UNIVERSIDADE FEDERAL DA BAHIA

Dérick Gabriel Fernandes Borges

Utilização de bases de dados para a vigilância epidemiológica combinando conceitos e técnicas de sistemas dinâmicos, inteligência artificial e ciência de rede.

Tese de Doutorado

Tese apresentada ao Programa de Pós–graduação em Física do Instituto de Física da UFBA como requisito parcial para obtenção do título de Doutor em Física

Orientador: Prof. Dr. Roberto Fernandes Silva Andrade
Coorientador: Profa. Dra. Suani Tavares Rubim de Pinho

Salvador
Julho de 2025

UNIVERSIDADE FEDERAL DA BAHIA

Dérick Gabriel Fernandes Borges

Utilização de bases de dados para a vigilância epidemiológica combinando conceitos e técnicas de sistemas dinâmicos, inteligência artificial e ciência de rede.

Tese apresentada ao Programa de Pós–graduação em Física do Instituto de Física da UFBA como requisito parcial para obtenção do título de Doutor em Física. Aprovada pela comissão examinadora abaixo assinada.

Prof. Dr. Roberto Fernandes Silva Andrade
Orientador
Instituto de Física — UFBA

Profa. Dra. Suani Tavares Rubim de Pinho
Corientadora
Instituto de Física — UFBA

Prof. Dr. José Garcia Vivas Miranda
Instituto de Física — UFBA

Prof. Dr. Rodrigo Fernando Lugon Cornejo von Marttens
Instituto de Física — UFBA

Prof. Dr. Sílvio da Costa Ferreira Junior
Dep. de Física — UFV

Prof. Dr. Leonardo Bacelar Lima Santos
Cemaden — MCTI

Profa. Dra. Suani Tavares Rubim de Pinho
Coordenadora do programa de Pós Graduação em Física do
Instituto de Física — UFBA

Salvador, 21 de Julho de 2025

Dérick Gabriel Fernandes Borges

Graduado e Mestre em Física pela Universidade Federal da Bahia. Tem experiência na área de Física, com ênfase em Física Estatística e Sistemas Complexos, atuando principalmente nos seguintes temas: Modelagem Computacional, Ciência de Redes, Inteligência Artificial e Dinâmica Epidemiológica.

Ficha Catalográfica elaborada pela Biblioteca Universitária de Ciências e Tecnologias Prof. Omar Catunda, SIBI - UFBA.

B732 Borges, Dérick Gabriel Fernandes

Utilização de bases de dados para a vigilância epidemiológica combinando conceitos e técnicas de sistemas dinâmicos, inteligência artificial e ciência de rede. / Dérick Gabriel Fernandes Borges. — Salvador 2025.

175 f.

Orientador: Prof. Dr. Roberto Fernandes Silva Andrade
Coorientadora: Prof.ª Dr.ª Suani Tavares Rubim de Pinho

Tese (Doutorado em Física) - Universidade Federal da Bahia, Instituto de Física, 2025.

1. Vigilância Epidemiologia. 2. Aprendizado de Máquina. 3. Número de Reprodução. 4. Ciência de Redes. I. Andrade, Roberto Fernandes Silva. II. Pinho, Suani Tavares Rubim. III. Universidade Federal da Bahia. IV. Título.

CDU: 004.658:616-036.22

Agradecimentos

Expresso meu profundo agradecimento à minha mãe, Maria José Fernandes Boa Sorte, pelo suporte incondicional e pela estrutura que me proporcionou, essenciais para concluir mais um nível em minha formação.

Um agradecimento especial é dedicado à memória do meu avô, Francisco de Assis Boa Sorte. Sua vida e seus valores continuam a me inspirar todos os dias.

Ao meu orientador, Roberto Fernandes Silva Andrade, exemplo de dedicação e inspiração no campo da pesquisa, manifesto minha profunda gratidão. Sua orientação, reflexões, discussões, presença constante e paciência foram pilares essenciais no meu desenvolvimento como pesquisador.

À minha coorientadora, Suani Tavares Rubim de Pinho, agradeço pela atenção e pelo suporte contínuo durante o desenvolvimento deste trabalho.

Agradeço à minha namorada, Raiana de Oliveira Castro, por estar ao meu lado em cada etapa desse processo me apoiando e incentivando, sua presença foi importante para mais essa conquista.

Agradecimento ao meu amigo Eluã Ramos Coutinho, pelas longas conversas, pelo apoio constante e, principalmente, pela ajuda inestimável na programação e na escrita dos artigos científicos derivados desse trabalho.

Por fim, sou grato ao CNPq pelo apoio financeiro.

Resumo

Este estudo explora a aplicação de sistemas dinâmicos, abordagens estatísticas, inteligência artificial e ciência de redes no contexto da vigilância epidemiológica, com ênfase na vigilância sindrômica de infecções respiratórias. Um primeiro estudo foi realizado utilizando dados da atenção primária à saúde de 27 regiões geográficas imediatas, correspondentes às capitais dos estados do Brasil. A integração de inteligência artificial e sistemas dinâmicos resultou na criação do Modelo Misto de Inteligência Artificial e Próxima Geração, que combina diferentes métodos para aprimorar a detecção precoce de surtos a partir de séries temporais. Em seguida, um segundo estudo foi realizado, aplicando um modelo metapopulacional e conceitos da ciência de redes. Utilizando informações de mobilidade, e dados da atenção primária de saúde de um dos maiores estados do Brasil, a Bahia, investigou-se a disseminação espacial de potenciais doenças respiratórias com a identificação de hubs de propagação, a partir de um índice sentinela. Este trabalho contribui diretamente para o projeto Sistema de Alerta Precoce para Surtos com Potencial Epi-Pandêmico (ÆSOP), demonstrando o potencial de novas ferramentas para mitigar o impacto de doenças emergentes e reemergentes no Brasil.

Palavras–chave

Vigilância Epidemiologia, Aprendizado de Máquina, Número de Reprodução, Ciência de Redes.

Abstract

This study explores the application of dynamical systems, statistical approaches, artificial intelligence and network science in the context of epidemiological surveillance, with an emphasis on syndromic surveillance of respiratory infections. A first study was carried out using primary health care data from 27 immediate geographic regions, corresponding to the capitals of the states of Brazil. The integration of artificial intelligence and dynamical systems resulted in the creation of the Mixed Model of Artificial Intelligence and Next Generation, which combines different methods to improve the early detection of outbreaks from time series. Then, a second study was carried out, applying a metapopulation model and concepts from network science. Using mobility information and primary health care data from one of the largest states in Brazil, Bahia, the spatial dissemination of potential respiratory diseases was investigated by identifying propagation hubs, based on a sentinel index. This work directly contributes to the project *Alert-Early System for Outbreaks with Pandemic Potential* (ÆSOP), demonstrating the potential of new tools to mitigate the impact of emerging and re-emerging diseases in Brazil.

Keywords

Surveillance Epidemiology, Machine Learning, Reproduction Number, Network Science.

