

Epidemiológica

e Janeiro (2010-2016)

is com Complexos Simpliciais

mbro 2025

/mei-the-dev/dengue



Análise Epidemiológica

Dengue no Rio de Janeiro (2010-2016)

Topological Data Analysis com Complexos Simpliciais

Novembro 2025



github.com/mei-the-dev/dengue

- 1 Introdução e Dados
- 2 Semanas Epidemiológicas
- 3 Normalização dos Dados
- 4 Matrizes de Distância
- 5 Complexos Simpliciais
- 6 Análise de Componentes Principais
- 7 Clusterização
- 8 KeplerMapper

Visão Geral do Projeto

Fonte dos Dados

- **Arquivo:**
Dengue_Brasil_2010-2016_RJ.xlsx
- **Período:** 7 anos (2010–2016)
- **Região:** Estado do Rio de Janeiro
- **Municípios:** 91 analisados
- **Granularidade:** Semana epidemiológica

Números-Chave

91

Municípios

52

Semanas/Ano

2013

Ano de Referência

Objetivo

Analisar a dinâmica da dengue no RJ usando métodos de Topological Data Analysis (TDA)

output/tarefa1_curvas_por_ano.pdf

Achados Principais

- **2013**: Maior surto epidêmico
- Picos: janeiro a abril
- Padrão sazonal consistente
- 2016: dados incompletos (32 sem.)

Seleção

Ano **2013** escolhido como referência por ter dados completos e maior dinâmica epidêmica.



Definição (OMS/CDC)

A **semana epidemiológica (SE)** é a unidade de tempo padrão para vigilância:

- SE 1 inicia no primeiro domingo $\geq 1^{\circ}$ janeiro
- Maioria dos anos: **52 semanas**
- Anos bissextos especiais: **53 semanas**

| Ano | Semanas |
|-----------|-----------|
| 2010–2013 | 52 |
| 2014 | 53 |
| 2015 | 52 |
| 2016 | 32* |

output/tarefa1_curva_epidemica_tipica.

Curva epidêmica típica (normalizada).

Taxa de Incidência

$$\text{Taxa} = \frac{\text{Casos}}{\text{Pop.}} \times 100.000$$

Objetivo: Comparar intensidade entre municípios de diferentes tamanhos.

Usa dados do Censo 2010 (IBGE)

Área Unitária

$$\tilde{x}_i = \frac{x_i}{\sum_{j=1}^{52} x_j}$$

Objetivo: Comparar a *forma* das curvas, independente da magnitude.

Soma de cada série = 1


Por que normalizar?

A normalização por área unitária permite identificar **municípios com dinâmicas sincronizadas**, mesmo com números absolutos muito diferentes (ex: Rio de Janeiro vs. pequeno município do interior).

Top Municípios — Taxa de Incidência vs. Casos Absolutos

% Por Taxa (100.000 hab.)


| # | Município | Taxa |
|---|----------------|---------|
| 1 | Aperibé | 2,847.3 |
| 2 | Italva | 2,156.8 |
| 3 | Cambuci | 1,923.4 |
| 4 | Laje do Muriaé | 1,845.2 |
| 5 | Varre-Sai | 1,678.9 |

 Municípios pequenos com alta transmissão

Por Casos Absolutos

| # | Município | Casos |
|---|-----------------|--------|
| 1 | Rio de Janeiro | 89,234 |
| 2 | Niterói | 12,456 |
| 3 | São Gonçalo | 8,932 |
| 4 | Duque de Caxias | 7,845 |
| 5 | Nova Iguaçu | 6,234 |

 Grandes centros urbanos

extcolordengueRed  **Insight:** Municípios pequenos podem ter taxas altíssimas com poucos casos absolutos

Distância L1 (Manhattan)

$$d_{L1}(x, y) = \sum_{i=1}^{52} |x_i - y_i|$$

- Mais **robusta** a outliers
- Soma das diferenças absolutas
- Range: $[0, 2]$ para séries unitárias

Distância L2 (Euclidiana)

$$d_{L2}(x, y) = \sqrt{\sum_{i=1}^{52} (x_i - y_i)^2}$$

- **Penaliza** grandes diferenças
- Sensível a picos isolados
- Range: $[0, \sqrt{2}]$

| oprul extbfMétrica | L1 | L2 |
|--------------------|---------|-------|
| Dimensão | 91 × 91 | |
| Mínima | 0.199 | 0.047 |
| Máxima | 1.883 | 0.805 |
| Média | 0.751 | 0.224 |
| Mediana | 0.687 | 0.197 |
| Desvio Padrão | 0.312 | 0.121 |

