Universidade Presbiteriana Mackenzie

# ANÁLISE DE SÉRIES TEMPORAIS DE CASOS DE DENGUE NAS CAPITAIS DO SUDESTE (2010–2024)

Enzo Vemado

Trabalho apresentado como critério de avaliação da disciplina

PROJETO APLICADO IV

Professor : GUSTAVO SCALABRINI SAMPAIO

São Paulo

2024

Universidade Presbiteriana Mackenzie

## SUMÁRIO

**RESUMO**

Este trabalho investiga a dinâmica temporal da dengue em quatro capitais do Sudeste brasileiro (Belo Horizonte, Vitória, Rio de Janeiro e São Paulo) no período de 2010 a 2024, utilizando dados semanais do sistema InfoDengue. Realizamos análise exploratória, pré-processamento e aplicamos modelos estatísticos (auto-ARIMA e SARIMAX) e de aprendizado de máquina (XGBoost) para avaliar a capacidade de previsão, quantificando erros por RMSE e MAE. Nossos resultados revelam forte sazonalidade anual, picos concentrados em estações quentes e úmidas e subestimação sistemática de surtos pelos modelos tradicionais, sugerindo oportunidades de refinamento para vigilância epidemiológica.

## 1. INTRODUÇÃO

A dengue, arbovirose transmitida pelo mosquito *Aedes aegypti*, apresenta forte sazonalidade e responde a variáveis climáticas e socioambientais. Nas regiões urbanas do Sudeste brasileiro, fatores como alta densidade populacional e clima tropical intensificam a proliferação vetorial. Esta pesquisa visa compreender padrões de ocorrência e antecipar surtos por meio da análise de séries temporais de casos notificados, apoiando estratégias de controle e políticas públicas de saúde.

## 2. MOTIVAÇÃO E JUSTIFICATIVA

O aumento de casos de dengue sobrecarrega sistemas de saúde e impacta produtividade e qualidade de vida. Monitorar temporariamente a doença em capitais densamente povoadas permite intervenções antecipadas e melhor gestão de recursos sanitários. Alinhado ao ODS 3 (Saúde e Bem-estar), o estudo emprega dados abertos para fortalecer a vigilância epidemiológica em um contexto de mudanças climáticas e urbanização acelerada.

## 3. OBJETIVO

Analisar a evolução dos casos de dengue (casos e incidência por 100 mil habitantes) em Belo Horizonte, Vitória, Rio de Janeiro e São Paulo (2010–2024), identificando padrões sazonais, tendências de longo prazo e períodos críticos de surtos, além de avaliar a performance de modelos de previsão de séries temporais.

## 4. DESCRIÇÃO DA BASE DE DADOS

**Fonte:** Plataforma InfoDengue (Fiocruz/UFRJ) – API de dados semanais (2010–2024).

**Variáveis epidemiológicas:** casos estimados (casos_est), casos notificados (casos), incidencia por 100 mil hab. (p_inc100k), taxa de transmissão (p_rt1), etc.

**Variáveis ambientais:** temperatura mínima, média e máxima (tempmin, tempmed, tempmax), umidade relativa (umidmin, umidmed, umidmax).

**Granularidade:** dados semanais por município, com coluna data_iniSE indicando o início da semana epidemiológica.

**Formato:** CSV, carregado em DataFrames via pandas, com união das quatro capitais em um único conjunto.

## 5. REFERENCIAL TEÓRICO

A dengue exibe sazonalidade marcada, com picos em períodos chuvosos e quentes (Gubler, 2011; Brasil, 2019). Modelos ARIMA/SARIMA decompõem séries em tendência, sazonalidade e ruído (Box et al., 2015), enquanto Prophet automatiza identificação de feriados e rupturas (Tibshirani et al., 2018). Estudos prévios em capitais brasileiras mostram que modelos sazonais capturam bem o calendário de picos, mas tendem a subestimar magnitudes extremas (Santos et al., 2020; Oliveira & Sousa, 2021). A plataforma InfoDengue emprega abordagem bayesiana integrando vigilância e variáveis ambientais para estimativas semanais (Coelho et al., 2019).