Publicações

Alguns trabalhos derivados desta tese já foram submetidos para publicação. Os detalhes das publicações são fornecidos abaixo:

(P1) *Combining machine learning and dynamic system techniques to early detection of respiratory outbreaks in routinely collected primary healthcare records*, submetido à *BMC Medical Research Methodology* (ISSN: 1471-2288) em 07 de Maio de 2024 e aceito em 26 de Março de 2025. DOI: `https://doi.org/10.1186/s12874-025-02542-0`

(P2) *An integrated framework for modeling respiratory disease transmission and designing surveillance networks using a sentinel index*, submetido à *Royal Society Open Science* (ISSN: 2054-5703) em 06 de Março de 2025 e aceito em 04 de Agosto de 2025. DOI: `https://doi.org/10.1098/rsos.251195`

Sumário

Lista de Abreviações

AIDS	Síndrome da Imunodeficiência Adquirida
APS	Atenção Primária à Saúde
AUC	Área Sob a Curva (AUC do inglês, area under the curve)
ÆSOP	Sistema de Alerta Precoce para Surtos com Potencial Epi-Pandêmico (ÆSOP do inglês, Alert-Early System of Outbreaks with Pandemic Potential)
ASMODEE	Seleção Automática de Modelos e Detecção de Outliers para Epidemias (ASMODEE do inglês, Automatic Selection of Models and Outlier Detection for Epidemics)
CIEVS	Centro de Informações Estratégicas em Vigilância em Saúde
CIDACS	Centro de Integração de Dados e Conhecimentos para Saúde
CID-10	Classificação Internacional de Doenças
CIAP-2	Classificação Internacional de Atenção Primária
COPOD	Detecção de Anomalia Baseado em Cópula (COPOD do inglês, Copula-Based Outlier Detection)
DQI	Índice de Qualidade de Dado (DQI do inglês, Data Quality Index)
EWS	Sinais de Alerta Precoce (EWS do inglês, Early Warning Signals)
EARS	Sistema de notificação antecipada de aberrações (EARS do inglês, Early Aberration Reporting System)
EVI	Índice de Volatilidade Epidêmica (EVI do inglês, Epidemic Volatility Index)
EDOs	Equações Diferenciais Ordinárias
EDPs	Equações Diferenciais Parciais
IA	Inteligência Artificial
IS	Índice Sentinela

ISF	Floresta de Isolamento (ISF do inglês, Isolation Forest)
IVAS	Infecções de Vias Aéreas Superiores
LOF	Fator de Anomalia Local (LOF do inglês, Local Outlier Factor)
ML	Aprendizado de Máquina (ML do inglês, Machine Learning)
MMAING	Modelo Misto de Inteligência Artificial e Próxima Geração (MMAING do inglês, Mixed Model of Artificial Inteligence and Next Generation)
NGM	Método de Próxima Geração (NGM do inglês, Next-Generation Model)
OCSVM	Máquina de Vetor de Suporte de Classe Única (OCSVM do inglês, One-class Support Vector Machine
RGI	Região Geográfica Imediata
ROC	Característica Operacional do Receptor (ROC do inglês, Receiver Operating Characteristic)
SEIR	Modelo compartimental Suscetível - Exposto - Infectado - Recuperado
SIR	Modelo compartimental Suscetível - Infectado - Recuperado
SISAB	Sistema de Informação em Saúde para a Atenção Básica
TDO	Técnica de Detecção de Outlier

1
Introdução

O que vias metabólicas, ecossistemas, o mercado financeiro e a propagação de doenças infecciosas têm em comum? Até algumas décadas atrás, a resposta teria sido "muito pouco". Os dois primeiros exemplos estão relacionados a problemas biológicos e moldados pela dinâmica evolutiva; o terceiro é uma criação humana inserida no sistema capitalista; e o quarto envolve uma complexa interação entre fatores biológicos e sociológicos. Contudo, atualmente, reconhece-se que todos esses sistemas compartilham características fenomenológicas semelhantes, nas quais a aparente imprevisibilidade tem origem em leis estatísticas e físicas bem definidas. Esses fenômenos não triviais caracterizam uma ampla gama de condições que levam à formação dos chamados sistemas complexos [1].

Os sistemas complexos constituem um tema central na física e em outras ciências naturais, nos quais a complexidade emerge da interação entre múltiplos componentes. Tais sistemas são frequentemente caracterizados por comportamento não linear, capacidade de auto-organização e padrões emergentes, embora a presença de apenas algumas dessas características já possa ser suficiente para classificar um sistema como complexo [1, 2]. Uma propriedade fundamental desses sistemas é a multiestabilidade, que permite transições súbitas e frequentemente imprevisíveis entre diferentes estados dinâmicos [3]. Tais transições, especialmente em sistemas naturais, podem levar a eventos catastróficos, tornando a previsão e a detecção dessas mudanças essenciais para mitigar impactos negativos e gerenciar riscos [4, 5].

A ciência de redes é uma ferramenta importante para entender e modelar sistemas complexos, permitindo a análise de estruturas e padrões nas interações entre seus componentes [6]. O estudo de redes facilita a identificação de componentes (representados por vértices) e interações (representadas por arestas) críticos, fundamentais para a dinâmica do sistema. Técnicas de análise estrutural, como a detecção de comunidades e a centralidade de intermediação, fornecem insights relevantes para compreender a organização, dinâmica e influência dos elementos dentro de um sistema complexo, permitindo a identificação de padrões, pontos críticos e relações que impactam o funcionamento global da rede. Essas abordagens são aplicáveis em diversos contextos, como redes sociais, biológicas, epidemiológicas e tecnológicas, auxiliando na tomada de decisões e na otimização de processos [7, 8].

No contexto epidemiológico, a emergência e reemergência de doenças infecciosas são exemplos clássicos de sistemas complexos. A dinâmica de transmissão de patógenos envolve fatores biológicos, ambientais e sociais, resultando em padrões de propagação altamente variáveis e previsão desafiadora [9, 10]. Modelos epidemiológicos tradicionais, tanto determinísticos quanto estocásticos, enfrentam dificuldades para capturar essas complexidades, especialmente quando se busca antecipar transições críticas que antecedem surtos e epidemias [4, 11]. No entanto, avanços recentes na ciência indicam que a proximidade de uma transição crítica, inclusive em sistemas epidêmicos, pode ser detectada por meio de sinais genéricos — padrões estatísticos universais que emergem independentemente dos detalhes específicos do sistema. Esses sinais podem ser utilizados como alertas precoces da iminência de surtos ou mudanças abruptas no regime transmissão [4, 5, 11–13].

O problema da detecção precoce de surtos epidemiológicos é um desafio constante para a saúde pública, dado que sistemas de alerta precoce ainda enfrentam limitações relacionadas à qualidade e à tempestividade dos dados, além de sensibilidade insuficiente para captar surtos emergentes [14–17]. Os estudos apresentados nesta tese se baseiam na hipótese de que a integração de técnicas de inteligência artificial com a modelagem epidemiológica, por meio de modelos de tempo de infecção, modelos metapopulacionais e métodos de ciência de redes, aplicada a dados, incluindo dados sindrômicos, podem fornecer alertas mais precoces do que a vigilância tradicional. Além disso, assumimos que a análise de redes de mobilidade permite identificar nós estratégicos para a disseminação de doenças infecciosas e revelar caminhos preferenciais de propagação, contribuindo para a compreensão da dinâmica epidêmica e no desenvolvimento de estratégias aplicadas à vigilância sentinela.

