



**UNIVERSIDADE FEDERAL DO CEARÁ**  
**FACULDADE DE MEDICINA**  
**DEPARTAMENTO DE MEDICINA CLÍNICA**  
**PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS MÉDICAS**  
**DOUTORADO EM CIÊNCIAS MÉDICAS**

**RÔNEY REIS DE CASTRO E SILVA**

**MODELAGEM INTEGRADA DE REDES MULTILAYER, ANÁLISES ESPACIAIS E  
APRENDIZADO DE MÁQUINA NA INVESTIGAÇÃO DA DISSEMINAÇÃO DE  
COVID-19 E DE ARBOVIROSES (DENGUE, ZIKA E CHIKUNGUNYA) EM  
FORTALEZA**

**FORTALEZA**

**2025**

RÔNEY REIS DE CASTRO E SILVA

MODELAGEM INTEGRADA DE REDES MULTILAYER, ANÁLISES ESPACIAIS E  
APRENDIZADO DE MÁQUINA NA INVESTIGAÇÃO DA DISSEMINAÇÃO DE COVID-19  
E DE ARBOVIROSES (DENGUE, ZIKA E CHIKUNGUNYA) EM FORTALEZA

Tese apresentada ao Programa de Pós-Graduação em Ciências Médicas da Faculdade de Medicina da Universidade Federal do Ceará, como requisito parcial à obtenção do título de doutor em Ciências Médicas. Área de Concentração: Medicina

Orientador: Prof. Dr. Miguel Ângelo Nobre e Souza

Coorientador: Prof. Dr. Ângelo Roncalli Alencar Brayner

FORTALEZA

2025

Dados Internacionais de Catalogação na Publicação  
Universidade Federal do Ceará  
Sistema de Bibliotecas  
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

---

- R312m Reis de Castro e Silva, Rôney.  
Modelagem Integrada de Redes Multilayer, Análises Espaciais e Aprendizado de Máquina na Investigação da Disseminação de COVID-19 e de Arboviroses (Dengue, Zika e Chikungunya) em Fortaleza / Rôney Reis de Castro e Silva. – 2025.  
117 f. : il. color.
- Tese (doutorado) – Universidade Federal do Ceará, Faculdade de Medicina, Programa de Pós-Graduação em Ciências Médicas, Fortaleza, 2025.  
Orientação: Prof. Dr. Miguel Ângelo Nobre e Souza.  
Coorientação: Prof. Dr. Ângelo Roncalli Alencar Brayner.
1. COVID-19. 2. Arboviroses. 3. Redes multilayer. 4. Análise espacial. 5. Aprendizado de máquina.  
I. Título.
-

RÔNEY REIS DE CASTRO E SILVA

MODELAGEM INTEGRADA DE REDES MULTILAYER, ANÁLISES ESPACIAIS E  
APRENDIZADO DE MÁQUINA NA INVESTIGAÇÃO DA DISSEMINAÇÃO DE COVID-19  
E DE ARBOVIROSES (DENGUE, ZIKA E CHIKUNGUNYA) EM FORTALEZA

Tese apresentada ao Programa de Pós-Graduação em Ciências Médicas da Faculdade de Medicina da Universidade Federal do Ceará, como requisito parcial à obtenção do título de doutor em Ciências Médicas. Área de Concentração: Medicina

Aprovada em:

BANCA EXAMINADORA

---

Prof. Dr. Miguel Ângelo Nobre e  
Souza (Orientador)  
Universidade Federal do Ceará (UFC)

---

Prof. Dr. Ângelo Roncalli Alencar  
Brayner (Coorientador)  
Universidade Federal do Ceará (UFC)

---

Prof. Dr. Ronaldo Menezes  
University of Exeter (UK)

---

Prof. Dr. Marcelo Alcantara Holanda  
Escola de Saúde Pública do Ceará

---

Prof. Dr. Geziel dos santos de Sousa  
Universidade Estadual do Ceará

À minha esposa, pelo amor, parceria e paciência inabaláveis. Aos meus pais, por todo cuidado, exemplo e apoio.

## **AGRADECIMENTOS**

Agradeço, em primeiro lugar, à minha esposa Ana Luisa Viana Barroso Bacelar, pelo amor, paciência e apoio incondicional ao longo desta jornada. Sua compreensão e incentivo foram essenciais para que eu pudesse me dedicar plenamente a esta pesquisa. Aos meus pais e irmãos, por sempre estarem ao meu lado, oferecendo suporte inestimável e me inspirando com seus exemplos de dedicação e resiliência.

Ao meu orientador, Prof. Dr. Miguel Ângelo Nobre e Souza, e ao meu coorientador, Prof. Dr. Ângelo Roncalli Alencar Brayner, expresso minha profunda gratidão pela orientação, pelo incentivo constante e pelo comprometimento com minha formação acadêmica. Suas contribuições foram fundamentais para a construção deste trabalho e para o meu crescimento como pesquisador.

Ao Prof. Dr. Ronaldo Menezes, agradeço pelo apoio e pela oportunidade de colaboração, além da recepção na Universidade de Exeter, um momento enriquecedor para minha trajetória acadêmica.

Agradeço, ainda, à CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, pelo financiamento da pesquisa de doutorado por meio da bolsa de estudos, possibilitando a realização deste trabalho e contribuindo para o avanço do conhecimento científico.

A todas as pessoas que, direta ou indiretamente, fizeram parte desta caminhada, minha sincera gratidão. Cada palavra de incentivo, cada troca de conhecimento e cada desafio superado contribuíram para que este momento fosse alcançado.

“O sonho é que leva a gente para frente. Se a gente for seguir a razão, fica aquietado, acomodado.”

(Ariano Suassuna)

## RESUMO

O avanço das arboviroses — e, em especial, da COVID-19 — demanda abordagens capazes de integrar múltiplas dimensões territoriais. Esta tese adota uma abordagem integrada com redes multilayer, análises estatísticas e espaciais e modelos de aprendizado de máquina interpretável para compreender padrões de difusão de COVID-19, dengue, zika e chikungunya em Fortaleza (CE), 2020–2024. As camadas incluem variáveis epidemiológicas, socioeconômicas/demográficas, de infraestrutura e políticas/IDH; as relações intracamadas são definidas por similaridade entre bairros e as intercamadas por associações entre variáveis. A base de COVID-19 foi construída a partir de Boletins Epidemiológicos semanais, agregados por mês e por bairro, priorizando padrões territoriais comparáveis entre doenças. No plano espacial, utilizamos autocorrelação global e local (Moran's I, LISA) para identificar clusters e padrões locais; no plano de redes, calculamos centralidades e aplicamos detecção de comunidades. Modelos preditivos (Random Forest, XGBoost) e interpretações com SHAP avaliaram o poder explicativo das variáveis e suas contribuições. Agrupamentos via K-means sintetizaram perfis territoriais ao combinar características estruturais e epidemiológicas. Os achados mostram heterogeneidade espacial e diferenças entre camadas, com convergências e divergências territoriais úteis para leitura de vulnerabilidades e prioridades. Como contribuição prática, o trabalho apresenta, em anexo, uma Proposta de Aprimoramento Metodológico (“Framework Integrado”) para orientar padronizações futuras e apoiar a tomada de decisão municipal.

**Palavras-chave:** COVID-19, Arboviroses, Redes multilayer, Análise espacial, Aprendizado de máquina

## ABSTRACT

The rise of arboviral diseases—and, in particular, COVID-19—calls for approaches that integrate multiple territorial dimensions. This thesis adopts an integrated approach combining multilayer networks, statistical and spatial analyses, and interpretable machine learning to understand diffusion patterns of COVID-19, dengue, zika, and chikungunya in Fortaleza (CE), 2020–2024. Layers comprise epidemiological, socioeconomic/demographic, infrastructure, and political/HDI variables; intralayer relations are defined by neighborhood similarity, and interlayer relations by associations between variables. The COVID-19 dataset was built from weekly epidemiological bulletins, aggregated by month and neighborhood, prioritizing territorial patterns comparable across diseases. Spatially, we use global and local autocorrelation (Moran’s I, LISA) to identify clusters and local patterns; in the network domain, we compute centralities and apply community detection. Predictive models (Random Forest, XGBoost) and SHAP-based explanations assess the variables’ predictive power and contributions. K-means clustering synthesizes territorial profiles by combining structural and epidemiological characteristics. Findings reveal spatial heterogeneity and cross-layer differences, with territorial convergences and divergences that inform vulnerability assessment and prioritization. As a practical contribution, the thesis includes an Appendix: Methodological Enhancement Proposal (“*Integrated Framework*”) to guide future standardization and support municipal decision-making.

**Keywords:** COVID-19, Arboviruses, Multilayer Networks, Spatial Analysis, Machine Learning

## LISTA DE ILUSTRAÇÕES

Figura 1 – Ilustração de LISA. . . . .	31
Figura 2 – Representação esquemática de uma rede <i>Dual-Layer</i> . . . . .	32
Figura 3 – Representação esquemática do funcionamento do algoritmo Random Forest. . . . .	36
Figura 4 – Ilustração do funcionamento do algoritmo K-means. . . . .	39
Figura 5 – Mapa dos 121 bairros de Fortaleza com identificadores numéricos. . . . .	60
Figura 6 – Diagramas de dispersão com as correlações altas. . . . .	62
Figura 7 – Mapas Comparativos de <i>clusters</i> Espaciais (LISA). . . . .	67
Figura 8 – Síntese da frequência de <i>hotspots</i> (Alto-Alto) por bairro nas análises LISA para as quatro doenças. . . . .	69
Figura 9 – Comunidades multicamadas detectadas (Louvain). . . . .	71
Figura 10 – Importância das variáveis (SHAP) nos modelos XGBoost para COVID-19, dengue, zika e chikungunya. . . . .	78
Figura 11 – Importância das variáveis (SHAP) nos modelos Random Forest para COVID-19, dengue, zika e chikungunya. . . . .	80
Figura 12 – Método do cotovelo aplicado à clusterização dos bairros de Fortaleza. . . . .	82
Figura 13 – Distribuição espacial dos <i>clusters</i> territoriais identificados pelo algoritmo K-means. . . . .	84

## LISTA DE TABELAS

Tabela 1 – Resumo dos clusters territoriais gerados via K-Means . . . . .	46
Tabela 2 – Desempenho dos modelos XGBoost na predição do risco relativo (RR) por doença. . . . .	76
Tabela 3 – Desempenho dos modelos Random Forest na predição do risco relativo (RR) por doença. . . . .	79
Tabela 4 – Resumo dos clusters territoriais gerados via K-means . . . . .	85
Tabela 5 – Resumo dos módulos do arranjo metodológico integrado (proposta futura) .	111

## LISTA DE ABREVIATURAS E SIGLAS

CAPS	Centro de Atenção Psicossocial
CEVEPI	Célula de Vigilância Epidemiológica
CORES	Coordenadorias Regionais de Saúde
DAG	Directed Acyclic Graph
IBGE	Instituto Brasileiro de Geografia e Estatística
IDH	Índice de Desenvolvimento Humano
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
LISA	<i>Local Indicators of Spatial Association</i>
ML	<i>machine learning</i> / aprendizado de máquina
PMF	Prefeitura Municipal de Fortaleza
PNUD	Programa das Nações Unidas para o Desenvolvimento
RF	Random Forest
RR	Risco Relativo
SESA	Secretaria da Saúde do Estado do Ceará
SHAP	<i>SHapley Additive exPlanations</i>
SMS	Secretaria Municipal da Saúde
TRE	Tribunal Regional Eleitoral
TSE	Tribunal Superior Eleitoral
UAPS	Unidade de Atenção Primária à Saúde
UPA	Unidade de Pronto Atendimento
XGBoost	Extreme Gradient Boosting

## LISTA DE SÍMBOLOS

$G = (V, E)$	Grafo (rede) com conjunto de nós $V$ e arestas $E$
$V$	Conjunto de nós (bairros)
$E$	Conjunto de arestas
$L$	Conjunto de camadas (multilayer)
$A^{(\ell)}$	Matriz de adjacência da camada $\ell$
$W$	Matriz de pesos espaciais (contiguidade entre bairros)
$w_{ij}$	Peso espacial entre os bairros $i$ e $j$
$d_{ij}$	Distância (ou dissimilaridade) entre os bairros $i$ e $j$
$\rho_s$	Correlação de Spearman
$r$	Correlação de Pearson
$RR$	Risco relativo
$I$	Índice de autocorrelação espacial de Moran (global)
LISA	Indicadores Locais de Associação Espacial (Moran local)
$Q$	Modularidade (particionamento por Louvain)
$k$	Número de <i>clusters</i> /grupos (ex.: K-means)
$\mathbf{X}$	Matriz de variáveis explicativas (atributos por bairro)
$\mathbf{y}$	Variável resposta (ex.: casos ou $RR$ )
$\beta$	Vetor de coeficientes de modelo
$\text{deg}(i)$	Grau (ponderado) do nó $i$
$C_{\text{eig}}(i)$	Centralidade de autovetor do nó $i$
$C_{\text{bet}}(i)$	Centralidade de intermediação ( <i>betweenness</i> ) do nó $i$
$t$	Tempo (mês/ano de referência)

## SUMÁRIO

<b>1</b>	<b>INTRODUÇÃO</b>	<b>19</b>
<b>1.1</b>	<b>Objetivos</b>	<b>21</b>
<i>1.1.1</i>	<i>Geral</i>	<i>21</i>
<i>1.1.2</i>	<i>Específicos</i>	<i>22</i>
<i>1.1.3</i>	<i>Contribuições Esperadas</i>	<i>22</i>
<i>1.1.4</i>	<i>Justificativa</i>	<i>22</i>
<i>1.1.5</i>	<i>Hipóteses</i>	<i>23</i>
<i>1.1.6</i>	<i>Estrutura da Tese</i>	<i>23</i>
<b>1.2</b>	<b>Resultados Prévios</b>	<b>24</b>
<b>2</b>	<b>FUNDAMENTAÇÃO TEÓRICA</b>	<b>26</b>
<b>2.1</b>	<b>Saúde urbana e determinantes territoriais</b>	<b>26</b>
<i>2.1.1</i>	<i>Mecanismos: exposição, suscetibilidade e capacidade de resposta</i>	<i>26</i>
<i>2.1.2</i>	<i>Implicações territoriais e orientação metodológica</i>	<i>26</i>
<i>2.1.3</i>	<i>Especificidades das Doenças Abordadas (COVID-19, dengue, zika e chikungunya)</i>	<i>27</i>
<b>2.2</b>	<b>Análise Estatística</b>	<b>28</b>
<i>2.2.1</i>	<i>Correlação de Pearson</i>	<i>28</i>
<i>2.2.2</i>	<i>Regressão Linear Múltipla (OLS)</i>	<i>28</i>
<b>2.3</b>	<b>Análise Espacial</b>	<b>29</b>
<i>2.3.1</i>	<i>Conceito de Autocorrelação Espacial</i>	<i>29</i>
<i>2.3.2</i>	<i>Moran's I e LISA (Local Indicators of Spatial Association)</i>	<i>29</i>
<b>2.4</b>	<b>Redes Multilayer</b>	<b>31</b>
<i>2.4.1</i>	<i>Fundamentos de Redes: nós, arestas e métricas estruturais</i>	<i>31</i>
<i>2.4.2</i>	<i>Conceitos e formalização multilayer</i>	<i>32</i>
<i>2.4.3</i>	<i>Estrutura multiplex e justificativa</i>	<i>33</i>
<i>2.4.4</i>	<i>Métricas multilayer e interpretação substantiva</i>	<i>34</i>
<i>2.4.5</i>	<i>Detecção de comunidades: por camada e na perspectiva multilayer</i>	<i>34</i>
<i>2.4.6</i>	<i>Boas práticas e cautelas analíticas</i>	<i>35</i>
<b>2.5</b>	<b>Aprendizado de Máquina</b>	<b>35</b>
<i>2.5.1</i>	<i>Modelos Baseados em Árvores: Random Forest e XGBoost</i>	<i>35</i>

2.5.2	<i>Explicabilidade de Modelos: SHAP (SHapley Additive exPlanations)</i> . . . . .	37
2.5.3	<i>Limitações e Vantagens da Abordagem Preditiva em Epidemiologia</i> . . . . .	37
2.6	<b>Agrupamento Espacial</b> . . . . .	38
2.6.1	<i>Fundamentos do Algoritmo K-means</i> . . . . .	38
3	<b>TRABALHOS RELACIONADOS</b> . . . . .	40
3.1	<b>Aprendizado de Máquina na Saúde Pública</b> . . . . .	40
3.1.1	<i>Perspectivas Nacionais</i> . . . . .	41
3.2	<b>Análise Espacial e Clusterização na Epidemiologia</b> . . . . .	43
3.3	<b>Redes Multilayer na Epidemiologia: Trabalhos Relacionados e Avanços Propostos</b> . . . . .	44
3.4	<b>Síntese Comparativa dos Trabalhos Relacionados</b> . . . . .	45
4	<b>METODOLOGIA</b> . . . . .	47
4.1	<b>Pré-processamento dos Dados</b> . . . . .	47
4.1.0.0.1	<i>Organização e créditos institucionais.</i> . . . . .	48
4.1.0.0.2	<i>Observações.</i> . . . . .	48
4.1.0.1	<i>Considerações Éticas</i> . . . . .	48
4.1.1	<b>Variáveis selecionadas, limitações e justificativas</b> . . . . .	49
4.1.1.0.1	<i>Variáveis utilizadas.</i> . . . . .	49
4.1.1.0.2	<i>Por que estas variáveis.</i> . . . . .	49
4.1.1.0.3	<i>Limitações e decisões de escopo.</i> . . . . .	50
4.1.2	<b>Preparação e Integração dos Dados</b> . . . . .	50
4.1.2.1	<i>Limpeza e Integração dos Dados</i> . . . . .	50
4.1.2.2	<i>Análise exploratória inicial e seleção de variáveis</i> . . . . .	51
4.1.2.2.1	<i>Regras aplicadas.</i> . . . . .	51
4.1.2.2.2	<i>Checagens exploratórias (enxutas).</i> . . . . .	51
4.1.2.3	<i>Categorização das Variáveis Políticas</i> . . . . .	52
4.1.2.4	<i>Escolha da Métrica de Risco Epidemiológico</i> . . . . .	53
4.2	<b>Análise Estatística</b> . . . . .	53
4.2.1	<b>Correlação de Pearson</b> . . . . .	54
4.2.2	<b>Regressão Linear Múltipla (OLS)</b> . . . . .	54
4.3	<b>Análise Espacial</b> . . . . .	54
4.3.1	<b>Autocorrelação Espacial Global (Índice de Moran)</b> . . . . .	55

4.3.2	<i>Autocorrelação Espacial Local (LISA)</i> . . . . .	55
4.3.3	<i>Análise de Hotspots Acumulados</i> . . . . .	55
4.3.4	<i>Considerações Técnicas</i> . . . . .	55
4.4	<b>Modelagem de Redes Multilayer</b> . . . . .	56
4.4.1	<i>Estrutura da rede</i> . . . . .	56
4.4.2	<i>Detecção de comunidades</i> . . . . .	57
4.4.3	<i>Análise de centralidade e estrutura multidimensional</i> . . . . .	57
4.5	<b>Modelos de Machine Learning para Predição e Clusterização</b> . . . . .	58
4.5.1	<i>Modelos Supervisionados para Predição do Risco Relativo</i> . . . . .	58
4.5.1.1	<i>Análise de Explicabilidade com SHAP</i> . . . . .	59
4.5.2	<i>Clusterização Territorial com K-means</i> . . . . .	59
5	<b>RESULTADOS</b> . . . . .	60
5.1	<b>Panorama Territorial dos Bairros de Fortaleza</b> . . . . .	60
5.2	<b>Análise Estatística</b> . . . . .	61
5.2.1	<i>Correlação de Pearson</i> . . . . .	61
5.2.2	<i>Modelagem por Regressão Linear (OLS)</i> . . . . .	64
5.3	<b>Análise Espacial</b> . . . . .	65
5.3.1	<i>Autocorrelação Espacial Global (Moran's I)</i> . . . . .	65
5.3.2	<i>Autocorrelação Espacial Local (LISA)</i> . . . . .	66
5.3.3	<i>Síntese dos Hotspots por Doença</i> . . . . .	68
5.4	<b>Redes Multilayer</b> . . . . .	70
5.4.1	<i>Comunidades por camada (Louvain)</i> . . . . .	74
5.4.1.0.1	<i>Camadas epidemiológicas (COVID-19, dengue, zika, chikungunya).</i> . . . . .	74
5.4.1.0.2	<i>Camada socioeconômica/demográfica.</i> . . . . .	74
5.4.1.0.3	<i>Camada de infraestrutura.</i> . . . . .	74
5.4.1.0.4	<i>Camadas políticas (municipal e presidencial) e IDH.</i> . . . . .	74
5.4.2	<i>Leitura multilayer: convergências e divergências</i> . . . . .	75
5.4.3	<i>Interpretação prática e apoio à decisão</i> . . . . .	75
5.4.4	<i>Síntese</i> . . . . .	75
5.5	<b>Machine Learning (ML)</b> . . . . .	76
5.5.1	<i>Aprendizado Supervisionado com XGBoost</i> . . . . .	76
5.5.1.1	<i>Análise de Importância das Variáveis via SHAP</i> . . . . .	77

5.5.2	<i>Aprendizado Supervisionado com Random Forest</i> . . . . .	79
5.5.2.1	<i>Análise de Importância das Variáveis via SHAP</i> . . . . .	80
5.5.2.2	<i>Comparativo XGBoost vs Random Forest</i> . . . . .	81
5.6	<b>Agrupamento através do algoritmo K-means</b> . . . . .	81
5.6.1	<i>Determinação do Número de Clusters</i> . . . . .	81
5.6.2	<i>Teste de Estabilidade dos Agrupamentos</i> . . . . .	82
5.6.3	<i>Síntese dos Resultados</i> . . . . .	83
6	<b>DISCUSSÃO</b> . . . . .	87
6.1	<b>Padrões de Risco Epidemiológico e Determinantes Urbanos</b> . . . . .	87
6.2	<b>Importância dos Fatores Socioeconômicos, Políticos e de Infraestrutura</b>	87
6.3	<b>Interpretação das Comunidades na Rede Multilayer</b> . . . . .	88
6.4	<b>Desempenho dos Modelos de Aprendizado de Máquina</b> . . . . .	89
6.5	<b>Agrupamentos Territoriais com K-means</b> . . . . .	89
6.6	<b>Implicações para Políticas Públicas e Planejamento Territorial</b> . . . . .	89
7	<b>CONCLUSÕES E TRABALHOS FUTUROS</b> . . . . .	91
7.1	<b>Conclusões</b> . . . . .	91
7.2	<b>Trabalhos Futuros</b> . . . . .	92
	<b>REFERÊNCIAS</b> . . . . .	93
	<b>ANEXOS</b> . . . . .	96
	<b>ANEXO A – Correlações Altas de Pearson</b> . . . . .	97
	<b>ANEXO B – Correlações Medianas de Pearson</b> . . . . .	98
	<b>ANEXO C – Matriz de Correlação de Pearson</b> . . . . .	100
	<b>ANEXO D – Distribuição dos clusters LISA por doença, com respectivos bairros e estatísticas de autocorrelação espacial (Moran’s I).</b> . . . . .	101
	<b>ANEXO E – Cálculo do Risco Relativo (RR)</b> . . . . .	102
	<b>ANEXO A – Proposta de Aprimoramento Metodológico (Framework Integrado)</b> . . . . .	103
A.1	<b>Curadoria e Integração de Dados (M1)</b> . . . . .	103
A.1.1	<i>Fontes dos Dados</i> . . . . .	103
A.1.2	<i>Seleção de Variáveis</i> . . . . .	104
A.1.2.1	<i>Categorização das Variáveis Políticas</i> . . . . .	104
A.1.3	<i>Limpeza e Integração dos Dados</i> . . . . .	105

<b>A.1.4</b>	<b><i>Análise Exploratória Inicial</i></b> . . . . .	105
<b>A.1.5</b>	<b><i>Métrica de Risco Epidemiológico (RR)</i></b> . . . . .	105
<b>A.1.6</b>	<b><i>Considerações Éticas</i></b> . . . . .	106
<b>A.2</b>	<b>Estatística Descritiva e Inferencial (M2)</b> . . . . .	106
<b>A.2.1</b>	<b><i>Correlação de Pearson</i></b> . . . . .	106
<b>A.2.2</b>	<b><i>Regressão Linear Múltipla (OLS)</i></b> . . . . .	107
<b>A.3</b>	<b>Dependência e Padrões Espaciais (M3)</b> . . . . .	107
<b>A.3.1</b>	<b><i>Autocorrelação Espacial Global (Moran's I)</i></b> . . . . .	107
<b>A.3.2</b>	<b><i>Autocorrelação Espacial Local (LISA)</i></b> . . . . .	108
<b>A.3.3</b>	<b><i>Hotspots Acumulados</i></b> . . . . .	108
<b>A.4</b>	<b>Rede Multilayer (M4)</b> . . . . .	108
<b>A.4.1</b>	<b><i>Estrutura da Rede</i></b> . . . . .	109
<b>A.4.2</b>	<b><i>Deteção de Comunidades</i></b> . . . . .	109
<b>A.4.3</b>	<b><i>Centralidade e Estrutura</i></b> . . . . .	109
<b>A.5</b>	<b>Modelagem Supervisionada e Tipificação (M5)</b> . . . . .	110
<b>A.5.1</b>	<b><i>Predição do RR (XGBoost/RF)</i></b> . . . . .	110
<b>A.5.2</b>	<b><i>Explicabilidade (SHAP)</i></b> . . . . .	110
<b>A.5.3</b>	<b><i>Tipificação Territorial (K-Means)</i></b> . . . . .	110
<b>A.6</b>	<b>Tabela-resumo do Framework (Proposta)</b> . . . . .	111
	<b>ANEXO B – Tabela com quantitativo de doenças</b> . . . . .	112

## 1 INTRODUÇÃO

Nas últimas décadas, as áreas urbanas vêm enfrentando desafios crescentes relacionados à propagação de doenças infecciosas emergentes e endêmicas, como COVID-19, dengue, zika e chikungunya. Este fenômeno é particularmente grave em países em desenvolvimento, onde fatores socioeconômicos, políticos, ambientais e estruturais frequentemente atuam como catalisadores para o aumento da vulnerabilidade populacional e a intensificação da dinâmica epidemiológica. A concentração populacional, a desigualdade territorial, a precariedade na infraestrutura de saneamento, o acesso desigual aos serviços de saúde e as desigualdades de participação e representação política moldam não apenas o risco de adoecimento, mas também a capacidade de resposta das populações afetadas. Neste contexto, compreender profundamente a complexidade dos processos que regem a difusão de doenças infecciosas em ambientes urbanos tornou-se fundamental não apenas para a ciência, mas também para subsidiar intervenções eficazes, ações intersetoriais e políticas públicas territorializadas.

Do ponto de vista da saúde pública, este trabalho usa informações que os municípios já registram para entender o que acontece em cada bairro. Trabalhamos com contagens mensais de casos (e, inicialmente, óbitos, que não entraram na modelagem final) de COVID-19, dengue, zika e chikungunya, além de dados demográficos (população total, densidade e número de domicílios), presença de equipamentos urbanos (hospitais, clínicas, escolas, universidades, mercados, feiras, centros culturais e esportivos), resultados de votação municipal e presidencial (2020 e 2022) e índices como o Índice de Desenvolvimento Humano (IDH). Com essas camadas de dados, buscamos responder perguntas simples e úteis para a gestão: onde o risco é maior, que características dos bairros estão mais associadas a mais casos e como priorizar ações no território.

Embora a literatura já traga avanços importantes em análises isoladas — como regressões espaciais, métodos preditivos e estudos de redes —, poucos trabalhos propõem uma integração metodológica abrangente que una essas dimensões sob um mesmo arcabouço teórico-computacional. A originalidade desta tese está na consolidação e validação de uma *abordagem integrada* multidimensional que combina ciência de redes, análise espacial e aprendizado de máquina, bem como na apresentação, Anexo A, de uma *Proposta de Aprimoramento Metodológico* (doravante, “*Framework Integrado*”), que sistematiza e generaliza os procedimentos aqui empregados. Ressalta-se que o *Framework Integrado* não foi implementado como produto final nesta tese e sim apresentado como diretriz metodológica complementar e passível de adoção

futura.

Uma das abordagens com potencial para este estudo é o uso de *redes multilayer* (múltiplas camadas). Tecnicamente, a rede adotada nesta tese é do tipo *multiplex*, pois o mesmo conjunto de nós — os bairros — aparece repetidamente em diferentes camadas (epidemiológica, socioeconômica/demográfica, de infraestrutura, política e IDH), e as conexões *intercamadas* ligam a mesma réplica de bairro entre essas camadas (com pesos derivados de associações entre variáveis). Para manter o texto simples, usamos o termo genérico *multilayer* ao longo da tese, preservando aqui a caracterização de que se trata de uma rede *multiplex*. Essa modelagem evita perder relações cruzadas que seriam ocultadas em um grafo único agregado.

A combinação sinérgica considerada ao longo deste trabalho contempla:

- Modelagem de redes *multilayer*, com camadas interligadas por métricas de associação entre variáveis;
- Técnicas de análise espacial, incluindo autocorrelação global e local (Moran's I, LISA) e agrupamentos espaciais via K-means;
- Algoritmos de aprendizado de máquina interpretável Random Forest (RF) Forest, Extreme Gradient Boosting (XGBoost), *SHapley Additive exPlanations* (SHAP), voltados à predição de risco e à identificação dos fatores com maior influência na propagação das doenças.

Este estudo empírico é centrado na cidade de Fortaleza, capital do estado do Ceará, Brasil — uma metrópole caracterizada por fortes desigualdades socioespaciais, alta densidade demográfica e recorrência de surtos infecciosos. A cidade conta com uma base sólida de dados georreferenciados, incluindo indicadores de infraestrutura, saúde, desenvolvimento humano, demografia e comportamento eleitoral, o que a torna um território ideal para a aplicação da abordagem proposta. Com 121 bairros e uma configuração urbana heterogênea, Fortaleza exemplifica os desafios enfrentados pelas grandes cidades latino-americanas no enfrentamento de epidemias.

**Intenção aplicada.** A intenção inicial desta pesquisa foi criar uma estrutura que auxilie os municípios na tomada de decisões, com foco em priorização territorial e alocação eficiente de recursos. A presente versão concentra-se na avaliação analítica e na demonstração de utilidade da abordagem; a formalização dessa estrutura em um artefato metodológico está sintetizada na Proposta de Aprimoramento Metodológico no Anexo A.

**Escopo e limitações.** Embora todos os testes tenham sido conduzidos também com dados de zika, a baixa disponibilidade e a reduzida série histórica para essa doença implica-

ram pouca potência analítica. Por esse motivo, resultados e subseções específicas sobre zika foram suprimidos ou não detalhados em diversas partes da tese, priorizando-se transparência metodológica e robustez estatística.

**Sobre as ondas de COVID-19.** Nesta tese optamos por não segmentar as análises por “ondas” de COVID-19. A justificativa é dupla: (i) o foco central é a compreensão de *padrões territoriais* e determinantes estruturais (socioeconômicos, de infraestrutura e político-territoriais), cuja identificação tende a ser mais estável quando se prioriza a variação espacial agregada em detrimento da cronologia das flutuações epidêmicas; e (ii) a delimitação de ondas depende de escolhas metodológicas sensíveis a mudanças exógenas (variantes, política de testagem/vigilância, retrospectiva de notificação e intervenções não farmacológicas), o que introduz heterogeneidade temporal difícil de comparar entre doenças com sazonalidades e dinâmicas distintas (como dengue, zika e chikungunya). Assim, privilegiamos janelas e agregações que aumentam a comparabilidade entre camadas e doenças, reduzindo o risco de confundir oscilações temporais de curto prazo com sinais estruturais de vulnerabilidade espacial.

