor atrás dos olhos e prurido (WHO, 2022), e a assistência em saúde visa apenas atenuar tais sintomas. Visto que não existe uma vacina efetiva, a estratégia mais adequada para mitigar os casos de dengue é o combate ao vetor do vírus (Rizzi et al., 2017).

Países tropicais são os mais afetados por pela dengue devido às suas características climáticas, ambientais e socioeconômicas (Ribeiro et al., 2006) que favorecem à proliferação do *Aedes* bem como a transmissão do vírus entre a população. No Brasil, foram registradas 554 mortes pela dengue em 2020, com incidência (casos prováveis) de 690,2 e apresentando uma taxa de mortalidade de 0,26 por 100.000 habitantes (PAHO, 2020). Cabe destacar que há um gasto para o sistema público de saúde, proporcional à incidência da doença, variando de US$ 413 a US$ 966, para cada caso hospitalizado (Laserna et al., 2018).

Um alerta precoce de surtos de dengue pode aumentar a eficácia de campanhas de controle ao vetor e orientar ações preventivas (Phung et al., 2015). Assim, intervenções prévias podem retardar e mitigar a intensidade de epidemias e, consequentemente, amenizar o impacto sanitário e social, além de reduzir a mortalidade por meio da resposta prévia adequada do sistema de saúde pública (Gharbi et al., 2011).

Ao possibilitar a tomada de ações de forma preventiva de controle da doença, a previsão de incidência de dengue adquire uma função social indispensável. Várias abordagens podem ser empregadas para tal, utilizando séries temporais e dados climáticos, focados em uma determinada região geográfica, aplicados em estratégias de aprendizagem de máquina e modelos matemáticos/estatísticos, como os trabalhos de Pham et al. (2016), Braga et al. (2017), Mittelmann and Soares (2017b) e Baquero et al. (2018), nos quais foram utilizados o modelo ARIMA, Árvores de decisão, Redes Neurais Artificiais e Rede Neural Recorrente, respectivamente, para realizar as predições de incidência e casos de dengue.

Dada a importância do combate efetivo ao vetor da dengue, uma preparação adequada de políticas públicas para a diminuição do número de Aedes e também para o tratamento de pacientes infectados pela doença. Neste trabalho realizou-se a aplicação e avaliação de várias técnicas para estimar o número de casos de dengue no município de Cascavel-PR. No estudo foram adotadas três abordagens estatísticas (Média Móvel, Suavização Exponencial e ARIMA) e quatro baseadas em aprendizagem de máquina (MLP, SVM, RF e RNN).

2 Fundamentação

A estimação do comportamento de uma série temporal pode ser realizada de diferentes formas e a partir de diversos paradigmas. Neste trabalho visamos comparar o desempenho, em termos de exatidão da estimação dos casos de dengue, de modelos matemáticos e baseados em aprendizagem de máquina.

A **Média Móvel (MM)** fornece um método simples para suavizar o passado histórico dos dados (Makridakis et al., 1997). Essa métrica considera apenas a variável objetivo para realizar sua predição, ou seja, utiliza os valores descritos pela série temporal. Médias móveis normalmente são calculadas para identificar a direção da tendência de um evento.

O processo de cálculo consiste, basicamente, em estimar o valor atual de uma série temporal com base na média dos n valores anteriores presentes na série. A Eq. (1) detalha como o processo é realizado para o instante t.

$$y_t = \frac{1}{n}\sum_{i=1}^{n} y_{t-i} \tag{1}$$

em que y_{t-i} são os elementos da série temporal e n corresponde ao período de tempo anterior à estimação. O valor de n incluído em uma média móvel afeta a suavidade da estimativa resultante (Makridakis et al., 1997), de forma que, quanto maior o valor de n mais suave será a curva de resultados.

A **Suavização Exponencial (SE)** produz previsões nas quais as médias ponderadas de observações passadas recebem pesos que decaem exponencialmente à medida que as observações envelhecem. Em outras palavras, quanto mais recente a observação, maior o peso associado (Hyndman and Athanasopoulos, 2018).

O valor do elemento previsto y_{t+1} é dado pela Eq. (2). Nela, o elemento α, com $\alpha \in [0..1]$, corresponde ao parâmetro de suavização, que controla a taxa na qual os pesos diminuem relacionados as observações $y_1, \dots y_t$.

$$\begin{aligned} y_{t+1} = \alpha y_t + \alpha(1-\alpha)y_{t-1} + \alpha(1-\alpha)^2 y_{t-2} + \\ \cdots + \alpha(1-\alpha)^{t-1} y_2 + \alpha(1-\alpha)^t y_1 \end{aligned} \tag{2}$$

Quando um pequeno valor de α é adotado, a previsão mais recente desempenha um peso maior em comparação à adoção de um α maior. Caso o parâmetro apresentar um valor próximo de 1, a influência do processo de inicialização torna-se rapidamente menos significante com o passar do tempo. No entanto, se α for próximo de zero, o processo de inicialização pode desempenhar um papel significativo por um período longo de tempo (Makridakis et al., 1997).

O modelo **ARIMA** (*Autoregressive Integrated Moving Average*) provê uma abordagem para previsão de séries temporais que integra o modelo autorregressivo e a média móvel. Em um modelo de autorregressão, a variável de interesse é predita utilizando uma combinação linear de valores anteriores desta variável. Portanto, um modelo autorregressivo $AR()$ de ordem p, é determinado pela Eq. (3).

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t \tag{3}$$

Um modelo de média móvel que utiliza erros de predi-

ções passadas para obter valores atuais *MA*() de ordem q é descrito na Eq. (4) a seguir. Nela, ε_t corresponde ao erro de cada predição anterior.

$$y_t = c + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots + \theta_q\varepsilon_{t-q} \quad (4)$$

Combinando, então, o modelo autorregressivo (Eq. (3)) e o modelo de média móvel (Eq. (4)), obtém-se o modelo ARIMA completo descrito na Eq. (5).

$$y_t = c + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \theta_1\varepsilon_{t-1} + \cdots + \theta_q\varepsilon_{t-q} + \varepsilon_t \quad (5)$$

Além dos modelos apresentados, os métodos de aprendizado de máquina podem ser eficazes em problemas de previsão de séries temporais mais complexos com múltiplas variáveis de entrada, relacionamentos não lineares complexos e dados ausentes (Brownlee, 2018). Considerando este fato, neste trabalho foram elencados quatro modelos de aprendizagem de máquina, que são descritos nas seguintes seções.

O ***Multilayer Perceptron* (MLP)** é uma arquitetura de redes neurais do tipo Perceptron que possui pelo menos uma camada intermediária. O MLP foi proposto para superar as limitações do Perceptron simples, que é incapaz de resolver problemas que não são linearmente separáveis (Tan et al., 2009).

O treinamento da rede normalmente é realizado através do algoritmo de retropropagação (*backpropagation*). A ideia da estratégia é fazer a alimentação dos pesos de cada conexão no sentido para frente, desde a primeira camada. Assim que o erro da estimação for calculado, a atualização dos pesos é dada no sentido contrário, desde a camada final até a inicial. A intensidade com que essa atualização é feita é baseada na taxa de aprendizagem (Tan et al., 2009).

O MLP aprende uma função $f : R^m \rightarrow R^o$ por treinamento de conjunto de dados, onde m é o número de dimensões para entrada, em que m refere-se ao número de atributos do conjunto e o corresponde ao número de dimensões para saída (Scikit-learn developers, 2021c). A partir de um conjunto $X = x_1, x_2, ..., x_n$ e um alvo y, o modelo visa aprender um aproximador de função não linear para a tarefa de classificação (o corresponde ao número de classes possíveis) ou regressão. Neste caso o algoritmo retorna um valor contínuo e não apenas um entre vários possíveis como na classificação.

O MLP, quando aplicado a problemas de regressão, pode ser treinado usando retropropagação tendo a função identidade como função de ativação. A função de perda utilizada é o erro quadrado e a saída é um conjunto de valores contínuos (Scikit-learn developers, 2021c).

***Suport Vector Machines* (SVM)** é um método de aprendizado supervisionado, usado tanto para problemas de classificação quanto regressão, construindo um hiperplano ou um conjunto de hiperplanos em um espaço dimensional mais elevado (Scikit-learn developers, 2021e) em relação ao conjunto de dados original.

Segundo Tan et al. (2009), a ideia do método é, a partir de um conjunto de dados de treinamento, encontrar um hiperplano ótimo que separe as classes presentes de forma a maximizar a margem de separação entre elas.

Na regressão, o objetivo é encontrar uma função $f(x)$ que tenha no máximo uma divergência ε dos alvos reais y_i para todo o conjunto de treinamento. Em outras palavras, erros não são considerados desde que sejam menores que ε, mas não se aceita qualquer desvio maior do que este (Smola and Scholkopf, 2004).