A justificativa para esta hipótese é a premissa de que métodos isolados tradicionalmente utilizados na vigilância epidemiológica, como os modelos compartimentais clássicos baseados em dados clínicos convencionais de casos diagnosticados, apresentam limitações importantes para a detecção precoce de surtos, comprometendo a capacidade de resposta oportuna [14–18]. Em contraste, abordagens de vigilância integrada que combinam modelos epidemiológicos com outras abordagens, como técnicas de inteligência artificial e métodos da ciência de redes, especialmente quando aplicadas a dados não convencionais, como informações sindrômicas, ambientais e sociais, têm demonstrado maior sensibilidade e proatividade na identificação antecipada de surtos [15–17]. Essas estratégias integradas permitem a construção de sistemas de vigilância mais robustos, com maior sensibilidade de detectar padrões emergentes em séries temporais.

Ao longo da história, as epidemias e pandemias tiveram um profundo impacto nas sociedades humanas, desde a Peste de Atenas em 430 a.C., passando pela Peste Negra em 1346, a Gripe Espanhola em 1918, e a epidemia da Síndrome da Imunodeficiência Adquirida (AIDS) nos anos 1980 [19, 20]. Nas últimas décadas, emergências de vírus respiratórios, como o SARS-CoV em 2003, a nova cepa do vírus da gripe H1N1 em 2009, o MERS-CoV em 2012 e o SARS-CoV-2, causador da pandemia de COVID-19 em 2020 [21], destacaram-se como exemplos de questões concretas de saúde. Cada uma dessas crises sanitárias trouxe não apenas uma significativa perda de vidas, mas também profundas transformações sociais, econômicas e políticas [22]. Diante desses cenários, torna-se imperativo o desenvolvimento e a aplicação de ferramentas mais precisas para a detecção de doenças emergentes e reemergentes, permitindo decisões baseadas em evidências e avaliações de risco cientificamente fundamentadas [23, 24].

Com o rápido avanço da ciência de dados, incluindo big data e inteligência artificial, e o aumento exponencial de dados relacionados à saúde, os sistemas de vigilância em saúde estão evoluindo rapidamente [25]. O campo abrange uma gama crescente de aplicações, nas quais novos métodos são aplicados tanto a dados de diagnósticos clínicos quanto a conjuntos de informações menos específicas, como dados de atendimento de emergência e publicações em mídias sociais [26, 25]. Isso torna possível explorar novas abordagens para identificar padrões epidemiológicos em estágios iniciais.

A vigilância sindrômica [27] surge como um instrumento promissor nesse contexto. Embora existam diversas definições, a maioria destaca o uso de dados não diagnosticados, informações sobre possíveis eventos de saúde antes ou sem identificação laboratorial definitiva do patógeno, podendo permitir a identificação de padrões emergentes antes mesmo da confirmação clínica dos casos [23, 24, 27, 28]. No entanto, o grande volume e a complexidade desses dados exigem abordagens computacionais desenvolvidas mais recentemente, tais como algoritmos de aprendizado de máquina, aprendizado profundo e processamento de linguagem natural para extrair informações relevantes de maneira eficiente [29, 30].

Neste contexto, o presente estudo tem como foco no uso de um conjunto de dados sindrômicos referentes a atendimentos da Atenção Primária à Saúde (APS) relacionados a infecções do trato respiratório. Esses dados são provenientes do Sistema de Informação em Saúde para a Atenção Básica (SISAB) e foram disponibilizados pelo projeto Sistema de Alerta Precoce para Surtos com Potencial Epi-Pandêmico (ÆSOP, do inglês *Alert-Early System of Outbreaks with Pandemic Potential*), que visa monitorar surtos com potencial

epi-pandêmico [31].

O objetivo geral desta tese é desenvolver e avaliar metodologias integrativas oriundas da inteligência artificial, modelagem epidemiológica e ciência de redes, aplicadas a dados, com o intuito de aprimorar a vigilância epidemiológica, especialmente no que se refere à detecção precoce de surtos. Os resultados obtidos foram discutidos em dois trabalho.

No primeiro trabalho, propomos o desenvolvimento e aplicação do Modelo Misto de Inteligência Artificial e Próxima Geração (MMAING, do inglês *Mixed Model Artificial Inteligence and Next Generation*) [32], que visa identificar surtos a partir de séries temporais provenientes dos dados sindrômicos da APS, de modo a estabelecer uma alerta precoce em caso de sinais que indiquem processos epi-pandêmicos. Esse modelo integra quatro técnicas de inteligência artificial a um modelo baseado no tempo de infecção, permitindo capturar padrões dinâmicos de propagação de doenças respiratórias.

No segundo trabalho, aplicamos análises de redes para compreender os padrões de mobilidade urbana e sua influência na disseminação de surtos entre um dado conjunto de aglomerações urbanas, juntamente com a análise de um modelo metapopulacional SIR (Suscetível-Infectado-Recuperado) numa sub-rede sentinela de síndrome gripal no estado da Bahia, otimizando e validando estruturas de redes de vigilância. Dessa forma, desenvolvemos um modelo integrado para redes de vigilância sentinela e propusemos um índice sentinela, visando estruturar redes mais eficientes para a detecção precoce de surtos epidemiológicos. Por fim, avaliamos a eficácia dessas abordagens no contexto do projeto ÆSOP, buscando validar sua aplicação em cenários reais de vigilância em saúde.

O trabalho está estruturado em dez capítulos. Neste Capítulo 1, apresenta-se a introdução do estudo. Os Capítulos 2 a 6 revisam os principais conceitos teóricos relacionados à modelagem epidemiológica, aprendizado de máquina, ciência de redes e vigilância em saúde. Nos Capítulos 7 e 8, são detalhados o desenvolvimento e a avaliação do Modelo Misto de Inteligência Artificial e Próxima Geração (MMAING), voltado para a detecção precoce de surtos por meio de sinais de alerta precoce. O Capítulo 9 trata do desenvolvimento de um modelo integrado para redes de vigilância sentinela. Por fim, o Capítulo 10 apresenta as conclusões gerais e as perspectivas para a continuidade e ampliação desta linha de pesquisa.

Dessa forma, este estudo busca contribuir para o aprimoramento da vigilância epidemiológica, promovendo a tomada de decisões baseada em evidências para a prevenção e mitigação de surtos epidemiológicos.

2
Dinâmica de Sistemas Epidemiológicos

A partir da aplicação de conceitos da dinâmica populacional, é comum introduzir modelos simplificadas para abordar problemas epidemiológicos sem comprometer os principais elementos que regem a dinâmica de uma doença. Desse modo, é possível desenvolver um conjunto de equações diferenciais ordinárias que capturem essa dinâmica. As abordagens tradicionais nesse contexto incluem a definição e caracterização de compartimentos formados por subconjuntos de indivíduos da população, em função dos quais são estabelecidos modelos para descrever a interação causal da doença e do tempo de infecção. Nesse capítulo, vamos apresentar conceitos importantes no âmbito da epidemiologia matemática.

