

Previsão de Surtos de Dengue em Nível de Unidade de Desenvolvimento Humano (UDH) no Brasil: Uma Abordagem Integrada e Explicável

Silmara Barnabé¹, Maximus Borges da Rosa¹, Carla Maria Dal Sasso Freitas¹,
Weverton Cordeiro¹, Mariana Recamonde-Mendoza¹

¹Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

{sgbarnabe, mbrosa, carla, weverton.cordeiro, mrmendoza}@inf.ufrgs.br

Abstract. *This work aims to predict dengue outbreaks at the Human Development Unit (UDH) level in Brazil by integrating epidemiological, climatic, and social vulnerability data. Data from SINAN were combined with climate information from BDMEP/INMET and indicators from the Social Vulnerability Atlas. The panel covers 240 UDHs in Porto Alegre and Rio de Janeiro between 2021 and 2025, totaling more than 4,000 monthly records. The current stage involved preprocessing and integrating the databases, resulting in a robust dataset that will be used for developing machine learning models to model time series and predict dengue outbreaks, with interpretability provided by SHAP.*

Resumo. *Este trabalho visa prever surtos de dengue em nível de Unidade de Desenvolvimento Humano (UDH) no Brasil, integrando dados epidemiológicos, climáticos e de vulnerabilidade social. Foram combinados casos do SINAN, informações climáticas do BDMEP/INMET e indicadores do Atlas de Vulnerabilidade Social. O painel cobre 240 UDHs de Porto Alegre e Rio de Janeiro entre 2021 e 2025, totalizando mais de 4.000 registros mensais. A etapa atual envolveu o pré-processamento e a integração das bases, resultando em um dataset robusto que será utilizado para o desenvolvimento de modelos preditivos com aprendizado de máquina para modelar séries temporais e prever surtos de dengue, com interpretabilidade por SHAP.*

1. Introdução

A dengue é uma arbovirose que representa um desafio significativo para a saúde pública no Brasil e no mundo, especialmente em áreas urbanas densamente povoadas [Wilder-Smith et al. 2022]. Além de sua relevância epidemiológica, os impactos econômicos e sociais da doença reforçam a necessidade de estratégias de antecipação de risco [Shepard et al. 2023]. A previsão de surtos é, portanto, essencial para orientar ações de controle vetorial e otimizar a alocação de recursos. A literatura recente destaca que variáveis climáticas — particularmente a temperatura — modulam a dinâmica de transmissão ao influenciar o número reprodutivo e os intervalos de geração [Siraj et al. 2022]. Nesse contexto, abordagens de explainable machine learning (ML) têm se mostrado promissoras por produzirem alertas interpretáveis e úteis para tomadores de decisão [Kon et al. 2022].

Diferentemente de estudos tradicionais em nível municipal, este trabalho propõe o desenvolvimento de modelos preditivos para casos de dengue em recortes com granularidade de Unidade de Desenvolvimento Humano (UDH) — equivalente a um sub-bairro

—, possibilitando análises e intervenções mais direcionadas. Serão realizadas etapas de coleta, pré-processamento e integração de diferentes bases de dados públicas, resultando em uma base harmonizada para a modelagem preditiva. Embora a coleta de dados abranja todo o território nacional, a modelagem inicial será aplicada aos municípios de Porto Alegre e Rio de Janeiro, utilizando métodos estatísticos e de aprendizado de máquina. Dessa forma, esta pesquisa visa contribuir em duas frentes: (i) a geração e o compartilhamento de um conjunto de dados abrangente e robusto para a análise da dinâmica e dos padrões de casos de dengue no Brasil; e (ii) o desenvolvimento de modelos preditivos capazes de estimar o número de casos e antecipar riscos de surtos da doença.

2. Materiais e Métodos

2.1. Dados Epidemiológicos — SINAN

Os dados epidemiológicos foram obtidos do Sistema de Informação de Agravos de Notificação (SINAN)¹. A granularidade inicial é em nível de estabelecimento de saúde (CNES)², cada um associado a coordenadas geográficas (latitude/longitude). A partir desses pontos, procedeu-se ao mapeamento para Unidades de Desenvolvimento Humano (UDHs), utilizando centróides como critério de atribuição, em processo coerente com a integração de fontes para vigilância baseada em dados [Li et al. 2022, Nsoesie et al. 2023].

2.2. Dados de Vulnerabilidade Social — AVS

Os indicadores socioeconômicos foram obtidos do Atlas de Vulnerabilidade Social (AVS) em nível de UDH³. Foram incorporadas variáveis como índice de saneamento básico, renda média e vulnerabilidade social, permitindo contextualização social dos perfis de risco em linha com melhores práticas de vigilância e planejamento [Nsoesie et al. 2023].

