Análise Epidemiológica de Dengue

Rio de Janeiro (2010-2016) — Topological Data Analysis

Análise Epidemiológica

Projeto de Ciência de Dados

Novembro 2025

Sumário

Objetivos do Estudo

Objetivos Principais

1. **Análise Temporal**: Padrões de semanas epidemiológicas
2. **Normalização**: Comparação ajustada por população
3. **Distâncias**: Identificar dinâmicas sincronizadas (L1/L2)
4. **Complexos Simpliciais**: Estrutura topológica
5. **Clusterização**: Agrupamento de municípios

Técnica Principal

Topological Data Analysis (TDA)

Revelando estruturas ocultas nos dados epidemiológicos através de complexos simpliciais e análise de forma.

Fonte dos Dados

Características do Dataset

- **Arquivo**: `Dengue_Brasil_2010-2016_RJ.xlsx`
- **Período**: 2010 a 2016 (7 anos)
- **Região**: Estado do Rio de Janeiro
- **Granularidade**: Casos por município e semana epidemiológica
- **Municípios**: 91 municípios analisados

Ano de Referência: 2013

Selecionado por apresentar:

- Maior número de casos no período
- Dados completos (52 semanas)
- Dinâmica epidêmica bem definida

Semanas Epidemiológicas

Definição

A **semana epidemiológica** (SE) é a unidade de tempo padrão da OMS/CDC para vigilância epidemiológica.

- Maioria dos anos: **52 semanas**
- Anos especiais: **53 semanas** (ex: 2014)
- SE 1 inicia no primeiro domingo $\geq$ 1º janeiro

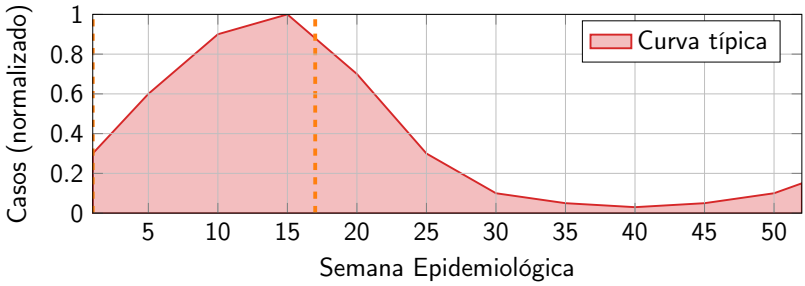
Ano	Semanas
2010	52
2011	52
2012	52
2013	52
2014	**53**
2015	52
2016	32*

Tabela: *Dados incompletos

Padrão Sazonal

Observações do Período Epidêmico

- **Picos epidêmicos**: Semanas 1–17 (janeiro a abril)
- **Padrão consistente**: Aumento no verão, queda no inverno
- **Maior surto**: 2013
- Correlação com período de chuvas e temperaturas elevadas



Métodos de Normalização

Normalização 1: Taxa de Incidência

Casos ajustados pela população (Censo 2010):

$$\text{Taxa} = \frac{\text{Casos}}{\text{População}} \times 100.000$$

Objetivo: Comparar intensidade da epidemia entre municípios de diferentes tamanhos.

Normalização 2: Área Unitária

Série temporal normalizada para soma = 1:

$$\tilde{x}_i = \frac{x_i}{\sum_j x_j}$$

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

Importância

A normalização por área unitária permite identificar municípios com **dinâmicas sincronizadas**, mesmo com números absolutos muito diferentes.

Taxa de Incidência — Resultados

Por Taxa (100.000 hab.)

Municípios com **maior risco relativo**:

1. Municípios pequenos
2. Alta densidade vetorial
3. Infraestrutura precária

Por Casos Absolutos

Municípios com **maior carga**:

1. Rio de Janeiro (capital)
2. Niterói
3. Grandes centros urbanos

Insight

Municípios pequenos podem ter taxas altíssimas com poucos casos absolutos — importante para políticas de saúde pública diferenciadas.

Distâncias L1 e L2

Para identificar municípios com dinâmicas **sincronizadas**, calculamos distâncias entre curvas normalizadas:

Distância L1 (Manhattan)

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

- Mais **robusta** a outliers
- Soma das diferenças absolutas
- Interpretação: "quanto difere em total"

Distância L2 (Euclidiana)

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

- **Penaliza** grandes diferenças
- Sensível a picos isolados
- Interpretação: "distância geométrica"

Matrizes de Distância — Estatísticas

Estatística	L1 (Manhattan)	L2 (Euclidiana)
Dimensões	91×91	91×91
Mínima (não-zero)	≈ 0.20	≈ 0.05
Máxima	≈ 1.88	≈ 0.80
Média	≈ 0.75	≈ 0.22

Tabela: Resumo das matrizes de distância (ano 2013)

Interpretação

- **Distância pequena** $\rightarrow$ curvas epidêmicas similares
- **Distância grande** $\rightarrow$ dinâmicas diferentes
- Matrizes simétricas: $d(A, B) = d(B, A)$

Pares Mais Sincronizados

Top Municípios com Dinâmicas Similares (L1)

Municípios conectados pela epidemia:

- Região metropolitana do Rio
- Municípios vizinhos geograficamente
- Padrões de mobilidade populacional

Aplicação

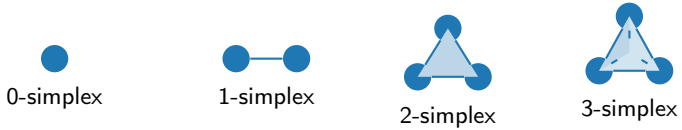
Identificar pares sincronizados permite:

- Ações coordenadas de controle vetorial
- Compartilhamento de recursos
- Modelos de propagação espacial

O que é um Complexo Simplicial?