✓ Interpretação

Distância **pequena** = curvas similares

Distância **grande** = dinâmicas diferentes

Distribuição das Distâncias L1 e L2 (2013)

output/tarefa3_distribuicao_distancias_2013.pdf

Pares Mais Sincronizados — Top 10

| oprul extbfMunicípio 1 | Município 2 | L1 |
|------------------------|----------------|-------|
| Niterói | Angra dos Reis | 0.199 |
| São Gonçalo | Angra dos Reis | 0.202 |
| Cabo Frio | Angra dos Reis | 0.206 |
| Volta Redonda | Angra dos Reis | 0.223 |
| Tanguá | Angra dos Reis | 0.230 |
| Pinheiral | Araruama | 0.261 |
| Rio de Janeiro | Angra dos Reis | 0.270 |
| Barra do Piraí | Angra dos Reis | 0.273 |
| Casimiro de Abreu | Angra dos Reis | 0.284 |
| Resende | Angra dos Reis | 0.301 |

Curvas dos Pares Mais Sincronizados

output/tarefa3_pares_similares_2013.pdf

Topological Data Analysis (TDA)

O que é um Complexo Simplicial?

Estrutura topológica que **generaliza grafos**, capturando relações de ordem superior entre dados.

| Dim. | Nome | Descrição |
|------|-----------|-------------------------------------|
| 0 | Vértice | Um município |
| 1 | Aresta | Par conectado ($d < \varepsilon$) |
| 2 | Triângulo | Trio completamente conectado |
| 3 | Tetraedro | Quatro todos conectados |

Regra de Conexão

$$\text{Aresta}(A, B) \iff d_{L1}(A, B) < \varepsilon$$

output/tarefa4_simplices_ilustracao

Ilustração dos tipos de simplices (0, 1, 2, 3-simplex).

output/tarefa4_analise_limiares_2013.pdf

Principal Component Analysis

Identifica as **direções de maior variância**:

$$Z = X \cdot W$$

onde W são os autovetores da covariância.

| PC | Var. Exp. | Cumulativa |
|-----|-----------|------------|
| PC1 | 35.2% | 35.2% |
| PC2 | 18.7% | 53.9% |
| PC3 | 9.4% | 63.3% |
| PC4 | 6.8% | 70.1% |
| PC5 | 5.1% | 75.2% |

output/tarefa5_pca_variancia.pdf

Variância explicada por componente principal (PCA).

Projeção PCA — Visualização 2D

output/kmapper_pca_2013.pdf

Algoritmos de Clusterização

🎯 K-Means

- Particiona em k grupos
- Minimiza variância intra-cluster
- Requer definir k a priori

Resultado: K=4 ótimo

🔍 DBSCAN

- Baseado em densidade
- Detecta outliers
- Formas arbitrárias

Resultado: 3 clusters + 8 outliers

🌳 Hierárquico

- Dendrograma
- Múltiplas resoluções
- Método Ward

Resultado: Corte em 4 clusters

| Algoritmo | Clusters | Silhouette | Calinski-Harabasz |
|-------------|-----------------|------------|-------------------|
| K-Means | 4 | 0.312 | 45.8 |
| DBSCAN | 3 (+8 outliers) | 0.287 | 38.2 |
| Hierárquico | 4 | 0.208 | 42.1 |

Perfis Epidêmicos por Cluster

`output/tarefa5_clusters_perfis.pdf`

O que é?

Implementação Python do algoritmo **Mapper** para TDA, gerando visualizações HTML interativas do espaço topológico dos dados.

Arquivos Gerados

- `kmapper_pca_2013.html`
- `kmapper_tsne_2013.html`
- `kmapper_l2norm_2013.html`
- `kmapper_distancia_2013.html`

`output/tarefa5_kmapper_grafo.pdf`

Representação esquemática do grafo Mapper. Números =

Principais Achados

Resultados Quantitativos

- **91 municípios** analisados
- **4 clusters** epidêmicos distintos
- **2013**: maior surto (185k+ casos)
- Silhouette Score: **0.312**
- 5 PCs capturam **75%** da variância

Insights Qualitativos

- Padrão sazonal **consistente**
- Picos: **janeiro–abril**
- Grupos **sincronizados** identificados
- Triângulos = corredores de transmissão
- Cluster 3: padrão **bimodal** único

Contribuição Metodológica

Demonstração pioneira do uso de **Topological Data Analysis** para epidemiologia da



Para Saúde Pública

- 1 Ações **coordenadas** entre municípios do mesmo cluster
- 2 Intensificar controle vetorial **pré-verão** (novembro–dezembro)
- 3 Alocar recursos conforme **perfil epidêmico**
- 4 Monitorar municípios do Cluster 3 (padrão atípico)



Trabalhos Futuros

- 1 Incluir dados **climáticos** (precipitação, temperatura)
- 2 Análise de **persistência** homológica
- 3 Modelos **preditivos** por cluster
- 4 Expandir para **outros estados**
- 5 Integrar dados de **mobilidade**

Obrigado!

⚙️ Análise Epidemiológica de Dengue



github.com/mei-the-dev/dengue



Python + NumPy + Pandas + Scikit-learn



LaTeX Beamer

Novembro 2025