Após a definição das funções, caso o problema não seja linearmente separável, realiza-se o mapeamento do *kernel* (núcleo) para transformação do conjunto no hiperplano superior. As funções mais comuns para o *kernel* do SVM são funções lineares, polinomiais, de bases radiais (rbf) ou sigmóides (Scikit-learn developers, 2021e).

Por fim, os valores são preditos pelo SVM com a regressão, encaixando-se à curva descrita na execução de sua primeira etapa.

Árvores de decisão são um método de aprendizado supervisionado não paramétrico, cujo objetivo é criar um modelo que preveja o valor de uma variável por meio de regras inferidas com base em recursos advindos dos próprios dados (Scikit-learn developers, 2021a). Na estrutura, cada nó interno corresponde a um teste de um atributo, cada ramo representa um valor que esse atributo pode apresentar e os nós folhas representam as classes do problema (Castro and Ferrari, 2016).

Considerando os valores de treinamento $[x_1, x_2, \ldots, x_l]$ e um valor objetivo y, uma árvore de decisão particiona recursivamente o espaço de recursos de forma que as amostras com os mesmos rótulos ou valores de destino semelhantes sejam agrupadas.

Pode-se destacar que as árvores de decisão apresentam um modelo simples de interpretar (dita caixa branca), que não requer muita preparação dos dados e é possível validar o modelo utilizando testes estatísticos. Entretanto, árvores podem ser instáveis pois uma pequena variação no conjunto de dados pode resultar em uma árvore completamente distinta. As árvores podem criar modelos muito específicos que não generalizam bem os dados (normalmente estruturas de grande profundidade), e este é o principal motivo para a utilização de um conjunto de árvores, ou seja, a floresta aleatória (Scikit-learn developers, 2021a).

Uma **Floresta Aleatória** (RF - *Random Forest*) é um meta estimador que ajusta várias árvores de decisão de classificação ou regressão em várias subamostras do conjunto de dados e utiliza um critério definido para melhorar a precisão preditiva e controlar o *overfitting* (Scikit-learn developers, 2021d).

Em florestas aleatórias, cada árvore no conjunto é construída a partir de uma amostra retirada do conjunto de treinamento. Além disso, ao dividir cada nó durante a construção de uma árvore, a melhor divisão é encontrada entre todos os recursos de entrada. O objetivo dessa fonte de aleatoriedade é diminuir a variância do estimador florestal (Scikit-learn developers, 2021b).

As **Redes Neurais Recorrentes** (RNN - *Recurrent Neural Networks*) são uma classe de redes neurais utilizadas para modelar dados em sequência, como séries temporais (Keras, 2022). Esse tipo de rede neural possui conexões entre suas passagens e através do tempo. Os neurônios são alimentados com informações da camada anterior à dele e informações de si próprios proveniente da passagem anterior. Com isso, a alimentação de uma RNN depende da

ordem de suas entradas (Van Veen and Leijnen, 219).

Um dos principais pontos negativos das RNN envolve a lentidão que os cálculos podem ser realizados, a não consideração de entradas futuras para tomada de decisões e o problema de gradiente de fuga, cujos gradientes usados para calcular a atualização dos pesos podem ficar muito próximo de zero, impedindo que a rede ajuste os pesos (Goodfellow et al., 2016). Para resolver este último problema, foram introduzidas as camadas *Long Short-Term Memory* (LSTM) à rede neural recorrente.

Uma estrutura LSTM é utilizada em arquiteturas de redes recorrentes em conjunto com um algoritmo de aprendizado baseado em gradiente apropriado. Este tipo de camada foi projetada para superar erros de fuga de gradiente, podendo aprender a preencher intervalos de tempo em até 1000 passos, mesmo no caso de sequências de entradas ruidosas e incompreensíveis, sem perda em curto espaço de tempo (Hochreiter and Schmidhuber, 1997).

Cada neurônio possui uma célula de memória e três conexões: entrada, saída e esquecimento. A função dessas conexões é resguardar a informação, interrompendo ou permitindo seu fluxo. A conexão de entrada determina quanto da informação da camada anterior é armazenada na célula. A ligação de saída determina quanto da próxima camada fica sabendo sobre o estado atual dessa célula. Já a conexão do esquecimento considera um possível esquecimento de uma informação que pode não ser mais útil ao longo do tempo (Van Veen and Leijnen, 219).

Combinando redes neurais recorrentes com células do tipo LSTM é possível criar modelos mais robustos e que permitem uma maior generalização dos dados, tornando a rede mais aplicável e mais eficiente (Brownlee, 2018).

2.1 Trabalhos Correlatos

Encontram-se, na literatura, diversas pesquisas cujos objetivos consistem em estimar o número de casos de dengue utilizando estratégias computacionais, com aprendizagem de máquina e inteligência artificial, estratégias baseadas na estatística, com regressão linear e modelos auto-regressivos. Dentre esses, pode-se destacar os trabalhos de Silva and Frances (2017) que adotaram redes neurais artificiais (MLP) aplicadas à estimação de epidemias na cidade de Manaus (zona metropolitana). Os trabalhos de Mittelmann and Soares (2017b,a) cujo foco deu-se nas cidades de Guarulhos-SP e Itajaí-SC, respectivamente. Em ambas as pesquisas foram empregadas redes NARX (*Nonlinear Auto-regressive with eXogenous inputs*) e MLPs.

Na pesquisa de Baquero et al. (2018) foram empregados dois tipos de redes neurais (MLP e LSTM), um modelo aditivo generalizado (GAM - *Generalized Additive Model*), um modelo ARIMA sazonal (SARIMA) e um modelo com análise bayesiana espaço-temporal. A estimação do número de casos da doença teve aplicação na cidade de São Paulo-SP.

Um estudo mais abrangente foi desenvolvido por Mussumeci and Codeço Coelho (2020), onde foram consideradas informações de 790 cidades brasileiras, distribuídas em cinco estados diferentes. Na pesquisa foram comparadas redes neurais do tipo LSTM frente a uma estratégia baseada em florestas aleatórias.

Destacam-se também estudos voltados a regiões estrangeiras, como a pesquisa de Guo et al. (2017) aplicado na província de Guangdong na China. Os autores utilizaram os números de casos ao longo de 4 anos em conjunto com informações climáticas. As conclusões indicaram que, dentre os modelos avaliados - *Gradient Boosted regression tree* (GBM), *Negative Binomial regression Model* (NBM), *Least Absolute Shrinkage and Selection Operator* (LASSO), SVR e GAM - o mais acurado foi o SVR.

No trabalho de Azhar et al. (2017) com foco em Denpasar, na Indonésia, aplicou-se um modelo de regressão linear visando avaliar várias combinações de fatores para estimar o número de casos de dengue na cidade. Além disso, podemos citar também os trabalhos de Phung et al. (2015) que avaliaram um modelo de regressão múltipla, o SARIMA e o *Poisson distributed lag model* (PDLM) aplicados em Can Tho no Vietnã e a pesquisa de Laureano-Rosario et al. (2018), realizada em Yucatán (México) e San Juan (Porto Rico) em que se adotou uma rede neural artificial para prever epidemias da doença.

Um ponto comum na maioria dos estudos realizados na área é a utilização de variáveis climáticas em conjunto com séries históricas do número de casos de dengue já registrados, para auxiliar na predição. Essas variáveis incluem, em grande parte, a precipitação, umidade relativa e temperatura. Trabalhos como os de Pham et al. (2015), Laureano-Rosario et al. (2018), Ahmad et al. (2018) e Zhu et al. (2016) entretanto, utilizam outros fatores ambientais e sociais em adição às variáveis climáticas e aos casos de dengue, como o índice de vegetação, temperatura da superfície terrestre e marítima, índice de poluição do ar, tamanho da população e densidade populacional.

Os estudos de Lee et al. (2015) e Carlos et al. (2017) realizaram predições utilizando *data mining* advindas de diversas bases de dados (*Big data*) e também de redes sociais, como o *Twitter*.

A utilização de variáveis climáticas no entanto, implica na restrição geográfica dos estudos, uma vez que em locais com características climáticas distintas pode-se não observar o mesmo comportamento da população do vetor. Os estudos levantados normalmente são realizados em uma cidade ou região metropolitana, de países como Brasil, China, Indonésia e Malásia. Tais países são denominados subdesenvolvidos economicamente e ambientalmente classificados como de clima tropical.

A comprovada influência de variáveis climáticas no impacto e abrangência do número de casos de dengue justificam sua adoção como fatores a serem considerados no presente estudo.