2.1
Modelos compartimentais clássicos

Na história da modelagem epidemiológica, um marco significativo é o trabalho de Kermack e McKendrick [33]. Nesse estudo os autores objetivaram criar uma representação matemática para a propagação de doenças transmissíveis em populações. Eles adotaram a divisão da população em compartimentos: a) os indivíduos susceptíveis, aqueles que não entraram em contato com o patógeno e podem se tornar infectados; b) os infectados, aqueles que carregam o agente infeccioso e podem vir a transmiti-lo para os indivíduos suscetíveis; c) os removidos, que englobam os indivíduos que não estão mais suscetíveis à infecção pelo patógeno, seja devido a imunidade ou morte. Assim, nesse sistema, temos um compartimento Suscetíveis (S), Infectados (I) e Removidos (R) formando o acrônimo SIR.

Ao considerar a dinâmica e as transições entre esses compartimentos, Kermack e McKendrick formularam um sistema de equações diferenciais ordinárias que descreviam como a doença se espalhava ao longo do tempo. Essa abordagem permitiu a quantificação dos padrões de propagação, a análise da influência de fatores como a taxa de transmissão e a duração da infecção. O modelo compartimental SIR pode ser descrito pelo conjunto de equações diferenciais ordinárias descritas a seguir:

$$\frac{dS}{dt} = -\beta\frac{SI}{N}, \tag{2-1}$$
$$\frac{dI}{dt} = \beta\frac{SI}{N} - \gamma I, \tag{2-2}$$
$$\frac{dR}{dt} = \gamma I, \tag{2-3}$$

sendo $N = S + I + R$ a população total, assumida constante, β a taxa de transmissão da doença, e γ a sua taxa de recuperação.

O sistema das equações 2-1, 2-2 e 2-3 é amplamente reconhecido na literatura como modelo SIR (Suscetível-Infectado-Removido). Devido à sua capacidade de abarcar os aspectos fundamentais da dinâmica de transmissão e devido à sua simplicidade, o modelo SIR é frequentemente considerado como o ponto de partida para as investigações de modelagem epidemiológica. No entanto, em muitas situações, torna-se necessário o desenvolvimento de modelos mais refinados, frequentemente construídos a partir da dinâmica estabelecida pelo modelo SIR.

Um outro modelo compartimental amplamente utilizado é o modelo SEIR (Suscetíveis-Expostos-Infetados-Removidos), o qual representa uma simples evolução do modelo SIR. Ambos modelos desempenham um papel crucial na compreensão e na modelagem da propagação de doenças infecciosas em populações. No entanto, o modelo SEIR introduz um compartimento adicional importante: o compartimento "Expostos"(E), que captura o período de incubação do patógeno. Essa inclusão torna-se fundamental para melhor descrever a dinâmica de diversas doenças, nas quais existe um intervalo de tempo em que os indivíduos carregam o patógeno, mas não estão contribuindo ativamente para a disseminação da doença. Essa distinção entre os estágios de incubação e de transmissividade é o que estabelece a distinção fundamental entre esses dois modelos.

O modelo SEIR pode ser representado pelo sistema de equações diferenciais ordinárias:

$$\frac{dS}{dt} = -\frac{\beta S}{N}\left[I + \varepsilon E\right], \tag{2-4}$$
$$\frac{dE}{dt} = \frac{\beta S}{N}\left[I + \varepsilon E\right] - \kappa E, \tag{2-5}$$
$$\frac{dI}{dt} = \kappa E - \gamma I, \tag{2-6}$$
$$\frac{dR}{dt} = \gamma I, \tag{2-7}$$

sendo o número de indivíduos expostos no tempo t denotado por E e $N =$

$S + E + I + R$ a população total, de modo que, quando $\varepsilon = 0$ os expostos não transmitem. Os parâmetros do modelo SEIR são definidos como no modelo SIR, com a adição do parâmetro κ, que representa a frequência com a na qual os indivíduos saem do compartimento dos expostos (E) e entram no compartimento dos infectados (I). Além disso, introduz-se o parâmetro ε, que representa a capacidade do indivíduo exposto em transmitir o patógeno para um indivíduo suscetível. Ele corresponde à probabilidade de que um individuo exposto possa transmitir o patógeno a uma taxa mais baixa do que os infectados.

Surtos epidêmicos podem apresentar características não previstas, como observado nos modelos mencionados anteriormente. Por exemplo, é amplamente reconhecido que muitos surtos epidêmicos incluem indivíduos assintomáticos, os quais também são contagiosos, ou seja, capazes de transmitir a doença para indivíduos suscetíveis. Também, é prudente considerar a inclusão de compartimentos no modelo que representem a evolução de casos hospitalizados e permitam a diferenciação entre óbitos e casos recuperados entre os removidos.

Além do mais, para fazer uma abordagem mais próxima à realidade, é importante incorporar elementos adicionais no modelo para representar a existência de diversos tipos de heterogeneidades, como abordados em trabalhos recentes [34–37].

2.2 Número de reprodução dependente do tempo

Apesar de receberem menos atenção em comparação com os modelos compartimentais simples, como SIR e SEIR, os modelos que levam em conta o tempo de infecção, apresentam uma abordagem mais intuitiva para modelar a dinâmica de agentes infecciosos. Adicionalmente, esses modelos oferecem uma vantagem significativa para essa aplicação pois não demandam a inclusão de muitos parâmetros. Entretanto, uma desvantagem é a potencial complexidade em incorporar fatores de heterogeneidade.

Apesar do fato de que ambos os modelos se originam do trabalho de [33], o modelo baseado no tempo de infecção não foi inicialmente descrito em termos do número de reprodução, conforme entendemos atualmente [38]. Na formulação original, Kermack e McKendrick [33] introduziram o conceito de um limiar crítico, indicando que uma epidemia ocorrerá apenas se a proporção inicial de indivíduos suscetíveis for suficientemente grande para permitir o aumento do número de infectados. Esse conceito é matematicamente equivalente ao que hoje chamamos de número básico de reprodução, R_0,

que posteriormente foi formalizado como um parâmetro limiar essencial para caracterizar a capacidade de invasão e disseminação de uma doença em uma população suscetível. A formulação atual também permite a definição de uma versão temporal, o número de reprodução dependente do tempo $R(t)$, que representa como essa capacidade de disseminação varia ao longo do tempo.

Dentro desse formalismo, o modelo prevê como a taxa de incidência, ou seja, novos casos, representada por $B(t)$, evolui ao longo do tempo calendário t em relação à função infectividade, denotada como $A(t, \tau)$. A função infectividade corresponde à taxa de novas infecções no tempo de calendário t causados por indivíduos infectados no tempo decorrido desde a infecção, τ , que corresponde à fase-infecciosa [39]. A função infectividade $A(t, \tau)$ normalmente expressa a carga do patógeno ou sua intensidade, frequentemente sendo uma função unimodal que descreve o crescimento do patógeno seguido pela supressão imunológica ou morte do hospedeiro. No entanto, essa função pode ser mais intrincada, como discutido em [40, 41]. Ademais, $A(t, \tau)$ também reflete a taxa efetiva de contato entre indivíduos suscetíveis e infecciosos.

Do ponto de vista matemático, a transmissão é descrita por um processo de infecção de Poisson, onde a probabilidade de que, entre o tempo t e $t + \delta$, um indivíduo infectado no tempo t possa eficazmente infectar outra pessoa é dada por $A(t, \tau)\,\delta$, sendo δ um intervalo de tempo muito pequeno.