Diante desse cenário, o objetivo geral desta tese é desenvolver e validar uma abordagem metodológica integrada que permita compreender, caracterizar e prever os padrões de difusão de doenças infecciosas em áreas urbanas complexas. Além disso, busca-se gerar subsídios analíticos concretos para orientar intervenções mais precisas, intersetoriais e territorialmente sensíveis. Espera-se que os resultados contribuam não apenas para o avanço científico na interface entre epidemiologia urbana e ciência de redes, mas também para o fortalecimento da capacidade de gestão pública em contextos vulneráveis e desiguais.

## 1.1 Objetivos

### 1.1.1 Geral

Desenvolver e validar uma *abordagem integrada* que combine redes *multilayer*, análise espacial e modelos de aprendizado de máquina para caracterizar e prever a difusão de COVID-19, dengue, zika e chikungunya em Fortaleza; e apresentar, em anexo, uma *Proposta de Aprimoramento Metodológico* (Framework Integrado) que sistematiza essa abordagem para uso futuro.

### 1.1.2 *Específicos*

- Modelar redes *multilayer* incorporando dados epidemiológicos, socioeconômicos, políticos e de infraestrutura;
- Avaliar padrões espaciais e *clusters* de risco utilizando técnicas como Moran's I, *Local Indicators of Spatial Association* (LISA) e K-means;
- Construir modelos preditivos interpretáveis (Random Forest, XGBoost, SHAP) para identificar bairros prioritários e variáveis mais relevantes na propagação das doenças.

### 1.1.3 *Contribuições Esperadas*

- **Metodológica:** Consolidação de uma abordagem integrada, robusta e replicável para análise epidemiológica urbana; e disponibilização, em anexo, de diretrizes (*Framework Integrado*) para sua padronização e futura adoção.
- **Teórica:** Avanço no uso de redes *multilayer* e aprendizado de máquina para compreensão de dinâmicas espaciais de doenças.
- **Aplicada:** Geração de evidências e recomendações práticas para políticas públicas de saúde orientadas ao território, com foco em apoio à tomada de decisão municipal.

### 1.1.4 *Justificativa*

O crescente desafio das doenças infecciosas em áreas urbanas brasileiras exige abordagens analíticas capazes de lidar com a complexidade dos fatores que influenciam sua propagação. Em cidades como Fortaleza, a combinação de desigualdades sociais, infraestrutura urbana precária e variações político-territoriais cria um cenário propício para a ocorrência e perpetuação de epidemias como COVID-19, dengue, zika e chikungunya (BRAGA *et al.*, 2020; MACCORMACK-GELLES *et al.*, 2018).

Apesar dos avanços nas análises espaciais e no uso de inteligência artificial aplicadas à saúde pública, ainda há uma lacuna significativa na integração de múltiplas dimensões (epidemiológica, socioeconômica, política e estrutural) em um modelo único e coerente (YE *et al.*, 2025; BEDSON *et al.*, 2021). As redes *multilayer*, aliadas a técnicas de aprendizado de máquina interpretável e análise espacial, oferecem um caminho promissor para representar de forma mais realista as interações que moldam a dinâmica urbana da saúde.

Esta tese se justifica, portanto, pela necessidade de desenvolver ferramentas meto-

dológicas que permitam compreender e antecipar padrões espaciais de risco, identificar áreas prioritárias e subsidiar ações públicas com base em evidências territoriais. Além disso, os resultados podem ser extrapolados para outras cidades latino-americanas com contextos similares, reforçando a relevância prática e científica da proposta.

### 1.1.5 Hipóteses

Com base nos objetivos e no referencial teórico, esta tese parte das seguintes hipóteses:

- **H1:** A difusão espacial de COVID-19, dengue, zika e chikungunya segue lógicas territoriais distintas, associadas a fatores específicos de cada doença.
- **H2:** Redes *multilayer* permitem identificar comunidades urbanas com características estruturais e epidemiológicas semelhantes, revelando padrões latentes de vulnerabilidade.
- **H3:** Variáveis socioeconômicas, políticas e de infraestrutura influenciam de maneira heterogênea o Risco Relativo (RR) das doenças, podendo ser identificadas por modelos preditivos interpretáveis.
- **H4:** A incorporação de múltiplas camadas de informação em modelos integrados contribui para uma compreensão mais precisa e territorializada da dinâmica epidemiológica em áreas urbanas.

Essas hipóteses orientam o desenvolvimento analítico apresentado e dialogam com a Proposta de Aprimoramento Metodológico anexada.

### 1.1.6 Estrutura da Tese

A estrutura desta tese está organizada da seguinte forma: primeiramente, são apresentados os trabalhos relacionados e os fundamentos teóricos e metodológicos das abordagens utilizadas, com ênfase em redes *multilayer*, análise estatística e espacial e aprendizado de máquina. Em seguida, detalha-se o processo metodológico adotado, desde a preparação dos dados até a execução das análises. Os capítulos subsequentes apresentam os resultados das análises espaciais, da modelagem de redes e dos modelos preditivos. Por fim, discute-se o valor agregado da abordagem proposta, suas implicações práticas, limitações — incluindo a restrição de dados para zika —, conclusões e sugestões para estudos futuros. Um anexo traz a *Proposta de Aprimoramento Metodológico* (Framework Integrado), visando orientar implementações e padronizações futuras.

## 1.2 Resultados Prévios

Conforme o cronograma planejado, as atividades desenvolvidas durante os anos anteriores do doutorado resultaram em importantes avanços nas pesquisas relacionadas à COVID-19 e ao desenvolvimento acadêmico. A seguir, são listados os principais resultados obtidos:

- **SCAD DIGITAL 2021, BR, prêmio de Melhor Pôster:** Colaboração ativa no projeto intitulado *Mecanismos associados à presença de sintomas digestivos na COVID-19* que resultou na publicação do pôster *Fatores relacionados à diarreia associada à COVID-19: um estudo prospectivo em um hospital terciário universitário* no Simpósio Nacional SCAD 21 DIGITAL e que foi premiado como Melhor Pôster.
- **DDW 2022, San Diego, EUA:** Submissão e aprovação do resumo *Atypical GERD symptoms may be associated with diaphragm strength in lean and older subjects* no Congresso Internacional DDW 2022 (San Diego, EUA), um evento de grande relevância na área de gastroenterologia, onde foi possível apresentar os achados iniciais das pesquisas realizadas.
- **IC2S2 2023, Copenhagen, DK:** Apresentação do trabalho *On the Impact of Political View to the Pandemic Outcomes* na conferência IC2S2 (9<sup>th</sup> International Conference on Computational Social Science IC<sup>2</sup>S<sup>2</sup>) em Copenhagen, Dinamarca, em julho de 2023, resultado das colaborações estabelecidas durante o período de pesquisa no Reino Unido.
- **SBBD 2024, Florianópolis, BR:** Publicação e apresentação do artigo *Multidimensional Analysis of the Influence of Socioeconomic and Political Indicators on the Spread of COVID-19: A Case Study of Brazilian Cities (2020-2024)* no 40º Simpósio Brasileiro de Banco de Dados, resultado dos estudos e análises realizadas durante o doutorado, consolidando os achados da pesquisa.
- **CompleNet 2025, Fortaleza, BR:** Aprovação de resumo e futura apresentação na modalidade *showcase* do trabalho *Dual-Layer Networks for the Analysis of Disease Diffusion and Its Socioeconomic, Political, and Demographic Determinants in the Neighborhoods of Fortaleza* na International Conference on Complex Networks (CompleNet).
- **JIDM 2025, BR:** Submissão do trabalho *Unraveling the Urban Complexity of Disease Risk: A Multilayer Spatial Analysis of COVID-19 and Arboviruses in Fortaleza, Brazil* ao Journal of Information and Data Management, 2025, após sermos convidados a estender o artigo apresentado no SBBD.
- **Colaboração com a Universidade de Exeter, UK:** Duração de seis meses (de agosto de 2022 a janeiro de 2023), através do programa CAPES Print na Universidade de Exeter,

no Reino Unido, sob a supervisão do Prof. Ronaldo Menezes. A imersão no ambiente acadêmico da Universidade de Exeter, em colaboração com o professor Ronaldo Menezes, foi crucial para a elaboração de modelos e metodologias de aprendizagem que estão desenvolvidas nesta tese.

- **Desenvolvimento de uma plataforma de coleta de dados:** Criação e implementação de uma plataforma para centralizar o armazenamento de dados das pesquisas, facilitando a gestão e análise das informações coletadas. Esse sistema teve desempenhado um papel fundamental no gerenciamento eficiente dos dados e na condução das análises posteriores.

## 2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo reúne, em linguagem acessível, os fundamentos que orientam a leitura territorial desta tese. Apresentamos noções estatísticas e espaciais, a estrutura de redes *multilayer*, princípios de aprendizado de máquina com explicabilidade e o uso de agrupamento.

### 2.1 Saúde urbana e determinantes territoriais

#### 2.1.1 *Mecanismos: exposição, suscetibilidade e capacidade de resposta*

A saúde nas cidades resulta da interação de fatores em múltiplas escalas. Em nível individual, contam idade, comorbidades e comportamentos; no domicílio, densidade de moradores, ventilação e manejo de água e resíduos; no entorno, mobilidade, uso do solo, áreas verdes e acesso a serviços; e, no plano institucional e político, a organização do sistema de saúde, a priorização orçamentária e a coordenação intersetorial. Essa visão é consistente com a literatura sobre determinantes sociais da saúde e epidemiologia urbana (WHO Commission on Social Determinants of Health, 2008; VLAHOV; GALEA, 2002).

Organizamos a análise em três mecanismos encadeados. Exposição: probabilidade de contato com o agente infeccioso ao longo do tempo e do espaço. Em arboviroses (dengue, zika, chikungunya), a ecologia do *Aedes* é fortemente condicionada por intermitência de abastecimento, armazenamento de água e manejo de resíduos; em doenças respiratórias (como COVID-19), mobilidade, aglomerações e ventilação dos ambientes são centrais. Suscetibilidade: diferenças na probabilidade de adoecer ou agravar sob exposições semelhantes, moduladas por renda, escolaridade, estrutura etária, comorbidades e possibilidade de adotar medidas protetivas. Capacidade de resposta: habilidade institucional e comunitária de detectar, atender e conter casos e surtos (atenção primária, vigilância, hospitalar, insumos, informação e coordenação).

#### 2.1.2 *Implicações territoriais e orientação metodológica*

Em cidades desiguais, esses mecanismos não se distribuem ao acaso. Segregação residencial, informalidade urbana e assimetrias de infraestrutura produzem gradientes geográficos em que riscos e recursos se acumulam (WHO Commission on Social Determinants of Health, 2008; ROUX, 2016). Bairros com precariedade de saneamento, menor renda média e menor cobertura de atenção básica tendem a combinar maior exposição, maior suscetibilidade e menor

capacidade de resposta, com trajetórias epidemiológicas desiguais — padrão também observado na COVID-19 no Brasil, em que desigualdades sociais e regionais se associaram a piores desfechos (BAQUI *et al.*, 2020).

Operacionalmente, isso orienta o restante da tese: padronizar variáveis e trabalhar com taxas para comparabilidade; usar diagnóstico espacial para identificar aglomerações, transições e outliers; representar interações entre dimensões por redes *multilayer*; aplicar modelos de aprendizado de máquina com validações espacial/temporal e técnicas de explicabilidade; e sintetizar perfis territoriais por agrupamento. Cuidados éticos são necessários: evitar a falácia ecológica, reconhecer o problema da unidade areal modificável e comunicar achados sem estigmatizar territórios, focando oportunidades de intervenção estrutural (WHO Commission on Social Determinants of Health, 2008; ROUX, 2016).

### **2.1.3 Especificidades das Doenças Abordadas (COVID-19, dengue, zika e chikungunya)**

As doenças analisadas neste trabalho — COVID-19, dengue, zika e chikungunya — compartilham a característica de serem fortemente influenciadas por determinantes sociais e territoriais, embora apresentem dinâmicas epidemiológicas distintas. A COVID-19, doença respiratória causada pelo coronavírus SARS-CoV-2, destacou as desigualdades estruturais das cidades brasileiras, com maior mortalidade observada em bairros com alta densidade populacional, habitações precárias e baixa renda (RODRIGUES *et al.*, 2021). Já as arboviroses — dengue, zika e chikungunya — são transmitidas pelo mosquito *Aedes aegypti*, cuja proliferação está associada à infraestrutura urbana inadequada, como acúmulo de água parada, falhas no saneamento e coleta de lixo (DONALISIO *et al.*, 2017).

A zika ganhou destaque pela associação com microcefalia congênita, exigindo articulação entre políticas reprodutivas e de saúde infantil. A chikungunya, por sua vez, provoca sintomas crônicos em parte dos infectados, afetando sua qualidade de vida por longos períodos. Já a dengue, com padrão endêmico e surtos sazonais, representa um desafio constante para o sistema de saúde pública. A análise integrada dessas doenças, à luz dos determinantes urbanos, permite compreender as interações entre condições estruturais e dinâmicas epidemiológicas locais.

## 2.2 Análise Estatística

A análise estatística desempenha papel central na investigação de relações entre variáveis epidemiológicas, socioeconômicas, políticas e estruturais. Neste trabalho, foram utilizadas abordagens estatísticas clássicas para avaliar tanto associações lineares bivariadas quanto relações multivariadas complexas.

### 2.2.1 Correlação de Pearson

O coeficiente de correlação de Pearson é uma medida estatística que quantifica a força e a direção da relação linear entre duas variáveis quantitativas (PEARSON, 1895). Representado por  $r$ , o coeficiente varia entre -1 e 1, onde valores próximos a 1 indicam forte correlação positiva, valores próximos a -1 indicam correlação negativa, e valores próximos a 0 sugerem ausência de relação linear. Sua fórmula é dada por:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (2.1)$$

onde  $x_i$  e  $y_i$  são os valores observados das variáveis  $X$  e  $Y$ , e  $\bar{x}$ ,  $\bar{y}$  representam suas médias. Essa métrica foi fundamental na fase exploratória do estudo, permitindo identificar pares de variáveis com alta associação linear como, por exemplo, entre risco relativo de COVID-19 e indicadores socioeconômicos ou de infraestrutura urbana.

### 2.2.2 Regressão Linear Múltipla (OLS)

A Regressão Linear Múltipla, também conhecida como *Ordinary Least Squares* (OLS), é uma técnica estatística utilizada para modelar a relação entre uma variável dependente contínua e um conjunto de variáveis independentes (MONTGOMERY *et al.*, 2012). O objetivo é estimar os coeficientes que minimizam a soma dos quadrados dos resíduos, ou seja, a diferença entre os valores observados e os valores preditos pelo modelo. A equação geral do modelo pode ser expressa como:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon \quad (2.2)$$

onde:

- $Y$  representa a variável dependente (ex: risco relativo de uma doença);

- $X_1, X_2, \dots, X_p$  são as variáveis independentes;
- $\beta_0$  é o intercepto;
- $\beta_1, \dots, \beta_p$  são os coeficientes estimados para cada preditor;
- $\varepsilon$  é o erro aleatório, assumido como normalmente distribuído com média zero e variância constante.

O modelo OLS foi utilizado neste estudo como linha de base para avaliar a significância estatística de variáveis explicativas selecionadas, além de permitir a comparação com modelos mais avançados, como algoritmos de aprendizado de máquina. Sua interpretação direta e capacidade de estimar efeitos marginais tornam-no uma ferramenta relevante para explorar relações entre múltiplos fatores determinantes da distribuição espacial das doenças.

## 2.3 Análise Espacial

### 2.3.1 Conceito de Autocorrelação Espacial

A autocorrelação espacial é um conceito fundamental na análise espacial que expressa o grau de similaridade entre os valores de uma variável em diferentes localidades geográficas, em função da proximidade espacial entre elas. Em termos simples, busca-se identificar se áreas geograficamente próximas tendem a apresentar valores semelhantes (autocorrelação positiva) ou diferentes (autocorrelação negativa) para determinada variável, como taxa de incidência de uma doença, renda ou acesso a serviços de saúde.

Essa dependência espacial entre observações viola os pressupostos dos modelos estatísticos tradicionais, que assumem independência entre unidades amostrais, tornando a análise espacial uma ferramenta imprescindível para estudos em epidemiologia urbana e saúde coletiva. Compreender os padrões de autocorrelação espacial permite identificar áreas críticas, revelar clusters e orientar políticas públicas com base em evidências territoriais (TOBLER, 1970; CHI, 2006; BAILEY; GATRELL, 1995).

### 2.3.2 Moran's I e LISA (Local Indicators of Spatial Association)

O índice de Moran (Moran's I) é uma das medidas globais mais utilizadas para quantificar a autocorrelação espacial. Ele fornece um valor único que indica a presença e intensidade da dependência espacial em todo o conjunto de dados (MORAN, 1950). Valores positivos de Moran's I sugerem que áreas com valores semelhantes estão agrupadas no espaço,

enquanto valores negativos indicam padrões de dispersão. Em geral, o valor do Moran's I varia entre -1 (dispersão perfeita) e +1 (aglomeração perfeita), com 0 indicando aleatoriedade espacial.

O cálculo do Moran's I envolve a comparação entre os valores observados em cada unidade espacial e os valores dos seus vizinhos, ponderados por uma matriz de pesos espaciais. Essa matriz pode ser construída com base em critérios de contiguidade (por exemplo, vizinhos que compartilham fronteiras) ou distância. O índice resultante é padronizado e testado estatisticamente (geralmente via permutação Monte Carlo) para verificar a significância da autocorrelação detectada. No entanto, como uma medida global, o Moran's I pode ocultar heterogeneidades locais importantes.

Para superar essa limitação, os *Local Indicators of Spatial Association* (LISA) foram propostos por Anselin (ANSELIN, 1995), permitindo avaliar a autocorrelação em nível local. O LISA identifica *clusters* espaciais (Alto-Alto, Baixo-Baixo) e *outliers* espaciais (Alto-Baixo, Baixo-Alto), possibilitando uma análise mais granular dos padrões de distribuição espacial. Em estudos de saúde, essas ferramentas são cruciais para detectar áreas com maior concentração de casos de doenças, zonas de risco e regiões que demandam intervenção urgente, especialmente quando aplicadas em dados urbanos altamente heterogêneos.

O algoritmo do LISA calcula, para cada unidade espacial  $i$ , uma estatística local baseada na autocorrelação espacial entre o valor observado  $x_i$  e os valores de suas unidades vizinhas  $x_j$ , ponderados por uma matriz de pesos espaciais  $w_{ij}$ . A fórmula básica é dada por:

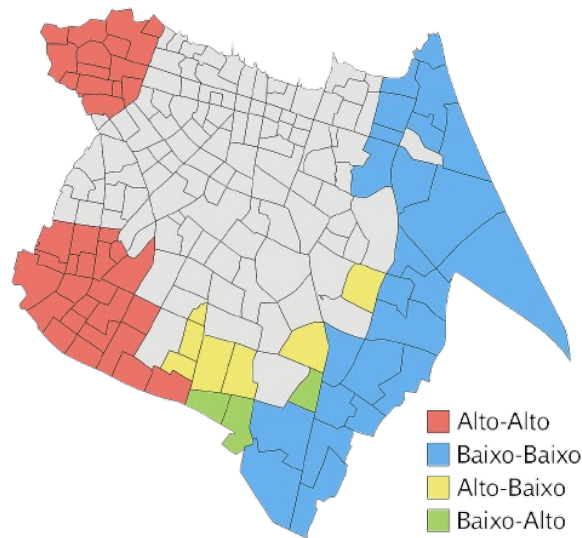
$$I_i = \frac{(x_i - \bar{x})}{m_2} \sum_j w_{ij}(x_j - \bar{x}) \quad (2.3)$$

onde  $\bar{x}$  representa a média global da variável de interesse,  $m_2$  é a variância total e  $w_{ij}$  são os pesos espaciais definidos a partir da matriz de contiguidade do tipo *queen* — critério em que dois bairros são considerados vizinhos se compartilham qualquer porção de fronteira ou vértice. Para cada unidade, o valor de  $I_i$  é comparado com uma distribuição de referência obtida por permutação aleatória (teste de permutação Monte Carlo com  $n = 999$ ), possibilitando a identificação de valores estatisticamente significativos.

Com base na combinação dos sinais dos termos  $(x_i - \bar{x})$  e  $(x_j - \bar{x})$ , os bairros são classificados em quatro categorias: **Alto-Alto** (valores altos cercados por valores altos), **Baixo-Baixo** (valores baixos cercados por baixos), **Alto-Baixo** (valores altos cercados por baixos) e **Baixo-Alto** (valores baixos cercados por altos). Essas categorias permitem identificar tanto

*clusters* espaciais homogêneos quanto *outliers* espaciais, sendo fundamentais para destacar padrões locais de concentração ou transição espacial da variável analisada.

Figura 1 – Ilustração de LISA.



Fonte: elaborado pelo autor.

A imagem acima ilustra a aplicação de LISA por meio de um mapa coroplético, onde cada região é colorida de acordo com o tipo de *cluster* identificado. As áreas em vermelho representam aglomerados do tipo Alto-Alto, sugerindo zonas críticas que concentram valores elevados da variável (por exemplo, alta incidência de doença), enquanto as regiões em azul indicam aglomerados Baixo-Baixo, possivelmente áreas de baixa vulnerabilidade. As cores amarelas e verdes destacam os *outliers* espaciais (Alto-Baixo e Baixo-Alto), que apontam regiões em transição ou com comportamentos atípicos frente ao entorno. Esse tipo de representação é crucial para a análise espacial detalhada e para a formulação de políticas públicas territorialmente direcionadas.

## 2.4 Redes Multilayer

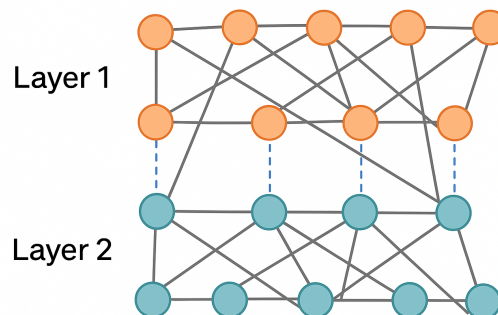
### 2.4.1 Fundamentos de Redes: nós, arestas e métricas estruturais

Redes *multilayer* representam sistemas compostos por múltiplas entidades interconectadas. Formalmente, uma rede é um grafo  $G = (V, E)$ , em que  $V$  é o conjunto de nós (vértices) e  $E$  o conjunto de arestas (ligações) que conectam pares de nós (NEWMAN, 2010).

Em contextos urbanos e epidemiológicos, nós podem ser bairros, instituições, agentes ou eventos de saúde; as arestas expressam fluxos, interações sociais, similaridade estatística ou proximidade espacial.

Para caracterizar a topologia e o funcionamento das redes, empregamos métricas estruturais clássicas: grau (número de conexões diretas), centralidade de intermediação (*betweenness*), centralidade de proximidade, centralidade por autovetor (influência considerando a importância dos vizinhos) e coeficiente de agrupamento (densidade de conexões no entorno imediato) (BARABÁSI, 2016). Essas métricas ajudam a identificar nós estratégicos para a conectividade global, potenciais disseminadores de informação (ou doenças) e regiões de maior coesão estrutural. Na aplicação territorial, a leitura das métricas deve ser ancorada no significado das arestas: por exemplo, um alto grau com arestas de “similaridade epidemiológica” comunica algo distinto de um alto grau com arestas de “mobilidade cotidiana”.

Figura 2 – Representação esquemática de uma rede *Dual-Layer*.



Fonte: elaborado pelo autor.

#### 2.4.2 Conceitos e formalização multilayer

A modelagem *multilayer* generaliza o conceito de rede ao incorporar múltiplas dimensões de relacionamento entre os mesmos nós (KIVELÄ *et al.*, 2014). Em vez de um único tipo de interação, organizamos os dados em camadas (*layers*) que representam domínios heterogêneos — epidemiológico, socioeconômico/demográfico, infraestrutura e político/IDH — com conexões *intra-camada* (entre bairros na mesma dimensão) e *inter-camada* (entre representações do mesmo bairro em camadas distintas), proporcionando uma visão mais realista e interdependente do sistema urbano (Figura 2).

Formalmente, uma rede *multilayer* pode ser descrita por uma matriz de *supra-*

*adjacência* que agrega, em blocos, as adjacências intra-camada e as conexões intercamadas (KIVELÄ *et al.*, 2014):

$$\mathbf{A}^{\text{supra}} = \begin{bmatrix} \mathbf{A}^{(1)} & \Omega^{(1,2)} & \dots & \Omega^{(1,L)} \\ \Omega^{(2,1)} & \mathbf{A}^{(2)} & \dots & \Omega^{(2,L)} \\ \vdots & \vdots & \ddots & \vdots \\ \Omega^{(L,1)} & \Omega^{(L,2)} & \dots & \mathbf{A}^{(L)} \end{bmatrix},$$

onde  $\mathbf{A}^{(l)}$  codifica arestas dentro da camada  $l$  e  $\Omega^{(l,m)}$  as conexões entre camadas  $l$  e  $m$ . Na prática, essa formalização se traduz em decisões operacionais: (i) seleção de variáveis por camada (p. ex., incidência/mortalidade na epidemiológica; densidade de equipamentos na infraestrutura); (ii) transformação de variáveis em *pesos* de arestas (correlações de Spearman, similaridades normalizadas, contiguidade espacial ponderada); (iii) direção das arestas (direcionadas quando há fluxo/influência assimétrica; não direcionadas quando há semelhança); e (iv) tratamento de redundâncias/colinearidade para evitar “arestas duplicadas”.

Para garantir comparabilidade, adotamos padronização de escalas entre variáveis (p. ex., *z-score* ou *ranks/quantis*), definição explícita de limiares de arestamento (por exemplo, manter apenas correlações  $\rho_s \geq 0,3$ ) e regras para dados ausentes (imputação ou exclusão). A incerteza dos pesos pode ser quantificada por *bootstrap* ou permutações, registrando intervalos de confiança e estabilidade das conexões.

### 2.4.3 Estrutura multiplex e justificativa

A estrutura adotada é *multiplex*: os mesmos bairros aparecem em todas as camadas e cada bairro conecta-se a si mesmo entre camadas (auto-conexões intercamadas) por meio de uma medida de associação. Usamos, por simplificação, o termo *multilayer* ao longo do texto, registrando aqui essa caracterização. O arranjo multiplex preserva o “retrato” de cada dimensão sem forçar fusão precoce de evidências — crucial em contextos urbanos onde mecanismos diferem entre camadas. Conexões por mobilidade (infraestrutura) podem não coincidir com conexões por perfil socioeconômico, e ambas interagem com padrões epidemiológicos de forma complementar (BOCCALETTI *et al.*, 2014).

Outra vantagem é a transparência: cada camada mantém sua semântica e seus critérios de arestamento, enquanto o acoplamento intercamadas (controlável por um parâmetro *omega*, se desejado) permite avaliar o quanto relações cruzadas contribuem para a estrutura global. Analiticamente, isso responde perguntas do tipo: “bairros que formam comunidades na

infraestrutura também se agrupam na epidemiologia?” Convergências sugerem reforço entre determinantes; divergências apontam especificidades que demandam políticas direcionadas.

#### **2.4.4 Métricas multilayer e interpretação substantiva**

Medidas clássicas de centralidade ganham nova leitura ao considerar múltiplas camadas (NEWMAN, 2010; BARABÁSI, 2016). Um bairro com *grau* alto simultaneamente nas camadas epidemiológica e de infraestrutura pode sinalizar um encontro entre risco biológico e condições materiais (acessos viários, densidade de serviços). A *proximidade multilayer* (menores distâncias médias considerando múltiplos mecanismos de conexão) destaca bairros estruturalmente “próximos” a muitos outros quando mais de uma dimensão é relevante. A *intermediação multilayer* identifica “pontes” entre subestruturas de camadas distintas, úteis para estratégias de vigilância, comunicação de risco e alocação de recursos.

Para evitar vieses, agregamos métricas por camada em escores compostos somente após padronização (p. ex., transformar cada centralidade em *z-score* e somar com pesos explícitos), impedindo que camadas com pesos naturalmente maiores dominem o índice final. Realizamos análises de sensibilidade variando limiares de arestamento e verificamos a estabilidade dos *rankings*. Quando possível, estimamos a variabilidade das centralidades por *bootstrap* de nós/arestas. A interpretação mantém foco no sentido territorial: “o que um alto valor de centralidade significa, concretamente, para o bairro X?”, sempre remetendo ao tipo de aresta e à camada.

#### **2.4.5 Detecção de comunidades: por camada e na perspectiva multilayer**

A detecção de comunidades busca identificar subgrupos de nós mais conectados entre si do que com o restante da rede (FORTUNATO, 2010). Empregamos o algoritmo de Louvain, eficiente e adequado para grafos grandes, otimizando a modularidade com pesos quando disponíveis (BLONDEL *et al.*, 2008). Nesta tese, optamos por detectar comunidades *separadamente por camada* e comparar resultados entre camadas. Essa escolha preserva a interpretabilidade: cada camada “conta sua história”; a comparação lado a lado e cruzamentos em tabelas simples revelam convergências (ou divergências) entre dimensões.

Em redes *multilayer*, também é possível buscar comunidades interdimensionais (partição conjunta) (KIVELÄ *et al.*, 2014; BOCCALETTI *et al.*, 2014). Ainda que não seja a estratégia principal aqui, empregamos ferramentas de *consenso* quando pertinente (p. ex.,

repetição de Louvain com sementes distintas, ajuste do parâmetro de resolução  $\gamma$ , matriz de coocorrência entre partições) e monitoramos a similaridade entre partições por NMI/ARI. Em contextos urbanos dinâmicos, avalia-se a persistência de comunidades em janelas temporais (meses/trimestres), útil para sazonalidades de arboviroses.

#### **2.4.6 Boas práticas e cautelas analíticas**

Duas cautelas são centrais. Primeiro, redes baseadas em correlações capturam *associações*, não causalidade; confundidores (como densidade populacional) podem induzir arestas espúrias. Mitigações: (i) considerar correlações parciais (controlando variáveis-chave) na criação de arestas; (ii) testar significância dos pesos por permutações, retendo apenas conexões acima de limiares estatísticos. Segundo, redes agregam realidades dinâmicas; “fotografias” de um período devem ser lidas à luz de mudanças temporais (sazonalidade de arboviroses, ondas epidêmicas).

Além disso: (a) efeito de escala e zonificação (MAUP) — resultados variam com o nível de agregação (bairro, setor, microárea); registre o nível espacial e discuta limitações; (b) incerteza e qualidade de dados — subnotificação, séries incompletas e mudanças de política pública afetam pesos e comunidades; documente fontes de incerteza e realize análises de sensibilidade (remoção de outliers, imputações alternativas). Por fim, explicithe escolhas (variáveis por camada, limiares de Spearman, janelas temporais) e reporte testes de robustez, fortalecendo a validade interna dos achados e sua utilidade para políticas públicas territorializadas (BOCCALETTI *et al.*, 2014).

## **2.5 Aprendizado de Máquina**

Modelos de aprendizado supervisionado baseados em árvores de decisão são amplamente utilizados por sua robustez, interpretabilidade e capacidade de capturar relações não lineares e interações complexas entre variáveis. Neste trabalho, utilizam-se dois algoritmos de referência: o *Random Forest* e o *Extreme Gradient Boosting* (XGBoost).