3 Metodologia

Nesta seção descreve-se como ocorre o fluxo de execução do projeto. Com uma abordagem metodológica de caráter mais experimental, o trabalho se divide em duas etapas. A primeira consiste na obtenção e tratamento dos dados que foram utilizados, enquanto a segunda abrange o treinamento dos modelos, sua avaliação e análise dos resultados.

Primeiramente, realizou-se a filtragem dos dados obtidos em relação a dengue e ao clima, incluindo neste processo a remoção de possíveis ruídos presentes em ambas as bases. Após a preparação dos dados, o conjunto foi divido

Figura 1: Visão geral da localização de Cascavel
Fonte:Wikipedia

em subconjuntos de treino e de teste, sendo o primeiro usado para a construção dos modelos ARIMA, Suavização Exponencial, Média Móvel, SVM, MLP, RF e RNN. Já o conjunto de teste, é utilizado para avaliar o desempenho dos modelos para assim concluir sobre a predição de casos de dengue a partir dos resultados apresentados. Maiores detalhes das etapas comentadas estão descritos nas subseções seguintes.

3.1 Caracterização da Área de Estudo

A cidade de Cascavel está localizada nas coordenadas 24°57'21"S / 53°27'19"W, no estado do Paraná, região sul do Brasil. Possui uma população estimada em 336.073 habitantes em uma área de 2.103,123 km² e densidade demográfica de 136,23 hab/km², segundo dados disponibilizados pelo IBGE (2021).

O território do município está situada no Terceiro Planalto do Paraná e no encontro de três unidades hidrográficas do Paraná, são elas a bacia do Rio Iguaçu, do Rio Piquiri e do Rio Paraná. Na Fig. 1 é apresentado um mapa do estado do Paraná com o município de Cascavel em destaque.

Segundo dados disponíveis no portal da Prefeitura Municipal de Cascavel (Prefeitura Municipal de Cascavel, 2021) o Instituto de Terras, Cartografia e Geociências (ITCG) identificou que a cidade se enquadra nas categorias Cfa (Clima Subtropical úmido), Cfb (Clima Oceânico Temperado) e Cfa/Cfb segundo a classificação de Köppen. Tais categorias indicam que o município não possui uma estação seca e que o verão pode ser quente ou fresco. A vegetação nativa que encobre a região de Cascavel é a Mata de Araucárias, com grande parte devastada devida à atividade intensa de agricultura e à expansão urbana (Prefeitura Municipal de Cascavel, 2021).

O índice de infestação de Dengue no município chegou a 5,3% em janeiro de 2022, segundo o Controle de Endemias da Secretaria de Saúde de Cascavel (G1, 2022). O valor indicado pela Organização Mundial de Saúde é de até 1%. Este índice provém da aplicação do Levantamento Rápido de Índices para Aedes aegypti (LIRAa), conforme detalhado em (Brasil, 2013). Além disso, a cidade apresentou 6.681 casos confirmados entre julho de 2019 e junho de 2020, com boletim epidemiológico que ultrapassava 10 mil notificações (Prefeitura Municipal de Cascavel, 2020), indicando que a presença do Aedes tem sido um problema recorrente no município.

Frequentemente, ações de combate são realizadas no município a fim de diminuir os níveis de infestação e prevenir novos casos. Geralmente, essas ações são centradas em bairros que apresentam maior índice de infestação (G1, 2021).

3.2 Obtenção dos dados

Os dados relativos aos casos de dengue foram levantados junto à Secretaria Municipal de Saúde de Cascavel - PR e ao Setor de Endemias, e são atendidos aos requisitos de ética quanto à manipulação de dados em arquivos, do parecer 261/2012-CEP, referente ao processo CAAE nº. 10726712.6.000.0107 vinculado ao projeto Sistema de Informações Geográficas Web (SIGAEDES). Além disso, não foram utilizadas informações específicas que possam identificar algum indivíduo conforme a Lei 13.709 de 14 de agosto de 2018. As informações dos pacientes consideradas nesta pesquisa envolvem apenas a data inicial dos primeiros sintomas da doença, referentes ao período de 2007 a 2014, sem distinção entre os tipos e sorotipos de dengue.

Além das informações acerca dos casos positivos da doença fez-se o levantamento dos dados climáticos, do mesmo período e localização. Tais informações envolvem as variáveis umidade relativa do ar (percentual), precipitação total (mm) e temperatura média (medida em graus Celcius). A escolha destas variáveis foi baseada nos estudos de Aburas et al. (2010), Guo et al. (2017) e Mittelmann and Soares (2017b). Estas informações foram obtidas junto ao SIMEPAR (Sistema de Tecnologia e Monitoramento Ambiental do Paraná).

3.3 Remoção de ruídos

Os ruídos encontrados na base adquirida foram descartados durante a consolidação dos dados. Tais elementos caracterizam-se por datas não contidas no período estudado (janeiro de 2007 a dezembro de 2014), e por classificações finais de dengue que não estão de acordo com a especificação do dicionário de dados (resultados inconclusivos ou negativos).

Além da remoção de amostras datadas fora do período de interesse, fez-se o descarte dos registros de casos suspeitos de dengue não confirmados. A base obtida junto à Secretaria de Saúde Municipal de Cascavel envolvia informações de todos os casos testados no município, uma vez que o protocolo de teste tem notificação compulsória. Visto que nosso interesse recai apenas sobre os casos positivos, todos os registros em que o resultado do exame laboratorial foi negativo foram removidos.

3.4 Tratamento dos dados

O conjunto dos dados climáticos possui uma amostragem diária. Contudo, para os experimentos realizados foi necessário agrupá-los em granularidades semanal e mensal.

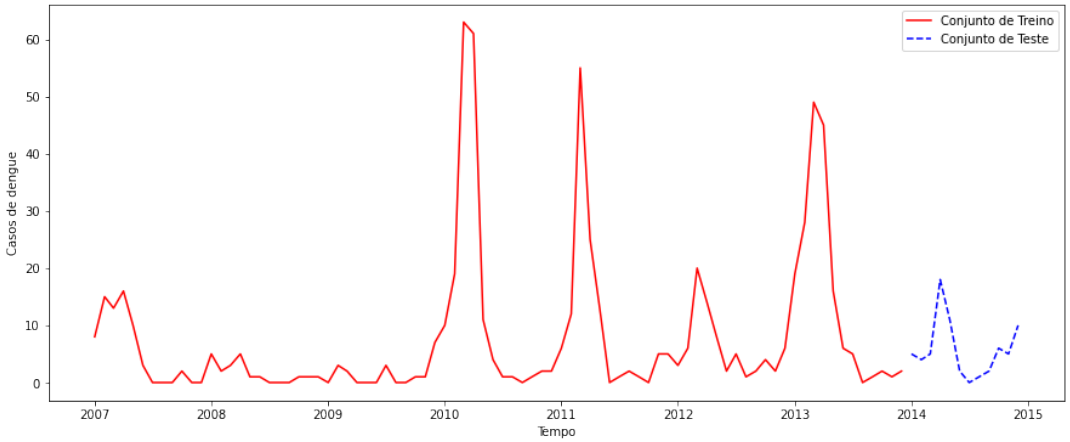

Figura 2: Curva ilustrando os casos positivos em sua distribuição temporal entre os anos de 2007 e 2014

Para as variáveis referentes à temperatura média e umidade relativa foi realizada a média aritmética simples dos dados. Em contrapartida, para a transformação da variável referente à precipitação foi realizado o somatório dos valores para obter a granularidade desejada.

Os dados históricos da dengue foram tratados em dois passos. Primeiramente, agrupou-se o número de casos para cada dia. Caso não houvesse casos registrados em uma determinada data, foi atribuído o valor zero. Após este processo, foi realizada a mesma estratégia aplicada à variável climática precipitação em que se fez a contagem de todos os casos ocorridos em períodos semanais e mensais. Além disso, considerou-se como um atributo o mês da ocorrência dos casos em análise. Esse dado foi utilizado nos modelos de aprendizagem de máquina com o objetivo de manter informações da dependência temporal.

3.5 Divisão dos dados

Após a consolidação dos dados, foi realizada a divisão da base em dois grupos: conjunto de treino e conjunto de teste. O primeiro tem como objetivo servir de base para o aprendizado e calibração dos modelos escolhidos para estimar o número de casos de dengue. Já o conjunto de teste é empregado como critério para a avaliação e análise dos modelos calibrados. Estes conjuntos correspondem à aproximadamente 85% e 15%, respectivamente, da base de dados consolidada. A divisão realizada implica no período de janeiro de 2007 a dezembro de 2013 na composição do conjunto de treino. Já o conjunto de teste contém todo ano de 2014, assim como ilustrado na Fig. 2, representados pelas linhas contínua na cor vermelha e tracejada na azul, respectivamente. A informação representada corresponde ao número de casos positivos observados ao longo dos oito anos.