A partir dessa premissa, é possível prever que a taxa média de incidência, $B(t)$, em função do do tempo t, segue o que é conhecido como a *equação de renovação*, definida como:

$$B(t) = \int_0^\infty A(t, \tau) B(t - \tau)\, d\tau. \tag{2-8}$$

Essa equação estabelece que o número de novos indivíduos infectados é proporcional ao número de casos prevalentes multiplicados por sua capacidade de infecção. Dessa forma, surge naturalmente o conceito do número de reprodução dependente do tempo, que representa aproximadamente a média de pessoas que um indivíduo infectado no momento t pode infectar durante todo o seu período infeccioso τ. Isso é expresso por:

$$R(t) = \int_0^\infty A(t, \tau)\, d\tau. \tag{2-9}$$

Assim, essa métrica está relacionada à habilidade de uma pessoa infectada infectar outras, caso as condições permaneçam inalteradas [42, 43]. Esse conceito geral pode ser incorporado a qualquer modelo matemático de compartimentos que descreva a propagação de doenças, permitindo a

interpretação de seus dados.

Tradicionalmente as propriedades derivadas desse conceito são aplicadas em dados de casos confirmados. Explora-se a capacidade do $R(t)$ em refletir as mudanças na infecciosidade de um patógeno, conferindo a ele um potencial significativo como índice em vigilância epidemiológica, aumentando à medida que a capacidade de transmissão se intensifica e declinando quando essa capacidade se reduz. Ademais, $R(t)$, também estabelece pontos de transição críticos: se o valor é maior que 1, indica um processo epidêmico cujo crescimento pode ser exponencial ou não; quando é menor que 1, sinaliza o declínio do processo epidêmico [42]; vale registrar que, no início de um surto, quando toda a população é suscetível, o crescimento é tipicamente exponencial, e o número de reprodução no tempo $t = 0$ corresponde ao número básico de reprodução, R_0 [39].

2.3
Método baseado em intervalos de geração

O conceito do $R(t)$ é importante no estudo de propagação de doenças infecciosas, e por isso, é fundamental saber como estimá-lo. Uma maneira de fazer isso é recorrendo à distribuição do intervalo de geração [44], também chamado de distribuição do tempo de geração, denotada por $g(\tau)$. Esta distribuição caracteriza o intervalo de tempo no qual um indivíduo infectado pode causar novas infecções. De acordo com Nishiura e Chowell [45], é possível definir essa distribuição fundamentada na função infectividade, como:

$$g(\tau) = \frac{A(t,\tau)}{\int_0^\infty A(t,\tau)d\tau}, \tag{2-10}$$

de tal forma que

$$\int_0^\infty g(\tau)d\tau = 1. \tag{2-11}$$

Com essa definição, a determinação de $g(\tau)$ também pode ser realizada pela análise de dados. Vários estudos têm extraído esta distribuição a partir de dados observacionais [46, 47] ou a partir de modelos e distribuições estabelecidas, conforme demonstrado por [45, 48–50].

Para estimar a expressão de $R(t)$, iremos analisar a situação em que $A(t,\tau)$ seja separável. Isso implica que a progressão relativa da infecciosidade, em relação ao tempo desde a infecção inicial, é constante ao longo do tempo. Neste contexto, $A(t,\tau)$ pode ser escrita como o produto de duas funções

distintas, $w_1(t)$ e $w_2(\tau)$:

$$A(t,\tau) = w_1(t)w_2(\tau). \tag{2-12}$$

Dado que $A(t,\tau)$ é produto dessas funções, podemos normalizá-las de forma arbitrária. Portanto, escolhendo $w_2(\tau)$ para ter uma integral total igual a 1, e substituindo na equação 2-9, temos:

$$R(t) = \int_0^\infty w_1(t)w_2(\tau)d\tau = w_1(t). \tag{2-13}$$

Assim, $w_1(t)$ é a representação de $R(t)$, enquanto $w_2(\tau)$ descreve como os eventos infecciosos são distribuídos ao longo do tempo τ desde a primeira infecção. Portanto, podemos concluir que a função infectividade é representada pelo produto entre o número de reprodução e a distribuição do tempo de geração, dado por:

$$A(t,\tau) = R(t)\ g(\tau). \tag{2-14}$$

É importante notar que, apesar de $g(\tau)$ traduzir uma ideia clara de medida da infecciosidade, a sua relação com os tempos de geração observados pode ser complexa devido a diversos fatores como mostrado em [42]. Inserindo 2-14 em 2-8, obtemos o número de reprodução dependente do tempo:

$$R(t) = \frac{B(t)}{\int_0^\infty g(\tau)\ \ B(t-\tau)\,d\tau}. \tag{2-15}$$

Geralmente a incidência $B(t)$ é disponibilizada como uma série temporal discreta na forma B_i, que contabiliza os casos incidentes relatados entre o tempo t_i e o tempo t_{i+1}. Nesse caso, a distribuição do tempo de geração deve ser adequadamente discretizada em uma forma g_i de modo que $\sum_{i=1}^n g_i = 1$. Logo, a partir de 2-15 obtemos a equação para $R(t)$ na forma discretizada, dada por:

$$R_i = \frac{B_i}{\sum_{j=1}^n g_j B_{i-j}}. \tag{2-16}$$

Esta expressão mostra que podemos realizar estimativas do R_i utilizando os dados reportados durante uma epidemia, desde que sejam conhecidos os g_j.

2.4 Método baseado na próxima geração

O método de Próxima Geração (NGM, do inglês Next-Generation Method) é uma abordagem predominantemente utilizada para estimar o número de reprodução e frequentemente aplicada a uma variedade de modelos dinâmicos, incluindo, entre outros, os modelos compartimentais [44].

Este método foi introduzido pela primeira vez por Diekmann et al., [51], em 1990, mas foi posteriormente detalhado por Van den Driessche e Watmough [52], em 2002, na qual se estimou o número básico de reprodução, R(0), usando a matriz de próxima geração dentro de um modelo dinâmico, incluindo alguns exemplos de modelos compartimentais.

Em 2022, Jorge et al. [50] propuseram uma extensão do NGM, estabelecendo um paralelo com a formulação original de van den Driessche e Watmough [52]. A ideia central dessa abordagem é escrever as equações de n compartimentos utilizando dois vetores principais. Assim, define-se $\mathbf{F}(t) = (F_1(t), \ldots, F_n(t))$ como a taxa de surgimento de novas infecções em cada instante de tempo t, e $\mathbf{V}(t,\tau) = (V_1(t,\tau), \ldots, V_n(t,\tau))$ como a taxa de transferência entre compartimentos infectados, para uma dada idade da infecção τ.

Para acessar informações sobre os n compartimentos, define-se o vetor $\mathbf{x}(t,\tau) = (x_1(t,\tau), \ldots, x_n(t,\tau))$, no qual t e τ são variáveis independentes que indicam, respectivamente, o tempo calendário e o intervalo de tempo da fase-infecciosa.