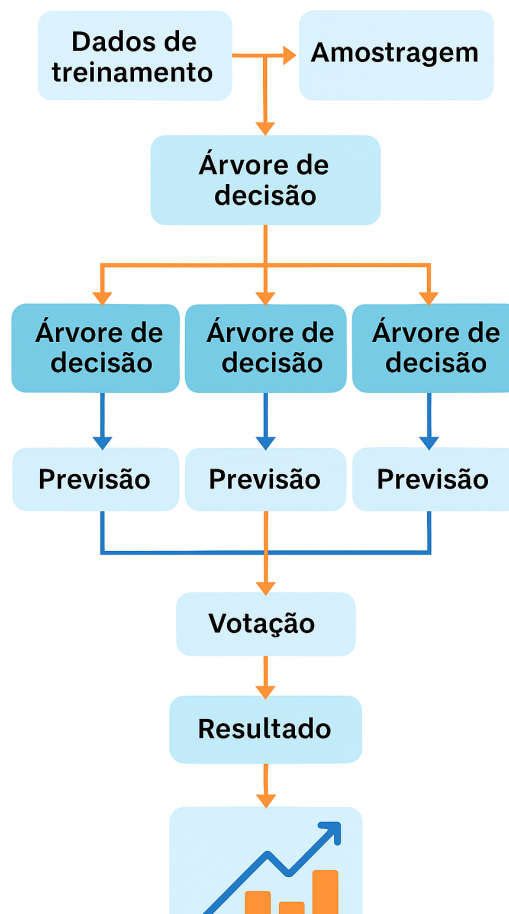
### **2.5.1 Modelos Baseados em Árvores: *Random Forest* e *XGBoost***

Modelos baseados em árvores de decisão têm se destacado em problemas supervisionados de classificação e regressão devido à sua robustez a dados ruidosos, capacidade de

modelar relações não lineares e facilidade de interpretação. Dentre esses, o *Random Forest* e o *Extreme Gradient Boosting* (XGBoost) representam duas abordagens amplamente consolidadas e complementares.

O **Random Forest** (BREIMAN, 2001) é um método de aprendizado de máquina baseado em *bagging* (*bootstrap aggregating*), no qual múltiplas árvores de decisão são treinadas de forma independente sobre subconjuntos aleatórios de observações e variáveis (veja Figura 3). As previsões são então agregadas (por média, no caso da regressão) para reduzir a variância do modelo e aumentar sua generalização. Esse processo aumenta a robustez e a estabilidade do modelo, sendo especialmente útil em conjuntos de dados com alta dimensionalidade e colinearidade entre variáveis. O algoritmo também permite estimar a importância das variáveis, seja pela redução da impureza (índice de Gini), seja pela queda na performance ao permutar uma variável, o que favorece interpretações mais claras sobre os fatores que influenciam a variável-alvo.

Figura 3 – Representação esquemática do funcionamento do algoritmo Random Forest.



Fonte: elaborado pelo autor.

Por outro lado, o **XGBoost** (CHEN; GUESTRIN, 2016) é uma implementação otimizada do algoritmo de *gradient boosting*, que constrói árvores de forma sequencial, em que cada novo modelo é treinado para minimizar os resíduos do modelo anterior. Essa abordagem permite capturar padrões complexos nos dados com alta acurácia preditiva, ao custo de maior complexidade computacional. O XGBoost incorpora regularização explícita, técnicas de *early stopping* e paralelização do treinamento, tornando-o eficiente mesmo em grandes volumes de dados.

### 2.5.2 *Explicabilidade de Modelos: SHAP (SHapley Additive exPlanations)*

Apesar do alto desempenho preditivo, modelos de aprendizado de máquina mais complexos são frequentemente criticados por sua falta de interpretabilidade, sendo considerados como "caixas-pretas". A interpretabilidade é especialmente crucial em domínios como saúde, onde é necessário justificar decisões baseadas em modelos algorítmicos.

Nesse contexto, o método SHAP (*SHapley Additive exPlanations*) (LUNDBERG; LEE, 2017) tem se destacado como uma abordagem consistente e teoricamente fundamentada para explicar as previsões de qualquer modelo de ML (LUNDBERG; LEE, 2017). Baseado na teoria dos valores de Shapley da teoria dos jogos, o SHAP atribui a cada variável uma contribuição marginal para a predição, permitindo compreender como diferentes características influenciam a saída do modelo, tanto globalmente quanto localmente (em instâncias específicas). O uso do SHAP em modelos epidemiológicos oferece uma camada adicional de transparência, facilitando a interpretação de fatores de risco e a comunicação de resultados para gestores públicos.

### 2.5.3 *Limitações e Vantagens da Abordagem Preditiva em Epidemiologia*

Embora o aprendizado de máquina ofereça vantagens importantes, como flexibilidade, capacidade de generalização e exploração de padrões não-lineares, sua aplicação em epidemiologia requer cautela. Uma das principais limitações diz respeito à qualidade e representatividade dos dados disponíveis, além do risco de sobreajuste e da dificuldade em generalizar modelos treinados em contextos específicos.

Outra limitação é a baixa interpretabilidade de modelos mais complexos, o que pode comprometer sua aceitação em ambientes clínicos e de políticas públicas. No entanto, o desenvolvimento de técnicas de explicabilidade tem atenuado esse problema. Por outro lado,

entre as principais vantagens da abordagem preditiva estão a possibilidade de antecipar cenários críticos, melhorar a alocação de recursos e integrar múltiplas fontes de informação em análises abrangentes (TOPOL, 2019a).

Neste trabalho, as técnicas de aprendizado de máquina são integradas à análise espacial e à modelagem em redes *multilayer*, permitindo não apenas prever a distribuição de doenças, mas também interpretar os mecanismos subjacentes à sua disseminação em ambientes urbanos desiguais como Fortaleza.

## 2.6 Agrupamento Espacial

### 2.6.1 Fundamentos do Algoritmo *K-means*

O algoritmo *K-means* é uma das abordagens mais tradicionais e amplamente utilizadas para agrupamento (*clustering*) de dados em contextos não supervisionados. Seu objetivo é particionar  $n$  observações em  $k$  grupos (ou clusters), de forma que cada observação pertença ao grupo cujo centróide (média do cluster) esteja mais próximo, minimizando a variabilidade interna de cada grupo (MACQUEEN, 1967).

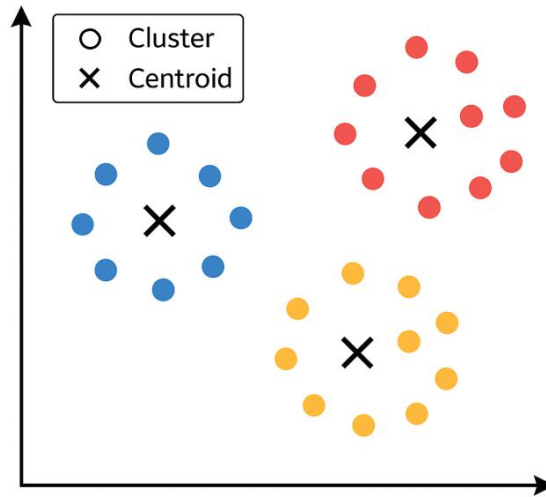
Matematicamente, o *K-means* busca minimizar a soma das distâncias quadradas entre cada ponto e o centróide do cluster ao qual ele pertence:

$$\arg \min_C \sum_{i=1}^k \sum_{x \in C_i} \|x - \mu_i\|^2 \quad (2.4)$$

onde  $C_i$  representa o conjunto de pontos atribuídos ao cluster  $i$  e  $\mu_i$  é o centróide desse cluster. O algoritmo opera de forma iterativa por meio dos seguintes passos:

1. **Inicialização:**  $k$  centróides são selecionados aleatoriamente no espaço de atributos;
2. **Atribuição:** cada observação é alocada ao cluster com centróide mais próximo (com base em distância euclidiana);
3. **Atualização:** os centróides de cada cluster são recalculados como a média dos pontos atribuídos;
4. **Repetição:** os passos de atribuição e atualização são repetidos até que os centróides não mudem significativamente ou atinja-se um critério de convergência.

Figura 4 – Ilustração do funcionamento do algoritmo K-means.



Fonte: elaborado pelo autor.

A Figura 4 apresenta um exemplo bidimensional do algoritmo K-means, com três *clusters* distintos representados por diferentes cores. Os pontos no plano correspondem às observações do conjunto de dados, enquanto os marcadores destacados em formato de “X” representam os centróides de cada grupo. O algoritmo, para chegar a essa configuração representada na imagem, iniciou com centróides aleatórios e, iterativamente, realizou dois passos principais: (1) atribuição de cada ponto ao centróide mais próximo, com base na distância euclidiana, e (2) atualização dos centróides como a média das coordenadas dos pontos atribuídos a cada grupo. Esse processo se repetiu até que os centróides não se movessem significativamente, indicando a convergência do algoritmo mostrado na imagem.

Apesar de sua simplicidade e eficiência computacional, o K-means possui limitações: assume que os *clusters* possuem forma aproximadamente esférica e similar em tamanho, é sensível a valores extremos e à inicialização dos centróides, e exige a definição prévia de  $k$ . Para selecionar o número ótimo de clusters, técnicas como o método do cotovelo (*Elbow Method*) e o índice de silhueta (*Silhouette Score*) são comumente utilizadas (ROUSSEEUW, 1987). Ainda assim, sua capacidade de identificar agrupamentos úteis em conjuntos multivariados faz do K-means uma ferramenta eficaz para análise exploratória e redução de complexidade em grandes bases de dados.

### 3 TRABALHOS RELACIONADOS

Nas últimas décadas, a epidemiologia tem se beneficiado substancialmente dos avanços computacionais e da disponibilidade de dados em grande escala. Técnicas como aprendizado de máquina, análise espacial e modelagem por redes complexas têm sido aplicadas a diversos contextos de saúde pública, especialmente em ambientes urbanos complexos. Este trabalho se insere nesse panorama metodológico, propondo uma abordagem integrada baseada em redes *multilayer* e aprendizado de máquina, aplicada à modelagem da disseminação de doenças infecciosas em Fortaleza. A seguir, discutimos os principais trabalhos relacionados, destacando como a presente tese avança em relação a essas contribuições.

#### 3.1 Aprendizado de Máquina na Saúde Pública

Nos últimos anos, o uso de *machine learning* / aprendizado de máquina (ML) em aplicações médicas cresceu exponencialmente, com uma ênfase marcante em ambientes clínicos e hospitalares. Estudos como os de Rajkomar et al.(RAJKOMAR *et al.*, 2018), Shickel et al.(SHICKEL *et al.*, 2018) e Beam e Kohane (BEAM; KOHANE, 2018) aplicam modelos supervisionados — como redes neurais profundas — na predição de desfechos clínicos, automação de diagnósticos e gerenciamento de prontuários eletrônicos. Essas contribuições representam avanços notáveis no contexto da medicina individualizada, mas compartilham uma limitação estrutural: operam quase exclusivamente sobre dados clínicos centrados no paciente, com pouca ou nenhuma consideração sobre os determinantes sociais, espaciais e territoriais que influenciam a saúde em escala coletiva.

Em contraste com esse paradigma centrado no indivíduo, esta tese adota uma abordagem orientada ao território, priorizando a análise de dados agregados por bairro e integrando múltiplas dimensões — epidemiológica, socioeconômica, política e de infraestrutura urbana — para compreender padrões de vulnerabilidade à saúde. Essa mudança de foco, do paciente para o território, exige não apenas outras fontes de dados e outras perguntas, mas também uma mudança epistemológica: a saúde deixa de ser tratada como um fenômeno meramente biomédico, e passa a ser entendida como resultado de processos estruturais, espaciais e sociais historicamente determinados.

Mesmo em estudos que reconhecem a importância de fatores contextuais, como Obermeyer e Emanuel (OBERMEYER; EMANUEL, 2016) ou Topol (TOPOL, 2019b), a

aplicação prática de ML segue limitada por uma lógica predominantemente individualizada. Pouco se discute, por exemplo, como desigualdades urbanas moldam a incidência de doenças infecciosas, ou como a ausência de infraestrutura básica (como saneamento, acesso à saúde ou equipamentos públicos) pode ser modelada computacionalmente como fator explicativo. Em outras palavras, mesmo quando reconhecem as limitações, esses trabalhos ainda partem de uma concepção restrita da saúde — como algo que ocorre no corpo, e não no território.

Dessa forma, esta pesquisa se diferencia não apenas por incorporar algoritmos modernos de predição, mas por aplicá-los em um arcabouço conceitual distinto. Ao invés de buscar soluções individualizadas para o cuidado clínico, investiga-se aqui a relação entre características estruturais dos bairros e a disseminação de doenças como COVID-19, dengue, zika e chikungunya. Isso implica considerar variáveis como densidade populacional, número de escolas, resultados eleitorais, infraestrutura de saúde, cultura e esporte, como possíveis explicadores de padrões epidêmicos.

Além disso, enquanto grande parte da literatura em ML para saúde ainda trata o território como um pano de fundo estático, esta tese mobiliza conceitos e técnicas da análise espacial para reconhecer a natureza geograficamente distribuída e socialmente produzida da vulnerabilidade à saúde. O uso de modelos interpretáveis, como SHAP (LUNDBERG; LEE, 2017) e LIME (RIBEIRO *et al.*, 2016), é aqui orientado não apenas à explicação técnica dos algoritmos, mas à construção de conhecimento substantivo sobre os mecanismos sociais e espaciais que produzem iniquidades em saúde.

Ao tensionar as abordagens tradicionais de ML aplicadas à medicina com uma perspectiva territorial e estrutural da saúde, esta tese busca ocupar um espaço ainda pouco explorado entre a epidemiologia computacional e os estudos urbanos. A proposta não é substituir os modelos clínicos existentes, mas complementá-los com uma visão crítica e geograficamente situada, mais adequada aos desafios enfrentados por cidades desiguais como Fortaleza.

### **3.1.1 Perspectivas Nacionais**

Trabalhos empíricos realizados no Brasil, como o de Santos *et al.* (SANTOS *et al.*, 2020), que analisam a vulnerabilidade à COVID-19 nas cidades brasileiras, revelam como a combinação de fatores socioeconômicos e demográficos pode influenciar o risco de contágio e a gravidade das infecções. A pesquisa de Santos *et al.* foca nas desigualdades sociais que tornam as populações em áreas periféricas mais suscetíveis a doenças infecciosas. Porém, esses estudos

não incorporam as abordagens de aprendizado de máquina explicável nem a análise de redes complexas, como proposto nesta tese.

Outro exemplo relevante é o trabalho de Ribeiro et al. (RIBEIRO *et al.*, 2021), que explora a relação entre desigualdade socioespacial e o impacto da COVID-19 no Rio de Janeiro. Embora a pesquisa ofereça uma importante análise dos fatores socioeconômicos que influenciam o cenário epidemiológico, ela se limita a abordagens estatísticas tradicionais e não propõe uma integração com modelos preditivos de aprendizado de máquina ou análises espaciais interligadas, como feito na presente tese.

Além disso, o estudo de Crepalde (CREPALDE, 2024), ao analisar a dinâmica da COVID-19 em diferentes áreas do Rio de Janeiro, particularmente nas favelas, mostra a relevância das desigualdades socioespaciais, mas não explora a combinação com abordagens de modelagem de redes *multilayer*, que permitem uma análise mais holística das interações entre variáveis socioeconômicas, políticas e de saúde. Esta tese propõe uma análise mais integrada ao utilizar aprendizado de máquina para prever o risco de doenças, enquanto explora a estrutura urbana através de redes *multilayer*, uma abordagem que visa compreender as interações entre diferentes camadas da vulnerabilidade.

Estudos realizados em Fortaleza destacam como as desigualdades sociais e espaciais influenciam os desfechos da COVID-19. Por exemplo, pesquisa de Câmara et al. (CÂMARA *et al.*, 2021) identificou a vulnerabilidade associada à saúde e às dimensões social e econômica dos moradores dos bairros de Fortaleza durante a pandemia. A análise revelou que bairros com maior pobreza extrema, menor renda anual e vínculos empregatícios precários apresentaram maior vulnerabilidade à COVID-19. No entanto, o estudo não utilizou técnicas de aprendizado de máquina explicável nem análise de redes complexas, abordagens que esta pesquisa incorpora para compreender melhor os determinantes estruturais da saúde urbana.

Outro estudo relevante é o de Silva et al. (SILVA *et al.*, 2021), que analisa a associação entre desigualdades sociais e sanitárias, condições socioeconômicas, segregação espacial e letalidade por COVID-19 em Fortaleza. A pesquisa identificou autocorrelações espaciais significativas para letalidade aparente e extrema pobreza, evidenciando a relação entre desigualdades sociais e desfechos de saúde em tempos de pandemia. Entretanto, o estudo se concentrou em análises estatísticas tradicionais e não explorou abordagens de aprendizado de máquina explicável ou modelagem de redes *multilayer*, como feito nesta tese.

Além disso, o trabalho de Freitas (FREITAS, 2021) utilizou Internet das Coisas e

aprendizado de máquina para prever casos de dengue em Fortaleza. Embora o estudo tenha aplicado técnicas de aprendizado de máquina, ele se concentrou na previsão de surtos de dengue e não abordou diretamente a COVID-19 nem integrou análise territorial baseada em redes complexas, como proposto nesta pesquisa.

### 3.2 Análise Espacial e Clusterização na Epidemiologia

A análise espacial tem sido uma ferramenta poderosa para compreender a distribuição geográfica de doenças e suas correlações com fatores sociais, econômicos e ambientais. A utilização de técnicas de clusterização espacial tem sido cada vez mais aplicada para identificar padrões de agregação e dispersão de casos, permitindo a detecção de áreas de risco e a implementação de intervenções direcionadas. Abaixo, destacam-se alguns trabalhos relevantes que exploram a análise espacial e a clusterização, tanto no contexto internacional quanto nacional.

Um exemplo clássico de análise espacial em epidemiologia é o estudo de Anselin et al. (ANSELIN *et al.*, 2006), que aplicou medidas de autocorrelação espacial, como o índice de Moran e a estatística LISA, para investigar a distribuição de doenças em áreas urbanas, destacando a influência de fatores sociais, ambientais e econômicos sobre a localização dos casos. De forma semelhante, este trabalho também adota Moran's I e LISA para detectar padrões espaciais e clusters locais de altas e baixas incidências. No entanto, nossa abordagem se diferencia ao incorporar métodos complementares de análise espacial e preditiva, como o algoritmo de clusterização KMeans, que permite a identificação de agrupamentos com base em múltiplas variáveis não necessariamente espaciais, e a Geographically Weighted Regression (GWR), que modela relações locais entre variáveis explicativas e os desfechos epidemiológicos. Essa integração metodológica permite uma compreensão mais abrangente e refinada da heterogeneidade espacial dos determinantes e da difusão de doenças em contextos urbanos complexos.

O trabalho de Rezende et al. (REZENDE *et al.*, 2023) analisou a distribuição espacial da COVID-19 na Amazônia Legal utilizando autocorrelação espacial, de forma semelhante à abordagem adotada neste estudo. No entanto, enquanto o trabalho de Rezende se concentrou em medidas globais e locais de associação espacial, nesta tese são incorporadas técnicas adicionais que permitem captar padrões mais complexos de agrupamento e variações espaciais nas relações entre variáveis, oferecendo uma caracterização mais detalhada das dinâmicas territoriais da doença.

No contexto brasileiro, o estudo de Oliveira (OLIVEIRA, 2021) analisou as desi-

gualdades sociais associadas à letalidade por COVID-19 em Fortaleza, utilizando estatísticas espaciais para identificar áreas com maior vulnerabilidade. A pesquisa evidenciou a importância da dimensão territorial na compreensão dos efeitos da pandemia e contribuiu para orientar políticas públicas locais. Neste trabalho, expande-se essa perspectiva ao explorar, além da autocorrelação espacial, outras abordagens que permitem uma análise mais aprofundada dos padrões de disseminação e suas relações com múltiplas variáveis contextuais.

Outro estudo interessante foi realizado por Magalhães (MAGALHÃES, 2015), que investigou a relação entre clima urbano e a ocorrência de surtos de dengue em Fortaleza. Através de análise de dados espaciais e variáveis climáticas, os autores mapearam áreas de maior incidência de dengue. A pesquisa não utilizou técnicas de clusterização, mas poderia se beneficiar de uma abordagem mais integrada com essas metodologias para detectar *clusters* de alta taxa de incidência.

Embora muitos dos estudos mencionados utilizem análise espacial e clusterização, poucos exploram a integração dessas abordagens com técnicas de aprendizado de máquina explicável ou modelagem de redes *multilayer*. A presente pesquisa visa avançar nesse campo, ao combinar análise espacial com modelos preditivos de aprendizado de máquina, oferecendo uma compreensão mais holística e integrada dos fatores estruturais que influenciam a saúde urbana.

### **3.3 Redes Multilayer na Epidemiologia: Trabalhos Relacionados e Avanços Propostos**

A teoria de redes *multilayer* tem sido amplamente utilizada para modelar a disseminação de doenças infecciosas, oferecendo ferramentas robustas para representar interações entre indivíduos, regiões geográficas ou fatores estruturais. Diferentes abordagens vêm sendo aplicadas, como redes baseadas em contatos, redes de mobilidade, redes espaço-temporais e redes sociais, cada uma com suas vantagens e limitações. Esta seção compara alguns dos principais trabalhos da literatura com a abordagem proposta neste estudo.

O trabalho de Pastor-Satorras e Vespignani (PASTOR-SATORRAS; VESPIGNANI, 2001) foi um dos marcos na modelagem de epidemias em redes complexas, ao mostrar que, em redes com distribuição de grau livre de escala, a ausência de limiar epidêmico torna as doenças mais difíceis de erradicar. Embora seminal, o modelo considera redes sintéticas e homogêneas em termos de contexto espacial e social, o que limita sua aplicação direta a territórios urbanos reais e heterogêneos.

Outra contribuição importante foi apresentada por Salathé e Jones (SALATHÉ;

JONES, 2010), que investigaram a disseminação de doenças em redes sociais reais baseadas em dados de proximidade entre indivíduos. A pesquisa reforçou o papel da estrutura da rede social na propagação de infecções, mas o foco está na escala individual (micro), sem considerar fatores estruturais como desigualdade social, infraestrutura ou variáveis políticas.

No contexto brasileiro, o estudo de Albuquerque et al. (ALBUQUERQUE *et al.*, 2021) propôs uma análise territorial da vulnerabilidade à COVID-19, construindo um índice de vulnerabilidade social com base em dados de infraestrutura urbana e indicadores socioeconômicos. Embora não tenha construído uma rede formal, a abordagem territorial multivariada antecipa a ideia de integração entre diferentes domínios estruturais, aproximando-se do conceito de redes *multilayer* adotado neste trabalho.

Outro exemplo é o trabalho de Scabini et al. (SCABINI *et al.*, 2021). Este estudo propõe uma modelagem da epidemia de COVID-19 utilizando uma rede complexa *multilayer*, onde os nós representam indivíduos, as arestas representam contatos sociais e as camadas representam diferentes atividades sociais. O modelo aprimora o tradicional SIR e é aplicado para estudar a epidemia brasileira, analisando possíveis ações futuras e suas consequências. Contudo, o modelo permanece centrado em uma estrutura unificada de rede, sem distinção temática entre camadas ou análise de modularidade entre domínios distintos.

Diferentemente desses trabalhos, a presente pesquisa propõe uma modelagem *multilayer* aplicada à cidade de Fortaleza, integrando diferentes domínios — epidemiológico, socioeconômico, político e de infraestrutura — por meio de redes *multilayer* em camadas interconectadas. São analisadas centralidades, conectividade, modularidade e detecção de comunidades, com foco na compreensão da difusão espacial de quatro doenças infecciosas (COVID-19, dengue, zika e chikungunya) em um contexto urbano real e desigual. Essa abordagem permite identificar comunidades estruturalmente semelhantes entre bairros, considerando múltiplas dimensões de vulnerabilidade e risco, o que representa um avanço metodológico frente aos modelos clássicos baseados em redes monocamada ou apenas de mobilidade.

### **3.4 Síntese Comparativa dos Trabalhos Relacionados**

A seguir (Tabela 1) compara os trabalhos relacionados com a proposta deste estudo, destacando as abordagens metodológicas e os avanços propostos.

Tabela 1 – Resumo dos clusters territoriais gerados via K-Means

<b>Estudo</b>	<b>Abordagem</b>	<b>Avanços Propostos</b>
Rajkomar et al. (2018)	Redes neurais profundas para predição de desfechos clínicos.	Foco clínico, não territorial. Não considera fatores socioeconômicos e espaciais.
Shickel et al. (2017)	Modelos supervisionados em dados clínicos, diagnóstico automatizado.	Aplicação em dados centrados no paciente, sem análise territorial ou contextual.
Beam e Kohane (2018)	Automação de diagnósticos médicos e gerenciamento de prontuários.	Abordagem individualizada, sem considerações territoriais ou sociais.
Obermeyer e Emanuel (2016)	Modelos preditivos para saúde, considerando fatores contextuais.	Limitação em dados clínicos, com pouca análise territorial ou espacial.
Topol (2019)	Análise de dados genômicos e predição clínica.	Foca em genética e dados individuais, sem considerar o contexto socioeconômico ou urbano.
Santos et al. (2020)	Análise de vulnerabilidade à COVID-19 com dados socioeconômicos.	Não utiliza aprendizado de máquina explicável nem redes <i>multilayer</i> .
Ribeiro et al. (2021)	Análise de desigualdade socioespacial no Rio de Janeiro.	Não integra aprendizado de máquina ou análise de redes <i>multilayer</i> .
Crepalde (2024)	Estudo de desigualdade e COVID-19 nas favelas do Rio de Janeiro.	Não aplica redes <i>multilayer</i> nem modelos explicáveis.
Câmara et al. (2021)	Análise de vulnerabilidade à COVID-19 em Fortaleza, considerando fatores sociais.	Não utiliza aprendizado de máquina explicável ou redes <i>multilayer</i> .
Silva et al. (2021)	Desigualdade social e letalidade por COVID-19 em Fortaleza.	Foca em estatísticas tradicionais, sem análise preditiva ou espacial avançada.
Freitas (2021)	Predição de surtos de dengue em Fortaleza com IoT e aprendizado de máquina.	Não integra análise territorial baseada em redes <i>multilayer</i> .

Fonte: elaborado pelo autor.

## 4 METODOLOGIA

Este capítulo apresenta detalhadamente a metodologia desenvolvida para alcançar os objetivos deste trabalho. A abordagem adotada integra procedimentos rigorosos desde a preparação e integração dos dados até técnicas analíticas mais avançadas, abrangendo análises estatísticas, espaciais, redes multilayer e técnicas de machine learning.

### 4.1 Pré-processamento dos Dados

A qualidade dos resultados em análises avançadas depende fundamentalmente da etapa inicial de pré-processamento dos dados. Nesta pesquisa, foram utilizados dados obtidos de fontes oficiais e confiáveis, visando assegurar a validade científica das análises posteriores.

O principal insumo deste estudo para COVID-19 foi extraído dos Boletins Epidemiológicos *semanais* de COVID-19 de cada ano (2020–2024). A partir desses boletins, consolidamos uma base de casos por bairro e por mês (agregação a partir da semana epidemiológica), mantendo a rastreabilidade por período e por bairro. Os óbitos por COVID-19 foram obtidos dos relatórios oficiais da Prefeitura de Fortaleza (Secretaria Municipal da Saúde) e unidades de vigilância indicadas abaixo. As demais dimensões (socioeconômica/demográfica, infraestrutura e política/Índice de Desenvolvimento Humano (IDH)) vieram de bases públicas oficiais. A seguir, listamos as fontes:

- **COVID-19 — Casos:** Secretaria da Saúde do Estado do Ceará (SESA) / IntegraSUS / Indicadores de COVID-19 (Boletins Epidemiológicos *semanais*, 2020–2024). A consolidação por mês e por bairro foi construída a partir das semanas epidemiológicas informadas nos boletins.
- **COVID-19 — Óbitos:** Prefeitura Municipal de Fortaleza (PMF) / Secretaria Municipal da Saúde (SMS) / Coordenadoria de Vigilância em Saúde (COVIS) / Célula de Vigilância Epidemiológica (CEVEPI) / Coordenadorias Regionais de Saúde (CORES).
- **Demografia e socioeconomia:** Instituto Brasileiro de Geografia e Estatística (IBGE) (Censo Demográfico 2022): população residente, densidade demográfica, domicílios (particulares e coletivos) e média de moradores por domicílio (por bairro).
- **Infraestrutura urbana:** Prefeitura de Fortaleza: localização e distribuição de equipamentos (hospitais, clínicas, Unidade de Atenção Primária à Saúde (UAPS), Unidade de Pronto Atendimento (UPA), Centro de Atenção Psicossocial (CAPS)), escolas (rede municipal

e estadual), universidades, mercados e feiras, centros culturais e esportivos (Rede Cuca, areninhas, praças, centros de juventude, Kite, Casas de Cultura Digital).

- **Educação (composição de infraestrutura):** Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP): presença de centros universitários públicos e privados por bairro.
- **Índices compostos:** Atlas do Desenvolvimento Humano (IBGE/Programa das Nações Unidas para o Desenvolvimento (PNUD), 2010) para o IDH por bairro (base mais recente disponível nessa escala).
- **Dados políticos:** Tribunal Regional Eleitoral (TRE)/Tribunal Superior Eleitoral (TSE) (2020 e 2022, 1º e 2º turnos) — distribuição de votos por candidato, agregada por bairro.

#### 4.1.0.0.1 Organização e créditos institucionais.

Relatórios e boletins municipais contaram com a coordenação e revisão técnica das seguintes instâncias e profissionais: *PMF, SMS, COVIS, CEVEPI, CORES*; Evandro Sá Barreto Leitão; Maria do Perpétuo Socorro Martins Breckenfeld; Josete Malheiro Tavares; Rui de Gouveia Soares Neto; Geziel dos Santos de Sousa. A extração, organização, normalização e revisão da base analítica a partir dos boletins semanais foram realizadas no âmbito desta pesquisa, preservando fontes e versões.

#### 4.1.0.0.2 Observações.

(i) Para dengue, zika e chikungunya, os dados seguem o mesmo procedimento de agregação por bairro e mês; entretanto, zika possui série pequena e menor disponibilidade, o que limita análises detalhadas. (ii) Quando necessário, foram aplicados ajustes de consistência temporal (mapeamento semana epidemiológica → mês) e padronização de nomes de bairros. (iii) O IDH utilizado é o de 2010, último censo disponível com divulgação e listagem por bairro pela Prefeitura de Fortaleza.

#### 4.1.0.1 Considerações Éticas

Este estudo não envolveu a coleta de dados diretamente de seres humanos, nem a realização de intervenções ou experimentos clínicos. Todas as informações utilizadas foram obtidas a partir de bases de dados públicas e oficiais, disponíveis em sites institucionais de acesso

livre. Por essa razão, não houve necessidade de submissão ou aprovação prévia por Comitê de Ética em Pesquisa, conforme as diretrizes da Resolução CNS *n* 510/2016.

#### **4.1.1 Variáveis selecionadas, limitações e justificativas**

Este estudo usa variáveis em camadas para descrever os bairros e suas relações. A seleção priorizou: (i) disponibilidade e qualidade dos dados por bairro; (ii) comparabilidade ao longo do tempo; (iii) utilidade para análise espacial e redes multilayer; e (iv) relevância para gestão pública.