2.3. Dados Climáticos — BDMEP/INMET

As variáveis climáticas foram obtidas do Banco de Dados Meteorológicos para Ensino e Pesquisa (BDMEP) do INMET⁴ (temperatura média, precipitação acumulada e umidade relativa). A granularidade destes dados é municipal, implicando replicação para todas as UDHs do mesmo município. A ênfase em temperatura e precipitação é consistente com evidências de que tais variáveis são determinantes do *timing* e magnitude de surtos [Kraemer et al. 2023, Siraj et al. 2022].

2.4. Integração e Estruturação do Dataset

A integração resultou em um dataset estruturado em nível UDH-mês, cobrindo 2021–2025, com três camadas de dados: epidemiológica (SINAN/CNES), climática (BDMEP/INMET) e social (AVS). As chaves de integração incluíram: (i) código CNES georeferenciado e mapeado para UDH e (ii) identificador de município para atribuição

¹ **SINAN**: <https://basedosdados.org/dataset/f51134c2-5ab9-4bbc-882f-f1034603147a?table=7d2d29c1-cfa8-470b-ab13-a83cf522c5ac>

² **CNES**: <https://cnes.datasus.gov.br/pages/downloads/arquivosBaseDados.jsp>

³ **AVS**: <https://basedosdados.org/dataset/1dd6317a-8e79-4a67-bb78-bc5fef214b4d?table=8dce525d-07f9-4d52-8ea0-36fa288d7df8>

⁴ **BDMEP/INMET**: <https://basedosdados.org/dataset/782c5607-9f69-4e12-b0d5-aa0f1a7a94e2?table=2c7fdc3d-f2ed-4c78-84b8-d9c792a06703>

de variáveis climáticas e socioeconômicas. O desenho segue princípios de *data engineering* para vigilância quase em tempo real e construção de *dashboards* operacionais [Li et al. 2022, Nsoesie et al. 2023] para permitir fácil visualização e exploração dos dados. A Figura 1 mostra a arquitetura do dataset gerado.

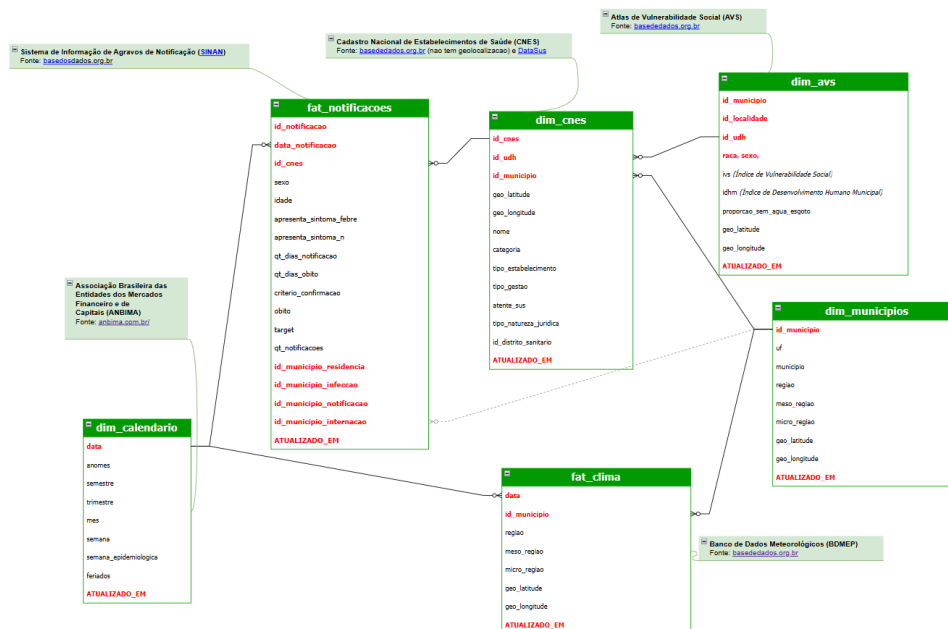


Figure 1. Arquitetura do dataset utilizado para a previsão de surtos de dengue (UDH-mês).

2.5. Desenvolvimento de Modelos Preditivos

Para a etapa de modelagem preditiva dos casos de dengue, serão comparados modelos sazonais (SARIMA/SARIMAX), modelos *ensemble* de gradiente (CatBoost) e modelos sequenciais. Será empregada análise de explicabilidade com SHAP, visando garantir transparência e auditabilidade dos resultados, em consonância com aplicações recentes na área [Kon et al. 2022]. Métricas de avaliação de regressão e de avaliação de classificação robustas ao desbalanceamento de casos, como área sobre a curva de precisão e recall (PR-AUC) e Matthews Correlation Coefficient (MCC), serão exploradas. Também serão priorizadas análises de calibração dos modelos, em conformidade com as recomendações atuais para avaliação de modelos em vigilância epidemiológica.