A randomização das amostras não foi realizada para manter a série histórica e a sincronia no qual os dados são correlatos uma vez que consideramos haver dependência temporal entre as amostras.

3.6 Treinamento e avaliação dos modelos

Os modelos matemáticos e estatísticos implementados foram a Média Móvel, a Suavização Exponencial e o ARIMA. Os modelos de aprendizagem de máquina foram o *Support Vector Machine* (SVM), *Multilayer Perceptron* (MLP), *Random Forest* (RF) e *Recurrent Neural Network* (RNN). As abordagens citadas foram calibradas de acordo com seus respectivos parâmetros, utilizando *grid search* orientado pela raiz do erro médio quadrático. Por fim, as técnicas foram avaliadas através da métrica do Erro Médio Absoluto (MAE), Erro Médio Quadrático (MSE) e Raíz do Erro Médio Quadrático (RMSE).

3.7 Ambiente de execução dos procedimentos e tecnologias utilizadas

Os procedimentos para realização deste trabalho foram implementados utilizando Python como linguagem de programação e executados em ambiente Google Colaboratory[1] (Colab). O Google Colab é um serviço gratuito de nuvem, desenvolvido para incentivar a pesquisa de Aprendizado de Máquina e Inteligência Artificial. Ao executar experimentos neste ambiente, os códigos ficam livres de contextos dependentes da máquina local e de sistemas operacionais.

Para o processamento dos dados, visualização dos dados e a avaliação das predições, foram utilizadas as bibliotecas Matplotlib[2], Numpy[3], Pandas[4] e Seaborn[5]. Já para o treinamento dos modelos, foram utilizadas as bibliotecas Keras[6], PmdARIMA[7], Sklearn[8] e Statsmodel[9].

[1]Disponível em: https://colab.research.google.com/?utm_source=scs-index
[2]Disponível em: https://matplotlib.org/stable/
[3]Disponível em: https://numpy.org/doc/stable/
[4]Disponível em: https://pandas.pydata.org/docs/
[5]Disponível em: https://seaborn.pydata.org/api.html
[6]Disponível em: https://keras.io/api/
[7]Disponível em: https://pypi.org/project/pmdarima/
[8]Disponível em: https://scikit-learn.org/0.21/documentation.html
[9]Disponível em: https://www.statsmodels.org/stable/user-guide.h

3.8 Parâmetros dos modelos preditivos

Para o modelo da MM, o único parâmetro a ser testado com diferentes valores é o número de amostras considerado para o cálculo da média. Este valor foi variado no conjunto $\{3, 4, 5, 6, 7, 8, 9, 10\}$.

A implementação do modelo de suavização exponencial simples, fornecida pela biblioteca Statsmodel, não possui nenhum parâmetro a ser variado e testado.

O modelo ARIMA conta com três principais parâmetros a serem testados, os quais p, d e q, que correspondem à ordem do modelo auto-regressivo, ao grau de diferenciação e à ordem da média móvel, respectivamente. Os valores testados estão representados na Tabela 1.

Tabela 1: Parâmetros do modelo ARIMA avaliados no *grid search*

ARIMA	
Parâmetro	**Valores testados**
p	$\{0, 1, 2, 3, 4, 5\}$
d	$\{0, 1, 2, 3, 4, 5\}$
q	$\{3, 4, 5, 6, 7, 8, 9, 10\}$

Ao utilizar a estrutura do SVM, implementada pela biblioteca Scikit-learn, pode-se fazer uso de vários parâmetros a fim de criar modelos bastante diversos. Neste estudo, foram variados os parâmetros correspondentes ao tipo de *kernel*, núcleo para a elevação do espaço dimensional, e ao parâmetro regularizador C (custo dos erros). Os valores utilizados na busca pelo melhor modelo do SVM estão representados na Tabela 2.

Tabela 2: Parâmetros do modelo SVM avaliados no *grid search*

SVM	
Parâmetro	**Valores testados**
C	$\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$
kernel	$\{linear, rbf, poly\}$

Como no SVM, existem vários parâmetros que permitem a criação de modelos de MLPs bastante diversos. Os parâmetros variados foram os correspondentes ao número de neurônios nas camadas escondidas (*hidden_layer_sizes*), à taxa de aprendizagem (*learning_rate*) e ao número máximo de iterações (*max_iter*) para convergência. O conjunto que contém os valores testados em cada parâmetro estão listados na Tabela 3.

Tabela 3: Parâmetros do modelo MLP avaliados no *grid search*

MLP	
Parâmetro	**Valores testados**
hidden_layer_sizes	$\{100, 200, 300, 400, 500\}$
learning_rate	$\{0.001, 0.002, 0.003, \ldots, 0.009, 0.01\}$
max_iter	$\{400, 600, 800, 1000\}$

tml

Ainda utilizando a biblioteca Sckit-learn, a implementação de florestas aleatórias conta com diversos parâmetros para criar seus modelos. Os parâmetros testados correspondem ao critério de escolha entre as árvores (*criterion*), o número de árvores na floresta (*n_estimators*) e ao número mínimo de amostras nas folhas (*min_samples_leaf*). Os valores variados para cada parâmetro do modelo são apresentados na Tabela 4.

Tabela 4: Parâmetros do modelo RF avaliados no *grid search*

RF	
Parâmetro	**Valores testados**
criterion	$\{squared\_error, absolute\_error\}$
n_estimators	$\{10, 20, 30, \ldots, 90, 100, 200, 300, \ldots, 500\}$
min_samples_leaf	$\{2, 3, 4\}$

Para o modelo RNR, não foram variados possíveis parâmetros fornecidos na implementação da biblioteca Keras. Contudo, foi utilizado um modelo fixo contendo quatro camadas LSTM sendo treinado por 100 épocas.

4 Protocolo experimental

Os dados processados, utilizados e analisados foram referentes ao período de 2007 a 2014. Essas informações incluem o número de casos de dengue e as variáveis climáticas de umidade relativa do ar, temperatura média e precipitação.

4.1 Distribuição do número de casos de dengue ao longo do período estudado

A exploração dos dados disponíveis sobre o fenômeno em questão possibilita a tomada de escolhas melhores. Dessa forma, foi realizada uma análise exploratória das informações levantadas visando conhecer o contexto da aplicação da pesquisa de forma mais aprofundada.

A Fig. 3 exibe a quantidade de casos em cada ano do período estudado. Nota-se que nos anos de 2010, 2011 e 2013 houve mais de 100 casos de dengue registrados no município, mais especificamente 175, 125 e 174, respectivamente. Um fato interessante foi o salto no número de registros do ano de 2009, quando foram observados apenas 17 ocorrências frente às 175 identificadas em 2010.

Ao analisarmos o número de casos de dengue distribuídos ao longo dos meses em uma representação circular, podemos perceber uma intensidade maior em alguns meses. Esta informação é representada na Fig. 4. Na distribuição circular dos casos, cada mês corresponde a uma faixa de 30^0. Assim, podemos pensar que cada mês do ano corresponde a uma direção específica. Como pode ser observado, o número de casos positivos ocorridos nos primeiros meses do ano (de janeiro a maio) correspondeu aos maiores valores mensais observados. Em janeiro foram somados 56 casos positivos, em fevereiro observou-se 89 ocorrências. Para os meses de março, abril e maio o número de casos confirmados foi de 210, 184 e 70, respectivamente. Por outro lado, no período entre junho e novembro, a frequência

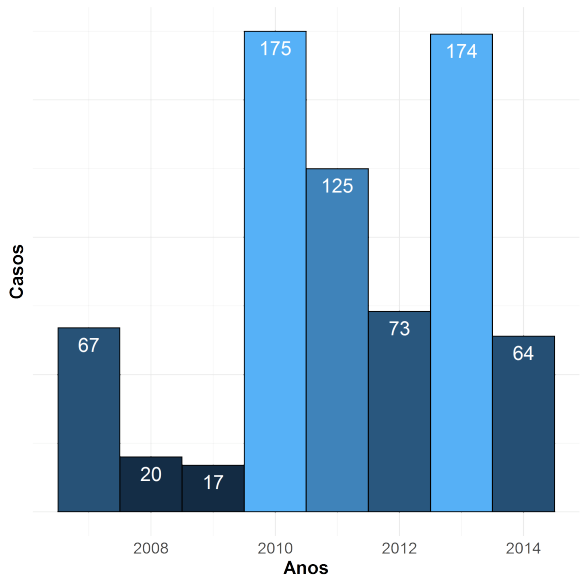

Figura 3: Número de casos positivos ao longo dos anos estudados

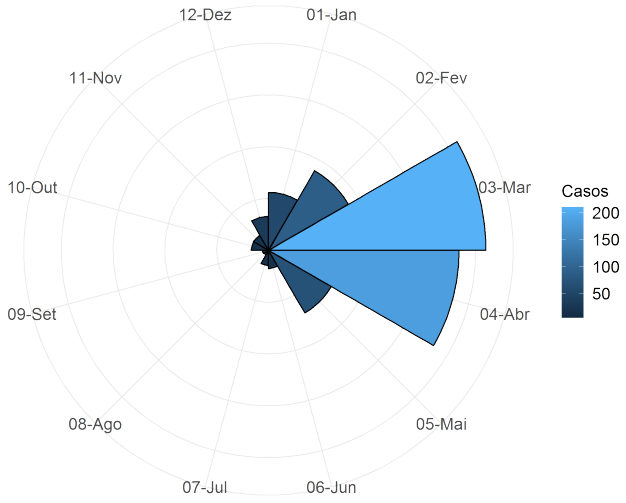

Figura 4: Distribuição dos casos ao longo dos doze meses para todos os anos estudados

de ocorrências da dengue foi bem menor, chegando aos menores valores nos meses de agosto (5) e setembro (6).