Neste contexto, o vetor $\mathbf{x}(t,\tau)$ representa a densidade de indivíduos infectados em cada compartimento no instante t, com uma fase-infeciosa τ. Os elementos de $\mathbf{x}$ podem ser interpretados como distribuições por idade da infecção, de modo que, para cada tempo t, $x_i(t,\tau)\,d\tau$ expressa o número de indivíduos infectados no compartimento i, com idade de infecção entre τ e $\tau + d\tau$. Deve-se garantir a condição $\mathbf{x}(t,\tau) = 0$ para $\tau < 0$.

O número total de indivíduos infectados em cada compartimento no tempo t, denotado por $\mathbf{X}(t)$, pode ser obtido por:

$$\mathbf{X}(t) = \int_0^\infty \mathbf{x}(t,\tau)\,d\tau, \tag{2-17}$$

onde $\mathbf{X}(t) = (X_1(t), \ldots, X_n(t))$. Dessa forma, um modelo típico baseado em idade da infecção pode ser descrito pelo seguinte sistema de equações diferenciais parciais (EDPs):

$$\left(\frac{\partial}{\partial t} + \frac{\partial}{\partial \tau}\right)\mathbf{x}(t,\tau) = -\mathbf{V}(t,\tau), \tag{2-18}$$

com condição inicial

$$\mathbf{x}(t, \tau = 0) = \mathbf{F}(t), \tag{2-19}$$

em que $\mathbf{F}(t)$ descreve o fluxo de indivíduos que entram nos compartimentos infectados a partir dos não infectados e depende de $\mathbf{X}(t)$. Por outro lado, $\mathbf{V}(t, \tau)$ está relacionada ao fluxo de indivíduos dentro dos compartimentos infectados, seja pela progressão da doença entre estágios, seja pela recuperação ou morte. Naturalmente, $\mathbf{V}(t, \tau)$ depende de $\mathbf{x}(t, \tau)$.

Assim, ao integrar a equação 2-18 de zero até o infinito em relação a τ, o sistema de EDPs em $\mathbf{x}(t, \tau)$ é convertido em um sistema de equações diferenciais ordinárias (EDOs) para $\mathbf{X}(t)$:

$$\frac{d}{dt}\mathbf{X}(t) = \mathbf{F}(t) - \int_0^\infty \mathbf{V}(t, \tau)\, d\tau. \tag{2-20}$$

A abordagem proposta é suficientemente geral para ser aplicada a qualquer modelo compartimental, bastando identificar os termos correspondentes a $\mathbf{F}$ e $\mathbf{V}$ nas equações do modelo. Com isso, torna-se possível derivar expressões tanto para $R(t)$ quanto para $g(\tau)$ a partir de variáveis e parâmetros do próprio modelo.

Considerando o modelo compartimental SEIR, composto pelas EDOs descritas pelas expressões 2-4 a 2-7, as transições entre compartimentos infectados, $\mathbf{V}(t, \tau)$, e as densidades dos compartimentos infectados $\mathbf{x}(t, \tau)$ apresentam dependência linear. Além disso, todos os parâmetros ($\beta, \varepsilon, \kappa$ e γ) do modelo são constantes.

Pode-se então ordenar os compartimentos infectados como $\mathbf{X}(t) = [E(t), I(t)]$. Desta forma as distribuições de indivíduos das fases infecciosas são definidas como $\mathbf{x}(t, \tau) = [i_e(t, \tau), i_i(t, \tau)]$. Sendo assim, define-se

$$\mathbf{F}(t) = \begin{pmatrix} \frac{\beta S}{N}[I(t) + \varepsilon E(t)] \\ 0 \end{pmatrix}, \tag{2-21}$$

e

$$\mathbf{V}(t, \tau) = \begin{pmatrix} \kappa i_e(t, \tau) \\ \gamma i_i(t, \tau) - \kappa i_e(t, \tau) \end{pmatrix}. \tag{2-22}$$

Procedendo com alguns cálculos matriciais, detalhados em [50], obtém-se

o número total de reprodução do sistema, como

$$R(t) = \frac{\beta S(t)}{N}\left(\frac{\varepsilon}{\kappa} + \frac{1}{\gamma}\right), \tag{2-23}$$

o que nos leva a distribuição do tempo de geração

$$g(\tau) = \frac{\varepsilon\ e^{-\kappa\ \tau}\ +\ \frac{\kappa}{\gamma-\kappa}[\ e^{-\kappa\ \tau} -\ e^{-\gamma\ \tau}]}{\frac{\varepsilon}{\kappa} + \frac{1}{\gamma}}. \tag{2-24}$$

Os cálculos detalhados podem ser obtidos em [50]. Para demonstrar a robustez da abordagem proposta, a partir da equação 2-23 pode-se recuperar o resultado do $R(t)$, para o modelo SIR (que não considera o estágio de infecção pré-sintomática). Ao fazer com que ε se aproxime de zero, obtém-se $R(t) = \frac{S(t)}{N}\frac{\beta}{\gamma}$, o qual está bem estabelecido na literatura [53]. De maneira análoga, a distribuição do intervalo de geração para o modelo SEIR, conforme descrito pela equação 2-24, pode ser simplificada para a forma que é reconhecida na literatura associada ao modelo SIR [45]; tomando-se o limite quando κ tende ao infinito e ε se aproxima de zero. Dessa forma, obtém-se a distribuição do intervalo de geração, $g(\tau) = \gamma e^{-\gamma\tau}$.

2.5
Modelo metapopulacional

O estudo da dinâmica de doenças infecciosas considerando cenários onde populações não estão isoladas, mas sim interligadas por sistemas de transporte e movimentação, demanda modelos que vão além dos tradicionais enfoques epidemiológicos compartimentais. Os modelos metapopulacionais representam um avanço importante nesse sentido, incorporando a estrutura espacial e o fluxo de indivíduos entre diferentes localidades, permitindo uma análise mais detalhada da dispersão espacial e temporal de patógenos [50, 54, 55]. Essa abordagem reflete de maneira mais realista as interações sociais e geográficas que influenciam a propagação de doenças transmissíveis em populações com distribuições não homogêneas [55].

No caso de populações espacialmente distribuídas, a abordagem metapopulacional combina duas técnicas de modelagem: i) o uso de EDOs para representar a dinâmica dos compartimentos dentro de cada população, ii) um modelo de difusão que descreve o fluxo populacional através de uma rede que pode ser regular ou não regular [50, 55]. A seguir, vamos empregar essa abordagem para elaborar um modelo meta-populacional adaptado do modelo epidemiológico SIR clássico.

2.5.1
Taxas de transmissão do modelo metapopulacional SIR

No cerne do modelo metapopulacional SIR, encontram-se os conceitos dos estados suscetíveis, (S), infecciosos (I) e recuperados (R) aplicados a múltiplas populações interconectadas. Assim, consideramos que qualquer situação em que a população esteja distribuída em um espaço geográfico pode ser interpretado como um sistema de metapopulação, com compartimentos e parâmetros próprios, representados por um vetor $\mathbf{y}_i(t)$, onde o índice i indica a metapopulação específica. No caso do modelo SIR, este vetor $\mathbf{y}_i(t)$ inclui os compartimentos $S_i(t)$, $I_i(t)$, e $R_i(t)$ para a metapopulação i, com a soma desses elementos refletindo o tamanho total da população $N_i(t)$.