Estrutura topológica que generaliza grafos, revelando conexões de ordem superior:

Dim.	Nome	Descrição
0	Vértice	Um ponto (município)
1	Aresta	Conexão entre 2 municípios
2	Triângulo	Trio completamente conectado
3	Tetraedro	Quatro municípios conectados



Construção do Complexo

Método

Dois municípios são **conectados** se sua distância é menor que um **limiar** ε:

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

ε pequeno

- Poucos vértices conectados
- Revela **núcleos** mais sincronizados
- Estrutura esparsa

ε grande

- Muitas conexões
- Complexo mais **denso**
- Pode perder informação local

Estratégia: Testar múltiplos limiares (percentis 10%, 20%, ..., 90% da distribuição de distâncias).

Resultados — Complexo Simplicial

Estruturas Identificadas

Ao variar o limiar ε, identificamos:

- **Componentes conexas**: Grupos de municípios sincronizados
- **Triângulos**: Trios com dinâmicas fortemente correlacionadas
- **Clusters topológicos**: Agrupamentos naturais da epidemia

Interpretação Epidemiológica

- Triângulos indicam **corredores de transmissão**
- Componentes isoladas sugerem **dinâmicas independentes**
- Evolução do complexo com ε revela **hierarquia de similaridade**

PCA — Redução Dimensional

Principal Component Analysis

Técnica que identifica as **direções de maior variância** nos dados:

$$Z = X \cdot W$$

onde W contém os autovetores da matriz de covariância.

Aplicação

- Reduzir 52 dimensões (semanas) para 2-3
- Visualizar municípios em espaço 2D
- Identificar padrões principais

Variância Explicada

- PC1: $\sim$30-40% da variância
- PC1 + PC2: $\sim$50-60%
- 5 componentes: $>$80%

Interpretação dos Componentes

Componentes Principais

- PC1 Intensidade geral da epidemia (pico vs. vale)
- PC2 Timing do pico (início vs. fim do período)
- PC3 Forma da curva (unimodal vs. bimodal)

Projeção 2D

No espaço PC1 $\times$ PC2:

- Municípios **próximos** $\rightarrow$ curvas similares
- **Clusters visuais** emergem naturalmente
- Outliers identificados facilmente

Métodos de Clusterização

K-Means

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

DBSCAN

- Baseado em densidade
- Detecta outliers
- Formas arbitrárias

Hierárquico

- Dendrograma
- Múltiplas resoluções
- Interpretável

Métricas de Validação

- **Silhouette Score**: Coesão vs. separação (-1 a 1)
- **Calinski-Harabasz**: Razão de variâncias (maior = melhor)

Resultados — Clusters Epidêmicos

Perfis Identificados

Os clusters revelam diferentes **padrões epidêmicos**:

- Cluster 0 Pico precoce (semanas 5-10), alta intensidade
- Cluster 1 Pico tardio (semanas 12-17), moderado
- Cluster 2 Baixa intensidade, curva achatada
- Cluster 3 Padrão bimodal (dois picos)

Implicações

- Municípios no mesmo cluster requerem **timing similar** de intervenções
- Recursos podem ser **compartilhados** entre municípios do mesmo cluster
- Modelos preditivos específicos por cluster

KeplerMapper — Visualização Interativa

O que é?

Implementação Python do algoritmo **Mapper** para TDA, gerando visualizações HTML interativas.

Arquivos Gerados

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

Funcionalidades

- Zoom e pan interativos
- Hover para ver municípios
- Cores por atributo
- Exportável

🌐 **Abra os arquivos HTML no navegador para explorar!**

Principais Achados

Resultados

1. **2013** foi o ano com maior surto epidêmico
2. **Padrão sazonal** consistente: picos em janeiro-abril
3. **Grupos de municípios** com dinâmicas sincronizadas identificados
4. **Complexos simpliciais** revelam estrutura topológica da epidemia
5. **Clusters** com perfis epidêmicos distintos

Contribuição Metodológica

Demonstração do uso de **Topological Data Analysis** para análise epidemiológica, indo além de métodos estatísticos tradicionais.

Recomendações

Para Saúde Pública

- Ações coordenadas entre municípios sincronizados
- Intensificar controle vetorial pré-verão
- Alocar recursos por cluster epidêmico

Trabalhos Futuros

- Incluir dados climáticos
- Análise de persistência
- Modelos preditivos por cluster
- Comparação com outros estados

Estrutura do Projeto

Módulos Python

`tarefa0_carregar_dados.py` Carregamento e preparação dos dados

`tarefa1_semanas_epidemiologicas.py` Calendário epidemiológico

`tarefa2_normalizacao.py` Normalização por população e área

`tarefa3_distancias.py` Cálculo de distâncias L1/L2

`tarefa4_complexo_simplicial.py` Construção de complexos

`tarefa5_kmapper.py` Visualizações KeplerMapper

Repositório

`https://github.com/mei-the-dev/dengue`

Obrigado!

Análise Epidemiológica de Dengue

Dúvidas e sugestões são bem-vindas

https://github.com/mei-the-dev/dengue