

Modelagem Preditiva Multi-step de Séries Temporais de Dengue no Brasil com Estratificação Espacial e Filtragem Multiescalar via Transformadas Wavelet

Marcos Paulo V. Pedrosa¹, Giovanni Zanardo¹,
Fabíola G. Nakamura¹, Eduardo F. Nakamura¹

¹Instituto de Computação (IComp) – Universidade Federal do Amazonas (UFAM)
Av. Gen. Rodrigo Octávio, 6200, Coroado I, Setor Norte
69080-900 – Manaus – AM – Brasil

{marcos.pedrosa,gio.zanardo,fabiola,nakamura}@icomp.ufam.edu.br

Abstract. *Dengue remains a persistent public health challenge in Brazil, requiring accurate predictive tools. This work in progress investigates the use of multiscale filtering techniques, particularly Wavelet Transforms, combined with machine learning for modeling weekly time series across Brazilian federative units (states). Using data from the InfoDengue platform covering all 27 states (2010–2026), preliminary experiments were conducted involving preprocessing, hierarchical clustering, and stability analysis via bootstrap. The results indicate that wavelet filtering improves series stability and enhances the identification of coherent epidemiological patterns, motivating the current stage of multi-step forecasting with both global and stratified approaches.*

Keywords: Time series; Wavelet; Machine learning; Dengue; Clustering.

Resumo. *A dengue é um desafio persistente à saúde pública no Brasil, exigindo ferramentas preditivas assertivas. Este trabalho em andamento investiga o uso de técnicas de filtragem multiescalar, especialmente Transformadas Wavelet, aliadas ao aprendizado de máquina para modelagem de séries temporais semanais nas unidades federativas (UFs). Utilizando dados da plataforma InfoDengue das 27 UFs (2010–2026), foram realizados experimentos preliminares de pré-processamento, clusterização hierárquica e análise de estabilidade via bootstrap. Os resultados indicam que a filtragem wavelet melhora a estabilidade das séries e a identificação de padrões epidemiológicos coerentes, motivando a etapa atual de previsão multi-step com abordagens globais e estratificadas.*

Palavras-chave: Séries temporais; Wavelet; Aprendizado de máquina; Dengue; Clusterização.

1. Introdução

As Doenças Tropicais Negligenciadas (DTNs) constituem um dos maiores desafios globais de saúde pública, impactando aproximadamente 1,65 bilhão de pessoas em todo o mundo [Nações Unidas 2023]. No Brasil, a dengue assume um papel central nesse cenário epidemiológico, apresentando um comportamento endêmico-epidêmico recorrente que sobrecarrega o sistema de saúde [Ross 2010]. A complexidade da doença é ampliada por fatores como mudanças climáticas, urbanização desordenada e a coexistência de múltiplos sorotipos do vírus DENV [Ross 2010, Silva and Angerami 2008].

Dados recentes indicam um aumento expressivo de casos, com mais de 1,5 milhão de ocorrências prováveis nos primeiros meses de 2024, reforçando a urgência de ferramentas preditivas robustas que auxiliem na tomada de decisão e na alocação antecipada de recursos [OPAS/OMS 2024]. A modelagem de séries temporais epidemiológicas, entretanto, apresenta desafios intrínsecos devido à natureza ruidosa, não-estacionária e à alta variabilidade dos dados de notificação [Shumway and Stoffer 2017]. Embora abordagens baseadas em aprendizado de máquina tenham demonstrado capacidade superior para capturar dependências temporais complexas em comparação a modelos estatísticos tradicionais [Benidis et al. 2022, Chae et al. 2018], a aplicação direta desses algoritmos em dados brutos pode ser prejudicada pela interferência de componentes estocásticos irrelevantes que obscurecem as tendências subjacentes [Percival and Walden 2000, Nason 2008].

Nesse contexto, este trabalho em andamento investiga como técnicas de processamento de sinais, especificamente a Transformada *Wavelet*, podem aprimorar a geometria do espaço de características antes da etapa de predição [Bishop 2006]. A Transformada *Wavelet* permite a decomposição multirresolução, oferecendo simultaneamente localização temporal e frequencial, o que é ideal para identificar surtos localizados e padrões sazonais recorrentes [Mallat 2009, Soares et al. 2018].