3. Resultados Parciais

O pré-processamento e a análise exploratória dos dados integrados evidenciam sazonalidade bem marcada das notificações (picos no verão), consistente com a literatura climato-epidemiológica da dengue [Kraemer et al. 2023]. Gráficos do *dashboard* desenvolvido em Power BI, apresentados na Figura 2 mostram o comportamento temporal em múltiplas escalas (Brasil, RS, RJ e Porto Alegre por UDH), reforçando a adequação do recorte espacial e temporal para modelagem prospectiva.

4. Conclusão

Os resultados parciais demonstram avanço na preparação de um dataset em granularidade UDH, passo crítico para intervenções mais focalizadas. O uso com-

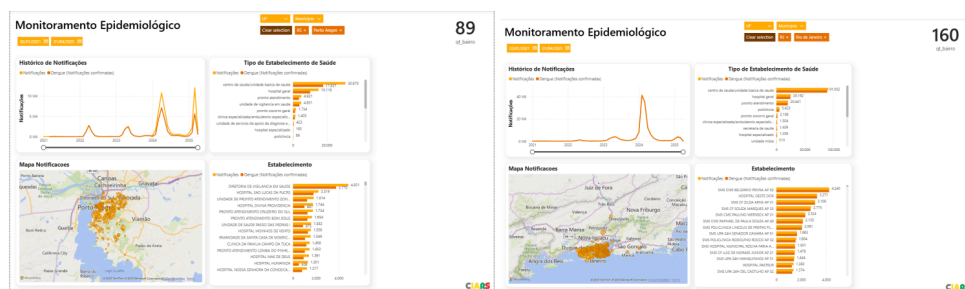


Figure 2. Histórico de notificações em Porto Alegre (RS) e Rio de Janeiro (RJ).

binado de notificações (SINAN/CNES), clima (BDMEP/INMET) e vulnerabilidade (AVS) alinha-se às prioridades de pesquisa e saúde pública para arboviroses [Wilder-Smith et al. 2022], maximizando a utilidade operacional de sistemas preditivos. A literatura sustenta que temperatura e precipitação são determinantes-chave do risco de dengue [Siraj et al. 2022, Kraemer et al. 2023]; assim, o plano de modelagem com explicabilidade [Kon et al. 2022] e métricas adequadas ao desbalanceamento favorece *insights* acionáveis para gestores. Os próximos passos da pesquisa incluem (i) definição de *thresholds* de surto por UDH com base em *baselines* sazonais, (ii) avaliação por *rolling origin* em múltiplos horizontes, (iii) explicabilidade global e local, e (iv) integração dos resultados em *dashboards* de vigilância [Li et al. 2022, Nsoesie et al. 2023].

Agradecimentos

Este estudo foi realizado com apoio da CAPES - Código de Financiamento 001, da FAPERGS [22/2551-0000390-7 (Projeto CIARS)] e do CNPq [308075/2021-8].

References

- Kon, F. et al. (2022). Data integration and machine learning approaches for improving dengue outbreak forecasts in brazil. In *Workshop AI4Health*.
- Kraemer, M. U. G., Pigott, D. M., and Brady, O. J. (2023). The global dengue dataset: modelled historical climate-driven dengue risk. *Nature Communications*, 14:2503.
- Li, J., Arsevska, E., and Fournié, G. (2022). Implementation of a data pipeline for near real-time disease surveillance dashboards. *International Journal of Environmental Research and Public Health*, 19(24):16468.
- Nsoesie, E. O., Kraemer, M. U. G., and Golding, N. (2023). A data-driven approach to building dashboards for infectious disease forecasts. *Epidemics*, 41:100618.
- Shepard, D. S., Undurraga, E. A., and Halasa, Y. A. (2023). Economic and disease burden of dengue in southeast asia. *PLoS Neglected Tropical Diseases*, 17(2):e0010111.
- Siraj, A. S., Santos-Vega, M., and Bouma, M. J. (2022). Temperature modulates dengue virus epidemic growth rates through its effect on reproduction number. *Scientific Reports*, 12:9590.
- Wilder-Smith, A., Scott, T. W., Rothman, A. L., and Stufkens, L. (2022). Epidemic arboviral diseases: priorities for research and public health. *The Lancet Infectious Diseases*, 22(9):e254–e262.