Visando avaliar se há um comportamento homogêneo na distribuição dos casos da doença ao longo do ano foi realizado o teste de periodicidade de Rayleigh (Brazier, 1994). Uma probabilidade menor que o nível de significância escolhido indica que os dados não têm uma distribuição uniforme (rejeita-se a hipótese nula que indica homogeneidade na distribuição dos dados). Neste caso, haverá

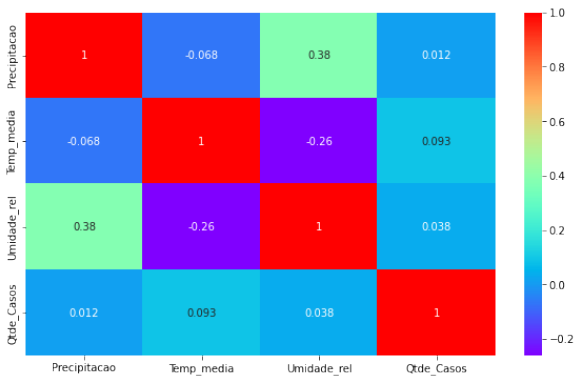

Figura 5: Matriz de correlação com granularidade diária (sem deslocamento)

evidências de que a distribuição é mais concentrada em determinadas direções.

O valor obtido pelo teste, $P - value = 0$, indica que a distribuição dos casos de dengue não segue um comportamento regular, considerando uma significância de 5%. Tal fato implica na existência de certa sazonalidade nos casos de dengue.

Visando avaliar se há uma concentração maior em alguma época do ano, calculamos o vetor circular médio P, em que $P \in [0, 1]$, cuja intensidade é usada para determinar o quão forte é a média da distribuição circular. Um valor igual a 1 indica que todas as amostras seguem a mesma direção indicando que todas os casos teriam ocorrido no mesmo mês, enquanto um valor próximo de zero indica que não há uma direção de maior dominância.

Ao analisarmos a intensidade do vetor médio obtido, $P = 0,6525$, percebe-se que há uma certa tendência em relação à direção leste que é onde estão representados os meses de março e abril, cujas quantidades de casos positivos foram as maiores do longo do ano.

A partir da análise conjunta do teste de Rayleigh e do valor de P pode-se afirmar se há sazonalidade na ocorrência dos casos da doença. Tal fato pode ser constatado na Fig. 4. Na ilustração nota-se um crescimento iniciado em janeiro, com pico de máximo nos meses de março e abril e, em seguida, uma diminuição no número de casos positivos iniciando já no mês de maio.

4.2 Estudo de correlação entre as variáveis climáticas e casos de dengue

Buscou-se analisar a relação entre as variáveis climáticas (umidade, precipitação e temperatura média) perante a quantidade de casos positivos ao longo de todo o período, com amostragem diária primeiramente. Para tal, utilizou-se o coeficiente de correlação de Pearson, que está representado na forma de matriz de correlação na Fig. 5. Entretanto, as correlações entre as variáveis climáticas e a quantidade de casos (última linha e última coluna) não mostrou uma intensidade significativa, a ponto de não superar um escalar de 0,1.

Levando em consideração o tempo de vida do mosquito,

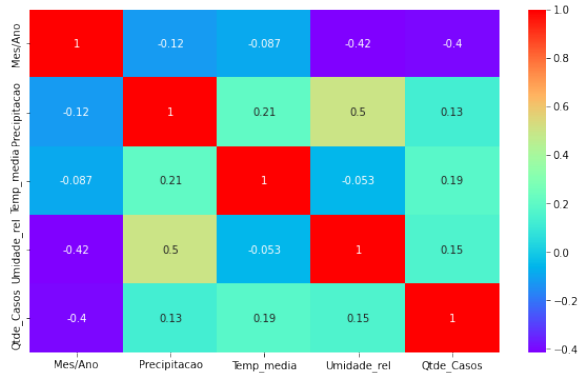

Figura 6: Correlação das variáveis com agrupamento mensal

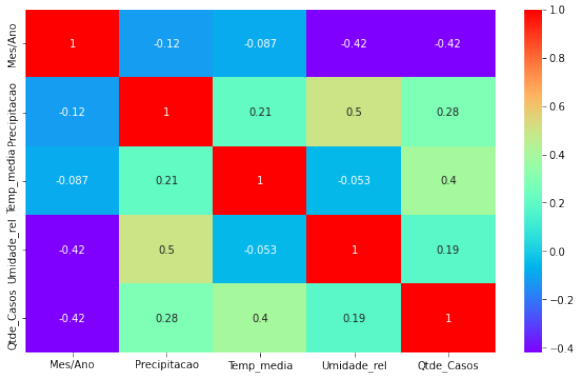

Figura 7: Correlação das variáveis com agrupamento mensal e deslocamento de um mês

que gira em torno de 20 a 30 dias para tornar-se adulto, infectar-se com o vírus e assim ser capaz de transmitir a doença, decidiu-se por deslocar as variáveis climáticas em retrocesso, isso é, acoplar a quantidade de casos de dengue com os dados climáticos referentes a 20 à 30 dias atrás. O objetivo foi o de avaliar se os fatores climáticos influenciam no ciclo de vida do mosquito. Acredita-se que com a incidência de chuvas e temperaturas altas, há uma maior eclosão dos ovos e dispara-se o início do ciclo (Nan et al., 2018). Dessa forma, optou-se por analisar essa relação considerando todo o possível ciclo de vida do Aedes, haja visto que não há uma concordância quanto à extensão do prazo entre a eclosão do ovo e o momento em que o mosquito se torna apto ao contágio.

Ao analisarmos os valores dos coeficientes apresentados, percebeu-se uma melhora significativa na correlação entre as variáveis estudadas. Os valores observados, no entanto, mostraram-se bastante reduzidos. Uma forma de buscar uma relação mais representativa foi agrupar os dados climáticos e dos casos positivos em amostragens mensais. Os níveis da correlação resultante dessa análise são exibidos na Fig. 6. Percebe-se que os valores realmente mostraram um aumento na correlação entre as variáveis de estudo.

Na Fig. 6 entretanto, não consideramos o deslocamento temporal de trinta dias necessários para que o vetor esteja apto à transmissão. Então, realizamos o deslocamento das variáveis climáticas retrocedendo o período de um mês, de forma que, se em um determinado período houve altas temperaturas e bastante chuva, a tendência é que seu impacto seja percebido a partir do mês seguinte, quando os mosquitos se tornam capazes da transmissão da dengue. Os valores obtidos por esta correlação são apresentados na Fig. 7.

Em comparação às correlações anteriores, os valores agrupados mensalmente com deslocamento de um mês frente às variáveis climáticas, mostrou um coeficiente de correlação ainda mais significativo, alcançando valores de até 0,4. A intensidade da correlação ainda é considerada fraca, mas mostra um certo grau de influência. Isso indica que os fatores climáticos ocorridos em um mês terão, efetivamente, impacto no mês seguinte, visto a duração de todo o ciclo do vetor.

Os experimentos apresentados anteriormente (nas Fig. 6 e Fig. 7) foram fundamentais pois, sabendo a forma mais eficiente de trabalhar os dados, é possível construir um modelo de predição mais eficaz e com resultados mais precisos. Desse modo, definiu-se que a granularidade que os dados seriam trabalhados seria a mensal.

4.3 Configuração de cada modelo e resultados obtidos

Com a granularidade mensal definida e com o aspecto do deslocamento de um mês para as variáveis climáticas em relação ao número de casos, realizou-se a modelagem das sete técnicas propostas para a predição de casos de dengue.

Após a execução do *grid search* para definição dos parâmetros para cada modelo de regressão obteve-se os valores apresentados na Tabela 5. Tais valores são aqueles que levaram o modelo a obter o menor erro quadrático médio.