A hipótese central desta pesquisa é que o pré-processamento via *wavelets*, aliado à estratificação espacial por meio de *clusters* epidemiologicamente coerentes, pode reduzir a variância dos modelos e aumentar sua capacidade de generalização [Benidis et al. 2022]. Nas etapas iniciais do trabalho, o foco concentrou-se na consolidação da base de dados (proveniente da plataforma InfoDengue [Codeço et al. 2018]), no desenvolvimento de *pipelines* de filtragem e na validação da estabilidade estrutural das séries [Hubert and Arabie 1985]. Atualmente, o projeto avançou para a aplicação de modelos supervisionados de *Ensemble Learning* e *Deep Learning*, incluindo arquiteturas baseadas em redes *Long Short-Term Memory* (LSTM) e *Transformers* [Hochreiter and Schmidhuber 1997, Vaswani et al. 2017]. O objetivo final é integrar processamento de sinais, análise espacial e aprendizado de máquina em uma abordagem metodológica para a predição *multi-step* da dengue no território brasileiro [Benidis et al. 2022].

2. Metodologia

A abordagem combina processamento de sinais epidemiológicos e análise estrutural para embasar a modelagem supervisionada.

2.1. Coleta e Curadoria de Dados

Os dados brutos foram obtidos a partir da plataforma InfoDengue [Codeço et al. 2018], compreendendo o número de notificações semanais das 27 unidades federativas (UFs) do Brasil no período de 2010 a 2026. O pré-processamento inicial incluiu a interpolação linear para preenchimento de lacunas temporais inerentes a sistemas de notificação em saúde. Em seguida, para estabilizar a variância da série e atenuar a disparidade extrema de magnitude entre períodos endêmicos e surtos epidêmicos agudos, aplicou-se a transformação logarítmica definida por $x' = \log(x + 1)$ [Shumway and Stoffer 2017].

2.2. Filtragem Multiescalar via Transformada Wavelet

Para a redução de ruídos estocásticos de alta frequência, utilizou-se a Transformada *Wavelet* Discreta [Percival and Walden 2000]. Adotou-se a família de *wavelets Daubechies* 4 (*db4*), cuja assimetria e suporte compacto assemelham-se morfológicamente à dinâmica de ascensão e queda dos picos epidêmicos de dengue [Daubechies 1992]. A série temporal foi decomposta em três níveis de resolução. Após a decomposição, aplicou-se a técnica de *soft-thresholding* [Donoho 1995] nos coeficientes de detalhe para reconstruir o sinal suavizado. Essa profundidade de filtragem (nível 3) provou-se eficaz para isolar flutuações irrelevantes preservando estritamente as componentes estruturais de tendência e sazonalidade da doença.

2.3. Pipelines Experimentais e Agrupamento

Para avaliar a influência da ordem e da necessidade das operações matemáticas de pré-processamento, a análise estrutural foi conduzida através de cinco *pipelines* metodológicos distintos:

1. **Wavelet** → **RobustScaler**: Aplicação primária da filtragem multiescalar seguida de normalização baseada no intervalo interquartil.
2. **RobustScaler** → **Wavelet**: Inversão da etapa anterior, com a normalização antecedendo a filtragem.
3. **Apenas RobustScaler**: Agrupamento sobre os dados submetidos unicamente à normalização.
4. **Apenas Wavelet**: Agrupamento sobre as séries suavizadas, sem normalização.
5. **Baseline**: Dados com a transformação inicial $\log(x + 1)$, sem processos de filtragem ou redimensionamento locais.

Para a identificação de padrões, empregou-se o algoritmo de *Hierarchical Clustering*. A matriz de dissimilaridade testou quatro abordagens: Euclidiana, Correlação de Pearson, Correlação de Spearman e *Dynamic Time Warping* (DTW) [Berndt and Clifford 1994]. O DTW destaca-se neste escopo devido ao seu alinhamento elástico, capaz de lidar computacionalmente com as defasagens temporais (*lags*) de contágio que ocorrem entre estados vizinhos.

2.4. Validação de Estabilidade

A robustez dos agrupamentos foi validada estatisticamente através do método *Moving Block Bootstrap* (MBB) [Künsch 1989], que permite a reamostragem preservando a autocorrelação de curto prazo típica das dinâmicas de infecção. Foram geradas $B = 50$ pseudomostras para cada *pipeline*, utilizando blocos temporais de $b = 8$ semanas epidemiológicas. O intervalo de busca para o número ótimo de *clusters* foi restrito a $k \in \{2, 3, 4, 5, 6, 7, 8\}$. A métrica principal de avaliação e seleção metodológica foi o *Adjusted Rand Index* (ARI) [Hubert and Arabie 1985], que quantifica a estabilidade das partições sob perturbações de reamostragem (penalizando agrupamentos formados ao acaso).

2.5. Modelagem Preditiva Multivariada

Após isso, foi feita a reformulação do *dataset* para tarefas supervisionadas focadas em predição *multi-step* (horizontes de 2 a 16 semanas de antecedência) [Benidis et al. 2022].