A dinâmica de movimentação entre metapopulações é capturada pelo fluxo de indivíduos, onde ϕ_{ij} representa o fluxo de i para j. Como cada metapopulação é descrita a partir de seus compartimentos, podemos descrever o fluxo usando $\mathbf{y}_i(t)$. Assim, o fluxo de i para j é descrito por $\phi_{ij}(t)\frac{\mathbf{y}_i}{N_i}$.

Portanto, podemos definir $\Phi_{ij}(t) = \frac{\phi_{ij}(t)}{N_i}$ como a densidade de fluxo. Desta forma, $\Phi_{ij}\mathbf{y}_i$ é o número de indivíduos de cada classe de compartimento que estão se movendo de i para j. É natural ver que a soma de $\Phi_{ij}\mathbf{y}_i$ para cada classe de compartimento é igual a ϕ_{ij}, ou seja, $\Phi_{ij}S_i + \Phi_{ij}I_i + \Phi_{ij}R_i = \phi_{ij}$.

Dado que as metapopulações estão conectadas através do fluxo, é necessário identificar como elas interagem. Então, a dinâmica de interação entre as metapopulações é essencialmente modelada por meio de redes ponderadas, onde a matriz de adjacência representa o fluxo de indivíduos entre elas. Cada elemento da matriz, denotado por $\Phi_{ij}(t)$, quantifica a densidade de fluxo de uma metapopulação para outra, refletindo o movimento de pessoas no espaço geográfico. Importante notar que o modelo exclui o auto-fluxo ($\phi_{ii} = 0$), indicando que a mobilidade é sempre entre metapopulações distintas. Esta abordagem permite uma representação da interconectividade entre diferentes espaços geográficos e a consequente propagação de doenças transmissíveis.

Prosseguimos agora para descrever como a transmissão de uma doença ocorre em uma metapopulação, levando em conta o fluxo de indivíduos na rede ponderada. Devemos então descrever a ocorrência de novas infecções na metapopulação $\mathbf{y}_i(t)$, que podem ocorrer dentro da própria metapopulação i usando

$$\frac{\beta_i}{N_i} \times \overbrace{S_i\left(1 - \sum_j \Phi_{ij}\right)}^{\text{Indivíduos Susceptíveis de } i \text{ em } i} \times \overbrace{I_{e,i}}^{\text{Indivíduos Infectados em } i}, \tag{2-25}$$

ou seja, podemos definir a função $F_i^{(i)}(t)$ como a quantidade de novos indivíduos suscetíveis infectados na metapopulação, devido à interação com os indivíduos infectados da própria metapopulação

$$F_i^{(i)}(t) = \frac{\beta_i}{N_i} S_i(t) \left(1 - \sum_j \Phi_{ij}\right) I_{e,i}(t), \tag{2-26}$$

onde, $S_i(t)\left(1 - \sum_j \Phi_{ij}\right)$ indica indivíduos suscetíveis de i em i, e $I_{e,i}(t)$ indivíduos Infectados em i. Similarmente, a ocorrência de novas infecções que podem ocorrer em outra metapopulação j, que são vizinhas da metapopulação i

$$\frac{\beta_j}{N_j} \times \overbrace{S_i \Phi_{ij}}^{\text{Indivíduos Susceptíveis de } i \text{ em } j} \times \overbrace{I_{e,j}}^{\text{Indivíduos Infectados em } j}, \tag{2-27}$$

ou seja,

$$F_i^{(j)}(t) = \frac{\beta_j}{N_j} S_i(t) \Phi_{ij} I_{e,j}(t), \tag{2-28}$$

onde, $S_i(t)\Phi_{ij}$ indica indivíduos suscetíveis de i que vieram de j, e $I_{e,j}(t)$ indivíduos infectados em j. Os parâmetros $\beta_i(t)$ e $\beta_j(t)$ são as taxas de transmissão dentro das metapopulações i e j, respectivamente. Como $\beta(t)$ e $\Phi(t)$ são dependentes do tempo, então podemos incorporar as mudanças no comportamento das populações nessas variáveis.

Sendo $y_{e,i} = (S_{e,i}(t), I_{e,i}(t), R_{e,i}(t))$ a população efetiva de indivíduos infectados, é possível expressá-los de acordo com a população infectada residente em cada metapopulação. As componentes da população de infectados efetivos das metapopulações i e j são descritas pelas expressões:

$$I_{e,i}(t) = I_i(t) \left(1 - \sum_j \Phi_{ij}\right) + \sum_j \Phi_{ji} I_j(t), \tag{2-29}$$

e

$$I_{e,j}(t) = I_j(t) \left(1 - \sum_k \Phi_{jk}\right) + \sum_k \Phi_{kj} I_k(t), \tag{2-30}$$

onde cada expressão reflete as dinâmicas de perda e ganho de infectados nas metapopulações. Substituindo a componente de infectados $I_{e,i}(t)$ na expressão

2-26 obtêm-se

$$F_i^{(i)}(t) = \frac{\beta_i}{N_i} S_i(t) \left(1 - \sum_j \Phi_{ij}\right) \left[I_i(t) \left(1 - \sum_j \Phi_{ij}\right) + \sum_j \Phi_{ji} I_j(t) \right]. \quad (2\text{-}31)$$

Agora expandimos os termos, aplicando a *distributividade* da multiplicação sobre a soma, e obtemos

$$F_i^{(i)}(t) = \frac{\beta_i}{N_i} S_i(t) \left[I_i(t) \left(1 - \sum_j \Phi_{ij}\right)^2 + \left(1 - \sum_j \Phi_{ij}\right) \sum_j \Phi_{ji} I_j(t) \right]. \quad (2\text{-}32)$$

O mesmo processo é realizado para a dinâmica de novas infecções em j, logo a expressão com termos expandidos é

$$F_i^{(j)}(t) = \frac{\beta_j}{N_j} \Phi_{ij} S_i(t) \left[I_j(t) \left(1 - \sum_k \Phi_{jk}\right) + \sum_k \Phi_{kj} I_k(t) \right]. \quad (2\text{-}33)$$

A taxa total de novas infecções na metapopulação i, $F_i(t)$, é então expressar por:

$$F_i(t) = \sum_j \lambda_{ij}(t) S_i(t) I_j(t), \quad (2\text{-}34)$$

onde a taxa de novas infecções de indivíduos que pertencem a i, é igual à soma entre $F_i^{(i)}(t)$ e todos os vizinhos j de i, $F_i^{(j)}(t)$. Logo,

$$F_i(t) = F_i^{(i)}(t) + \sum_j F_i^{(j)}(t), \quad (2\text{-}35)$$

e substituindo as expressões, obtêm-se

$$\begin{aligned} \sum_j \lambda_{ij}(t) S_i(t) I_j(t) = {} & \frac{\beta_i}{N_i} S_i(t) \left[I_i(t) \left(1 - \sum_j \Phi_{ij}\right)^2 + \left(1 - \sum_j \Phi_{ij}\right) \sum_j \Phi_{ji} I_j(t) \right] \\ & + \sum_j \frac{\beta_j}{N_j} \Phi_{ij} S_i(t) \left[I_j(t) \left(1 - \sum_k \Phi_{jk}\right) + \sum_k \Phi_{kj} I_k(t) \right]. \end{aligned} \quad (2\text{-}36)$$