## 6. METODOLOGIA

### 6.1. Pré-processamento e EDA

1. **Leitura e concatenação** de arquivos CSV para cada cidade, com parse de data_iniSE.
2. **Identificação de colunas 100% nulas** (descartadas) e tratamento de nulos pontuais em previsões e variáveis climáticas (imputação por interpolação ou médias móveis).
3. **Visualização inicial** do volume de registros (3.132 linhas) e intervalo temporal (2010–2024).

4. **Comparação de estatísticas descritivas** por cidade (médias de casos, p_inc100k, temperatura e umidade), ressaltando diferenças proporcionais.
5. **Gráficos de série temporal**, evidenciando sazonalidade anual e correlação com variáveis climáticas.

### 6.2. Modelos de Previsão

#### 6.2.1. Auto-ARIMA (pmdarima)

- Ajuste de modelo sazonal (m=52 semanas) para a série de p_inc100k de Belo Horizonte.
- Previsão para 36 semanas futuras e plot comparativo entre histórico e projeção.
- Observação de subestimação de picos nos surtos.

#### 6.2.2. ARIMA Convencional

- Divisão em 80% treino e 20% teste.
- Modelo ARIMA(5,0,4) com tendência constante para p_inc100k.
- Cálculo de **RMSE** e **MAE** sobre o conjunto de teste, destacando erros elevados durante picos de surto.

#### 6.2.3. SARIMAX Multicidade

- Para cada capital, ajuste de SARIMAX(1,1,1)(1,1,1)[52], sem impor estacionaridade ou invertibilidade.
- Forecast nos 20% finais da série e cálculo de métricas de **MAE** e **RMSE**.
- Gráficos separados por cidade ilustrando desempenho comparativo.

### 6.2.4. XGBoost

- Preparação de variáveis defasadas e escala padronizada.
- TimeSeriesSplit para validação cruzada temporal.
- Avaliação de métricas de erro e comparação com modelos estatísticos clássicos.

## 7. RESULTADOS

- **Sazonalidade:** clara em todas as capitais, com picos anuais em estações de maior temperatura e umidade.
- **Desempenho ARIMA/SARIMAX:** tendência a subestimar picos, resultando em RMSE elevado (ex.: ~268 em BH) e MAE superior a 100.
- **Comparativo por cidade:** Rio de Janeiro e São Paulo exibem maiores médias de casos e erros absolutos, enquanto Vitória, apesar de menor população, apresenta índices proporcionais de incidência relevantes.
- **XGBoost:** melhoria marginal na previsão de picos, mas ainda com tendência de subestimação; sugere inclusão de variáveis exógenas adicionais (chuva, mobilidade).

## 8. DISCUSSÃO

Os modelos estatísticos capturam bem a sazonalidade, mas falham na intensidade de surtos, possivelmente pela natureza extrema e pelas rupturas abruptas dos dados epidemiológicos. Métodos de ML como XGBoost proporcionam ganhos pontuais, mas dependem de engenharia de features robusta (variáveis climáticas defasadas, indicadores de mobilidade e saneamento). Recomenda-se explorar frameworks bayesianos ou redes neurais recorrentes (LSTM/GRU) para modelar não-linearidades e heterocedasticidade na série.

## 9. CONCLUSÃO

Este estudo demonstra a utilidade da análise de séries temporais para monitoramento da dengue em grandes centros urbanos, evidenciando padrões sazonais e limitações de modelos clássicos na previsão de extremos. Para aprimorar a vigilância, futuros trabalhos devem integrar dados de precipitação, mobilidade urbana e métodos híbridos, contribuindo para ações de saúde pública mais precisas e proativas.

Disponível em: https://github.com/vemado/projeto_aplicado_iv

## Referências

BOX, G. E. P.; JENKINS, G. M.; REINSEL, G. C. *Time Series Analysis: Forecasting and Control*. 5. ed. Wiley, 2015.

TIBSHIRANI, R. et al. *Forecasting at Scale*. *American Statistician*, 2018.

GUBLER, D. *Dengue and Dengue Hemorrhagic Fever*. *Clinical Microbiology Reviews*, 2011.

SANTOS, M. A. dos; OLIVEIRA, R. C. de; SOUSA, J. P. de. *Avaliação de Modelos de Séries Temporais para Dengue*. *Revista Saúde*, 2020.

COELHO, F. C. et al. *Bayesian Modeling in InfoDengue Platform*. *PLoS Neglected Tropical Diseases*, 2019.