##### 4.1.1.0.1 Variáveis utilizadas.

- **Epidemiológicas:** casos confirmados mensais por bairro (COVID-19, dengue, zika e chikungunya). Óbitos foram coletados, mas não entraram na modelagem final.
- **Socioeconômicas e demográficas:** população total, densidade demográfica, número de domicílios particulares e coletivos, média de moradores por domicílio.
- **Infraestrutura urbana:** contagem de hospitais, clínicas, escolas, universidades, mercados, feiras livres, centros culturais e esportivos por bairro.
- **Políticas:** distribuição de votos por candidato (eleições municipais de 2020 e presidenciais de 2022, 1º e 2º turnos).
- **Índices compostos:** Índice de Desenvolvimento Humano (IDH).

##### 4.1.1.0.2 Por que estas variáveis.

- **Aderência ao objetivo:** permitem observar como características dos bairros se associam à difusão de doenças ao longo do território.
- **Integração multilayer:** cada grupo forma uma camada; as relações entre camadas (p. ex., epidemiológica ↔ socioeconômica/infraestrutura/política/IDH) são usadas para construir conexões intercamadas.
- **Comparabilidade e replicação:** são dados oficiais e/ou consolidados por bairro, favorecendo replicação e uso por gestores.
- **Interpretação:** variáveis simples de explicar (contagens, proporções e índices conhecidos) aumentam a utilidade prática.

#### 4.1.1.0.3 Limitações e decisões de escopo.

- **Zika:** a série é pequena e com baixa disponibilidade; por isso, resultados específicos foram suprimidos ou pouco detalhados.
- **Óbitos:** não utilizados na modelagem final para evitar ruído por subregistro e baixa frequência mensal por bairro.
- **Agregação temporal mensal:** facilita comparação entre doenças, mas suaviza picos curtos e pode perder detalhes de sazonalidade.
- **Infraestrutura por contagem:** número de equipamentos não mede qualidade, capacidade ou cobertura de fato; é um proxy estrutural.
- **Dados políticos:** resultados eleitorais resumem preferências territoriais, mas não capturam motivação do voto nem mudanças entre pleitos.
- **Vizinhança espacial:** a estrutura intracamadas usa contiguidade por limites oficiais de bairro; isso não representa fluxos reais de mobilidade.
- **Escolhas metodológicas:** não segmentamos por “ondas” de COVID-19 para manter o foco em padrões territoriais estáveis e comparáveis entre doenças.

Em resumo, estas variáveis equilibram disponibilidade, comparabilidade e valor explicativo. Elas permitem construir a rede multilayer por bairro e investigar como diferentes dimensões do território se relacionam com a difusão de *COVID-19*, *dengue*, *zika* e *chikungunya*.

#### 4.1.2 Preparação e Integração dos Dados

A preparação e integração dos dados consistiu em etapas essenciais para garantir a qualidade e consistência necessárias para as análises subsequentes. Estas etapas foram realizadas com rigor metodológico, visando assegurar resultados confiáveis e reprodutíveis.

##### 4.1.2.1 Limpeza e Integração dos Dados

Nesta etapa inicial, os dados brutos provenientes das diferentes fontes foram submetidos a um processo minucioso de limpeza e padronização, incluindo:

- **Verificação de consistência:** identificação e correção de erros, valores ausentes ou inconsistentes;
- **Padronização de formatos:** Ajuste das unidades de medida, formatos de data e categorização das variáveis para permitir integração eficiente;

- **Tratamento de dados faltantes e outliers:** aplicação de técnicas estatísticas e critérios objetivos para lidar com valores ausentes ou extremos;
- **Harmonização de nomenclaturas de bairros:** nos últimos anos, alguns bairros foram criados, subdivididos ou renomeados, e diferentes fontes adotam variações ortográficas (como "Rachel" e "Raquel"). Foi necessário realizar um processo manual de correspondência e unificação para garantir consistência e precisão na integração territorial;
- **Construção de uma base unificada:** integração das diversas fontes em um banco de dados consolidado e estruturado por bairro e por período temporal.

#### 4.1.2.2 *Análise exploratória inicial e seleção de variáveis*

Após a limpeza e integração por bairro/mês, conduzimos uma análise exploratória com foco em significado substantivo, parcimônia e utilidade prática. A seleção combinou julgamento empírico (o que faz sentido para o problema) com regras simples e verificações gráficas/numéricas. Como princípios gerais, (i) priorização de variáveis com interpretação direta e relevância para leitura territorial; (ii) remoção de redundância e sobreposição de informação; (iii) manutenção de apenas o que é replicável com dados oficiais.

##### 4.1.2.2.1 Regras aplicadas.

- **Completeness:** exclusão de variáveis com > 20% de ausências.
- **Redundância:** remoção de colunas fortemente colineares ou duplicadas em sentido (ex.: votos no 1º e 2º turno quando traziam a mesma informação prática).
- **Baixa utilidade:** exclusão de campos pouco informativos para a pergunta de pesquisa (ex.: votos nulos), ou cuja relação com o desfecho não era plausível/útil no contexto.
- **Decisões específicas:** óbitos de COVID-19 foram retirados da modelagem final (baixa frequência mensal por bairro); alguns itens de infraestrutura foram descartados quando não apresentavam relação plausível com os desfechos analisados.

##### 4.1.2.2.2 Checagens exploratórias (enxutas).

- **Visual:** histogramas/boxplots para entender escala, assimetria e valores extremos; painéis comparativos (*heatmaps* simples) para enxergar padrões gerais.
- **Numérica:** correlações iniciais (Pearson/Spearman) apenas para sinalizar redundâncias

evidentes e orientar escolhas; quando havia pares muito próximos, manteve-se a variável mais simples de interpretar e com melhor cobertura.

O conjunto original (quase 100 atributos) foi reduzido a pouco mais de 20 variáveis, preservando representatividade entre camadas (epidemiológica, socioeconômica/demográfica, infraestrutura, política/IDH) e favorecendo modelos mais claros, comparáveis e úteis para leitura territorial.

#### 4.1.2.3 *Categorização das Variáveis Políticas*

Para analisar a influência das variáveis políticas na propagação da COVID-19 e demais doenças investigadas, organizamos os dados eleitorais conforme os critérios a seguir:

- **Classificação dos Candidatos:** Selecionamos apenas os dois principais candidatos a prefeito e presidente de cada eleição (municipal de 2020 e presidencial de 2022).
- **Categorização Política:** Os candidatos foram agrupados em três categorias: direita, esquerda e centro.
- **Votos do Primeiro Turno:** Foram considerados apenas os votos do primeiro turno, com o objetivo de simplificar as análises e evitar interferências decorrentes de alianças políticas formadas no segundo turno. Ademais, observou-se uma alta correlação entre os resultados de ambos os turnos, reforçando essa escolha metodológica.
- **Variável de Controle:** As candidaturas de centro foram utilizadas exclusivamente como variável de controle, com o intuito de isolar os efeitos da polarização política (direita versus esquerda).

Diferentemente da eleição presidencial de 2022, que apresentou uma clara polarização entre candidatos de direita e esquerda, as eleições municipais de 2020 exibiram maior complexidade, com muitos partidos e candidatos não alinhados às divisões ideológicas tradicionais. Para contornar essa limitação, adotamos a classificação proposta pelo Observatório das Eleições de 2020 (BOLOGNESI *et al.*, 2020), elaborada com base em estudos de universidades brasileiras. Essa abordagem posiciona os partidos em um espectro político que varia da esquerda à direita, incluindo uma posição intermediária de centro. Adicionalmente, realizamos verificações manuais no site do TSE para ajustar alterações nos nomes e siglas dos partidos entre 2020 e 2024.

#### 4.1.2.4 Escolha da Métrica de Risco Epidemiológico

A padronização do risco epidemiológico foi uma etapa essencial para garantir a comparabilidade entre bairros com diferentes tamanhos populacionais. Para isso, adotou-se o *Risco Relativo* (RR) como principal métrica de normalização dos dados de incidência. O RR é um indicador amplamente utilizado na epidemiologia por expressar a razão entre a incidência observada de uma doença em um determinado grupo (neste caso, um bairro) e a incidência média na população de referência (o conjunto dos bairros de Fortaleza), conforme ilustrado na Equação 4.1 e detalhado no Anexo E.

$$RR = \frac{\text{Incidência no bairro}}{\text{Incidência média na cidade}} \quad (4.1)$$

Essa abordagem permite mitigar o efeito da variação populacional entre bairros, destacando áreas com risco desproporcionalmente alto ou baixo. Além de sustentar as análises estatísticas e de aprendizado de máquina, os valores de RR foram fundamentais na construção da rede *multilayer*, servindo como base para a definição de conexões intra-layer e identificação de padrões espaciais de risco.

Com o objetivo de facilitar a interpretação dos resultados, os valores de RR foram categorizados em três níveis com base nos quartis: **alto** ( $\geq Q3$ , 25% superior), **médio** (entre  $Q1$  e  $Q3$ , 50% central) e **baixo** ( $\leq Q1$ , 25% inferior). Essa classificação permitiu detectar outliers espaciais e controlar a influência de valores extremos, contribuindo para uma análise robusta e interpretável do padrão territorial das doenças.

## 4.2 Análise Estatística

A etapa de análise estatística teve como objetivo explorar relações entre variáveis epidemiológicas, demográficas, socioeconômicas, políticas e de infraestrutura urbana. Essa fase inicial permitiu identificar associações relevantes e orientar a seleção de variáveis para as demais etapas da pesquisa, como a modelagem preditiva, a construção da rede *multilayer* e a análise espacial.

### 4.2.1 *Correlação de Pearson*

Foi utilizada a correlação de Pearson (PEARSON, 1895) para mensurar a associação linear entre pares de variáveis contínuas. As variáveis epidemiológicas consideradas foram os valores de risco relativo (RR) de COVID-19, dengue, zika e chikungunya por bairro. Essas variáveis foram correlacionadas com indicadores estruturais como densidade populacional, número de domicílios, infraestrutura de saúde, educação e cultura, indicadores de desenvolvimento humano (IDH), e variáveis de comportamento eleitoral (percentual de votos para a esquerda/direita em eleições municipais e presidenciais).

A análise de correlação foi conduzida com o auxílio da biblioteca pandas em Python, com testes de significância aplicados ao nível de 5%. Foram geradas matrizes de correlação e gráficos de dispersão para melhor visualização dos padrões observados (mais detalhes no Capítulo 5).

### 4.2.2 *Regressão Linear Múltipla (OLS)*

Para investigar a influência conjunta de variáveis explicativas sobre os níveis de risco relativo de cada doença, foi utilizado modelos de regressão linear múltipla (OLS — Ordinary Least Squares) (MONTGOMERY *et al.*, 2012). Cada modelo teve como variável dependente o RR de uma das doenças, enquanto as variáveis independentes foram selecionadas a partir das correlações mais relevantes, com atenção à ausência de multicolinearidade.

A especificação dos modelos seguiu os critérios de parcimônia e significância estatística, e os testes de diagnóstico incluíram análise de resíduos e avaliação do coeficiente de determinação ( $R^2$ ). Os modelos foram implementados em Python, utilizando as bibliotecas statsmodels e scikit-learn.

## 4.3 **Análise Espacial**

A análise espacial foi incorporada à metodologia com o objetivo de capturar a dependência espacial entre bairros e identificar padrões geográficos de risco para as quatro doenças analisadas. Essa abordagem permitiu avaliar a existência de clusters espaciais, bem como a variação local na influência de fatores explicativos, contribuindo para uma compreensão territorialmente sensível da dinâmica epidemiológica.

#### **4.3.1 Autocorrelação Espacial Global (Índice de Moran)**

Inicialmente, foi aplicado o Índice de Moran Global (Global Moran's I) (MORAN, 1950) para cada uma das doenças, utilizando valores de risco relativo (RR) calculados previamente. Essa métrica quantifica a autocorrelação espacial global, avaliando se os valores observados em uma unidade espacial (bairro) estão correlacionados com os de seus vizinhos.

A matriz de vizinhança foi construída com base no critério de contiguidade (queen), e os testes de significância foram realizados por permutação Monte Carlo (999 permutações), conforme implementado na função `Moran_Local` do subpacote `esda` (REYES *et al.*, 2020), parte da biblioteca `PySAL` (REY *et al.*, 2020) em Python.

#### **4.3.2 Autocorrelação Espacial Local (LISA)**

Para identificar agrupamentos locais de alto e baixo risco, foi utilizada a técnica LISA (ANSELIN, 1995), que fornece medidas locais de autocorrelação espacial. Os resultados da LISA permitem classificar os bairros em categorias como Alto-Alto, Baixo-Baixo, Alto-Baixo e Baixo-Alto, sinalizando áreas com risco elevado (*hotspots*), risco reduzido (*coldspots*) ou discrepâncias em relação ao entorno.

#### **4.3.3 Análise de Hotspots Acumulados**

Como etapa complementar, foi realizada uma contagem da frequência com que cada bairro foi classificado como hotspot nas análises LISA das quatro doenças. Essa síntese gerou um índice de risco acumulado que reflete o número de vezes que um bairro apresentou agrupamento estatisticamente significativo de alto risco, fornecendo uma visão integrada da carga epidemiológica territorial.

#### **4.3.4 Considerações Técnicas**

Todas as análises espaciais foram realizadas em ambiente Python, com suporte das bibliotecas `geopandas`, `pysal`, `esda`, `libpysal` e `mgwr`. Os mapas foram produzidos com o auxílio da biblioteca `matplotlib`, respeitando projeções cartográficas adequadas ao município de Fortaleza (EPSG: 31985).

A incorporação da análise espacial à metodologia da tese permite uma abordagem mais sensível ao território, fundamental para o entendimento da dinâmica de doenças infecciosas

em contextos urbanos marcados por desigualdade estrutural.

#### 4.4 Modelagem de Redes Multilayer

As redes *multilayer* (multicamadas) foram utilizadas como abordagem principal para organizar e analisar, em camadas distintas, as dimensões epidemiológica, socioeconômica/demográfica, de infraestrutura e política/IDH. Nota: tecnicamente, a estrutura adotada repete o mesmo conjunto de nós (bairros) em todas as camadas e conecta a mesma réplica de bairro entre camadas; isso caracteriza uma rede do tipo *multiplex*. Para manter o texto simples e direto, empregamos ao longo da tese o termo genérico *multilayer*, registrando aqui essa caracterização técnica.

Essa organização em camadas permite representar de forma integrada múltiplos fatores associados à difusão de doenças, preservando (i) relações *intracamadas* (dentro de cada dimensão) e (ii) relações *intercamadas* (entre dimensões). O objetivo é aumentar a comparabilidade territorial e apoiar leituras que possam ser úteis à gestão pública.

##### 4.4.1 Estrutura da rede

Para construir a rede, os dados foram pré-processados por bairro (padronização/normalização quando necessário) e organizados em nove camadas temáticas: quatro epidemiológicas (COVID-19, dengue, zika e chikungunya) e cinco estruturais (IDH, infraestrutura urbana, perfil político municipal, perfil político presidencial e indicadores demográficos). Em todas as camadas, os nós são os mesmos bairros de Fortaleza.

As conexões (arestas) foram definidas e ponderadas conforme a natureza da camada:

- **Nós:** bairros de Fortaleza, presentes em todas as camadas temáticas da rede *multilayer*.
- **Arestas intracamadas — camadas epidemiológicas:** conectam pares de bairros dentro de cada doença, a partir da similaridade de risco relativo (RR) mensal agregado. Bairros com perfis semelhantes de RR (altos entre si ou baixos entre si) são conectados. (Observação: quando usada uma medida de diferença como peso, menores diferenças indicam maior similaridade; em análises que exigem pesos diretamente proporcionais à força do vínculo, essa diferença é convertida em uma medida de similaridade ou reescalada.)
- **Arestas intracamadas — camadas política, socioeconômica/demográfica e infraestrutura:** definidas por similaridade entre vetores de atributos do bairro. Utilizou-se uma regra

de esparsidade por limiar percentílico: calcula-se a distância (por exemplo, euclidiana) entre todos os pares de bairros; apenas os pares abaixo de um limiar (p.ex., percentil 20) são conectados, preservando os vínculos mais fortes e evitando uma rede excessivamente densa.

- **Arestas intercamadas:** conectam a mesma representação do bairro entre camadas distintas (por exemplo, epidemiológica  $\leftrightarrow$  infraestrutura). Os pesos refletem associação entre variáveis temáticas (por exemplo, correlação de Pearson), considerando apenas associações positivas e/ou estatisticamente consistentes conforme definido na etapa de pré-análise.

A rede final resultou em 1,089 nós e 21,666 arestas, distribuídas entre ligações intra e intercamadas.

#### 4.4.2 Detecção de comunidades

A identificação de comunidades foi realizada com o algoritmo *Louvain* (BLONDEL *et al.*, 2008) aplicado por camada, de forma independente em cada dimensão (epidemiológica, socioeconômica/demográfica, infraestrutura e política/IDH), com ponderação quando os pesos de aresta estavam definidos. Não executamos otimização de modularidade em um supra-grafo multilayer.

A leitura multilayer das comunidades decorre da comparação das partições obtidas em cada camada, por meio de: (i) visualização cartográfica lado a lado das partições por camada, destacando convergências/divergências territoriais; e (ii) tabelas simples de cruzamento (contagens/percentuais) quando necessário para apoiar a interpretação. Assim, preservamos o uso adequado do Louvain *por camada* e mantemos a noção de multilayer para organizar e comparar resultados entre dimensões, sem otimização conjunta de modularidade.

#### 4.4.3 Análise de centralidade e estrutura multidimensional

Para avaliar a importância estrutural dos bairros, utilizamos métricas clássicas:

- **Grau ponderado (Weighted Degree):** intensidade total das conexões de um nó.
- **Centralidade de autovetor (Eigenvector):** influência de um bairro considerando a influência de seus vizinhos.
- **Centralidade de intermediação (Betweenness):** papel de “ponte” de um bairro em caminhos mínimos entre outros bairros.

Essas métricas foram calculadas *por camada* (quando pertinente) e interpretadas em

conjunto com os resultados de comunidades e os vínculos intercamadas, mantendo o foco em padrões territoriais comparáveis entre dimensões.

#### 4.5 Modelos de Machine Learning para Predição e Clusterização

Com o objetivo de prever padrões epidemiológicos e identificar agrupamentos territoriais com perfis similares, esta pesquisa incorporou técnicas de *Machine Learning* supervisionadas e não supervisionadas. O uso dessas abordagens permite não apenas a identificação de relações não lineares entre variáveis socioestruturais e risco de doenças, como também a síntese de tipologias urbanas com base em múltiplas dimensões.

A integração de técnicas de *Machine Learning* supervisionadas e não supervisionadas permitiu combinar previsão e segmentação territorial de forma robusta. Os modelos preditivos forneceram insights sobre os principais determinantes das doenças, enquanto a clusterização permitiu sintetizar a heterogeneidade urbana em perfis acionáveis para a gestão pública. A abordagem adotada reforça o potencial de métodos computacionais avançados para subsidiar políticas de saúde baseadas em evidências e sensíveis ao território.

##### 4.5.1 Modelos Supervisionados para Predição do Risco Relativo

Modelos supervisionados foram aplicados para estimar o risco relativo (RR) das quatro doenças analisadas com base em variáveis explicativas relacionadas à infraestrutura urbana, indicadores socioeconômicos, densidade populacional e perfil político dos bairros. Os algoritmos utilizados foram:

- **XGBoost (Extreme Gradient Boosting)** (CHEN; GUESTRIN, 2016): modelo baseado em árvores de decisão, eficiente para capturar interações complexas e efeitos não lineares entre variáveis. Seu desempenho foi avaliado com base em métricas como erro quadrático médio (RMSE) e coeficiente de determinação ( $R^2$ ).
- **Random Forest** (BREIMAN, 2001): algoritmo baseado na combinação de múltiplas árvores de decisão. Além da capacidade preditiva, permitiu avaliar a importância relativa das variáveis no modelo.

#### 4.5.1.1 *Análise de Explicabilidade com SHAP*

Para interpretar os modelos supervisionados utilizados na predição do risco relativo das doenças, foi empregada a técnica SHAP (SHapley Additive exPlanations), proposta por (LUNDBERG; LEE, 2017). Essa abordagem, baseada na teoria dos jogos de Shapley, permite decompor a predição de um modelo em contribuições atribuíveis a cada variável explicativa, oferecendo uma explicação local e global do comportamento do modelo.

A aplicação do SHAP possibilitou a identificação das variáveis com maior influência na predição do risco relativo da COVID-19. Entre os principais preditores identificados destacam-se o Índice de Desenvolvimento Humano voltado à educação (IDH-Educação), o IDH geral e variáveis políticas, como o percentual de votos para candidatos de direita nas eleições presidenciais. Além disso, fatores demográficos — como a média de moradores por domicílio e a presença de domicílios coletivos — também contribuíram de forma relevante para as decisões dos modelos. Essa etapa foi fundamental para a validação interpretativa das relações detectadas pelos algoritmos de aprendizado de máquina.

#### 4.5.2 *Clusterização Territorial com K-means*

Como etapa complementar à modelagem supervisionada, foi empregada uma abordagem de aprendizado não supervisionado com o objetivo de agrupar bairros com perfis socioestruturais e epidemiológicos semelhantes. Para isso, foi utilizado o algoritmo K-means (LLOYD, 1982), amplamente adotado para particionamento de dados com base em similaridade multivariada. Neste estudo, o número de *clusters* foi definido como  $k = 4$ , de modo a captar diferentes perfis territoriais sem comprometer a interpretabilidade dos grupos formados.

O algoritmo K-means envolve etapas aleatórias, especialmente na inicialização dos centróides, o que pode gerar variações nos resultados a cada execução. Para garantir a reprodutibilidade das análises e a consistência dos agrupamentos obtidos, foi fixado o valor `random_state = 42`, que define uma semente pseudoaleatória. Essa prática assegura que o processo de inicialização e, conseqüentemente, os resultados finais do algoritmo, sejam determinísticos e replicáveis por outros pesquisadores com os mesmos dados de entrada.

## 5 RESULTADOS

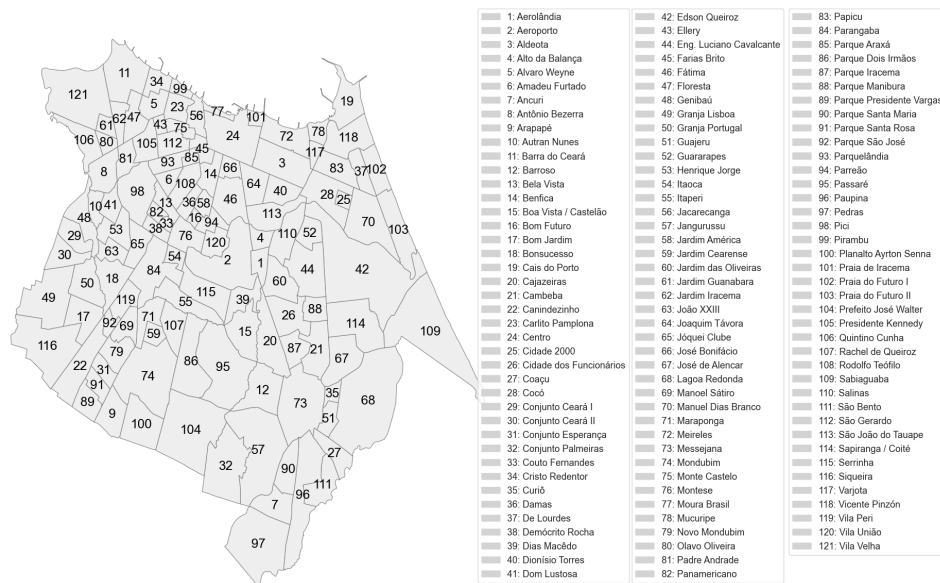
Este capítulo apresenta os principais resultados obtidos por meio da aplicação das metodologias descritas anteriormente. As análises foram conduzidas com o objetivo de caracterizar padrões territoriais de risco epidemiológico, identificar relações estruturais entre variáveis urbanas e doenças, e propor classificações espaciais úteis para o planejamento de políticas públicas. Os resultados estão organizados nas seguintes seções: análise estatística, análise espacial, modelagem de redes *multilayer*, modelagem preditiva com aprendizado de máquina e análise de agrupamentos espaciais (clustering).

### 5.1 Panorama Territorial dos Bairros de Fortaleza

Antes de apresentar os resultados das análises é importante oferecer uma visão geral da divisão territorial de Fortaleza. A capital do estado do Ceará é composta por 121 bairros oficialmente reconhecidos, distribuídos entre regiões com características socioeconômicas e urbanas bastante heterogêneas.

A Figura 5 apresenta a delimitação espacial dos bairros utilizada neste estudo, com os respectivos identificadores numéricos e nomes. Essa codificação foi empregada ao longo de toda a pesquisa para mapear os padrões territoriais de risco, infraestrutura, indicadores sociais e comportamento político.

Figura 5 – Mapa dos 121 bairros de Fortaleza com identificadores numéricos.



Fonte: elaborado pelo autor.

A área central concentra bairros históricos e comerciais como Centro, Meireles, Aldeota e Benfica, que apresentam maior densidade populacional, melhores índices de desenvolvimento humano (IDH) e infraestrutura urbana consolidada. Em contraste, as áreas periféricas localizadas nos extremos oeste, sul e leste da cidade — como Conjunto Palmeiras, Ancuri, Pirambu, Canindezinho e Siqueira — são marcadas por padrões de ocupação mais recentes e menor oferta de equipamentos públicos e serviços essenciais. Entre esses dois polos, encontram-se bairros intermediários com características mistas, como Parangaba, José Walter, Messejana e Montese. Essa diversidade territorial configura um mosaico urbano complexo, no qual coexistem diferentes níveis de vulnerabilidade social, acesso a serviços, e exposição ao risco de doenças, justificando abordagens analíticas integradas e espacialmente sensíveis.

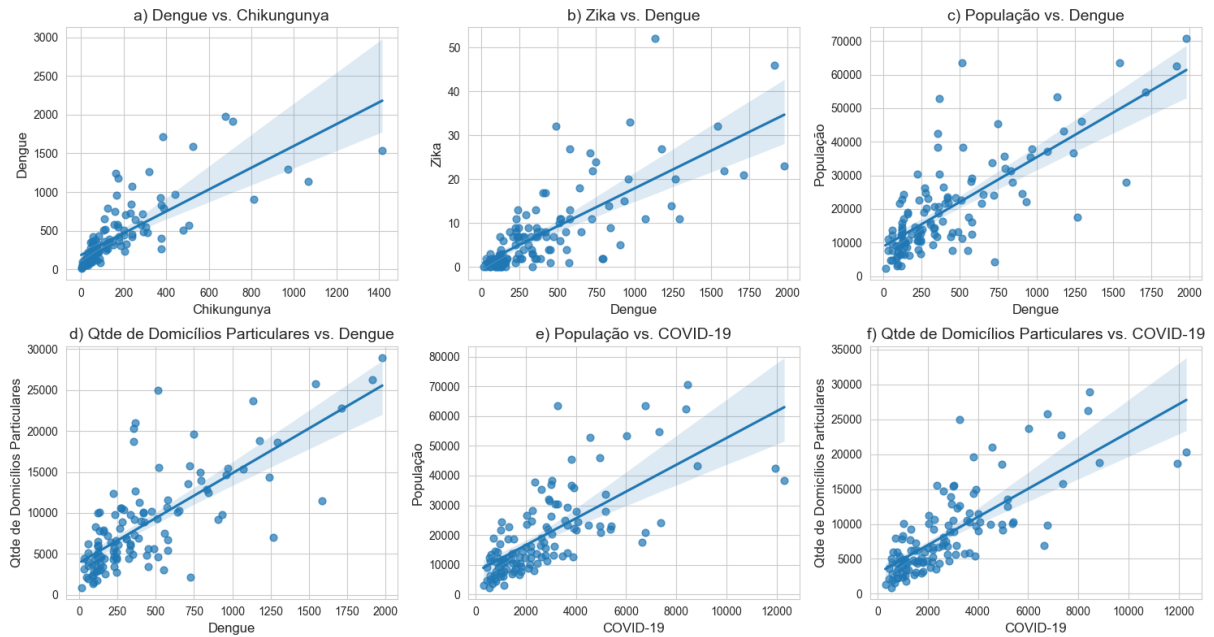
## **5.2 Análise Estatística**

Nesta seção, são apresentados os resultados da análise estatística descritiva e inferencial aplicada às variáveis epidemiológicas e estruturais dos bairros de Fortaleza. O objetivo principal foi explorar possíveis associações entre os fatores socioeconômicos, demográficos e políticos com os padrões de incidência de COVID-19, dengue, zika e chikungunya.

### **5.2.1 Correlação de Pearson**

A análise de correlação de Pearson permitiu identificar relações lineares significativas entre as variáveis investigadas, revelando padrões estruturais e comportamentais relevantes no contexto urbano de Fortaleza. Tais padrões podem ser visualizados nos diagramas de dispersão apresentados na Figura 6, que ilustram claramente as tendências positivas entre essas variáveis. Dentre os principais achados, destacam-se as fortes correlações entre as arboviroses, especialmente entre dengue e chikungunya ( $r = 0,76$ , Fig. 6a), e entre dengue e zika ( $r = 0,74$ , Fig. 6b). Esses resultados sugerem não apenas uma possível sobreposição de áreas de transmissão, mas também indicam a atuação de determinantes vetoriais e ambientais comuns, como a presença do mosquito *Aedes aegypti*, infraestrutura urbana precária e condições climáticas favoráveis à reprodução do vetor.

Figura 6 – Diagramas de dispersão com as correlações altas.



Fonte: elaborado pelo autor.

Em relação às variáveis demográficas, observou-se uma correlação positiva e expressiva entre o número de habitantes e os casos de dengue ( $r = 0,76$ , Fig. 6c), bem como entre o número de domicílios particulares e os casos de dengue ( $r = 0,75$ , Fig. 6d). Esses achados evidenciam o papel crítico da densidade populacional e da configuração espacial dos domicílios na propagação de doenças infecciosas, especialmente em contextos urbanos densamente ocupados.

A COVID-19, especificamente, também apresentou associação com a população total do bairro ( $r = 0,71$ , Fig. 6e) e entre o número de domicílios particulares ( $r = 0,76$ , Fig. 6f), o que sugere que bairros mais populosos estiveram mais expostos ao vírus, provavelmente por concentrarem maior fluxo de pessoas, comércio essencial, transporte coletivo e serviços de saúde, fatores que aumentam o risco de contágio em ambientes urbanos. Esse resultado destaca a vulnerabilidade estrutural dos grandes centros populacionais frente a pandemias, exigindo estratégias específicas de mitigação baseadas em planejamento territorial e monitoramento geoespacial. Resultado completo pode ser visto nos Anexos A, B e C.

Além das correlações entre variáveis de dimensões diferentes, observou-se correlações altas em indicadores semelhantes. Por exemplo, no campo dos indicadores socioeconômicos, observou-se correlação significativa entre os componentes do Índice de Desenvolvimento Humano (IDH), como entre IDH-Educação e IDH-Longevidade ( $r = 0,72$ ), e entre IDH-Renda e o IDH geral ( $r = 0,87$ ). Tais correlações confirmam a natureza multidimensional do desenvolvi-

mento humano e sua interdependência. As variáveis político-eleitorais também apresentaram correlações notáveis. Houve forte associação entre os votos em candidatos de esquerda para prefeito e presidente ( $r = 0,85$ ), bem como entre os votos nos candidatos da direita nas duas esferas ( $r = 0,81$ ), sugerindo uma consistência ideológica espacialmente distribuída nas preferências eleitorais da população. A correlação negativa entre votos à esquerda e à direita nas eleições municipais ( $r = -0,89$ ) reforça esse padrão. Tal divisão pode estar relacionada a fatores socioeconômicos, identitários ou históricos, e pode influenciar tanto a adesão a políticas públicas quanto a percepção de risco frente às doenças.