Tabela 5: Parâmetros selecionados pelo *Grid search*

Método	Parâmetro	Melhor valor
Média Móvel	*n*	3
ARIMA	*p*	0
	d	1
	q	3
SVM	*kernel*	*rbf*
	C	0.1
MLP	*hidden_layer_size*	300
	learning_rate	0.002
	max_iter	1000
RF	*criterion*	*squared_error*
	n_estimators	10
	min_samples_leaf	2

5 Resultados e Discussão

Em posse dos modelos devidamente calibrados, foi realizada a predição do número de casos de dengue no ano de 2014 (que corresponde ao nosso conjunto de teste) e a posterior comparação com os dados reais. Para avaliar os resultados obtidos foram empregadas as métricas MAE,

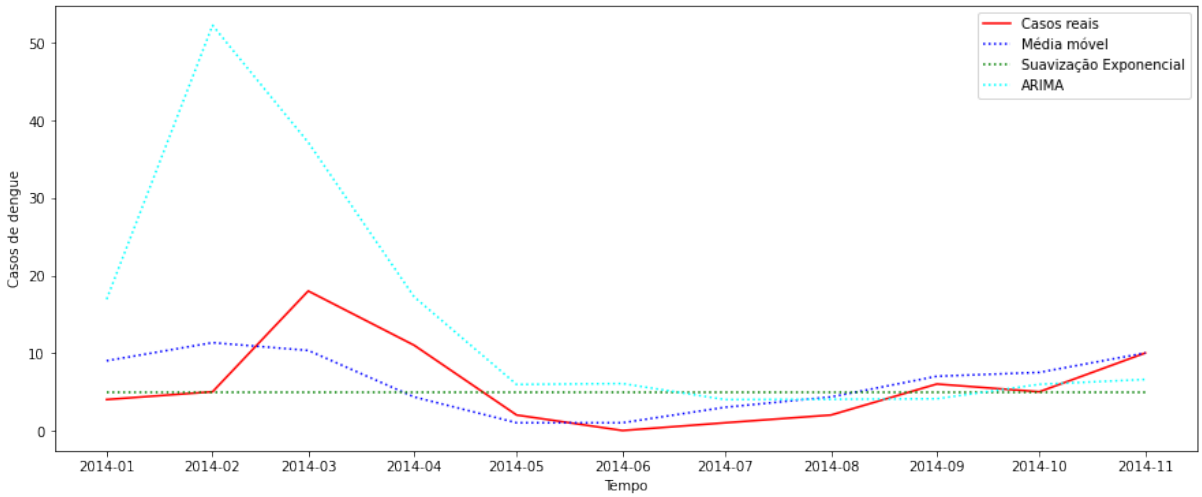

Figura 8: Comportamento dos métodos estatísticos sobre o conjunto de teste em comparação aos dados reais

MSE e RMSE.

Ao testar as predições realizadas pelos modelos estatísticos (MM, SE e ARIMA), a média móvel obteve maior performance dentre as três técnicas, apresentando um erro médio absoluto de 3,227, erro médio quadrático de 17,002 e a raiz do erro médio quadrático de 4,123. A Tabela 6 apresenta os resultados obtidos desses modelos estatísticos.

Tabela 6: Resultados obtidos pelos modelos estatísticos

	MM	SE	ARIMA
MAE	3,277	3,728	9,718
MSE	17,002	26,455	262,639
RMSE	4,123	5,143	16,206

O desempenho desses métodos é apresentado na Fig. 8. A curva na cor vermelha corresponde aos dados reais observados durante o ano. Já a curva na cor azul escuro identifica o número de casos de dengue estimado pelo modelo MM para cada mês do período. As curvas representadas nas cores azul claro e verde correspondem ao desempenho dos métodos ARIMA e Suavização Exponencial, respectivamente.

Ao analisarmos o gráfico presente na Fig. 8 nota-se que a média móvel manteve sua curva de predição próxima à curva real dos dados, acompanhando suas ascensões e declínios. A suavização exponencial considerou que todos os valores preditos seriam iguais, criando assim uma reta de tendência constante dos casos de dengue. O modelo ARIMA por sua vez, previu uma alta no número de casos maior do que realmente aconteceu e só posteriormente apresentou uma aproximação à curva real. Isso contribuiu para que suas métricas de erro fossem bastante elevadas em comparação às demais técnicas.

Além das três estratégias discutidas avaliou-se o comportamento de quatro abordagens usando técnicas de aprendizagem de máquina. O *Multilayer Perceptron* apresentou as melhores métricas dentre os quatro modelos testados. Seu MAE foi de 2,014, apresentou um MSE de 10,007 e RMSE de 3,163. A Tabela 7 apresenta os resultados obtidos. Pode-se perceber que o método de Florestas Aleatórias (RF), apesar de não ter tido os melhores resultados, apresentou desempenho bastante similar ao MLP.

Tabela 7: Resultados alcançados pelos modelos de aprendizagem de máquina

	SVM	MLP	RF	RNN
MAE	2,860	2,014	2,159	6,013
MSE	21,768	10,007	10,652	59,717
RMSE	4,666	3,163	3,264	7,728

O desempenho dos métodos SVM, MLP, RF e RNN são apresentados na Fig. 9. A curva na cor vermelha corresponde ao número real de casos ocorridos em cada mês. Já as curvas rosa, azul escuro, verde e azul claro, correspondem aos valores estimados pelos modelos SVR, MLP, Floresta Aleatória e RNN, respectivamente.

Comparando as curvas de predição construídas para cada modelo nota-se que todos acompanharam a curva do número real de casos com exceção do RNN que previu um pico de casos de dengue quando, na realidade, ocorreu um declínio no número de casos, no mês de junho. Tal fato levou ao aumento suas métricas de erros.

Analisando-se o comportamento dos modelos percebe-se que pequenos detalhes fizeram a diferença em seus desempenhos. O SVM não previu o pico ocorrido em março e predisse um aumento no número de casos em junho. O restante do gráfico, porém, está bem próximo à curva real. A floresta aleatória previu o aumento no número de casos em março, porém não um aumento tão significativo quanto o real observado. Já para os meses de junho e outubro o modelo apresentou uma superestimação dos casos. Podemos dizer que o modelo MLP "errou menos"em suas predições, fazendo com que sua curva ficasse próxima à curva real em grande parte do tempo, especialmente após o mês de maio.

Analisando-se o comportamento de todas as técnicas aplicadas, os modelos de aprendizagem de máquina apresentaram melhores performances em comparação aos mo-

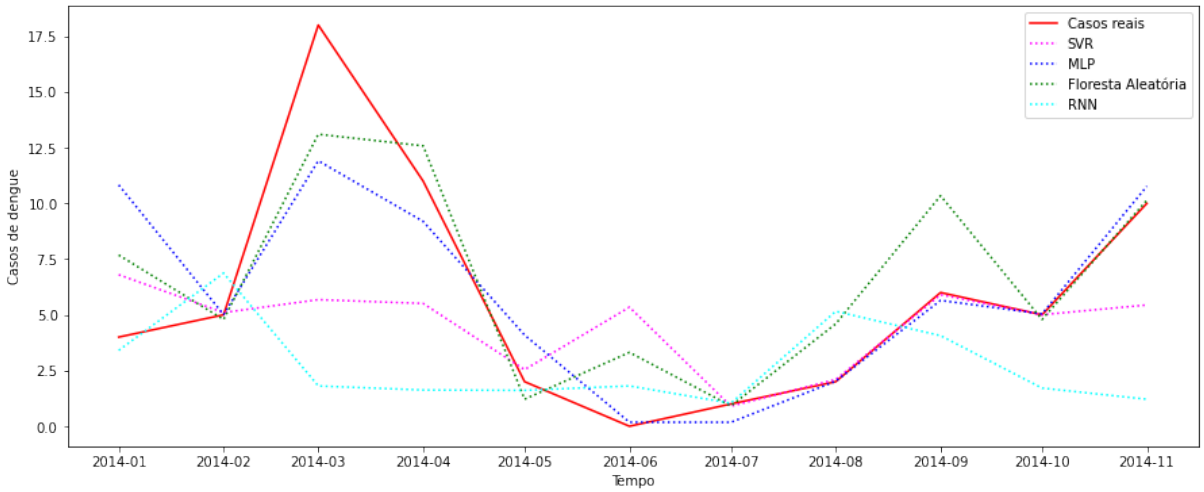

Figura 9: Comportamento dos métodos de aprendizagem de máquina comparados ao valor real

delos estatísticos. Isso se dá ao fato de serem modelos que podem ser escalados para mais de uma variável de influência, não somente aos dados referentes à própria série. Entretanto, modelos robustos como o ARIMA e o RNN, diferentemente do esperado, apresentaram as piores métricas durante a execução dos experimentos.