Assim, na etapa atual desta pesquisa, estão sendo investigadas duas abordagens de treinamento: (i) modelos globais, utilizando as 27 UFs simultaneamente; e (ii) modelos estratificados espacialmente, nos quais um modelo especialista é treinado exclusivamente com as unidades federativas pertencentes ao mesmo *cluster* estrutural. Esta estratificação tem como objetivo principal reduzir a variância do algoritmo por meio da homogeneidade regional imposta na entrada. O desempenho de arquiteturas *Ensemble* [Breiman 2001, Chen and Guestrin 2016] e *Deep Learning* [Hochreiter and Schmidhuber 1997, Vaswani et al. 2017] será comparado e validado via *Leave-One-Out* espacial, RMSE e MAE.

3. Resultados

Os experimentos conduzidos evidenciaram a capacidade das técnicas de filtragem multiescalar em revelar estruturas latentes e estáveis nos dados epidemiológicos semanais de dengue. A avaliação quantitativa da estabilidade dos *clusters* demonstrou que o pré-processamento afeta diretamente a geometria do espaço de características. A Tabela 1 apresenta as cinco configurações metodológicas que obtiveram a maior estabilidade estrutural sob as perturbações de reamostragem do *Moving Block Bootstrap* (MBB).

Tabela 1. Melhores configurações de *pipeline* na clusterização hierárquica, ordenadas pelo *Adjusted Rand Index* (ARI) médio.

<i>Pipeline</i>	Métrica de Distância	<i>k</i> ótimo	ARI Médio (↑)
<i>Wavelet</i> → <i>Robust</i>	<i>Dynamic Time Warping</i> (DTW)	2	0,885
Apenas <i>Wavelet</i>	<i>Dynamic Time Warping</i> (DTW)	5	0,866
Apenas <i>Wavelet</i>	Euclidiana	3	0,837
<i>Baseline</i> (Dados Brutos)	Euclidiana	2	0,749
<i>Wavelet</i> → <i>Robust</i>	Pearson	5	0,746

A combinação primária, composta por filtragem *Wavelet* (*db4*) seguida de normalização via *RobustScaler* e acoplada à métrica *Dynamic Time Warping* (DTW) [Berndt and Clifford 1994], obteve a maior estabilidade geral do estudo. Esta configuração alcançou um ARI [Hubert and Arabie 1985] médio de 0,885, indicando a consistência das partições encontradas. Constatou-se também que a ordem das operações matemáticas é um fator relevante. O *pipeline* inverso (*RobustScaler* → *Wavelet*) apresentou desempenho inferior. Ao aplicar o escalonamento baseado no intervalo interquartil diretamente sobre o sinal estocástico bruto, as flutuações de alta frequência são amplificadas desproporcionalmente. Esse fenômeno ofusca a componente estrutural do sinal, dificultando a distinção algorítmica entre o comportamento epidemiológico de base e o ruído dos sistemas de notificação.

No quesito da análise espacial, os agrupamentos ótimos obtidos levantaram discussões sobre as dinâmicas geográficas de disseminação da doença, evidenciadas na Figura 1.

Enquanto os dados brutos (*Baseline*) refletem apenas macro-tendências climáticas que dividem o país em dois blocos ($k = 2$), a filtragem *Wavelet* revela cinco padrões regionais específicos ($k = 5$). Ao limpar os ruídos de notificação sem perder a sazonalidade, a *Wavelet* permite capturar detalhes finos, como as defasagens de contágio entre estados

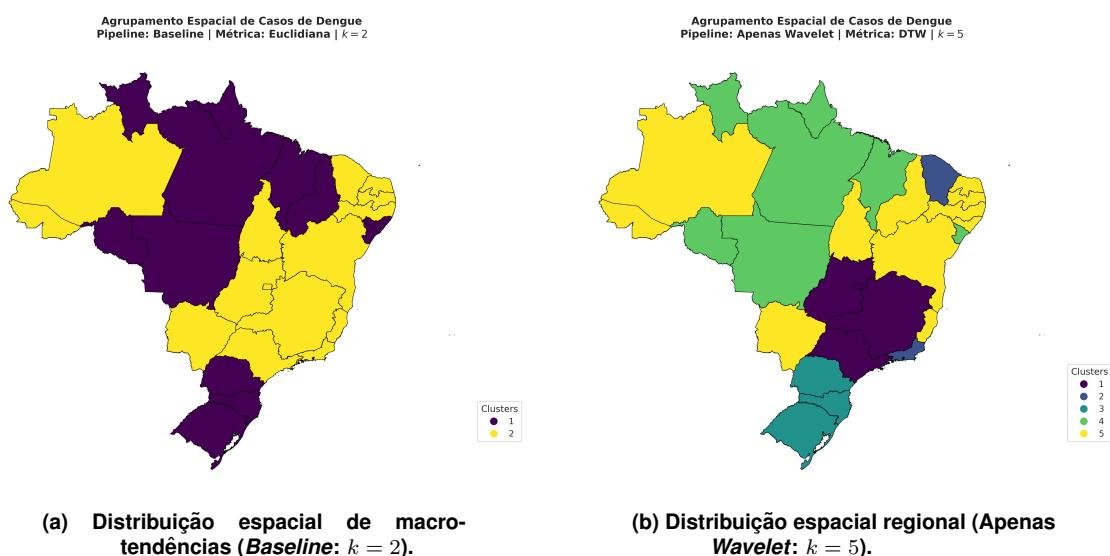


Figura 1. Comparativo da projeção espacial dos agrupamentos de dengue no Brasil.