Além das correlações fortes apresentadas anteriormente, um conjunto expressivo de associações de intensidade moderada também foi identificado (correlação de Pearson em módulo entre 0,3 e 0,7). Esses resultados intermediários, listados no Anexo B, revelam padrões complementares que ajudam a aprofundar a compreensão das dinâmicas sociais, demográficas e estruturais associadas à distribuição das doenças. Destaca-se, por exemplo, a correlação entre chikungunya e variáveis de infraestrutura urbana como saúde ( $r = 0,45$ ), cultura ( $r = 0,53$ ) e mercado ( $r = 0,43$ ), sugerindo que a presença de equipamentos urbanos pode estar associada à maior exposição da população ao vetor, seja por aumento da mobilidade, aglomeração de pessoas, ou condições ambientais favoráveis. De forma similar, zika e COVID-19 também apresentaram correlações moderadas com essas variáveis, indicando que fatores de urbanização e ocupação do espaço público podem afetar múltiplas doenças de maneira semelhante.

Entre os fatores socioeconômicos e políticos, observou-se correlação negativa entre o número médio de moradores por domicílio e os componentes do IDH, especialmente com o IDH de longevidade ( $r = -0,62$ ) e educação ( $r = -0,57$ ), apontando que bairros com maior adensamento intradomiciliar tendem a apresentar menores indicadores de desenvolvimento. No campo político, destaca-se a correlação negativa entre renda e votação em candidatos de esquerda para presidente ( $r = -0,35$ ), sugerindo um padrão ideológico associado a desigualdades econômicas. Já variáveis institucionais, como saúde e educação, mantêm correlações positivas entre si e com outros elementos da infraestrutura (cultura e mercado), indicando uma possível concentração de recursos e serviços em determinadas áreas da cidade.

Esses resultados fornecem evidências iniciais sobre as relações entre variáveis, oferecendo uma base sólida para a modelagem estatística e de redes nas etapas subsequentes. A identificação dessas correlações permite construir hipóteses causais mais refinadas, bem como identificar potenciais mediadores e moderadores na difusão espacial das doenças em Fortaleza.

Além disso, reforça a importância de abordagens interdisciplinares que combinem saúde pública, geografia urbana, ciência política e análise de dados para a compreensão aprofundada da vulnerabilidade urbana em contextos de crise sanitária.

### 5.2.2 Modelagem por Regressão Linear (OLS)

A regressão linear múltipla foi aplicada com o intuito de quantificar a influência de variáveis explicativas sobre a incidência relativa das doenças. Para a COVID-19, o modelo apresentou o melhor ajuste ( $R^2 = 0,808$ ,  $p < 0,001$ ), sendo a variável Educação a única estatisticamente significativa ( $p = 0,048$ ). Este resultado sugere uma possível associação entre a presença de instituições de ensino e fatores como mobilidade urbana, redes de contato social e centralidade funcional dos bairros com maior oferta educacional.

No caso da dengue, o modelo também apresentou bom desempenho ( $R^2 = 0,638$ ,  $p < 0,001$ ), sendo os equipamentos culturais (centros culturais, praças e areninhas) o único preditor significativo ( $p = 0,025$ ). Esse resultado reforça a hipótese de que espaços de convivência e lazer público podem funcionar como locais de exposição ao vetor *Aedes aegypti*, em razão do fluxo intenso de pessoas e da possível presença de criadouros em ambientes abertos e mal monitorados.

Para a chikungunya, o modelo de regressão apresentou desempenho intermediário ( $R^2 = 0,524$ ,  $p < 0,001$ ), indicando que aproximadamente 52% da variância na incidência da doença pode ser explicada pelas variáveis selecionadas. No entanto, a maioria dos preditores não apresentou significância estatística individual. A exceção foi a variável cultura, que se destacou como único preditor estatisticamente significativo ( $p < 0,001$ ), com coeficiente positivo ( $\beta = 16,09$ ), sugerindo que o número de equipamentos culturais está associado ao aumento da incidência da doença. Este achado é coerente com o padrão observado para a dengue e reforça a ideia de que áreas com maior vitalidade urbana, embora potencialmente associadas a benefícios sociais, também podem estar mais expostas ao risco de transmissão vetorial. As demais variáveis, incluindo indicadores demográficos, socioeconômicos e políticos, não demonstraram associação significativa com a chikungunya, o que pode refletir a complexidade da dinâmica de transmissão ou a limitação das variáveis consideradas para capturar plenamente os determinantes dessa arbovirose.

Esse resultado aponta para a necessidade de abordagens que considerem a interação entre fatores ambientais e padrões de mobilidade urbana, além de possíveis análises multiescalares e não-lineares para compreender melhor o comportamento espacial da chikungunya. Também

é relevante considerar que, embora o modelo não apresente alta precisão preditiva, ele revela indícios importantes para políticas de vigilância e intervenção focalizadas em regiões com maior densidade de infraestrutura cultural.

Os resultados estatísticos também sugerem que, embora algumas variáveis estruturais influenciem o padrão espacial das doenças, a magnitude dessas influências varia entre os tipos de infecção. A forte correlação entre arboviroses reforça a hipótese de coexistência de fatores ambientais e infraestruturais comuns. Por outro lado, a COVID-19 apresentou comportamento mais associado a dinâmicas sociais e funcionais centrais da cidade. Estes achados subsidiam as análises subsequentes de natureza espacial e em rede.

### **5.3 Análise Espacial**

Nesta seção, são apresentados os resultados das análises espaciais desenvolvidas com o objetivo de identificar padrões de autocorrelação espacial e áreas prioritárias de intervenção a partir da distribuição territorial dos casos de COVID-19, dengue, zika e chikungunya nos bairros de Fortaleza.

#### **5.3.1 Autocorrelação Espacial Global (Moran's $I$ )**

A estatística de Moran Global ( $I$ ) foi utilizada para medir a presença de autocorrelação espacial entre os valores do risco relativo das doenças nos bairros de Fortaleza. Esse indicador avalia se áreas geograficamente próximas apresentam valores semelhantes (positiva) ou diferentes (negativa) de uma variável, permitindo identificar padrões espaciais agrupados ou dispersos.

Os resultados revelaram autocorrelação espacial positiva e estatisticamente significativa para os casos de COVID-19 ( $I = 0,2558$ ,  $p = 0,001$ ) e chikungunya ( $I = 0,1602$ ,  $p = 0,007$ ), evidenciando a existência de padrões geográficos agrupados na distribuição dessas doenças. No caso da COVID-19, o índice relativamente mais alto sugere uma tendência mais forte à concentração de casos em bairros vizinhos, o que pode estar associado à alta conectividade entre regiões centrais, fluxo intenso de pessoas, e à presença de polos comerciais, educacionais e de saúde. A significativa autocorrelação observada para a chikungunya, embora mais discreta, indica que a doença também tende a se concentrar em determinados territórios, especialmente aqueles com condições socioambientais favoráveis à proliferação do vetor, como densidade

populacional elevada, saneamento precário e presença de criadouros.

Por outro lado, as arboviroses dengue e zika não apresentaram autocorrelação espacial global significativa. Para a dengue, o índice foi positivo, porém próximo de zero ( $I = 0,0225$ ,  $p = 0,217$ ), sugerindo uma distribuição mais aleatória, possivelmente influenciada por surtos localizados ou variações temporais de incidência que não resultaram em agregados espaciais consistentes no período analisado. Já para a zika, o índice foi negativo e não significativo ( $I = -0,0564$ ,  $p = 0,118$ ), indicando uma possível tendência à dispersão ou ausência de padrão espacial definido.

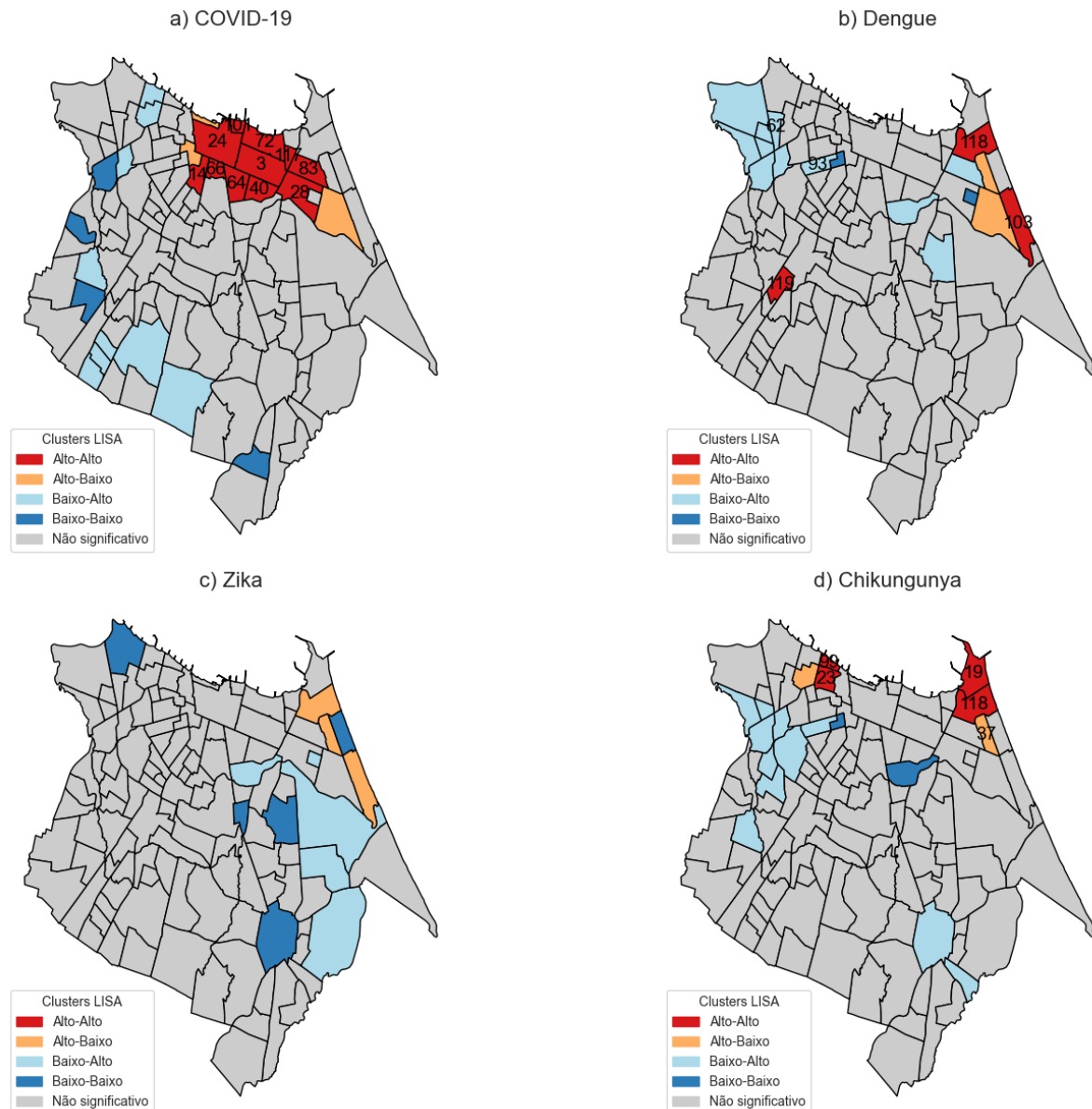
Esses achados reforçam a ideia de que a dinâmica espacial das doenças não é uniforme e responde a fatores específicos de cada enfermidade, como modo de transmissão, suscetibilidade populacional, nível de exposição, infraestrutura urbana e estratégias de vigilância epidemiológica. A presença ou ausência de autocorrelação espacial global fornece um primeiro indicativo sobre a existência de clusters relevantes, que são analisados em maior detalhe na subseção seguinte, por meio da estatística de Moran Local (LISA).

### 5.3.2 Autocorrelação Espacial Local (LISA)

A análise de autocorrelação espacial local (LISA) permitiu identificar agrupamentos espaciais (*clusters*) específicos, classificados como Alto-Alto (*High-High*), Baixo-Baixo (*Low-Low*), Alto-Baixo (*High-Low*) e Baixo-Alto (*Low-High*), indicando, respectivamente, áreas com alta incidência cercadas por áreas também de alta incidência, baixa incidência cercada por áreas de baixa incidência, alta incidência isolada e baixa incidência em meio a áreas com maiores valores. Essa abordagem possibilitou mapear padrões espaciais relevantes para a formulação de estratégias de intervenção mais precisas.

A Figura 7 apresenta os *clusters* formados por LISA, identificados para os casos de COVID-19 (a), dengue (b), zika (c) e chikungunya (d) nos bairros de Fortaleza, permitindo uma comparação visual dos padrões espaciais entre as diferentes doenças. O resultado completo encontra-se no Anexo D. A seguir, descrevemos os resultados obtidos para cada doença.

Figura 7 – Mapas Comparativos de *clusters* Espaciais (LISA).



Fonte: elaborado pelo autor.

Para a COVID-19, os *hotspots* identificados (Alto-Alto) se concentraram em bairros centrais e historicamente mais desenvolvidos de Fortaleza, como Meireles, Aldeota, Centro, Varjota e Praia de Iracema. Esses territórios são marcados por elevada densidade de serviços, maior mobilidade urbana e melhor acesso à testagem, fatores que podem explicar a maior notificação de casos. Bairros como Joaquim Távora, Benfica e Papicu também apareceram como áreas de alta prioridade, reforçando o padrão de concentração em regiões economicamente centrais. Em contraposição, bairros como Álvaro Weyne, Prefeito José Walter e Mondubim apresentaram o padrão Baixo-Alto, sugerindo regiões de menor incidência em meio a áreas com maiores taxas, o que pode indicar zonas de transição territorial ou barreiras estruturais à difusão.

No caso da dengue, observou-se um Moran's I não significativo, o que aponta para uma distribuição difusa ou instável. Ainda assim, a análise local permitiu identificar alguns *clusters* pontuais. Praia do Futuro II e Vicente Pinzón foram classificados como Alto-Alto, enquanto Vila Peri, apesar de aparecer como Alto-Alto e possuir alta incidência, não apresentou contiguidade espacial suficiente para formar conglomerados amplos. Já bairros como Jardim Iracema, Papicu, Parquelândia e Padre Andrade foram classificados como Baixo-Alto, o que significa que, apesar de apresentarem baixa incidência da doença, estão geograficamente próximos a áreas de maior risco. Essa configuração pode indicar episódios pontuais de contenção da transmissão ou refletir fatores microambientais e estruturais que oferecem certa proteção local, como ações de controle vetorial, menor densidade populacional, ou menor presença de criadouros. Alternativamente, esses bairros podem estar sob risco iminente de expansão do contágio a partir dos vizinhos com alta incidência.

Para o zika, a ausência de *clusters* do tipo Alto-Alto reforça a inexistência de um padrão territorial definido (índice global foi negativo e não significativo). Os *clusters* detectados foram pontuais e dispersos, com bairros como De Lourdes e Praia do Futuro II classificados como Alto-Baixo (alta incidência isolada) e Cidade 2000 como Baixo-Alto (baixa incidência cercada por áreas com alta ocorrência). A distribuição fragmentada sugere que a transmissão do vírus pode estar mais associada a condições locais específicas, como microambientes propícios à proliferação do vetor ou à subnotificação dos casos.

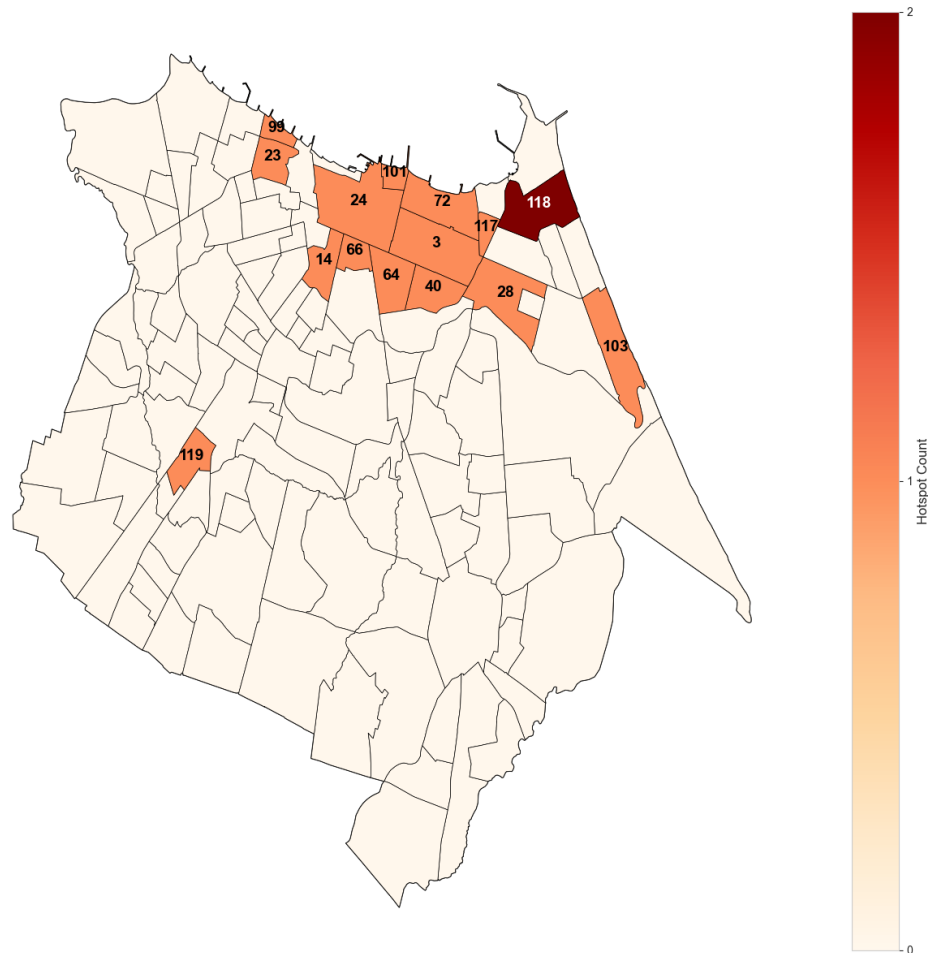
A chikungunya apresentou padrão intermediário, com autocorrelação espacial positiva e estatisticamente significativa ( $I = 0,1602$ ,  $p = 0,007$ ), indicando agrupamentos relevantes de risco. Bairros periféricos como Carlito Pamplona, Pirambu, Vicente Pinzón e Cais do Porto configuraram-se como *hotspots* (Alto-Alto), evidenciando uma territorialidade distinta das demais doenças, concentrada em áreas historicamente vulneráveis. Esses territórios apresentam infraestrutura urbana precária, altas densidades populacionais e condições ambientais favoráveis à reprodução do vetor. Em contraste, bairros como Álvaro Weyne e De Lourdes foram classificados como Alto-Baixo, revelando incidência elevada isolada, cercada por regiões de menor risco, o que pode indicar surtos localizados ou flutuações temporárias de incidência.

### 5.3.3 Síntese dos Hotspots por Doença

Foi realizada uma agregação dos resultados da análise LISA para verificar a frequência com que cada bairro foi classificado como *hotspot* (Alto-Alto) entre as quatro doenças

analisadas. Essa abordagem permitiu identificar áreas com possível acúmulo de risco epidemiológico, bem como avaliar o grau de sobreposição territorial entre os padrões espaciais das doenças.

Figura 8 – Síntese da frequência de *hotspots* (Alto-Alto) por bairro nas análises LISA para as quatro doenças.



Fonte: elaborado pelo autor.

A Figura 8 sintetiza graficamente o número de vezes que cada bairro foi identificado como *hotspot* nas análises LISA das quatro doenças. O mapa acumulado evidencia uma concentração de risco em zonas centrais e litorâneas, tanto na faixa oeste (Pirambu e adjacências) quanto leste (Praia do Futuro II), ainda que com perfis distintos de doença. Os valores variam de 0 a 2, onde 0 representa bairros não classificados como *hotspot* para nenhuma doença, 1 corresponde a *hotspot* para uma única doença, e 2 indica ocorrência de múltiplas doenças. Apenas um bairro, Vicente Pinzón, apresentou valor 2, indicando sua classificação como *hotspot* tanto para dengue quanto para chikungunya, sendo, portanto, um território de atenção prioritária.

Portando, os resultados mostram que a maioria dos bairros apareceu como *hotspot*

em apenas uma doença, e nenhum bairro foi classificado como *hotspot* simultaneamente para todas as quatro enfermidades. Também não foram identificados bairros compartilhados entre COVID-19 e dengue, COVID-19 e zika, ou dengue e zika, o que indica ausência de sobreposição direta entre essas doenças em nível local. Essa fragmentação territorial reforça a hipótese de que os determinantes espaciais de risco são específicos a cada patógeno, refletindo diferenças nos modos de transmissão, infraestrutura, mobilidade urbana, condições socioambientais e acesso a serviços de saúde.

Bairros como Aldeota, Meireles, Varjota e Centro foram classificados como *hotspots* apenas para a COVID-19. Já Praia do Futuro II e Vila Peri apareceram como *hotspots* exclusivamente para a dengue, indicando possíveis focos de proliferação do vetor em áreas específicas do litoral e da periferia. No caso da chikungunya, os *hotspots* foram restritos a bairros com vulnerabilidades sociais mais acentuadas, como Carlito Pamplona, Pirambu e Cais do Porto, situados em áreas periféricas ou de baixa infraestrutura urbana. Nenhum bairro foi classificado como *hotspot* exclusivo para zika, e essa doença não apresentou sobreposição territorial com nenhuma das demais.

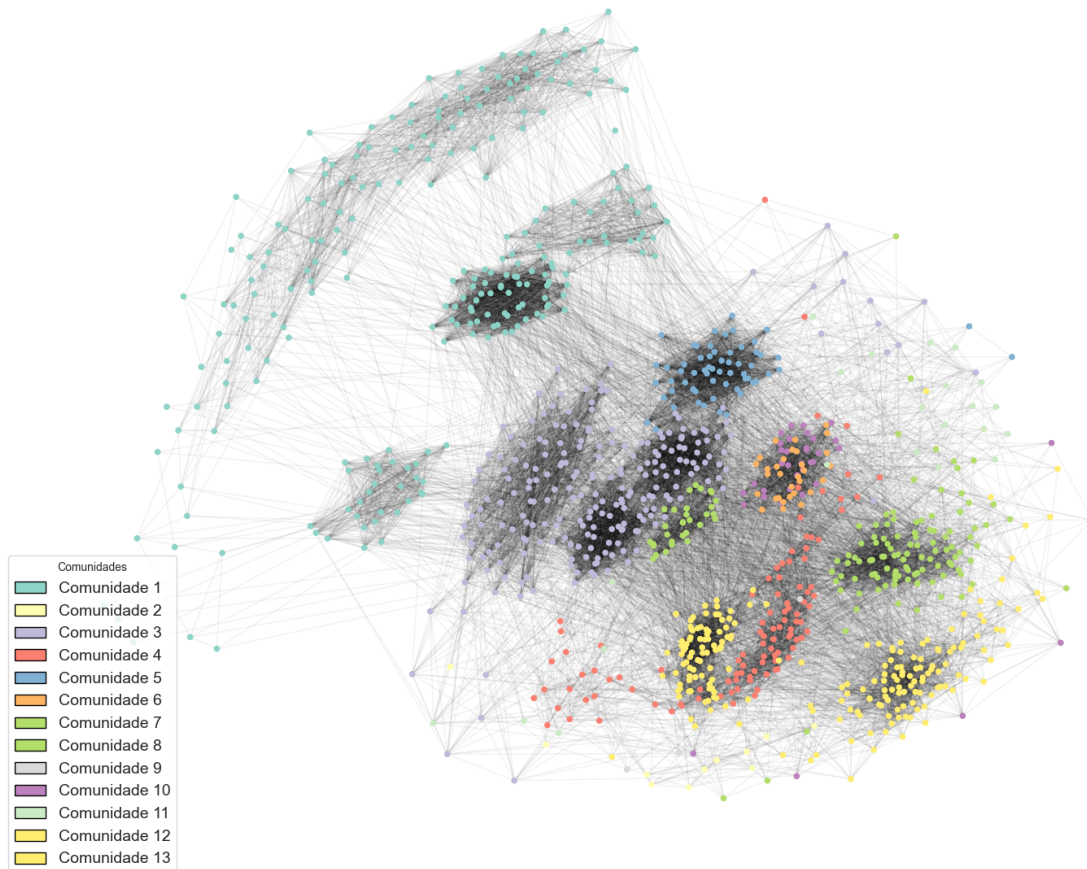
Esses resultados apontam para a necessidade de estratégias territoriais diferenciadas por tipo de doença, considerando as especificidades de cada contexto local. Áreas com múltiplas ocorrências de *hotspots* devem ser tratadas como zonas prioritárias de vigilância integrada, enquanto bairros com padrões isolados podem demandar intervenções direcionadas a fatores ambientais ou estruturais pontuais.

#### 5.4 Redes Multilayer

Nesta seção, são apresentados os resultados da construção e análise da rede *multilayer* que integra múltiplas dimensões — epidemiológica, socioeconômica, política e de infraestrutura — com o objetivo de identificar padrões estruturais e agrupamentos territoriais de risco.

A rede foi construída a partir da agregação de múltiplas camadas temáticas (*layers*), representando diferentes tipos de variáveis associadas aos bairros de Fortaleza. A estrutura resultante, Figura 9, incluiu **1.089 nós** e **21.666 arestas**, compreendendo conexões intra-layer e inter-layer. Essa modelagem permitiu representar, de forma integrada, os vínculos entre fatores epidemiológicos, políticos, estruturais e socioeconômicos, criando um espaço relacional onde padrões territoriais complexos emergem a partir das interações multilaterais.

Figura 9 – Comunidades multicamadas detectadas (Louvain).



Fonte: elaborado pelo autor.

A centralidade dos nós foi avaliada por meio de três métricas complementares: grau ponderado, centralidade de intermediação (*betweenness*) e centralidade por autovetor (*eigenvector centrality*). Essas medidas permitem identificar bairros e dimensões temáticas que desempenham papéis estratégicos na conectividade geral da rede, seja por sua força de ligação (grau), por seu papel como ponte entre diferentes regiões da rede (*betweenness*), ou por sua influência estrutural (*eigenvector*).

#### *Grau Ponderado (Weighted Degree)*

O grau ponderado mede a força total de conexão de cada nó, considerando o peso das arestas (correlação, similaridade, coocorrência, etc.). Na rede *multilayer*, isso reflete o quanto um bairro se conecta fortemente a outros em uma ou mais camadas. No ranking por **grau ponderado**, o bairro Praia do Futuro I aparece como o nó mais conectado da rede, sobretudo na camada epidemiológica de zika (*weighted degree* = 319,46), seguido pelas conexões com

dengue e chikungunya. Esse destaque sugere uma condição de risco acumulado e centralidade funcional no processo de disseminação das arboviroses. Outros bairros notáveis incluem Vila União (zika), Boa Vista/Castelão, Sapiranga/Coité e Aerolândia (camada política presidencial), indicando que variáveis político-territoriais também concentram forte conectividade. Vejamos mais detalhes de cada bairro citado:

- **Praia do Futuro I (zika, dengue, chikungunya):** Com o maior grau ponderado da rede (319,46 em zika), Praia do Futuro I se destaca como um epicentro estrutural e epidemiológico. Isso se deve provavelmente à sua alta incidência de arboviroses, aliada a características urbanas que favorecem a conectividade com outros bairros vulneráveis, como proximidade geográfica com bairros litorâneos, fluxo populacional, e urbanização precária em algumas partes.
- **Vila União (zika):** Alto grau ponderado (125,49) indica que compartilha forte similaridade com outros bairros em termos de risco de zika. Pode funcionar como um *hub* secundário, possivelmente por reunir características de transição entre periferia e zonas de fluxo (próxima a corredores de transporte, como avenidas principais, e próxima ao Centro).
- **Boa Vista/Castelão, Coaçu, Sapiranga/Coité, Aerolândia (política – presidente):** Esses bairros aparecem com graus ponderados elevados na camada política. Isso pode refletir perfis de votação altamente semelhantes entre si e com outros bairros. A presença de valores agregados altos indica que eles são representativos de padrões de comportamento político, conectando-se com diversos outros bairros que compartilham escolhas eleitorais semelhantes. Além disso, esses bairros estão entre os mais periféricos ou semiurbanizados da cidade, sugerindo um agrupamento político-territorial nas extremidades urbanas.

#### *Centralidade de Intermediação (Betweenness Centrality)*

Essa métrica revela nós que atuam como pontes estruturais na rede. Eles não necessariamente têm o maior número de conexões, mas estão no caminho mais curto entre muitas outras regiões, o que os torna estratégicos para difusão ou transição de fenômenos. O bairro Cocó apresentou a maior centralidade de intermediação da rede (*betweenness* = 49,53), funcionando como ponto crítico de travessia entre diferentes partes da rede *multilayer*. Outros bairros como Conjunto Ceará II, Couto Fernandes, Guararapes, Curió e Parque Araxá também apareceram com alta *betweenness*, refletindo possíveis posições-chave de transição entre grupos políticos, estruturais ou epidemiológicos. Vejamos mais detalhes de cada bairro citado:

- **Cocó (política – prefeito):** Com a maior *betweenness* (49,53), Cocó aparece como elo entre comunidades urbanas distintas, talvez por ser uma zona de transição entre bairros centrais e periféricos, com forte presença de áreas verdes, condomínios e vias expressas. Politicamente, isso pode indicar um comportamento eleitoral menos homogêneo, funcionando como “fronteira” entre regiões ideologicamente polarizadas.
- **Conjunto Ceará II, Couto Fernandes, Guararapes, Curió, Parque Araxá (política – prefeito):** Esses bairros, também na camada de eleição para prefeito, possuem valores altos de intermediação. Sua importância pode estar relacionada ao fato de conectarem áreas com perfis muito distintos — bairros altamente votantes em um espectro político e outros com comportamento oposto. Assim, funcionam como pontes políticas ou socioterritoriais na rede.
- **Pirambu (política – prefeito):** Surge como importante conector em áreas de vulnerabilidade. Mesmo que não tenha o maior grau, sua posição pode unir regiões de baixa infraestrutura ou alta incidência de arboviroses com outras áreas mais estruturadas.
- **Aeroporto (zika):** Apesar de estar numa camada epidemiológica, aparece com alta intermediação, o que sugere que tem padrões de risco intermediários entre regiões de alta e baixa incidência — sendo, portanto, estruturalmente estratégico para propagação ou mitigação do vírus.

#### *Centralidade por Autovetor (Eigenvector Centrality)*

A métrica Centralidade por Autovetor revela nós que estão conectados a outros nós influentes. Ou seja, não basta estar conectado: o que importa é com quem você está conectado. Em redes urbanas, isso tende a capturar zonas estruturais importantes, onde há reforço de centralidade por vizinhança. A análise destacou novamente a camada de zika como dominante: Praia do Futuro I concentrou a maior influência estrutural na rede *multilayer* (eigenvector = 0,6042), com diversos outros bairros — como Vila União, Manoel Sátiro, Barra do Ceará, Fátima, Monte Castelo, Genibaú, Rodolfo Teófilo e Conjunto Ceará I — figurando entre os mais influentes. Esses resultados reforçam a importância dessa dimensão epidemiológica e sua conexão com a estrutura global de risco da cidade. Vejamos mais detalhes dos bairros citados:

- **Praia do Futuro I (zika):** Aparece novamente como mais influente (0,6042), o que reforça seu papel de nó dominante estruturalmente, conectado a outros bairros também relevantes na disseminação de zika.