Em relação aos modelos de aprendizagem de máquina, é possível notar um bom comportamento das redes neurais, como observado neste trabalho e nos trabalhos de Aburas et al. (2010), Mittelmann and Soares (2017b), Mittelmann and Soares (2017a) e Silva and Frances (2017). Isso se deve, sobretudo, às semelhanças nas escolhas das variáveis climáticas estudadas e na divisão entre conjuntos de treino e teste dos modelos AM.

Analisando-se os trabalhos na literatura, no entanto, percebe-se uma grande variedade de técnicas empregadas na tarefa de estimar o número de casos de dengue. Fato que evidencia a o conjunto de possibilidades de extensão desta pesquisa.

Ademais, em várias das pesquisas levantadas foram empregadas informações além de dados climáticos e do número de casos confirmados. Os autores utilizaram informações sociais como densidade demográfica, tamanho da população, temperatura superficial, índice de poluição do ar. Tais informações, segundo os autores, podem influenciar no desenvolvimento e comportamento do Aedes. Assim, pode ser proveitoso agregar novas informações sobre o município de Cascavel-PR de forma a refinar o processo de estimação do número de casos de dengue.

5.1 Aplicações da metodologia desenvolvida

Uma possível aplicação do presente trabalho é aquele em que o Setor de Controle de Endemias do Município de Cascavel o utilizaria em conjunto com o Levantamento Rápido de Índices para Aedes aegypti (LIRAa). O LIRAa é uma metodologia do Ministério da Saúde que busca identificar o percentual de infestação pelo mosquito Aedes em determinadas regiões, já que para efeitos de ações de controle vetorial um município é subdividido em regiões, denominadas de Localidades (Brasil, 2013). Para tal controle, um percentual na Localidade menor de 1% é considerado satisfatório, mas se entre 1% e 3,9% já se tem uma situação de alerta, e acima de 4% tem-se uma situação com risco de surto de dengue.

Nesse sentido, uma proposta é utilizar inicialmente um LIRAa para selecionar as Localidades do município com maiores índices de infestação, sendo elas designadas no contexto por Regiões de Atenção (RA). Essas RAs podem ser clusterizadas randomicamente em dois grupos, o RA1 que é o agrupamento de RAs em que são aplicadas as políticas de controle vetorial preconizadas pelo Ministério da Saúde e pelas experiências operacionais dos gestores e agentes de saúde do Município, e o RA2, em que tais ações seriam antecipadamente aplicadas, em decorrência de os métodos desenvolvidos neste trabalho preverem que em determinadas Localidades podem ocorrer casos de dengue além daqueles considerados satisfatórios pela Saúde Pública Municipal.

Isso implica que os métodos apresentados neste trabalho devem considerar os casos de dengue observados, as variáveis meteorológicas e os índices de infestação prévios por agrupamentos de Localidades, para então serem empregados como uma ferramenta à geração de tais cenários. Como cada LIRAa tem duração de 3 meses, um estudo estatístico e comparativo entre os resultados obtidos, com e sem a aplicação prévia das políticas de controle vetorial em pelo menos 4 LIRAa, poderá servir de elemento importante para aprimorar e validar a metodologia desenvolvida.

Tais estratégias poderiam otimizar a destinação de recursos e a efetividade no combate ao vetor da dengue, pois ao identificar os padrões indicadores de possibilidade de surtos elas poderiam ser aplicadas com antecedência o suficiente para minorar os custos operacionais e de insumos usados no combate ao vetor, além de majorar o controle vetorial.

6 Conclusões e Trabalhos Futuros

Neste trabalho foi realizada a aplicação de várias técnicas para estimar o número de casos de dengue no município de Cascavel-PR. Três abordagens estatísticas (Média Móvel, Suavização Exponencial e ARIMA) e quatro estratégias baseadas em aprendizagem de máquina (MLP, SVM, RF e RNN) foram utilizadas.

Os dados empregados no estudo estão dispostos ao longo do período de 2007 a 2014 e correspondem ao número de casos positivos da doença divididos em granularidade mensal. Também foram utilizadas informações climáticas de Temperatura média (ºC), Umidade Relativa do Ar (%) e Índice Pluviométrico (mm) do município referentes ao mesmo período.

Para treinamento dos modelos de regressão foram usadas as informações referentes aos anos de 2007 até 2013. Já a avaliação da qualidade dos métodos foi realizada sobre o ano de 2014. Como critério de análise foram empregados as medidas MAE, MSE e RMSE.

Os resultados apontaram a Média Móvel como melhor método dentre os estatísticos. Já o perceptron de múltiplas camadas apresentou o melhor desempenho entre as estratégias de aprendizagem de máquina. O MLP obteve o melhor desempenho entre todas as abordagens avaliadas.

Entretanto, os desempenhos apresentados pelos métodos ARIMA e RNN ficaram aquém do esperado pois, dadas suas características robustas, esperava-se que obtivessem estimações mais precisas.

Os desempenhos alcançados se mostraram promissores como estratégias para estimação dos casos de dengue no município, podendo servir como um bom critério para definição de políticas de combate prévio ao vetor e também de preparação para o melhor atendimento dos contaminados.

Pretende-se, em um próximo momento, consolidar um conjunto de dados mais abrangente, cobrindo o período de 2007 até 2020. Acredita-se que com um conjunto maior de dados o processo de treinamento e avaliação fiquem mais robustos. Além disso, será testada a granularidade semanal dos dados, fazendo com que o número de amostras presentes no conjunto seja aumentada consideravelmente.

Os valores obtidos pela correlação linear entre as variáveis climáticas e o número de casos positivos mostraram-se fracas. Pretende-se analisar a correlação multivariada dos fatores de forma conjunta para tentar identificar relações mais relevantes.

Agradecimentos

Ao MEC-SESU, pelo financiamento parcial desta pesquisa a partir do Programa de Educação Tutorial (PET).

Referências

Aburas, H. M., Cetiner, B. G. and Sari, M. (2010). Dengue confirmed-cases prediction: A neural network model, *Expert Systems with Applications* **37**: 4257–4260. https://doi.org/10.1016/j.eswa.2009.11.077.

Ahmad, R., Suzilah, I., Najdah, W. M. A. W., Topek, O., Mustafakamal, I. and Lee, H. L. (2018). Factors determining dengue outbreak in malaysia, *PLoS ONE* **13**(2). https://doi.org/10.1371/journal.pone.0193326.

Azhar, K., Marina, R. and Anwar, A. (2017). A prediction model of dengue iincidence using climate variability in denpasar city, *Health Science Journal of Indonesia* **8**(2): 68–73. https://doi.org/10.22435/hsji.v8i2.6952.68-73.

Baquero, O. S., Santana, L. M. R. and Neto, F. C. (2018). Dengue forecasting in são paulo city with generalized additive models, artificial neural networks and seasonal autoregressive integrated moving average models, *PLoS ONE* **13**(4). https://doi.org/10.1371/journal.pone.0195065.

Braga, O. C., Fonsêca, O. C., Moreira, M. W. L., Rodrigues, J. J. P. C., Silveira, F. R. V., Oliveira, A. M. B. and Neto, A. J. V. (2017). A mobile health solution for diseases control transmitted by aedes aegypti mosquito using predictive classifiers, *CoUrb*, Porto Alegre, pp. 144–156. Disponível em https://sol.sbc.org.br/index.php/courb/article/view/2570.

Brasil (2013). *Levantamento rápido de índices para Aedes aegypti: metodologia para avaliação dos índices de Breteau e Predial e tipo de recipientes*, Ministério da Saúde. Secretaria de Vigilância em Saúde. Departamento de Vigilância das Doenças Transmissíveis. Disponível em http://bvsms.saude.gov.br/bvs/publicacoes/manual_liraa_2013.pdf.

Brazier, K. T. S. (1994). Confidence intervals from the Rayleigh test, *Monthly Notices of the Royal Astronomical Society* **268**(3): 709–712. https://doi.org/10.1093/mnras/268.3.709.

Brownlee, J. (2018). *Deep Learning for Time Series Forecasting*, 1.4 edn, Machine Learning Mastery.

Carlos, M. A., Nogueira, M. and Machado, R. J. (2017). Analysis of dengue outbreaks using big data analytics and social networks, *The 2017 4th International Conference on Systems and Informatics*, pp. 1592–1597. Disponível em https://doi.org/10.1109/ICSAI.2017.8248538.

Castro, L. N. and Ferrari, D. G. (2016). *Introdução à Mineração de dados: Conceitos básicos, algoritmos e aplicaçções*, 1 edn, Saraiva, São Paulo, SP, Brasil.

G1 (2021). *Prefeitura de Cascavel realiza ação de combate à dengue em bairros com maior índice de infestação*. Disponível em https://tinyurl.com/2p9dvf5v.