vizinhos (especialmente quando alinhada ao DTW), resultando em um mapeamento da heterogeneidade da dengue no Brasil. A obtenção de *clusters* geometricamente estáveis valida a hipótese de que a filtragem multiescalar atua como um mecanismo alternativo de estratificação espacial. Estes resultados estabelecem a base para a etapa de predição *multi-step*, na qual o treinamento de modelos supervisionados multivariados será isolado dentro de cada subconjunto estruturalmente homogêneo visando a redução da variância.

4. Conclusão

Os resultados deste trabalho em andamento mostram que a filtragem multiescalar aprimora a geometria do espaço de características das séries temporais de dengue. O aumento consistente da estabilidade estrutural — medida pelo Adjusted Rand Index — indica que a Transformada Wavelet (db4) atenuou os ruídos estocásticos de notificação, preservando a sazonalidade e as tendências agudas da doença. A formação de agrupamentos espaciais estáveis confirma que a disseminação da dengue no Brasil não é globalmente homogênea. Combinada ao alinhamento elástico (DTW) sobre o sinal processado, a filtragem revelou nuances regionais ($k=5$), capturando as defasagens temporais de contágio entre estados vizinhos. Atualmente, esta estrutura geométrica está sendo explorada em tarefas de aprendizado supervisionado para a previsão *multi-step* (horizontes de 2 a 16 semanas). Modelos especialistas baseados em *Ensemble Learning* [Breiman 2001, Chen and Guestrin 2016] e *Deep Learning* (como LSTMs [Hochreiter and Schmidhuber 1997] e *Transformers* [Vaswani et al. 2017]) serão treinados dentro de cada subconjunto estruturalmente homogêneo (*cluster*). A hipótese principal, sustentada pelos resultados atuais, é que esta homogeneidade regional imposta reduzirá a variância dos algoritmos, maximizando a sua capacidade de generalização espacial.

Referências

Benidis, K. et al. (2022). Deep learning for time series forecasting: tutorial and literature survey. *ACM Computing Surveys*, 55(6):1–36.

- Berndt, D. J. and Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370.
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
- Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.
- Chae, S. et al. (2018). Predicting infectious disease using deep learning and big data. *International Journal of Environmental Research and Public Health*, 15(8):1596.
- Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.
- Codeço, C. T. et al. (2018). Infodengue: a nowcasting system for the surveillance of arboviruses in brazil. *Revue d'Épidémiologie et de Santé Publique*, 66:S386.
- Daubechies, I. (1992). *Ten Lectures on Wavelets*. SIAM.
- Donoho, D. L. (1995). De-noising by soft-thresholding. *IEEE transactions on information theory*, 41(3):613–627.
- Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2(1):193–218.
- Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *The annals of Statistics*, pages 1217–1241.
- Mallat, S. (2009). *A Wavelet Tour of Signal Processing: The Sparse Way*. Academic Press, 3rd edition.
- Nason, G. P. (2008). *Wavelet Methods in Statistics with R*. Springer.
- Nações Unidas (2023). Mais de 1,6 bilhão precisam de cuidados contra doenças tropicais negligenciadas. ONU News.
- OPAS/OMS (2024). *Atualização Epidemiológica: aumento de casos de dengue nas Américas*. Organização Pan-Americana da Saúde.
- Percival, D. B. and Walden, A. T. (2000). *Wavelet Methods for Time Series Analysis*. Cambridge University Press.
- Ross, T. M. (2010). Dengue virus. *Clinics in Laboratory Medicine*, 30(1):149–160.
- Shumway, R. H. and Stoffer, D. S. (2017). *Time Series Analysis and Its Applications*. Springer.
- Silva, L. J. and Angerami, R. N. (2008). *Viroses emergentes no Brasil*. Editora Fiocruz.
- Soares, A. P. M. R. et al. (2018). Prognóstico da incidência de casos de dengue na cidade de salvador-bahia, utilizando a transformada wavelet discreta em conjunção com redes neurais artificiais. *Interfaces Científicas - Saúde e Ambiente*, 6(3):53–62.
- Vaswani, A. et al. (2017). Attention is all you need. In *Advances in neural information processing systems*, volume 30.