- **Vila União, Manoel Sátiro, Barra do Ceará, Fátima, Monte Castelo, Genibaú, Rodolfo Teófilo, Conjunto Ceará I:** Todos esses bairros aparecem com alto *eigenvector*. Isso indica que fazem parte de uma rede coesa de influência epidemiológica, na qual cada bairro contribui para amplificar a importância dos demais. Muitos deles estão localizados em áreas de transição entre zonas centrais e periféricas, com histórico de surtos e elevada densidade.

#### 5.4.1 Comunidades por camada (Louvain)

##### 5.4.1.0.1 Camadas epidemiológicas (COVID-19, dengue, zika, chikungunya).

A aplicação do Louvain em cada doença gerou partições com número de comunidades compatível com a estrutura espacial de Fortaleza. Em geral, observam-se blocos territoriais coesos, com bairros adjacentes compartilhando perfis semelhantes de risco relativo. A leitura destaca: (i) áreas com *hotspots* persistentes; (ii) faixas de transição entre comunidades; e (iii) regiões de baixa coesão epidemiológica. *Nota:* chikungunya apresenta série pequena e menos informativa; por isso, sua leitura é restrita e, em diversas partes, foi suprimida.

##### 5.4.1.0.2 Camada socioeconômica/demográfica.

A partição evidencia gradientes associados a densidade, tamanho dos domicílios e estrutura populacional. Comunidades socioeconômicas tendem a ser mais extensas e estáveis, servindo como “fundo” territorial para comparar as demais.

##### 5.4.1.0.3 Camada de infraestrutura.

A comunidade de infraestrutura reflete a distribuição de equipamentos (hospitais, clínicas, escolas, mercados etc.). A rede resultante é mais esparsa (regra de limiar), ressaltando núcleos com maior proximidade funcional entre bairros.

##### 5.4.1.0.4 Camadas políticas (municipal e presidencial) e IDH.

As partições políticas capturam territorialidades eleitorais; o IDH agrega padrão mais suave, com comunidades que acompanham gradientes socioeconômicos. Essas camadas funcionam como referenciais para confrontar com os padrões epidemiológicos.

### 5.4.2 *Leitura multilayer: convergências e divergências*

A integração entre camadas foi feita de forma **comparativa**, sem otimização conjunta de modularidade. Dois resultados são enfatizados:

- **Convergências:** grupos de bairros que aparecem na mesma comunidade em mais de uma camada (por exemplo, epidemiológica e infraestrutura) indicam alinhamento estrutural e maior probabilidade de padrão consistente no território.
- **Divergências:** quando comunidades epidemiológicas cruzam fronteiras das comunidades socioeconômicas ou de infraestrutura, isso sugere influência de fatores adicionais (mobilidade, dinâmica local, sub-registros, especificidades do serviço), ou limiares de similaridade que geraram redes mais/menos esparsas.

A leitura cartográfica lado a lado facilita identificar onde as comunidades “batem” entre camadas e onde não batem, guiando hipóteses e verificações pontuais.

### 5.4.3 *Interpretação prática e apoio à decisão*

Os resultados por camada, combinados de forma comparativa, permitem:

1. **Priorizar territórios:** convergência entre comunidades epidemiológicas e de infraestrutura sugere locais onde ações sobre equipamentos e acesso podem ter efeito mais direto.
2. **Ler padrões estáveis:** a camada socioeconômica/demográfica e o IDH oferecem pano de fundo para entender persistência/recorrência de padrões.
3. **Focar transições:** fronteiras de comunidades (especialmente onde centralidades são altas) sinalizam áreas de atenção para evitar “espraiamento” entre blocos.

### 5.4.4 *Síntese*

O uso de Louvain *por camada* mantém a consistência técnica da detecção de comunidades. A leitura *multilayer* — entendida aqui como comparação entre partições e métricas calculadas por dimensão — oferece uma visão mais clara das convergências e das transições territoriais relevantes para a gestão. Embora a estrutura seja tecnicamente *multiplex*, adotar o termo *multilayer* no texto facilita a comunicação e não altera a interpretação dos resultados.

## 5.5 Machine Learning (ML)

Nesta seção, são apresentados os resultados obtidos por meio da aplicação de técnicas de aprendizado de máquina. Foram utilizados modelos supervisionados *XGBoost* e *Random Forest* para estimar o risco relativo das doenças COVID-19, dengue, zika e chikungunya. As variáveis preditoras incluíram indicadores socioeconômicos, de infraestrutura, demográficos e políticos.

### 5.5.1 Aprendizado Supervisionado com XGBoost

Nesta etapa, foram ajustados modelos de regressão baseados em *eXtreme Gradient Boosting* (XGBoost) para estimar o risco relativo (RR) de cada doença a partir de variáveis explicativas estruturais, políticas e socioeconômicas. A avaliação de desempenho foi realizada por meio das métricas de erro quadrático médio (RMSE), erro absoluto médio (MAE) e coeficiente de determinação ( $R^2$ ), cujos resultados são apresentados a seguir.

O modelo de melhor desempenho foi o da COVID-19, representado na Tabela 2, com  $RMSE = 0,2065$ ,  $MAE = 0,3728$  e  $R^2 = 0,1214$ , indicando capacidade limitada, mas superior às demais doenças, de explicar a variabilidade do risco relativo a partir das variáveis independentes. O desempenho relativamente melhor sugere que a disseminação da COVID-19 está mais associada a fatores estruturais e sociais capturados no modelo, como densidade populacional, IDH e mobilidade.

Tabela 2 – Desempenho dos modelos XGBoost na predição do risco relativo (RR) por doença.

Doença	RMSE	MAE	$R^2$
COVID-19	0,2065	0,3728	0,1214
dengue	0,2271	0,3905	0,0553
zika	0,6598	0,5829	-0,2519
chikungunya	1,1378	0,7963	-0,6564

Fonte: elaborado pelo autor.

Para a dengue, o modelo apresentou  $RMSE = 0,2271$ ,  $MAE = 0,3905$  e  $R^2 = 0,0553$ , revelando baixo poder explicativo, embora os erros absolutos e quadráticos estejam próximos aos observados para a COVID-19. Isso indica que, apesar de algum nível de predição, a ocorrência de dengue está associada a fatores adicionais não contemplados no modelo — como variações microambientais e condições climáticas.

Os modelos para zika e chikungunya apresentaram desempenho insatisfatório, com  $R^2$  negativos, indicando que os modelos ajustados foram piores do que a simples média dos dados. O RMSE do modelo de zika foi de 0,6598, com MAE = 0,5829 e  $R^2 = -0,2519$ . Já o modelo de chikungunya obteve o pior desempenho, com RMSE = 1,1378, MAE = 0,7963 e  $R^2 = -0,6564$ . Esses resultados sugerem elevada variabilidade e imprevisibilidade dos padrões territoriais dessas doenças, possivelmente influenciados por fatores contextuais mais complexos, como práticas locais de combate ao vetor, ciclos sazonais e subnotificação.

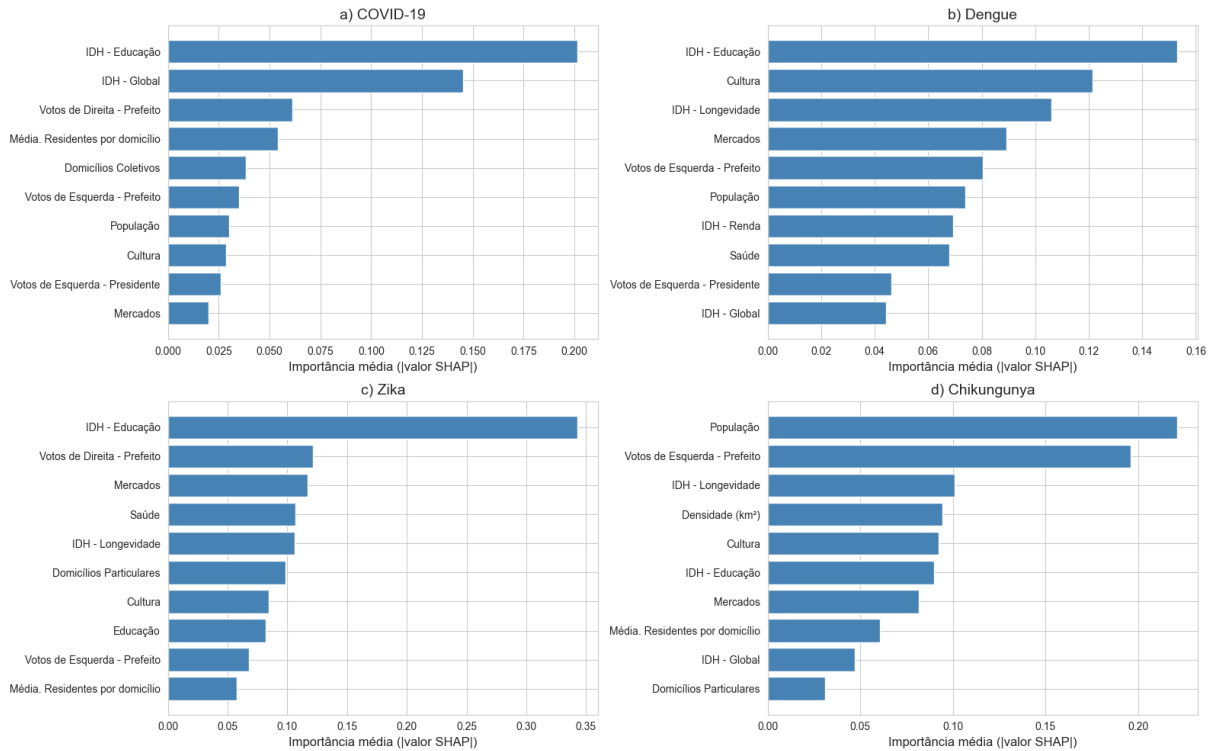
Em conjunto, os resultados indicam que, embora os modelos supervisionados consigam captar parcialmente os padrões de risco da COVID-19 e, em menor grau, da dengue, eles não são eficazes na predição de zika e chikungunya com as variáveis atualmente disponíveis, reforçando a necessidade de incorporar novas fontes de dados e estratégias analíticas mais robustas, como abordagens espaciais ou modelos híbridos.

#### 5.5.1.1 *Análise de Importância das Variáveis via SHAP*

Para avaliar a contribuição individual de cada variável na predição dos casos de COVID-19, dengue, zika e chikungunya, utilizamos o algoritmo SHAP (SHapley Additive exPlanations) aplicado aos modelos preditivos treinados com o XGBoost. O SHAP fornece uma decomposição local e global da importância de cada atributo, fundamentada em teoria dos jogos, permitindo interpretar como cada variável impacta positivamente ou negativamente a predição.

A Figura 10 apresenta os gráficos de barras com as dez variáveis mais relevantes para cada uma das doenças analisadas. Observa-se que os fatores associados ao Índice de Desenvolvimento Humano (IDH), especialmente o IDH - Educação, destacam-se com maior influência na predição de COVID-19, dengue e zika. Para a COVID-19, por exemplo, as variáveis IDH - Educação, IDH Global e o número de votos para o candidato de direita para prefeito aparecem como as mais impactantes, sugerindo uma forte correlação entre níveis educacionais, participação política e a disseminação da doença.

Figura 10 – Importância das variáveis (SHAP) nos modelos XGBoost para COVID-19, dengue, zika e chikungunya.



Fonte: elaborado pelo autor.

No contexto desta análise, maior influência em SHAP refere-se ao tamanho médio da contribuição ( $\text{mean}(|\text{SHAP}|)$ ) das variáveis para a predição — isto é, à importância em magnitude, e não ao sinal do efeito. Valores positivos de SHAP aumentam a predição do risco e valores negativos a reduzem, podendo variar por faixas do preditor. O IDH-Educação utilizado corresponde à dimensão educacional do IDH por bairro (2010), disponibilizada pela Prefeitura de Fortaleza, que sintetiza condições educacionais (escolaridade/fluxo educacional) em um índice composto. Assim, o fato de o IDH-Educação aparecer entre as variáveis mais influentes para COVID-19, dengue e zika não implica que “mais educação cause mais casos” (ou o contrário); indica, sim, que essa dimensão educacional é um marcador estrutural fortemente associado ao padrão predito no nosso conjunto de dados — possivelmente por vias como mobilidade, padrões de interação, acesso à testagem/notificação ou outras variáveis correlacionadas. Por isso, reportamos a importância em magnitude e avaliamos a direção caso a caso via gráficos de dependência, evitando inferências causais além do escopo desta tese.

No caso da dengue, além do IDH educacional, cultura, longevidade, mercados e votos de esquerda para prefeito são evidenciados como relevantes, apontando para a importância de aspectos socioculturais e de infraestrutura urbana na propagação da doença.

Para a zika, o destaque fica com o IDH - Educação, seguido novamente por fatores políticos e de infraestrutura urbana, como votos para o candidato da direita para prefeito e mercados. Já para a chikungunya, os determinantes mais relevantes incluem a população, votos de esquerda para prefeito, e indicadores espaciais como densidade populacional e média de residentes por domicílio.

### 5.5.2 *Aprendizado Supervisionado com Random Forest*

A aplicação do algoritmo Random Forest permitiu avaliar a capacidade preditiva de um modelo baseado em árvores de decisão para estimar o risco relativo (RR) das quatro doenças analisadas, com base nas variáveis estruturais, epidemiológicas, socioeconômicas, políticas e espaciais dos bairros.

Os resultados indicaram que, entre as doenças, a COVID-19 foi a que apresentou o melhor desempenho preditivo, Tabela 3, com  $RMSE = 0,1572$ ,  $R^2 = 0,3313$ , evidenciando uma explicação moderada da variância do risco relativo a partir das variáveis disponíveis. Para as demais doenças, os modelos apresentaram desempenho mais limitado. A predição do RR da dengue teve  $RMSE = 0,3934$  e  $R^2 = -0,6363$ , enquanto o modelo de zika obteve  $RMSE = 0,9116$ , com um  $R^2$  ainda mais negativo ( $-0,7296$ ). O modelo para chikungunya também apresentou baixo desempenho, com  $RMSE = 0,9516$  e  $R^2 = -0,3854$ .

Tabela 3 – Desempenho dos modelos Random Forest na predição do risco relativo (RR) por doença.

<b>Doença</b>	<b>RMSE</b>	<b><math>R^2</math></b>
COVID-19	0,1572	0,3313
dengue	0,3934	-0,6363
zika	0,9116	-0,7296
chikungunya	0,9516	-0,3854

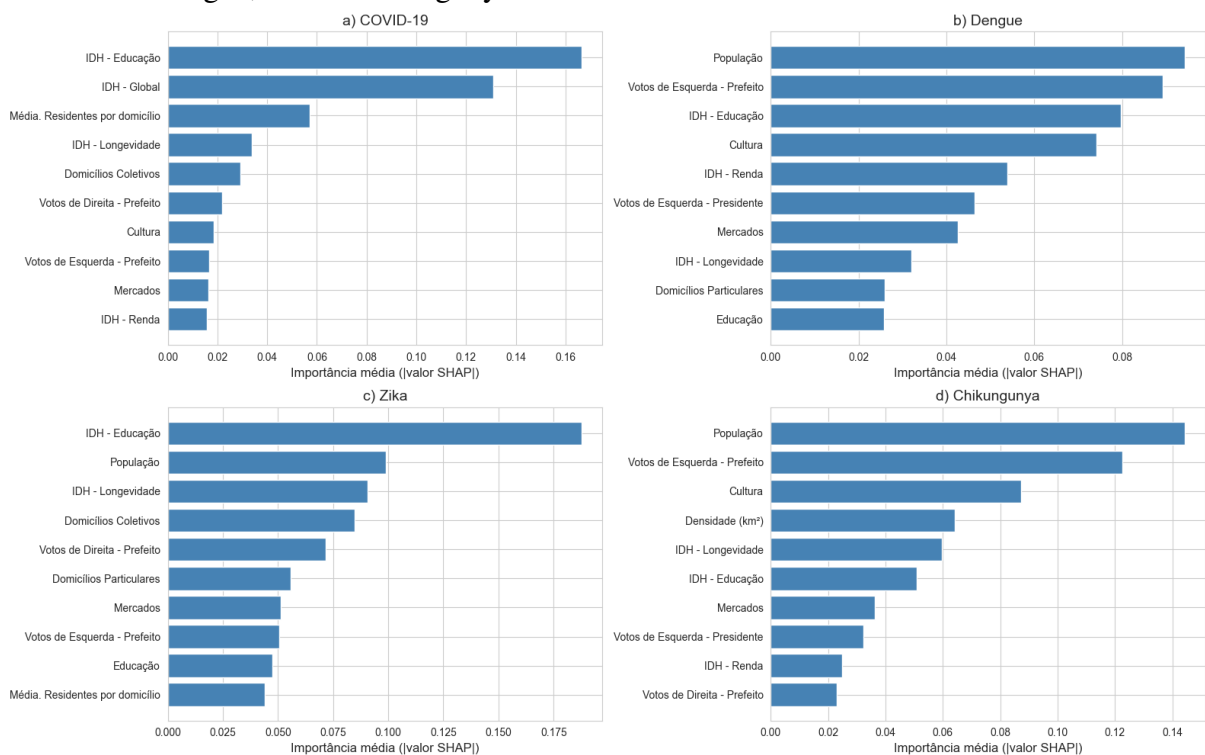
Fonte: elaborado pelo autor.

O desempenho negativo de  $R^2$  para dengue, zika e chikungunya indica que, nesses casos, o modelo foi menos eficaz do que uma simples média global, o que sugere uma complexidade maior na predição desses padrões ou ausência de variáveis explicativas determinantes no modelo atual.

### 5.5.2.1 Análise de Importância das Variáveis via SHAP

A Figura 11 apresenta o gráfico de importância das variáveis para o modelo Random Forest treinado com os mesmos dados e variáveis da análise anterior com XGBoost. Observa-se que, de modo geral, há consistência nos fatores mais relevantes, como as variáveis relacionadas ao IDH - Educação, IDH Global, média de residentes por domicílio, e fatores políticos, como votos por orientação ideológica.

Figura 11 – Importância das variáveis (SHAP) nos modelos Random Forest para COVID-19, dengue, zika e chikungunya.



Fonte: elaborado pelo autor.

A análise revelou que os bairros com maior peso nos modelos foram, em sua maioria, regiões centrais e litorâneas, como **Meireles, Aldeota, Praia de Iracema, Centro, Varjota**, além de áreas periféricas com destaque epidemiológico, como **Vila Peri, Vicente Pinzón, Jardim Iracema, Pirambu e Carlito Pamplona**. Esses bairros concentram características estruturais contrastantes, sendo alguns marcados por alta densidade e desenvolvimento urbano, enquanto outros refletem vulnerabilidades sociais e sanitárias.

Esses resultados indicam que a modelagem baseada em Random Forest é capaz de captar, com maior precisão, os padrões estruturais associados ao risco de COVID-19, mas encontra limitações relevantes para explicar a variabilidade espacial das arboviroses. A inclusão

de variáveis ambientais, climáticas ou de mobilidade urbana pode representar uma via de aprimoramento futuro dos modelos para essas doenças.

#### 5.5.2.2 *Comparativo XGBoost vs Random Forest*

Enquanto o modelo XGBoost atribuiu maior peso a variáveis diretamente ligadas a indicadores socioeconômicos e educacionais — como IDH - Educação e IDH Global, o modelo Random Forest demonstrou um padrão levemente distinto, em que a média de residentes por domicílio, população e densidade populacional ganharam mais importância relativa. Isso pode indicar uma maior sensibilidade do Random Forest a variáveis com distribuição mais ampla ou variância maior, especialmente aquelas relacionadas à estrutura demográfica.

Além disso, o XGBoost apresentou uma hierarquia mais clara e concentrada nas primeiras variáveis, enquanto o Random Forest apresentou uma distribuição mais "plana" da importância — com menos discrepância entre as variáveis do top 10. Essa diferença está relacionada à natureza dos algoritmos: o XGBoost, por ser um método de *boosting*, tende a ajustar-se mais fortemente a padrões específicos nos dados; já o Random Forest, por ser baseado em *bagging*, é mais robusto a *overfitting*, mas menos "agressivo" na seleção de variáveis dominantes.

## 5.6 Agrupamento através do algoritmo K-means

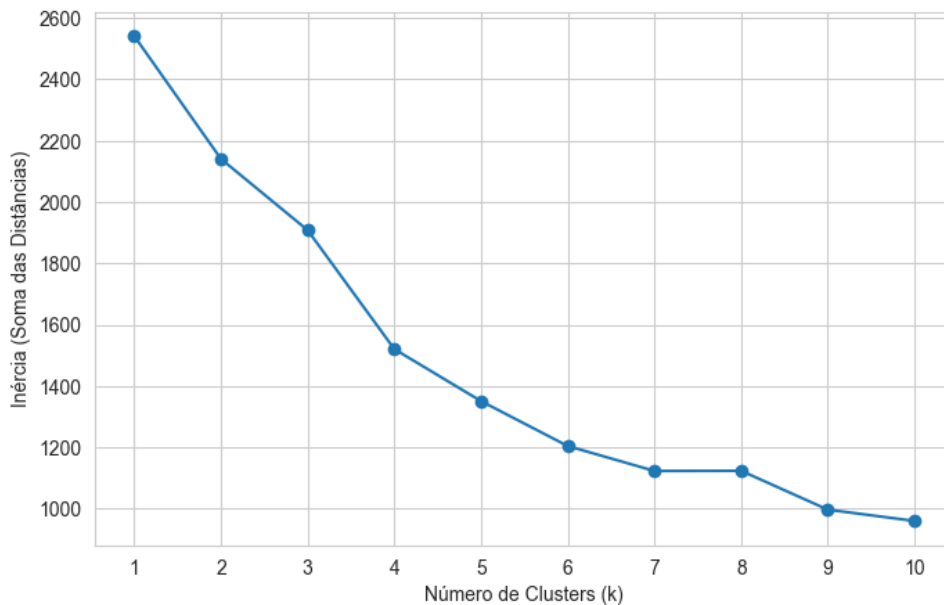
Esta seção apresenta os resultados da aplicação do algoritmo K-means para a classificação dos bairros de Fortaleza em quatro *clusters* distintos, com base em variáveis estruturais e epidemiológicas. O algoritmo K-means é uma técnica de aprendizagem não supervisionada amplamente utilizada para segmentação de dados multivariados. O método agrupa observações (neste caso, bairros) em  $k$  grupos que minimizam a variância intra-cluster, alocando cada observação ao centróide mais próximo no espaço das variáveis consideradas.

### 5.6.1 *Determinação do Número de Clusters*

A definição do número ideal de grupos ( $k$ ) para o algoritmo K-means foi realizada por meio do método do cotovelo (LLOYD, 1982), que consiste em observar o ponto de inflexão da curva que relaciona o número de clusters à soma das distâncias intra-cluster (inércia). Como ilustrado na Figura 12, observa-se uma redução expressiva da inércia entre  $k = 1$  e  $k = 4$ , seguida

por uma queda mais gradual nos valores subsequentes. Esse comportamento sugere que a escolha de quatro *clusters* ( $k = 4$ ) representa um equilíbrio adequado entre a complexidade do modelo e a capacidade de explicação da variabilidade dos dados.

Figura 12 – Método do cotovelo aplicado à clusterização dos bairros de Fortaleza.



Fonte: elaborado pelo autor.

### 5.6.2 Teste de Estabilidade dos Agrupamentos

Para avaliar a robustez dos agrupamentos obtidos com o algoritmo, foi realizado um teste de estabilidade com múltiplas inicializações aleatórias, variando o parâmetro *random state* do algoritmo K-means. Os resultados de cada segmentação foram comparados com a clusterização base (*seed* = 42), valor explicado na Seção 4.5.2, por meio do índice Adjusted Rand Index (ARI) (HUBERT; ARABIE, 1985).

O Adjusted Rand Index (ARI) é uma métrica utilizada para quantificar a similaridade entre duas partições de um mesmo conjunto de dados, ajustando a pontuação esperada ao acaso. A métrica varia de -1 a 1, onde valores próximos de 1 indicam perfeita concordância entre os agrupamentos comparados, valores próximos de 0 indicam que a concordância não é melhor que o acaso, e valores negativos indicam pior que o acaso. O ARI leva em conta todas as combinações possíveis de pares de elementos, analisando se cada par foi colocado no mesmo cluster ou em clusters diferentes nas duas segmentações analisadas. Assim, trata-se de uma ferramenta útil para validar a consistência de algoritmos de agrupamento não supervisionado,

como o K-means, especialmente quando submetido a diferentes inicializações. Os valores obtidos foram heterogêneos, variando de 0,2162 a 1,0000, como mostrado abaixo:

- Seed 0: ARI = 0,8354
- Seed 1: ARI = 0,5040
- Seed 10: ARI = 1,0000
- Seed 99: ARI = 0,2162
- Seed 123: ARI = 0,6247
- Seed 2024: ARI = 0,5762

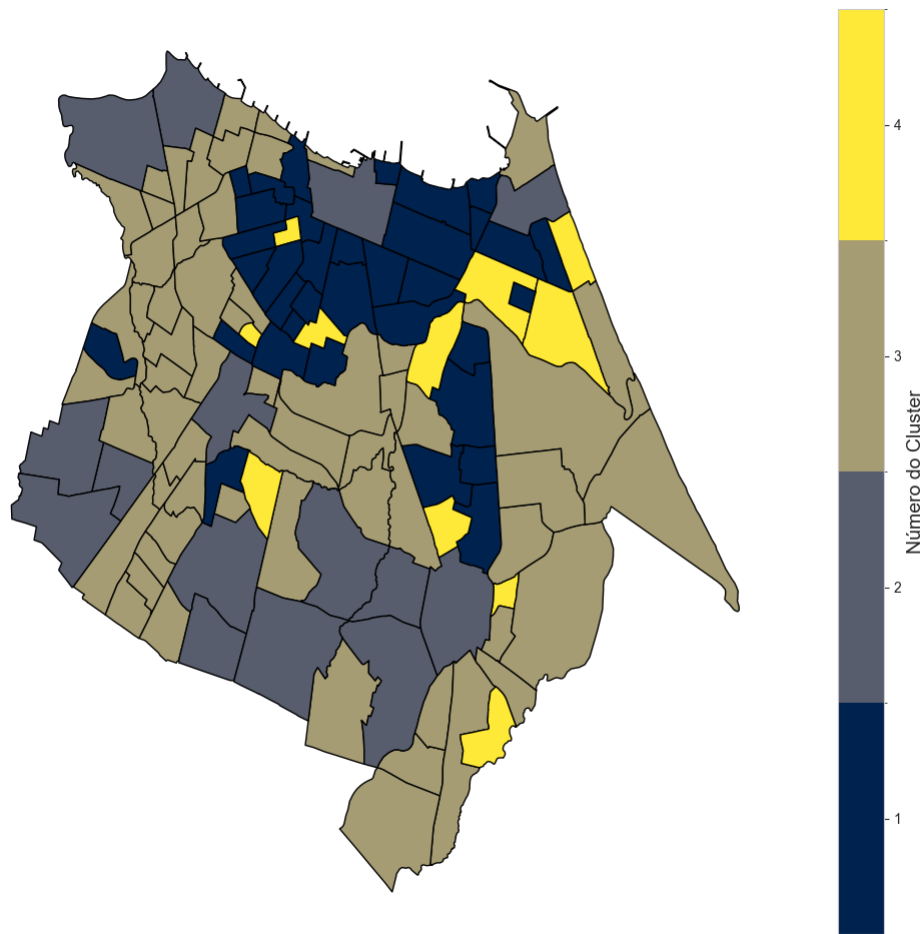
Esses resultados indicam que, embora em alguns casos a segmentação tenha sido altamente consistente (como em `seed=10` com ARI = 1,0), em outras inicializações a variação foi considerável, apontando para alguma sensibilidade à aleatoriedade do algoritmo. Isso sugere que a topologia dos dados pode conter fronteiras difusas entre grupos, ou ainda que há bairros com características intermediárias, suscetíveis a serem classificados em diferentes clusters dependendo da inicialização.

Ao comparar os clusters formados por `random state=42` e `random state=10` (ARI = 1.0), revelou-se uma correspondência perfeita entre os grupos gerados, confirmando a estabilidade estrutural dos agrupamentos. Com base nisso, manteve-se a clusterização original para interpretação dos perfis territoriais, por sua confiabilidade e reprodutibilidade.

### 5.6.3 *Síntese dos Resultados*

A incorporação simultânea de dimensões epidemiológicas, estruturais (quantidade de equipamentos de saúde, educação, cultura e mercados), sociodemográficas (densidade populacional, domicílios) e políticas (perfil de votação para prefeitos e presidentes) permitiu a construção de uma tipologia multifatorial dos bairros. O resultado é uma classificação que reflete diferentes padrões de vulnerabilidade e exposição ao risco sanitário no espaço urbano de Fortaleza. A Figura 13 ilustra a distribuição espacial dos *clusters* resultantes. Cada cor representa um perfil territorial distinto, que será detalhado a seguir em termos de características médias e implicações para a gestão da saúde pública.

Figura 13 – Distribuição espacial dos *clusters* territoriais identificados pelo algoritmo K-means.



Fonte: elaborado pelo autor.

Os *clusters* apresentam padrões contrastantes. Um dos grupos, concentrado nas áreas centrais e litorâneas, reúne bairros com alto IDH, elevada densidade de infraestrutura urbana e maior risco relativo de COVID-19, sugerindo maior exposição associada à circulação e à conectividade urbana. Outro *cluster*, localizado em zonas periféricas, apresenta baixos índices de desenvolvimento e maior incidência de arboviroses, especialmente chikungunya, refletindo um padrão de vulnerabilidade estrutural clássica. Há ainda grupos intermediários, com níveis mistos de risco e estrutura, indicando territórios de transição ou com dinâmicas híbridas de vulnerabilidade. A Tabela 4 lista os bairros formados por cada *cluster*:

Tabela 4 – Resumo dos clusters territoriais gerados via K-means

Cluster	Exemplos de bairros	Nº de bairros
1	Aldeota, Amadeu Furtado, Benfica, Bom Futuro, Cais do Porto, Cambeba, Cidade 2000, Cidade dos Funcionários, Conjunto Ceará I, Conjunto Esperança, Damas, De Lourdes, Demócrito Rocha, Dionísio Torres, Ellery, Eng. Luciano Cavalcante, Farias Brito, Fátima, Guararapes, Jacarecanga, Jardim América, Joaquim Távora, José Bonifácio, Maraponga, Meireles, Monte Castelo, Montese, Mucuripe, Papicu, Parque Manibura, Parquelândia, Praia de Iracema, Rodolfo Teófilo, São Gerardo, São João do Tauape, Varjota, Vila União	37
2	Aerolândia, Aeroporto, Alto da Balança, Ancuri, Antônio Bezerra, Aracape, Autran Nunes, Bela Vista, Boa Vista / Castelão, Bonsucesso, Cajazeiras, Canindezinho, Carlito Pamplona, Coacu, Conjunto Ceará II, Conjunto Palmeiras, Cristo Redentor, Dias Macedo, Dom Lustosa, Edson Queiroz, Floresta, Genibaú, Granja Portugal, Guajeru, Henrique Jorge, Itaoca, Itaperi, Jardim Cearense, Jardim Guanabara, Jardim Iracema, Jardim das Oliveiras, Joquei Clube, José de Alencar, João XXIII, Lagoa Redonda, Manoel Sátiro, Moura Brasil, Novo Mondubim, Olavo Oliveira, Padre Andrade, Panamericano, Parque Dois Irmãos, Parque Pres. Vargas, Parque Santa Maria, Parque São José, Pici, Praia do Futuro II, Sabiaguaba, Sapiranga / Coité, Serrinha, Vila Peri, Álvaro Weyne	52
3	Barra do Ceará, Boa Vista, Cajazeiras do Norte, Couto Fernandes, Granja Lisboa, Itaperi, Messejana, Moura Brasil, Parque Araxá, Praia do Futuro I	10
4	Coco, Couto Fernandes, Curio, Manuel Dias Branco, Parque Araxá, Parque Iracema, Parreao, Praia do Futuro I, Rachel De Queiroz, Salinas, São Bento	11

Fonte: elaborado pelo autor.