G1 (2022). *Cascavel tem alto risco de infestação de dengue, aponta levantamento*. Disponível em https://tinyurl.com/y79cxv99.

Gharbi, M., Quenel, P., Gustave, J., Cassadou, S., Ruche, G. L., Girdary, L. and Marrama, L. (2011). Time series analysis of dengue incidence in guadeloupe, french west indies: Forecasting models using climate variables as predictors, *BMC infectious diseases* **11**: 166. https://doi.org/10.1186/1471-2334-11-166.

Goodfellow, I., Yoshua, B. and Courville, A. (2016). *Deep Learning (Adaptive Computation and Machine Learning series)*, The MIT Press.

Guo, P., Liu, T., Zhang, Q., Wang, L., Xiao, J., Zhang, Q., Luo, G., Li, Z., He, J., Zhang, Y. and Ma, W. (2017). Developing a dengue forecast model using machine learning: A case study in China, *PLoS Neglected Tropical Diseases* **11**(10). https://doi.org/10.1371/journal.pntd.0005973.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory, *Neural Computation* **9**(8): 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Hyndman, R. J. and Athanasopoulos, G. (2018). *Forecasting: Principles and Practice*, 2 edn, Texts, Monash University, Australia.

IBGE (2021). *Brasil/Paraná/Cascavel - Panorama*. Disponível em https://cidades.ibge.gov.br/brasil/pr/cascavel/panorama.

Keras (2022). *Redes Neurais Recorrentes (RNN) com Keras*, TensorFlow. Disponível em https://www.tensorflow.org/guide/keras/rnn.

Laserna, A., Barahona-Correa, J., Baquero, L., Castañeda-Cardona, C. and Rosselli, D. (2018). Economic impact of dengue fever in latin america and the caribbean: a systematic review, *Revista Panamericana de Salud Pública* **42**(17): 1–9. https://doi.org/10.26633/RPSP.2018.111.

Laureano-Rosario, A. E., Duncan, A. P., Mendez-Lazaro, P. A., Garcia-Rejon, J. E., Gomez-Carro, S., Farfan-Ale, J., Savic, D. A. and Muller-Karger, F. E. (2018). Application of artificial neural networks for dengue fever outbreak predictions in the northwest coast of yucatan, mexico and san juan, puerto rico, *Tropical Medicine and Infectious Disease* **3**(1). https://doi.org/10.3390/tropicalmed3010005.

Lee, C., Yang, H. and Lin, S. (2015). Incorporating big data and social sensors in a novel early warning system of dengue outbreaks, *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, Paris, França, pp. 1428–1433. Disponível em https://doi.org/10.1145/2808797.2808883.

Makridakis, S. G., Wheelwright, S. C. and Hyndman, R. J. (1997). *Forecasting: Methods and Applications*, 3 edn, Wiley, Nova Iorque, USA.

Mittelmann, M. and Soares, D. G. (2017a). Previsão de casos de dengue em itajaí - sc por meio de redes neurais artificiais multicamadas e recorrentes, *Computer on the Beach*, Florianópolis, pp. 130–139. Disponível em https://doi.org/10.14210/cotb.v0n0.p130-139.

Mittelmann, M. and Soares, D. G. (2017b). Previsão de casos de dengue no município de guarulhos com redes neurais artificiais multicamadas e recorrentes, *Revista de Informática Aplicada* **13**(2): 68–74. https://doi.org/10.13037/ria.vol13n2.200.

Mussumeci, E. and Codeço Coelho, F. (2020). Large-scale multivariate forecasting models for dengue - lstm versus random forest regression, *Spatial and Spatio-temporal Epidemiology* **35**: 100372. https://doi.org/10.1016/j.sste.2020.100372.

Nan, J., Liao, X., Chen, J., Chen, X., Chen, J., Dong, G., Liu, K. and Hu, G. (2018). Using climate factors to predict the outbreak of dengue fever, *2018 7th International Conference on Digital Home (ICDH)*, pp. 213–218. Disponível em https://doi.org/10.1109/ICDH.2018.00045.

PAHO (2020). *Mortalidad por Dengue - Número y Tasa*, Pan American Health Organization. Disponível em https://www3.paho.org/data/index.php/es/temas/indicadores-dengue/dengue-nacional/238-dengue-mortalidad-tasa.html.

Pham, D. N., Aziz, T., Kohan, A., Nellis, S., binti Abd. Jamil, J., Khoo, J. J., Lukose, D., bin Abu Bakar, S. and Sattar, A. (2015). An efficient method to predict dengue outbreaks in kuala lumpur, *3rd International Conference on Artificial Intelligence and Computer Science*, Penang, Malaysia, pp. 169–178. Disponível em https://www.researchgate.net/publication/309738064.

Pham, D. N., Nellisy, S., Sadanand, A. A., binti Abd. Jamily, J., Khooy, J. J., Aziz, T., Lukose, D., bin Abu Bakary, S. and Sattarz, A. (2016). A literature review of methods for dengue outbreak prediction, *The Eighth International Conference on Information, Process, and Knowledge Management*, Veneza, Itália, pp. 7–13. Disponível em https://www.researchgate.net/publication/318316853_A_Literature_Review_of_Methods_for_Dengue_Outbreak_Prediction/stats.

Phung, D., Hauang, C., Rutherford, S., Chu, C., Wang, X., Nguyen, M., Nguyen, N. H. and Manh, C. D. (2015). Identification of the prediction model for dengue incidence in can tho city, a mekong delta area in vietnam, *Acta Tropica* **141**: 88–96. https://doi.org/10.1016/j.actatropica.2014.10.005.

Prefeitura Municipal de Cascavel (2020). *Boletim da dengue: Cascavel registra 6.681 casos da doença*. Disponível em https://tinyurl.com/mr2n73ae.

Prefeitura Municipal de Cascavel (2021). *Clima - Mapa do Paraná*. Disponível em https://cascavel.atende.net/cidadao/pagina/mapas.

Ribeiro, A. F., Marques, G. R. A. M., Voltolini, J. C. and Condino, M. L. F. (2006). Associação entre incidência de dengue e variáveis climáticas, *Revista de Saúde Pública* **40**(4): 671–676. https://doi.org/10.1590/S0034-89102006000500017.

Rizzi, C. B., Rizzi, R. L., Pramiu, P. V., Hoffmann, E. and Codeço, C. T. (2017). Considerações sobre a dengue e variáveis de importância à infestação por aedes aegypti, *Revista Brasileira de Geografia Médica e da Saúde* **13**(24): 24–40. https://seer.ufu.br/index.php/hygeia/article/view/35133.

Scikit-learn developers (2021a). *Decision Trees*, Scikit-learn Development. Disponível em https://scikit-learn.org/stable/modules/tree.html#.

Scikit-learn developers (2021b). *Ensemble Methods*, Scikit-learn Development. Disponível em https://scikit-learn.org/stable/modules/ensemble.html#forest.

Scikit-learn developers (2021c). *Neural Network Models (supervised)*, Scikit-learn Development. Disponível em https://scikit-learn.org/stable/modules/neural_networks_supervised.html.

Scikit-learn developers (2021d). *sklearn.ensemble.RandomForestRegressor*, Scikit-learn Development. Disponível em https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html.

Scikit-learn developers (2021e). *Support Vector Machines*, Scikit-learn Development. Disponível em https://scikit-learn.org/stable/modules/svm.html.

Silva, W. and Frances, R. (2017). Monitoramento de epidemia de dengue na amazônia usando redes neurais artificiais, *in* L. Martí and N. Sãnchez Pi (eds), *Anais do 13 Congresso Brasileiro de Inteligência Computacional*, ABRICOM, Curitiba, PR, pp. 1–12. Disponível em https://doi.org/10.21528/CBIC2017-50.

Smola, A. J. and Scholkopf, B. (2004). A tutorial on support vector regression, *Statistics and Computing* **14**(3): 199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88.

Tan, P. N., Steinbach, M. and Kumar, V. (2009). *Introdução do Data Mining: Mineração de dados*, 1 edn, Ciência Moderna Ltda, Rio de Janeiro, RJ, Brasil.

Van Veen, F. and Leijnen, S. (219). *The Neural Network Zoo*, The Asimov Institute. Disponível em https://www.asimovinstitute.org/neural-network-zoo/.

WHO (2022). *Dengue and severe dengue*, World Health Organization. Disponível em https://www.who.int/en/news-room/fact-sheets/detail/dengue-and-severe-dengue.

Zhu, G., Hunter, J. and Jiang, Y. (2016). Improved prediction of dengue outbreak using the delay permutation entropy, *2016 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, Chengdu, China, pp. 828–832. Disponível em https://doi.org/10.1109/iThings-GreenCom-CPSCom-SmartData.2016.172.