O **Cluster 1** engloba majoritariamente bairros centrais e historicamente mais desenvolvidos da cidade, com elevada densidade populacional, bons indicadores de infraestrutura (média de 2 equipamentos de educação, 5 de cultura e 1 de saúde por bairro), alto IDH (0,58) e menor número médio de moradores por domicílio (2,70). Politicamente, apresenta equilíbrio nas votações entre candidatos de esquerda e direita, e se destaca pelo maior risco relativo de COVID-19 ( $RR = 1,65$ ), além de valores elevados de zika ( $RR = 1,32$ ). Esses bairros representam zonas de centralidade funcional e mobilidade urbana intensa, refletindo vulnerabilidades associadas ao contágio por contato interpessoal.

Já o **Cluster 2** compreende bairros com a maior concentração populacional (média de 47 mil habitantes) e infraestrutura urbana elevada (média de 4 escolas, 4,7 unidades de saúde e 12 equipamentos culturais). Apesar dos bons indicadores de acesso, o IDH médio é o mais baixo entre todos os grupos (0,27), sugerindo que a densidade pode estar associada a condições precárias de habitação e sobrecarga dos serviços. Esse grupo apresenta os maiores valores de risco relativo para chikungunya ( $RR = 1,22$ ) e dengue ( $RR = 1,12$ ), além de forte apoio a candidatos de esquerda, tanto para prefeito quanto para presidente. Trata-se de territórios densos, populares e estruturalmente vulneráveis ao risco vetorial.

O **Cluster 3** é composto por bairros de densidade intermediária, infraestrutura urbana modesta e indicadores de desenvolvimento ainda mais baixos, com o menor IDH médio entre os grupos (0,28) e baixa presença de equipamentos públicos. Apresenta a menor média de moradores por domicílio e um perfil político de apoio majoritário à esquerda nas eleições presidenciais. Epidemiologicamente, apresenta risco relativo mais baixo para COVID-19 ( $RR = 0,81$ ) e zika ( $RR = 0,71$ ), indicando menor exposição urbana, embora mantenha níveis moderados para chikungunya e dengue. Esses bairros configuram territórios periféricos com infraestrutura insuficiente, mas sem elevada concentração de risco sanitário.

Por último, o **Cluster 4** agrupa bairros mais isolados ou com menor presença do Estado, refletindo os piores indicadores de infraestrutura urbana entre todos os grupos (média de 0,55 unidades de saúde, 0,27 de educação e apenas 0,18 mercados por bairro). Apesar de apresentar IDH um pouco superior ao Cluster 2 (0,41), sua ausência total de dados eleitorais e o baixo número de domicílios coletivos sugerem um padrão territorial singular, possivelmente ligado a regiões novas, informais ou de difícil acesso. Este grupo apresentou os maiores valores de risco relativo para dengue ( $RR = 1,20$ ) e zika ( $RR = 1,49$ ), o que reforça sua vulnerabilidade ao contágio por arboviroses em ambientes urbanos precários.

Essa classificação fornece uma base empírica para o planejamento territorial de ações em saúde, permitindo que políticas públicas sejam direcionadas conforme os perfis específicos de risco e infraestrutura de cada grupo de bairros. Além disso, a metodologia é replicável e pode ser atualizada periodicamente à medida que novas variáveis forem incorporadas ou o cenário epidemiológico se modificar.

## 6 DISCUSSÃO

Este capítulo tem como objetivo interpretar criticamente os resultados obtidos com base nos fundamentos teóricos e metodológicos apresentados, conectando as evidências empíricas às discussões contemporâneas sobre saúde urbana, desigualdades territoriais e análise computacional de dados socioespaciais.

### 6.1 Padrões de Risco Epidemiológico e Determinantes Urbanos

A utilização do risco relativo (RR) permitiu normalizar os dados epidemiológicos por população, possibilitando comparabilidade entre bairros com tamanhos populacionais distintos. Os resultados revelaram que a COVID-19 apresentou padrão espacial distinto das arboviroses, concentrando-se em bairros centrais com maior densidade e infraestrutura urbana, como Meireles, Centro e Aldeota. Em contrapartida, as arboviroses (dengue, zika e chikungunya) evidenciaram incidência mais elevada em bairros periféricos e com menor infraestrutura, como Bom Jardim, Siqueira e Curió.

Essa não sobreposição entre hotspots reforça a hipótese de que diferentes patógenos são sensíveis a determinantes estruturais específicos. A análise de Moran's I e LISA evidenciou clusters espaciais significativos e revelou a existência de autocorrelação espacial positiva para todas as doenças, com destaque para padrões de alto-alto e baixo-baixo em regiões historicamente vulneráveis.

### 6.2 Importância dos Fatores Socioeconômicos, Políticos e de Infraestrutura

As análises de correlação de Pearson e os modelos de regressão linear (OLS) mostraram relações estatisticamente significativas entre os índices de infraestrutura e socioeconômicos com a incidência de doenças, especialmente para COVID-19. A ausência de correlação forte para zika e chikungunya sugere que outros fatores contextuais (ambientais ou institucionais) também estão em jogo.

As variáveis políticas, categorizadas com base na orientação ideológica do voto (esquerda, centro, direita) nas eleições municipais (2020) e presidenciais (2022), foram integradas à rede *multilayer* como camadas distintas. A análise estatística revelou que a proporção de votos em candidatos de direita se correlacionou negativamente, ainda que de forma fraca, com indicadores de risco relativo para arboviroses, enquanto o voto à esquerda apresentou correlações

positivas, sobretudo com variáveis socioeconômicas associadas à vulnerabilidade urbana.

Do ponto de vista estrutural, a camada de votos para prefeito (2020) evidenciou maior homogeneidade, com predominância de comunidades formadas por bairros majoritariamente alinhados à direita, especialmente nas regiões centrais e de maior IDH. Em contraste, os bairros com maior proporção de votos à esquerda — com destaque para áreas periféricas e de menor infraestrutura — apresentaram maior fragmentação e interconectividade com camadas epidemiológicas e socioeconômicas, atuando como elos estruturais entre diferentes domínios da rede.

No caso da COVID-19, observou-se um padrão distinto em relação às demais doenças analisadas. Bairros com maior proporção de votos em candidatos de direita apresentaram, em média, maiores valores de risco relativo (RR) para COVID-19, especialmente nas zonas centrais da cidade. Essa associação, ainda que moderada, sugere uma possível relação entre comportamento político e adesão às medidas de prevenção, como uso de máscara, distanciamento social e vacinação, frequentemente mediadas por discursos político-ideológicos durante a pandemia.

Por outro lado, bairros com predominância de votos à esquerda, embora concentrados em áreas com maior vulnerabilidade estrutural, apresentaram menor RR para COVID-19, possivelmente refletindo maior adesão às recomendações sanitárias, conforme sugerido por outros estudos nacionais e internacionais.

Esse contraste reforça a hipótese de que a variável política, no contexto da pandemia de COVID-19, atua não apenas como um reflexo da desigualdade territorial, mas também como um fator comportamental que influencia diretamente os desfechos epidemiológicos, tornando-se particularmente relevante para a formulação de políticas públicas em situações de crise sanitária com alta polarização ideológica.

### 6.3 Interpretação das Comunidades na Rede Multilayer

A estruturação de uma rede *multilayer*, integrando camadas epidemiológicas, estruturais, socioeconômicas e políticas, revelou a existência de 13 comunidades detectadas via algoritmo de Louvain. Tais comunidades apresentaram perfis distintos, como:

- **Comunidade 1:** Dominada por nós da COVID-19 e IDH, com forte coesão interna e representação de bairros centrais;
- **Comunidade 2:** Exclusiva de zika, indicando comportamento autônomo da doença;
- **Comunidade 4:** Mistura camadas epidemiológicas e estruturais, com bairros periféricos e

centrais conectados;

- **Comunidade 13:** Predominância de nós estruturais e socioeconômicos, com fraca participação de dimensões políticas.

Esses agrupamentos refletem não apenas similaridades internas, mas também a ausência de conexão com outras camadas, indicando nichos territoriais específicos ou zonas de isolamento estrutural.

#### 6.4 Desempenho dos Modelos de Aprendizado de Máquina

Os modelos Random Forest e XGBoost apresentaram desempenhos variáveis, com melhor acurácia preditiva para COVID-19 (Random Forest:  $R^2 = 0.33$ ) e desempenho mais fraco para chikungunya e zika (com  $R^2$  negativos no XGBoost). Isso indica que o risco relativo dessas arboviroses é menos previsível com base nas variáveis estruturais analisadas. A interpretação com SHAP evidenciou que bairros específicos, como Meireles e Aldeota, têm peso desproporcional na predição da COVID-19, provavelmente por concentrarem alta renda e densidade de serviços.

#### 6.5 Agrupamentos Territoriais com K-means

A aplicação do algoritmo de agrupamento K-means permitiu sintetizar padrões estruturais e epidemiológicos em perfis territoriais distintos. A técnica identificou grupos de bairros com características convergentes de risco relativo, infraestrutura urbana e indicadores socioeconômicos, revelando blocos urbanos com condições compartilhadas de vulnerabilidade ou resiliência. Embora o K-means não seja supervisionado, sua capacidade de segmentação espacial se mostrou coerente com os clusters detectados pelas análises de autocorrelação (LISA) e pelas comunidades da rede *multilayer*, reforçando sua utilidade como ferramenta exploratória e de apoio à estratificação territorial. A interpretação dos grupos obtidos evidenciou a existência de zonas com risco múltiplo (para mais de uma doença) e perfis urbanos críticos, importantes para priorização de políticas públicas intersetoriais.

#### 6.6 Implicações para Políticas Públicas e Planejamento Territorial

Os achados desta pesquisa reforçam a necessidade de abordagens integradas e multiescalares para o planejamento em saúde pública urbana. A análise espacial de risco, combinada à modelagem de redes *multilayer*, permite identificar territórios prioritários para

intervenções, assim como compreender como diferentes fatores estruturais e institucionais moldam a vulnerabilidade às doenças.

Os resultados sugerem que políticas uniformes de enfrentamento às epidemias podem ser ineficazes diante da heterogeneidade dos territórios. A implementação de planos locais integrados, que articulem variáveis ambientais, estruturais e políticas, é essencial para aumentar a efetividade das ações em saúde.

Por fim, a abordagem proposta demonstra o potencial da análise computacional de redes para subsidiar decisões baseadas em evidências espaciais, especialmente em contextos urbanos marcados por desigualdade e complexidade sistêmica.

## 7 CONCLUSÕES E TRABALHOS FUTUROS

### 7.1 Conclusões

Esta tese adotou uma abordagem integrada que combina redes multilayer, análise espacial e modelos de aprendizado de máquina interpretável para compreender padrões de difusão de COVID-19, dengue, zika e chikungunya em Fortaleza. A integração de dimensões epidemiológica, socioeconômico/demográfica, de infraestrutura e política/IDH permitiu comparar resultados por camada e interpretar convergências e divergências territoriais relevantes para a gestão. Tecnicamente, a rede utilizada é do tipo multiplex (mesmos bairros em todas as camadas), embora se use o termo multilayer ao longo do texto por simplicidade. A detecção de comunidades foi realizada por camada, sem otimização conjunta em supra-grafo, e optou-se por não segmentar análises por ondas de COVID-19 para priorizar comparabilidade territorial entre doenças. Resultados específicos de chikungunya foram suprimidos ou pouco detalhados devido à baixa disponibilidade e curta série temporal.

Principais achados:

- Padrões espaciais distintos entre doenças: COVID-19 apresenta maior concentração em áreas centrais e com melhor estrutura; dengue e zika exibem distribuição mais heterogênea e periférica. A baixa sobreposição entre hotspots sugere determinantes estruturais específicos por patógeno.
- Comunidades por camada evidenciam sobreposição parcial entre fatores estruturais e epidemiológicos e permitem identificar fronteiras territoriais de transição; a leitura lado a lado das partições foi útil para destacar convergências e divergências.
- Modelos preditivos tiveram melhor desempenho para COVID-19 do que para as arboviroses, indicando que variáveis estruturais explicam com mais estabilidade a variação de COVID-19 no recorte analisado.
- O agrupamento por k-means sintetizou perfis territoriais com base em características estruturais e epidemiológicas, oferecendo uma tipologia útil para priorização territorial.
- Nas análises de importância, dimensões ligadas a IDH (especialmente IDH-Educação) apareceram com influência elevada nas previsões de COVID-19, dengue e zika; a interpretação foi feita em termos de associação, com direção avaliada em gráficos de dependência do SHAP, evitando extrapolações causais.

Em síntese, os resultados reforçam o valor de uma leitura integrada por camadas,

territorialmente sensível e comparável entre doenças, para apoiar diagnósticos situados e indicar prioridades de intervenção. Como contribuição prática adicional, apresenta-se em anexo uma proposta de aprimoramento metodológico para orientar padronizações futuras e facilitar a adoção dessa leitura integrada por equipes técnicas.

## 7.2 Trabalhos Futuros

Com base nas limitações e oportunidades identificadas, destacam-se caminhos de continuidade:

- Incorporar dinâmica temporal explícita nas redes (janelas móveis e testes de sensibilidade de limiares) para avaliar estabilidade dos achados ao longo do tempo.
- Integrar dados ambientais e climáticos (temperatura, umidade, precipitação) e mobilidade/fluxo quando disponíveis, buscando melhorar a predição para arboviroses.
- Ampliar a validação externa em outras cidades, avaliando transportabilidade dos modelos e da leitura por camadas.
- Explorar modelos espaciais e em grafos adicionais (por exemplo, modelos baseados em grafos e variações de centralidade), mantendo comparabilidade com os resultados atuais.
- Detalhar a Directed Acyclic Graph (DAG) conceitual para guiar conjuntos de ajuste e documentar de forma mais sistemática as decisões de seleção de variáveis e limiares de rede.
- Desenvolver painéis de visualização voltados a gestores, com mapas de comunidades por camada, indicadores de centralidade e resumos interpretáveis de importância de variáveis.

Espera-se que essas extensões consolidem a utilidade prática da abordagem e incentivem aplicações comparativas em diferentes contextos territoriais.

## REFERÊNCIAS

- ALBUQUERQUE, M. d. G.; SANTOS, J. R. d.; SILVEIRA, A. F.; JUNIOR, D. L. B.; NUNES, R. B.; GANDRA, T. B. R.; SILVA, A. L. C. d.; MIRANDA, G. B. d.; TREVISOL, A. R. Influence of socio-economic indicators and territorial networks at the spatiotemporal spread dynamics of covid-19 in brazil. **Sociedade & Natureza**, v. 33, 2021. Disponível em: <<https://seer.ufu.br/index.php/sociedadnatureza/article/view/59688>>.
- ANSELIN, L. Local indicators of spatial association—lisa. **Geographical analysis**, Wiley Online Library, v. 27, n. 2, p. 93–115, 1995.
- ANSELIN, L.; SYABRI, I.; KHO, Y. Geoda: An introduction to spatial data analysis. **Geographical Analysis**, v. 38, p. 5–22, 2006.
- BAILEY, T. C.; GATRELL, A. C. **Interactive spatial data analysis**. [S.l.]: Prentice Hall, 1995.
- BAQUI, P. *et al.* Ethnic and regional variations in hospital mortality from covid-19 in brazil: a cross-sectional observational study. **The Lancet Global Health**, v. 8, n. 8, p. e1018–e1026, 2020.
- BARABÁSI, A.-L. **Network Science**. [S.l.]: Cambridge University Press, 2016.
- BEAM, A. L.; KOHANE, I. S. Big data and machine learning in health care. **JAMA**, American Medical Association, v. 319, n. 13, p. 1317–1318, 2018.
- BEDSON, J.; SKRIP, L. A.; PEDI, D.; ABRAMOWITZ, S.; CARTER, S.; JALLOH, M. F.; FUNK, S.; GOBAT, N.; GILES-VERNICK, T.; GILES-VERNICK, T. *et al.* A review and agenda for integrated disease models including social and behavioural factors. **Nature Human Behaviour**, v. 5, n. 6, p. 675–688, 2021.
- BLONDEL, V. D.; GUILLAUME, J.-L.; LAMBIOTTE, R.; LEFEBVRE, E. Fast unfolding of communities in large networks. **Journal of Statistical Mechanics: Theory and Experiment**, IOP Publishing, v. 2008, n. 10, p. P10008, 2008.
- BOCCALETTI, S.; BIANCONI, G.; CRIADO, R.; GENIO, C. I. del; GÓMEZ-GARDEÑES, J.; ROMANCE, M.; SENDIÑA-NADAL, I.; WANG, Z.; ZANIN, M. The structure and dynamics of multilayer networks. **Physics Reports**, Elsevier, v. 544, n. 1, p. 1–122, 2014.
- BOLOGNESI, B.; RIBEIRO, E.; CODATO, A. Esquerda, centro ou direita? Como classificar os partidos no Brasil. **Observatório das Eleições**, 2020.
- BRAGA, J. U.; JR, A. N. R.; FERREIRA, A. F.; LACERDA, V. M.; FREIRE, R. M. C.; BERTONCINI, B. V. Propensity for covid-19 severe epidemic among the populations of the neighborhoods of fortaleza, brazil, in 2020. **BMC Public Health**, v. 20, p. 1486, 2020.
- BREIMAN, L. Random forests. **Machine learning**, Springer, v. 45, n. 1, p. 5–32, 2001.
- CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: **Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining**. [S.l.: s.n.], 2016. p. 785–794.
- CHI, G. Spatial regression models for demographic analysis. **Population Research and Policy Review**, Springer, v. 25, n. 4, p. 361–375, 2006.

CREPALDE, M. M. **Desigualdades socioespaciais e a dinâmica da COVID-19 no município do Rio de Janeiro, no período de 2020 - 2022**. Dissertação (Mestrado) — Fiocruz, 2024. Disponível em: <<https://www.arca.fiocruz.br/handle/icict/63583>>.

CÂMARA, S. F.; CARVALHO, H. J. B. d.; GERHARD, F.; SILVA, F. R.; PAULA, T. M. D. Vulnerabilidade socioeconômica da cidade de fortaleza ao covid-19: o epicentro da pandemia na região nordeste do brasil. **Caminhos de Geografia**, v. 22, n. 80, p. 133–152, 2021. Disponível em: <<https://www.redalyc.org/journal/2410/241064529028/html/>>.

DONALISIO, M. R.; FREITAS, A. R. L.; ZUBEN, A. P. B. V. Arboviruses causing dengue, zika and chikungunya in brazil: epidemiological situation, entomological surveillance and challenges ahead. **Revista do Instituto de Medicina Tropical de São Paulo**, v. 59, p. e48, 2017.

FORTUNATO, S. Community detection in graphs. **Physics reports**, Elsevier, v. 486, n. 3-5, p. 75–174, 2010.

FREITAS, N. **Predição de casos de dengue na cidade de Fortaleza-CE utilizando internet das coisas e aprendizado de máquina**. Dissertação (Mestrado) — Universidade Federal do Ceará, 2021. Disponível em: <<https://repositorio.ufc.br/handle/riufc/74332>>.

HUBERT, L.; ARABIE, P. Comparing partitions. **Journal of Classification**, Springer, v. 2, n. 1, p. 193–218, 1985.

KIVELÄ, M.; ARENAS, A.; BARTHELEMY, M.; GLEESON, J. P.; MORENO, Y.; PORTER, M. A. Multilayer networks. **Journal of complex networks**, Oxford University Press, v. 2, n. 3, p. 203–271, 2014.

LLOYD, S. P. Least squares quantization in pcm. **IEEE Transactions on Information Theory**, IEEE, v. 28, n. 2, p. 129–137, 1982.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. **Advances in neural information processing systems**, v. 30, 2017.

MACCORMACK-GELLES, B.; NETO, A. S. L.; SOUSA, G. S.; NASCIMENTO, O. J.; MACHADO, M. M. T.; WILSON, M. E.; CASTRO, M. C. Epidemiological characteristics and determinants of dengue transmission during epidemic and non-epidemic years in fortaleza, brazil: 2011–2015. **PLoS Neglected Tropical Diseases**, v. 12, n. 12, p. e0006990, 2018.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: UNIVERSITY OF CALIFORNIA PRESS. **Proceedings of the fifth Berkeley symposium on mathematical statistics and probability**. [S.l.], 1967. v. 1, p. 281–297.

MAGALHÃES, G. B. **Comportamento espaço-temporal da dengue e sua relação com os elementos atmosféricos e socioeconômicos em Fortaleza/CE**. Tese (Doutorado) — Universidade Federal do Ceará, 2015. Disponível em: <<https://repositorio.ufc.br/handle/riufc/17636>>.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. **Introduction to Linear Regression Analysis**. 5th. ed. [S.l.]: Wiley, 2012. ISBN 978-0470542811.

MORAN, P. A. P. Notes on continuous stochastic phenomena. **Biometrika**, Oxford University Press, v. 37, n. 1/2, p. 17–23, 1950.

NEWMAN, M. E. J. **Networks: an introduction**. [S.l.]: Oxford University Press, 2010.

OBERMEYER, Z.; EMANUEL, E. J. Predicting the future — big data, machine learning, and clinical medicine. **The New England Journal of Medicine**, Massachusetts Medical Society, v. 375, n. 13, p. 1216–1219, 2016.

OLIVEIRA, L. S.

**Análise da autocorrelação espacial dos casos e óbitos confirmados de Covid-19 em Fortaleza, Ceará** — Universidade Federal do Ceará, 2021. Disponível em: <<https://repositorio.ufc.br/handle/riufc/63003>>.

PASTOR-SATORRAS, R.; VESPIGNANI, A. Epidemic spreading in scale-free networks. **Physical review letters**, APS, v. 86, n. 14, p. 3200, 2001.

PEARSON, K. Note on regression and inheritance in the case of two parents. **Proceedings of the Royal Society of London**, v. 58, p. 240–242, 1895.

RAJKOMAR, A.; OREN, E.; CHEN, K.; DAI, A. M.; HAJAJ, N.; HARDT, M.; LIU, P. J.; LIU, X.; MARCUS, J.; SUN, M. *et al.* Scalable and accurate deep learning with electronic health records. **NPJ Digital Medicine**, Nature Publishing Group, v. 1, n. 1, p. 1–10, 2018.

REY, S. J.; ANSELIN, L.; LI, X.; SUN, Y.; KANG, W.; STEPHENS, M.; WOLF, L. J.; LIU, Z.; REY, A. M.; FURTADO, B. The pysal ecosystem: Philosophy and implementation. **Geographical Analysis**, Wiley Online Library, v. 52, n. 4, p. 527–550, 2020.

REYES, P. A.; LIU, Z.; WOLF, L. J.; REY, S. J. esda: A python package for exploratory spatial data analysis (esda). **Journal of Open Source Software**, The Open Journal, v. 5, n. 52, p. 2104, 2020.

REZENDE, A. A. B.; SILVA, R. P. d.; PEDROSA, N. L.; LUZ, R. A. d.; PAIXÃO, A. N. d.; RODRIGUES, W.; SILVA, M. A. d. R.; CAMPOS, A. d. R. Distribution of covid-19 cases and health resources in brazil's amazon region: a spatial analysis. **Ciência & Saúde Coletiva**, SciELO Public Health, v. 28, p. 131–141, 2023.

RIBEIRO, D. A.; BRAGA, A. F. D.; TEIXEIRA, L. Desigualdade socioespacial e o impacto da covid-19 na população do rio de janeiro: análises e reflexões. **Metrópole: Revista de Ciências Urbanas**, v. 23, n. 2, 2021. Disponível em: <<https://revistas.pucsp.br/index.php/metropole/article/view/51191>>.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. "why should i trust you?": Explaining the predictions of any classifier. In: **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. [S.l.: s.n.], 2016. p. 1135–1144.

RODRIGUES, M.; PRATA, D.; CAMARGO, L. Spatial modeling of covid-19 incidence and determinants in brazil. **Public Health**, Elsevier, v. 190, p. 87–94, 2021.

ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. **Journal of computational and applied mathematics**, Elsevier, v. 20, p. 53–65, 1987.

ROUX, A. V. D. **Neighborhoods and Health: What Do We Know? What Should We Do?** [S.l.]: American Journal of Public Health, 2016. Review on neighborhood effects and health.

SALATHÉ, M.; JONES, J. H. Dynamics and control of diseases in networks with community structure. **PLoS Computational Biology**, Public Library of Science, v. 6, n. 4, p. e1000736, 2010.

SANTOS, J. P. C. d.; LIMA, A. F.; SILVA, R. P. Vulnerabilidade a formas graves de covid-19: uma análise intramunicipal na cidade do rio de janeiro, brasil. **Cadernos de Saúde Pública**, v. 36, n. 4, p. e00031720, 2020. Disponível em: <<https://www.scielo.br/j/csp/a/FQPRkjPcCkkXX5qWsNgRpWp/?lang=pt>>.

SCABINI, L. F.; RIBAS, L. C.; NEIVA, M. B.; JUNIOR, A. G.; FARFAN, A. J.; BRUNO, O. M. Social interaction layers in complex networks for the dynamical epidemic modeling of covid-19 in brazil. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 564, p. 125498, 2021.

SHICKEL, B.; TIGHE, P. J.; BIHORAC, A.; RASHIDI, P. Deep learning in medical diagnosis: a survey. **Journal of biomedical informatics**, Elsevier, v. 83, p. 168–185, 2018.

SILVA, R. P. d.; LIMA, A. F.; TEIXEIRA, L. Desigualdades sociais associadas com a letalidade por covid-19 em fortaleza, ceará, brasil. **Revista Brasileira de Epidemiologia**, v. 24, n. 3, p. e21000317, 2021. Disponível em: <<https://www.scielo.br/j/ress/a/X9JPKqGqnV9ZyXWYDx4KVNWd/>>.

TOBLER, W. R. A computer movie simulating urban growth in the detroit region. **Economic geography**, Taylor & Francis, v. 46, n. sup1, p. 234–240, 1970.

TOPOL, E. **Deep medicine: how artificial intelligence can make healthcare human again**. [S.l.]: Basic Books, 2019.

TOPOL, E. **Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again**. [S.l.]: Basic Books, 2019.

VLAHOV, D.; GALEA, S. Urbanization, urbanicity, and health. **Journal of Urban Health**, v. 79, n. S1, p. S1–S12, 2002.

WHO Commission on Social Determinants of Health. **Closing the Gap in a Generation: Health equity through action on the social determinants of health**. 2008. World Health Organization Report.

YE, Y.; PANDEY, A.; BAWDEN, C.; SUMSUZZMAN, D. M.; RAJPUT, R.; SHOUKAT, A.; SINGER, B. H.; MOGHADAS, S. M.; GALVANI, A. P. Integrating artificial intelligence with mechanistic epidemiological modeling: a scoping review of opportunities and challenges. **Nature Communications**, v. 16, p. 581, 2025.

## ANEXO A – CORRELAÇÕES ALTAS DE PEARSON

<b>Variável 1</b>	<b>Variável 2</b>	<b>Correlação de Pearson</b>
chikungunya	dengue	0.765356
dengue	zika	0.743578
dengue	população	0.758837
dengue	domic_partic	0.750357
covid	população	0.706807
covid	domic_partic	0.759039
população	domic_partic	0.990809
idh_educ	idh_longev	0.721557
idh_educ	idh	0.827998
idh_longev	idh	0.814345
idh_renda	idh	0.874572
vot_esq_pref	vot_dir_pref	0.894835
vot_esq_pref	vot_esq_pres	0.854928
vot_esq_pref	vot_dir_pres	0.740325
vot_dir_pref	vot_esq_pres	0.798211
vot_dir_pref	vot_dir_pres	0.813246
vot_esq_pres	vot_dir_pres	0.785879

Fonte: elaborado pelo autor.

## ANEXO B – CORRELAÇÕES MEDIANAS DE PEARSON

Variável 1	Variável 2	Correlação de Pearson
chikungunya	dengue	0.765356
dengue	zika	0.743578
dengue	população	0.758837
dengue	domic_partic	0.750357
covid	população	0.706807
covid	domic_partic	0.759039
população	domic_partic	0.990809
idh_educ	idh_longev	0.721557
idh_educ	idh	0.827998
idh_longev	idh	0.814345
idh_renda	idh	0.874572
vot_esq_pref	vot_dir_pref	0.894835
vot_esq_pref	vot_esq_pres	0.854928
vot_esq_pref	vot_dir_pres	0.740325
vot_dir_pref	vot_esq_pres	0.798211
vot_dir_pref	vot_dir_pres	0.813246
vot_esq_pres	vot_dir_pres	0.785879
chikungunya	covid	0.442761
chikungunya	zika	0.653797
chikungunya	mercado	0.427819
chikungunya	saude	0.451013
chikungunya	cultura	0.529709
chikungunya	população	0.638673
chikungunya	domic_partic	0.626364
dengue	covid	0.610047
dengue	mercado	0.475443
dengue	saude	0.522390
dengue	cultura	0.492925
dengue	idh	-0.340966
covid	zika	0.567022
covid	mercado	0.343728
covid	saude	0.512883
covid	educacao	0.515383
covid	cultura	0.541412
covid	domic_colet	0.495330
covid	vot_dir_pres	0.303349
zika	mercado	0.370545
zika	saude	0.465327
zika	cultura	0.371659
zika	população	0.635274
zika	domic_partic	0.646322

Fonte: elaborado pelo autor.

<b>Variável 1</b>	<b>Variável 2</b>	<b>Correlação de Pearson</b>
mercado	saude	0.484237
mercado	educacao	0.449815
mercado	cultura	0.451803
mercado	população	0.467532
mercado	domic_partic	0.491129
saude	educacao	0.497079
saude	cultura	0.632929
saude	população	0.651257
saude	domic_partic	0.669001
educacao	cultura	0.559059
educacao	população	0.307612
educacao	domic_partic	0.367912
educacao	domic_colet	0.472866
cultura	população	0.509523
cultura	domic_partic	0.534289
domic_colet	avg_morarador	-0.488716
domic_colet	idh_renda	0.359403
domic_colet	idh	0.360937
avg_morarador	idh_educ	-0.574021
avg_morarador	idh_longev	-0.618086
avg_morarador	idh_renda	-0.442089
avg_morarador	idh	-0.637110
idh_educ	idh_renda	0.633230
idh_longev	idh_renda	0.466469
idh_renda	vot_esq_pres	-0.350075

Fonte: elaborado pelo autor.

## ANEXO C – MATRIZ DE CORRELAÇÃO DE PEARSON

	chikungunya	dengue	covid	zika	mercado	saude	educacao	cultura	densidade	população	domic_partic	domic_colet	avg_morador	idh_educ	idh_longev	idh_renda	idh	vot_esq_pref	vot_dir_pref	vot_esq_pres	vot_dir_pres
chikungunya	1.000000	0.765356	0.442761	0.653707	0.427819	0.451013	0.215381	0.529709	0.117121	0.638673	0.626364	0.035350	0.190334	-0.200823	-0.174413	-0.207399	-0.229473	0.192391	0.173075	0.212192	0.116849
dengue	0.765356	1.000000	0.610047	0.743879	0.475443	0.522390	0.257987	0.492225	0.094535	0.758837	0.756357	0.065812	0.199411	-0.245507	-0.239989	-0.239170	-0.346986	0.244178	0.230984	0.261112	0.154330
covid	0.442761	0.610047	1.000000	0.567022	0.343728	0.512883	0.151583	0.541412	0.066994	0.706607	0.759039	0.495330	-0.243724	0.254088	0.175889	0.259193	0.258667	0.234151	0.271013	0.141667	0.303349
zika	0.653707	0.743879	0.567022	1.000000	0.370545	0.465327	0.279892	0.371659	0.045511	0.635274	0.646322	0.108496	0.041081	-0.081969	-0.113551	-0.135746	-0.136650	0.180624	0.197847	0.180924	0.124443
mercado	0.427819	0.475443	0.343728	0.370545	1.000000	0.484237	0.449815	0.451803	0.108044	0.467532	0.491129	0.235918	0.075370	-0.207112	-0.172886	-0.239131	-0.240273	0.251797	0.224345	0.261095	0.112951
saude	0.451013	0.522390	0.512883	0.465327	0.484237	1.000000	0.497079	0.632929	0.031185	0.651257	0.669001	0.194386	0.027766	-0.061872	-0.100954	-0.241428	-0.210475	0.185465	0.211676	0.207467	0.171972
educacao	0.215381	0.257987	0.515383	0.279892	0.449815	0.497079	1.000000	0.559059	0.064765	0.307612	0.367912	0.472866	-0.199919	0.221584	0.264942	0.055007	0.178001	0.139967	0.168858	0.182387	0.239480
cultura	0.529709	0.492225	0.541412	0.371659	0.451803	0.632929	0.559059	1.000000	0.062740	0.509523	0.534289	0.299141	-0.088011	0.148941	0.065937	-0.029016	0.034476	0.143797	0.152239	0.137789	0.179071
densidade	0.117121	0.094535	0.066994	0.045511	0.108044	0.031185	0.064765	0.062740	1.000000	0.113154	0.091967	-0.079931	0.017591	0.026198	0.269350	-0.188032	-0.030372	0.034851	0.143103	0.236444	0.097045
população	0.638673	0.758837	0.706607	0.635274	0.467532	0.651257	0.307612	0.509523	0.113154	1.000000	0.999809	0.129069	0.212543	-0.197619	-0.280172	-0.171873	-0.270394	0.222952	0.246055	0.242291	0.216529
domic_partic	0.626364	0.750387	0.759039	0.646322	0.491129	0.669001	0.367912	0.534289	0.091967	0.999809	1.000000	0.238247	0.116275	-0.147342	-0.224528	-0.121165	-0.207874	0.227445	0.252611	0.231260	0.225609
domic_colet	0.035350	0.065812	0.495330	0.108496	0.235918	0.194386	0.472866	0.299141	-0.079931	0.129069	0.238247	1.000000	-0.488716	0.222253	0.265768	0.359403	0.360937	0.022640	0.085979	-0.052615	0.139649
avg_morador	0.180334	0.199411	0.243724	0.041081	0.075370	0.027766	-0.199919	-0.088011	0.017591	0.212543	0.116275	-0.488716	1.000000	-0.574021	-0.618086	-0.442089	-0.637110	0.035045	0.050931	0.175079	0.056934
idh_educ	-0.200823	-0.245507	0.254088	-0.081969	-0.207112	-0.061872	0.221584	0.148941	0.026198	-0.197619	-0.147342	0.222253	-0.574021	1.000000	0.721557	0.633230	0.627998	-0.029436	0.034532	-0.143270	0.167627
idh_longev	-0.174413	-0.200699	0.175889	-0.113551	-0.172886	-0.100954	0.264942	0.065937	0.269350	-0.280172	-0.224528	0.265768	-0.618086	0.721557	1.000000	0.466469	0.814345	0.060297	0.073991	0.017445	0.144394
idh_renda	-0.207399	-0.239170	0.259193	-0.135746	-0.239131	-0.241428	0.055007	-0.029016	-0.188032	-0.171873	-0.121795	0.359403	-0.442089	0.633230	0.466469	1.000000	0.874572	-0.208354	-0.186023	-0.350075	0.078967
idh	-0.229473	-0.346986	0.258667	-0.136650	-0.240273	-0.210475	0.178001	0.034476	-0.030372	-0.307612	-0.360937	-0.637110	0.627998	0.814345	0.874572	1.000000	0.084867	-0.069756	-0.223796	0.118541	
vot_esq_pref	0.192391	0.244178	0.234151	0.180624	0.251797	0.185465	0.139967	0.143797	0.024651	0.222952	0.227445	0.022640	0.035045	-0.029436	0.060297	-0.208354	-0.084867	1.000000	0.894835	0.854928	0.740925
vot_dir_pref	0.173075	0.230984	0.271013	0.197847	0.224345	0.211676	0.168858	0.152239	0.143103	0.246055	0.252611	0.085979	0.050931	0.034532	0.073991	-0.186023	-0.069756	0.894835	1.000000	0.798211	0.813246
vot_esq_pres	0.212192	0.261112	0.141667	0.180924	0.261095	0.207467	0.182387	0.137789	0.231644	0.242291	0.231260	-0.052615	0.175079	-0.143270	0.017445	-0.350075	-0.223796	0.854928	0.798211	1.000000	0.785879
vot_dir_pres	0.116849	0.154330	0.303349	0.124443	0.112951	0.171972	0.239480	0.179671	0.097045	0.218529	0.225609	0.139649	0.056934	0.167627	0.144394	0.078967	0.118541	0.740925	0.813246	0.785879	1.000000

Fonte: elaborado pelo autor.

**ANEXO D – DISTRIBUIÇÃO DOS CLUSTERS LISA POR DOENÇA, COM  
RESPECTIVOS BAIRROS E ESTATÍSTICAS DE AUTOCORRELAÇÃO ESPACIAL  
(MORAN’S I).**

<b>Análise</b>	<b>Clusters e Bairros</b>
<b>COVID-19</b>	<b>Moran’s I:</b> 0.0225 <b>p-valor:</b> 0.2530
<b>Alto-Alto (11)</b>	aldeota, benfica, centro, coco, dionisio torres, joaquim tavora, jose bonifacio, meireles, papicu, praia de iracema, varjota
<b>Alto-Baixo (3)</b>	farias britto, manuel dias branco, moura brasil
<b>Baixo-Alto (9)</b>	alvaro weyne, conjunto esperanca, cristo redentor, granja portugal, mondubim, padre andrade, parque presidente vargas, parque santa rosa, prefeito jose walter ancuri, antonio bezerra, bom jardim, conjunto ceara i
<b>Baixo-Baixo (4)</b>	
<b>Dengue</b>	<b>Moran’s I:</b> 0.0225 <b>p-valor:</b> 0.2530
<b>Alto-Alto (3)</b>	praia do futuro ii, vicente pinzon, vila peri
<b>Alto-Baixo (2)</b>	de lourdes, manuel dias branco
<b>Baixo-Alto (11)</b>	antonio bezerra, engenheiro luciano cavalcante, jardim guanabara, jardim iracema, olavo oliveira, padre andrade, papicu, parquelandia, quintino cunha, sao joao do tauape, vila velha
<b>Baixo-Baixo (2)</b>	cidade 2000, parque araxa
<b>Zika</b>	<b>Moran’s I:</b> -0.0564 <b>p-valor:</b> 0.1260
<b>Alto-Alto (0)</b>	—
<b>Alto-Baixo (3)</b>	de lourdes, praia do futuro ii, vicente pinzon
<b>Baixo-Alto (5)</b>	cidade 2000, edson queiroz, lagoa redonda, sao joao do tauape, sapiranga / coite
<b>Baixo-Baixo (5)</b>	aerolandia, barra do ceara, engenheiro luciano cavalcante, messejana, praia do futuro i
<b>Chikungunya</b>	<b>Moran’s I:</b> 0.1602 <b>p-valor:</b> 0.0050
<b>Alto-Alto (4)</b>	cais do porto, carlito pamplona, pirambu, vicente pinzon
<b>Alto-Baixo (2)</b>	alvaro weyne, de lourdes
<b>Baixo-Alto (11)</b>	antonio bezerra, coacu, dom lustosa, granja portugal, henrique jorge, messejana, olavo oliveira, padre andrade, parquelandia, pici, quintino cunha
<b>Baixo-Baixo (2)</b>	parque araxa, sao joao do tauape

Fonte: elaborado pelo autor.

## ANEXO E – CÁLCULO DO RISCO RELATIVO (RR)

O risco relativo (RR) é uma métrica amplamente utilizada em epidemiologia para comparar a probabilidade de ocorrência de uma determinada doença entre diferentes populações ou regiões. Neste estudo, o RR foi utilizado como base para normalizar a distribuição dos casos de doenças por bairro, permitindo comparações mais justas entre áreas com diferentes tamanhos populacionais.

O processo de cálculo do RR foi realizado em duas etapas principais:

1. Primeiro, foi calculada a incidência da doença em cada bairro, representando a proporção de casos em relação à população local. A fórmula é:

$$\text{Incidência}_i^{(d)} = \frac{\text{Casos}_i^{(d)}}{\text{População}_i} \quad (\text{E.1})$$

Onde:

- $\text{Casos}_i^{(d)}$  é o número total de casos da doença  $d$  no bairro  $i$ ;
  - $\text{População}_i$  é a população total do bairro  $i$ ;
  - $\text{Incidência}_i^{(d)}$  é a taxa de incidência da doença  $d$  no bairro  $i$ .
2. Em seguida, foi calculado o **risco relativo** propriamente dito, dividindo a incidência local pela média de incidência da cidade como um todo:

$$RR_i^{(d)} = \frac{\text{Incidência}_i^{(d)}}{\text{Incidência}^{(d)}} = \frac{\frac{\text{Casos}_i^{(d)}}{\text{População}_i}}{\frac{\sum_{j=1}^n \text{Casos}_j^{(d)}}{\sum_{j=1}^n \text{População}_j}} \quad (\text{E.2})$$

Onde:

- $\text{Incidência}^{(d)}$  representa a incidência média da doença  $d$  em Fortaleza;
- $RR_i^{(d)} > 1$  indica que o bairro possui risco superior à média da cidade;
- $RR_i^{(d)} < 1$  indica risco inferior à média;
- $RR_i^{(d)} \approx 1$  indica risco próximo à média municipal.

Essa metodologia foi essencial para garantir comparabilidade e robustez nas análises espaciais, na modelagem de redes multilayer e nos algoritmos de aprendizado de máquina.

## ANEXO A – PROPOSTA DE APRIMORAMENTO METODOLÓGICO (FRAMEWORK INTEGRADO)

Este anexo apresenta uma *proposta de aprimoramento metodológico* a ser incorporada em versões futuras da pesquisa. Entendemos *framework*, no sentido metodológico, como um *arranjo integrado de procedimentos* que articula: (i) curadoria e integração de dados; (ii) análises estatísticas; (iii) análises espaciais; (iv) modelagem por redes multilayer; e (v) aprendizado de máquina para predição e tipificação territorial. A adoção dessa estrutura visa ampliar reprodutibilidade, comparabilidade entre cidades/períodos e transparência de decisões analíticas. *Observação:* As diretrizes abaixo foram refinadas durante a tese, mas sua implementação completa é proposta como trabalho futuro, sem alterar os resultados consolidados no corpo do texto.

### A.1 Curadoria e Integração de Dados (M1)

**Entradas:** bases oficiais (epidemiologia, demografia/socioeconômico, infraestrutura, política), malha territorial e metadados.

**Procedimentos:** verificação de consistência; padronização de formatos; tratamento de faltantes/outliers; harmonização toponímica; integração temporal/espacial.

**Saídas:** base unificada por bairro e ano; dicionário de dados.

**Dependências:** alimenta M2 (estatística), M3 (espaço), M4 (rede) e M5 (ML).

#### A.1.1 Fontes dos Dados

Identificar e documentar as bases que sustentarão as análises, assegurando: (i) cobertura temporal; (ii) granularidade espacial compatível (p.ex., bairro); (iii) rastreabilidade (metadados, versão, data de extração); (iv) licenças/ética de uso. Objetivo: formar um inventário mínimo reproduzível de insumos.

Procedimentos essenciais.

- (1) Delimitar unidade espacial e período de análise.
- (2) Especificar requisitos por dimensão (epidemiológica, demográfica/socioeconômica, infraestrutura, política, malha geográfica).

- (3) Levantar e priorizar fontes que cumpram os requisitos.
- (4) Registrar metadados (órgão, URL, versão, data/hora de coleta, licença).
- (5) Verificar integridade (cobertura, formatos, chaves territoriais) e aderência a padrões.

Aplicação ao estudo de caso (Fortaleza) — referência.

Prefeitura (equipamentos e rede escolar); SMS-Fortaleza (séries mensais 2020–2024 de casos/óbitos<sup>1</sup>); outras fontes municipais (areninhas, Rede Cuca etc.); INEP (infraestrutura educacional); IBGE/Censo 2022 (demografia); Atlas do Desenvolvimento Humano 2010 (IDH); TRE/TSE (eleições 2020/2022). No framework futuro, a seleção e documentação seguirão o inventário reprodutível acima.

### ***A.1.2 Seleção de Variáveis***

Definir o conjunto mínimo suficiente para sustentar análises estatísticas, espaciais, em rede e preditivas, alinhando variáveis a *constructos* (demografia, infraestrutura, IDH, política), reduzindo redundâncias e garantindo comparabilidade.

Procedimentos essenciais.

- (1) Avaliar candidatos a variáveis por dimensão.
- (2) Mapear cada variável a um *constructo*/hipótese.
- (3) Verificar completude, variância e redundância (correlações/colinearidade).
- (4) Padronizar unidades/escala e codificação.
- (5) Montar dicionário de dados versionado.

#### ***A.1.2.1 Categorização das Variáveis Políticas***

Resultados eleitorais convertidos em variáveis comparáveis por unidade e pleito, com tipificação ideológica rastreável e foco no 1º turno para evitar viés de alianças do 2º turno.

Operacionalização proposta.

- Pleitos e escopo: 2020 (prefeito) e 2022 (presidente), por bairro.
- Seleção: dois mais votados por pleito, somente 1º turno.

---

<sup>1</sup> Óbitos não utilizados na modelagem final.

- Mapeamento ideológico: {esquerda, centro, direita} com base em fonte acadêmica (p.ex., Observatório das Eleições 2020), com conferências no TSE (2020–2024).
- Construção: proporções por espectro; indicador simples de polarização (%direita–%esquerda).
- Compatibilização territorial: alinhamento de chaves e log de harmonização.

### ***A.1.3 Limpeza e Integração dos Dados***

Garantir coerência, completude e integrabilidade entre fontes; unificar por chaves territoriais e temporais; produzir base unificada por unidade e período, com versionamento.

Procedimentos essenciais.

- (1) Normalizar tipos/formatos (datas, números, categorias).
- (2) Harmonizar toponímia e chaves territoriais.
- (3) Tratar faltantes e valores extremos (regras claras e log).
- (4) Integrar por chaves (espaciais/temporais).
- (5) Versionar artefatos (dataset final, dicionário e log de alterações).

### ***A.1.4 Análise Exploratória Inicial***

Entender distribuições e relações dominantes para orientar seleção, transformações e hipóteses, sem antecipar inferências causais.

Procedimentos essenciais.

- (1) Histogramas/boxplots por variável.
- (2) Dispersões e *heatmaps* de correlação.
- (3) Tabelas-resumo por estratos (quartis, regiões).
- (4) Mapas rápidos de RR/níveis.
- (5) Registro de achados e anomalias.

### ***A.1.5 Métrica de Risco Epidemiológico (RR)***

Escala comparável entre unidades:  $RR = \frac{\text{incidência local}}{\text{incidência média de referência}}$ . Mitiga variações populacionais e fundamenta análises estatísticas, espaciais e em rede.

Procedimentos essenciais.

- (1) Definir área/período de referência.
- (2) Calcular incidências e RR por doença/unidade.
- (3) Tratar zeros/valores muito pequenos com regra explícita.
- (4) Classificar RR em faixas (quartis) para leitura.
- (5) Anexar RR ao dataset integrado.

### ***A.1.6 Considerações Éticas***

Uso preferencial de bases públicas e agregadas; checagem de licenças e ausência de dados pessoais identificáveis; nota ética conforme normativas.

Procedimentos essenciais.

- (1) Verificar natureza pública/oficial e licença.
- (2) Garantir ausência de dados pessoais identificáveis.
- (3) Registrar origem, data de extração e versão.
- (4) Avaliar necessidade de CEP segundo normativas.
- (5) Redigir nota ética/LGPD.

## **A.2 Estatística Descritiva e Inferencial (M2)**

**Entradas:** base M1, RR e variáveis selecionadas.

**Procedimentos:** correlações de Pearson (e Spearman, se aplicável); OLS por doença.

**Saídas:** matrizes de correlação; gráficos de dispersão; modelos OLS com diagnósticos.

### ***A.2.1 Correlação de Pearson***

Triagem de relações e detecção de redundâncias; insumo para seleção de preditores e especificação de modelos.

Procedimentos essenciais.

- (1) Padronizar variáveis quando necessário.
- (2) Calcular matriz  $r$  e valores- $p$ .
- (3) Visualizar por *heatmap*/dispersões.
- (4) Registrar correlações fortes e moderadas.

### A.2.2 *Regressão Linear Múltipla (OLS)*

Efeitos ajustados por doença: a OLS complementa a correlação ao estimar o efeito de cada variável mantendo as demais constantes.

Procedimentos essenciais.

- (1) Especificar um modelo por doença (dependente: RR).
- (2) Escolher preditores por teoria e correlação, evitando multicolinearidade.
- (3) Estimar OLS parcimoniosa.
- (4) Verificar resíduos e  $R^2$ .
- (5) Reportar coeficientes, sinais, IC e valores- $p$ .

### A.3 *Dependência e Padrões Espaciais (M3)*

**Entradas:** RR por unidade, malha e matriz de vizinhança.

**Procedimentos:** Moran Global; LISA; síntese de hotspots multi-doença.

**Saídas:** estatísticas de autocorrelação; mapas de clusters; índice acumulado de hotspots.

#### A.3.1 *Autocorrelação Espacial Global (Moran's I)*

Checar dependência global; orientar aplicação de análises locais (LISA) e interpretação.

Procedimentos essenciais.

- (1) Definir matriz de vizinhança (queen/ $k$ -vizinhos/distância).
- (2) Calcular Moran's I com permutações.

- (3) Reportar  $I$ ,  $z$  e valor- $p$ .
- (4) (Opcional) Sensibilidade a escolhas de vizinhança.

### **A.3.2 Autocorrelação Espacial Local (LISA)**

Detectar clusters locais (Alto–Alto, Baixo–Baixo) e outliers (Alto–Baixo, Baixo–Alto) para priorização territorial.

Procedimentos essenciais.

- (1) Calcular LISA por doença sobre RR (mesma vizinhança do global).
- (2) Definir significância ( $\alpha$ ) e permutações.
- (3) Classificar e mapear categorias.
- (4) Registrar agrupamentos significativos.

### **A.3.3 Hotspots Acumulados**

Sintetizar frequência de Alto–Alto entre doenças por unidade; mapear sobreposições de risco.

Procedimentos essenciais.

- (1) Recuperar rótulos LISA por doença/unidade.
- (2) Definir critério de inclusão (p.ex.,  $p < 0,05$ ).
- (3) Somar frequências e mapear.
- (4) (Opcional) Classificar faixas de prioridade.

## **A.4 Rede Multilayer (M4)**

**Entradas:** RR (M1), vetores estruturais (IDH, demografia, infraestrutura, política), regras de pesos.

**Procedimentos:** arestas intra (similaridade RR; p20 por distância em camadas estruturais); arestas inter (acoplamento/afinidade); Louvain; centralidades.

**Saídas:** grafo ponderado; comunidades; grau/eigenvector/betweenness; mapas por camada.

#### **A.4.1 Estrutura da Rede**

Organizar dimensões em espaço relacional multilayer; nós por unidade/camada; arestas como afinidades intra/inter.

Procedimentos essenciais.

- (1) Definir camadas e variáveis; unidade e período.
- (2) Padronizar variáveis por camada.
- (3) Arestas intra: epidemiologia (similaridade de RR); estruturais (distância com limiar percentílico e conversão em similaridade).
- (4) Arestas inter: acoplamento por identidade e, se pertinente, modulação por afinidade temática.
- (5) Normalizar pesos; remover auto-laços; checar densidade/componentes; exportar *edgelist/nodelist*.

#### **A.4.2 Detecção de Comunidades**

Partição coesa para reduzir complexidade e revelar blocos temáticos/territoriais; modularidade  $Q$  como indicador de estrutura.

Procedimentos essenciais.

- (1) Executar Louvain no grafo ponderado.
- (2) Registrar  $Q$ , número e tamanhos de comunidades.
- (3) Mapear rótulos por camada/território e verificar coerência.

#### **A.4.3 Centralidade e Estrutura**

Identificar nós influentes e pontes (grau ponderado, eigenvector, betweenness) para qualificar interpretação das comunidades/conectividade.

Procedimentos essenciais.

- (1) Definir cálculo no multilayer e/ou por camada.
- (2) Calcular e ranquear métricas.

- (3) Confrontar com comunidades e contexto territorial.

## A.5 Modelagem Supervisionada e Tipificação (M5)

**Entradas:** *features* estruturais/demográficas/infraestrutura/política; alvo RR.

**Procedimentos:** treino/validação (CV  $k$ -fold); XGBoost e Random Forest; SHAP; K-Means (tipificação).

**Saídas:** RMSE/MAE/ $R^2$ ; explicabilidade; rótulos de clusters.

### A.5.1 Predição do RR (XGBoost/RF)

Quantificar capacidade preditiva em comparação a um baseline simples e gerar artefatos reproduzíveis.

Procedimentos essenciais.

- (1) Definir alvo (RR por doença) e *features* padronizadas quando necessário.
- (2) Treinar XGBoost/RF com CV  $k$ -fold e `random_state`.
- (3) Avaliar RMSE, MAE,  $R^2$  vs. baseline (média).
- (4) Registrar transformações e checar *leakage*.
- (5) Salvar parâmetros, sementes e métricas.

### A.5.2 Explicabilidade (SHAP)

Valores de Shapley para interpretar contribuições por variável (global/local) e verificar plausibilidade.

Procedimentos essenciais.

- (1) Calcular SHAP por doença no modelo final.
- (2) Produzir *summary* e *dependence plots*.
- (3) Confrontar ranking com hipóteses e estatística clássica.

### A.5.3 Tipificação Territorial (K-Means)

Agrupar unidades em perfis multivariados para facilitar leitura e ação.

Procedimentos essenciais.

- (1) Padronizar variáveis de entrada.
- (2) Escolher  $k$  (cotovelo) e fixar `random_state`.
- (3) Estimar clusters, extrair centróides e mapear rótulos.
- (4) (Opcional) Testar estabilidade com sementes distintas.

## A.6 Tabela-resumo do Framework (Proposta)

Tabela 5 – Resumo dos módulos do arranjo metodológico integrado (proposta futura)

<b>Módulo</b>	<b>Entradas</b>	<b>Saídas</b>
M1 Curadoria	Bases oficiais; malha; metadados	Base unificada; dicionário; log
M1.2 RR	Casos; população de referência	RR por doença; classes de risco
M2 Estatística	Base unificada; RR; variáveis	Matrizes; OLS + diagnósticos
M3 Espacial	RR; malha; vizinhança	Moran/LISA; mapas; hotspots
M4 Rede	RR; vetores estruturais; regras de peso	Grafo; comunidades; centralidades
M5 Predição	Features; RR (alvo)	RMSE/MAE/ $R^2$ ; SHAP
M5.1 K-Means	Variáveis padronizadas	Rótulos; centróides; estabilidade

Fonte: elaborado pelo autor.

*Síntese:* Este anexo descreve um roteiro metodológico unificado que pode ser adotado e ampliado em trabalhos futuros para generalizar a aplicação a outras cidades/doenças, fortalecer a reprodutibilidade e facilitar atualizações periódicas sem impacto nos resultados já publicados nesta tese.

## ANEXO B – TABELA COM QUANTITATIVO DE DOENÇAS

bairro	COVID	chikungunya	dengue	zika
0 aerolandia	1807	150.0	321.0	7.0
1 aeroporto	454	53.0	103.0	0.0
2 aldeota	11936	126.0	354.0	7.0
3 alto da balanca	1042	85.0	138.0	2.0
4 alvaro weyne	2888	180.0	425.0	7.0
5 amadeu furtado	970	8.0	58.0	3.0
6 ancuri	1501	175.0	550.0	8.0
7 antonio bezerra	4503	56.0	419.0	17.0
8 aracape	695	26.0	156.0	0.0
9 autran nunes	1508	51.0	360.0	2.0
10 barra do ceara	6760	1415.0	1541.0	32.0
11 barroso	2936	377.0	833.0	14.0
12 bela vista	2181	55.0	206.0	7.0
13 benfica	2624	61.0	187.0	8.0
14 boa vista / castelao	2187	80.0	223.0	5.0
15 bom futuro	758	30.0	82.0	1.0
16 bom jardim	5168	207.0	712.0	26.0
17 bonsucesso	3905	125.0	788.0	2.0
18 cais do porto	1323	481.0	508.0	10.0
19 cajazeiras	1595	64.0	252.0	2.0
20 cambeba	2102	32.0	137.0	3.0
21 canindezinho	2994	241.0	1069.0	11.0
22 carlito pamplona	2019	238.0	420.0	8.0
23 centro	7374	289.0	720.0	11.0
24 cidade 2000	2334	66.0	225.0	2.0
25 cidade dos funcionarios	2791	185.0	305.0	5.0
26 coacu	1000	6.0	95.0	1.0

Continua na próxima página

	bairro	COVID	chikungunya	dengue	zika
27	coco	4942	30.0	122.0	2.0
28	conjunto ceara i	6618	318.0	1268.0	20.0
29	conjunto ceara ii	995	19.0	103.0	2.0
30	conjunto esperanca	1730	52.0	245.0	6.0
31	conjunto palmeiras	2264	505.0	574.0	1.0
32	couto fernandes	439	16.0	45.0	1.0
33	cristo redentor	2224	812.0	909.0	5.0
34	curio	736	30.0	137.0	1.0
35	damas	2156	37.0	144.0	6.0
36	de lourdes	554	1.0	13.0	0.0
37	democrito rocha	2459	32.0	222.0	1.0
38	dias macedo	1416	82.0	438.0	4.0
39	dionisio torres	3647	29.0	123.0	3.0
40	dom lustosa	861	17.0	121.0	1.0
41	edson queiroz	3542	119.0	273.0	4.0
42	ellery	1443	118.0	242.0	7.0
43	engenheiro luciano cavalcante	4010	375.0	400.0	17.0
44	farias brito	1489	62.0	128.0	4.0
45	fatima	5417	73.0	289.0	12.0
46	floresta	1045	49.0	131.0	0.0
47	genibau	2623	166.0	959.0	20.0
48	granja lisboa	3279	104.0	512.0	11.0
49	granja portugal	3826	162.0	1240.0	14.0
50	guajeru	587	19.0	109.0	0.0
51	guararapes	1795	4.0	29.0	0.0
52	henrique jorge	4069	109.0	652.0	8.0
53	itaoca	1608	82.0	225.0	8.0
54	itaperi	3901	372.0	931.0	15.0
55	jacarecanga	3096	239.0	451.0	4.0

Continua na próxima página

	bairro	COVID	chikungunya	dengue	zika
56	jangurussu	8456	679.0	1979.0	23.0
57	jardim america	1760	83.0	244.0	9.0
58	jardim cearense	1127	44.0	163.0	2.0
59	jardim das oliveiras	3010	375.0	271.0	7.0
60	jardim guanabara	1964	34.0	118.0	2.0
61	jardim iracema	2607	124.0	266.0	2.0
62	joao xxiii	2588	68.0	384.0	2.0
63	joaquim tavora	6758	75.0	336.0	3.0
64	joquei clube	3048	44.0	281.0	7.0
65	jose bonifacio	1596	30.0	99.0	2.0
66	jose de alencar	1859	39.0	120.0	0.0
67	lagoa redonda	3264	131.0	363.0	8.0
68	manoel satiro	2187	171.0	579.0	27.0
69	manuel dias branco	813	34.0	57.0	0.0
70	maraponga	3888	156.0	577.0	13.0
71	meireles	12306	108.0	356.0	13.0
72	messejana	8837	175.0	1176.0	27.0
73	mondubim	8371	711.0	1914.0	46.0
74	monte castelo	2838	201.0	512.0	6.0
75	montese	5368	254.0	642.0	18.0
76	moura brasil	304	43.0	90.0	0.0
77	mucuripe	2428	204.0	238.0	13.0
78	novo mondubim	1302	58.0	156.0	1.0
79	olavo oliveira	581	6.0	30.0	2.0
80	padre andrade	1205	31.0	129.0	2.0
81	panamericano	1239	34.0	153.0	1.0
82	papicu	4978	148.0	358.0	9.0
83	parangaba	5170	236.0	843.0	9.0
84	parque araxa	1105	54.0	167.0	2.0

Continua na próxima página

	bairro	COVID	chikungunya	dengue	zika
85	parque dois irmaos	2923	389.0	795.0	2.0
86	parque iracema	1460	8.0	57.0	1.0
87	parque manibura	987	90.0	87.0	1.0
88	parque presidente vargas	812	49.0	242.0	3.0
89	parque santa maria	770	57.0	331.0	0.0
90	parque santa rosa	1107	43.0	348.0	2.0
91	parque sao jose	1123	201.0	448.0	1.0
92	parquelandia	3469	52.0	228.0	11.0
93	parreao	913	34.0	82.0	2.0
94	passare	7307	383.0	1712.0	21.0
95	paupina	2049	133.0	393.0	8.0
96	pedras	963	28.0	132.0	0.0
97	pici	3042	115.0	518.0	11.0
98	pirambu	860	214.0	330.0	4.0
99	planalto ayrton senna	2374	441.0	971.0	33.0
100	praia de iracema	1135	75.0	115.0	1.0
101	praia do futuro i	814	232.0	726.0	22.0
102	praia do futuro ii	1333	58.0	334.0	3.0
103	prefeito jose walter	6010	1067.0	1137.0	52.0
104	presidente kennedy	2805	73.0	235.0	7.0
105	quintino cunha	3155	76.0	221.0	9.0
106	rachel de queiroz	619	39.0	123.0	3.0
107	rodolfo teofilo	3124	97.0	308.0	9.0
108	sabiaguaba	661	36.0	96.0	0.0
109	salinas	580	18.0	59.0	0.0
110	sao bento	554	22.0	60.0	2.0
111	sao gerardo	2001	15.0	120.0	3.0
112	sao joao do tauape	3648	312.0	474.0	7.0
113	sapiranga / coite	3668	283.0	579.0	11.0

Continua na próxima página

	bairro	COVID	chikungunya	dengue	zika
114	serrinha	4023	527.0	1588.0	22.0
115	siqueira	3806	158.0	749.0	24.0
116	varjota	1736	46.0	121.0	7.0
117	vicente pinzon	4954	973.0	1295.0	11.0
118	vila peri	2721	304.0	554.0	4.0
119	vila uniao	2765	294.0	488.0	32.0
120	vila velha	4557	179.0	366.0	7.0

Fonte: elaborado pelo autor.