Agrupando os termos dependentes de $I_j(t)$ e $I_i(t)$ para derivar as expressões de $\lambda_{ij}(t)$, as expressões resultantes para os casos onde i é igual

a j e onde i é diferente de j, são respectivamente

$$\lambda_{ii} = \frac{\beta_i}{N_i}\left(1 - \sum_j \Phi_{ij}\right)^2 + \sum_j \frac{\beta_j}{N_j}\Phi_{ij}^2, \tag{2-37}$$

e

$$\lambda_{ij} = \frac{\beta_i}{N_i}\Phi_{ji}\left(1 - \sum_k \Phi_{ik}\right) + \frac{\beta_j}{N_j}\Phi_{ij}\left(1 - \sum_k \Phi_{jk}\right) + \sum_k \frac{\beta_k}{N_k}\Phi_{ik}\Phi_{jk}. \tag{2-38}$$

Os coeficientes $\lambda_{ii}(t)$ e $\lambda_{ij}(t)$ capturam as taxas de transmissão de um patógeno dentro de uma metapopulação e entre diferentes metapopulações. Eles são importantes porque refletem tanto o comportamento de transmissão local (dentro de uma metapopulação) quanto a influência das interações entre diferentes metapopulações.

2.5.2
Número de reprodução e distribuição de probabilidade do intervalo de geração

Consideramos a existência de n metapopulações com dinâmicas acopladas. Devido ao movimento de indivíduos entre meta-populações, um compartimento infectado em uma metapopulação pode influenciar o processo de transmissão de doenças de todos os outros. O acoplamento das equações ocorre pelas taxas de transmissão λ_{ij} relacionadas ao processo de contaminação que emerge do fluxo de indivíduos.

Dessa forma, o modelo meta-populacional do tipo SIR para n metapopulações, pode ser escrito incorporando as taxas de transmissão emergentes do fluxo, conforme o seguinte conjunto de EDOs:

$$\frac{dS_i}{dt} = -\sum_{j=1}^{n} \lambda_{ij}(t) I_j(t) S_i(t), \tag{2-39}$$

$$\frac{dI_i}{dt} = \sum_{j=1}^{n} \lambda_{ij}(t) I_j(t) S_i(t) - \gamma I_i(t), \tag{2-40}$$

$$\frac{dR_i}{dt} = \gamma I_i(t). \tag{2-41}$$

Em síntese, o modelo meta-populacional do tipo SIR apresentado descreve a evolução dos compartimentos $S_i(t)$, $I_i(t)$, e $R_i(t)$ em cada metapopulação i com características próprias (β_i e N_i), incorporando a dinâmica de transmissão cruzada entre todas as metapopulações, definida nas equações 2-37 e 2-38. Fundamentalmente, o modelo presume uniformidade na

taxa de recuperação γ através de todas as metapopulações.

O número de reprodução dependente do tempo, $R_{ij}(t)$, que representa o número médio de novas infecções causadas por um indivíduo infectado em j ao longo de seu período infeccioso na metapopulação i, é dado por:

$$R_{ij}(t) = \frac{S_i(t)\lambda_{ij}(t)}{\gamma}. \tag{2-42}$$

A distribuição de probabilidade do intervalo de geração, $g_{ij}(\tau)$, permite estimar o tempo médio entre a infecção de um indivíduo e a infecção secundária, assumida ser a mesma para todas as metapopulações e dada por:

$$g_{ij}(\tau) = g(\tau) = \gamma e^{-\gamma\tau}. \tag{2-43}$$

A relação $g_{ij}(t) = g(t)$ aparece naturalmente da suposição de que todas as metapopulações têm a mesma dinâmica de recuperação.

O modelo meta-populacional SIR descrito aqui incorpora as dinâmicas de transmissão de doenças dentro de cada metapopulação e entre elas, considerando os fluxos periódicos pendulares de indivíduos. As taxas de transmissão $\lambda_{ii}(t)$ e $\lambda_{ij}(t)$ capturam as interações internas e externas das metapopulações, e o sistema de equações diferenciais descreve a evolução temporal dos compartimentos de suscetíveis, infectados e recuperados em cada metapopulação. A inclusão do número de reprodução $R_{ij}(t)$ e do intervalo de geração $g_{ij}(\tau)$ fornece uma base sólida para a análise da propagação de doenças em uma rede de populações interconectadas.

Mais detalhes das derivações das equações 2-42 e 2-43 são apresentados em [50].

3
Anomalias em Séries Temporais

Séries temporais são sequências de observações registradas de forma ordenada ao longo do tempo, frequentemente com intervalos regulares, e aparecem em diversas áreas como finanças, meteorologia, indústria e saúde [56]. A análise de séries temporais visa entender e modelar padrões temporais (tendências, sazonalidades) para tarefas como previsão ou detecção de comportamentos anômalos. A detecção de anomalias (ou *outliers*) em séries temporais é particularmente importante, pois valores atípicos podem representar erros ou eventos de interesse [56–59]. Por exemplo, uma queda abrupta em sensores industriais pode sinalizar uma falha iminente no sistema, enquanto um aumento inesperado no número de casos por síndrome gripal pode indicar o início de um surto epidemiológico. Devido à relevância desses eventos, a identificação de anomalias em dados temporais tem sido objeto de pesquisas desde meados do século XX, com métodos evoluindo de técnicas estatísticas clássicas para abordagens modernas de *Machine Learning* e *deep learning* [58, 59]. A seguir, exploramos conceitos fundamentais de séries temporais, técnicas de detecção de anomalias (das regras estatísticas a técnicas de inteligência artificial), aplicações práticas, desafios comuns e direções futuras de pesquisa.

3.1
Conceitos fundamentais de séries temporais

Séries temporais referem-se a sequências de observações $x(t)$, registradas em momentos discretos t, geralmente igualmente espaçados. Diferentemente de conjuntos de dados independentes e identicamente distribuídos, observações em séries temporais costumam apresentar dependência temporal, ou seja, valores próximos no tempo tendem a ser correlacionados. Três componentes estruturais clássicos são comumente analisados: a tendência, que representa variações sistemáticas de longo prazo; a sazonalidade, que envolve padrões recorrentes em intervalos regulares (como ciclos diários, semanais ou anuais); e o componente aleatório, que inclui flutuações não explicadas (ruído branco).

Séries temporais podem ser univariadas, quando consistem em apenas uma variável por ponto no tempo, ou multivariadas, quando múltiplas variáveis são observadas simultaneamente. Cada variável é então modelada como uma série temporal própria, potencialmente correlacionada com as demais.

Outro conceito central é a estacionaridade. Diz-se que uma série temporal é estacionária quando suas propriedades estatísticas, como média, variância e autocorrelação, permanecem aproximadamente constantes ao longo do tempo. Muitos métodos tradicionais de análise, como os modelos auto-regressivos [60], pressupõem essa propriedade; por isso, transformações como a diferenciação ($\Delta x(t) = x(t) - x(t-k)$ com $k \geqslant 1$) são frequentemente aplicadas para remover tendências e componentes sazonais, aproximando a série de um comportamento estacionário.

No contexto da detecção de anomalias, em séries temporais, o objetivo é identificar observações que se desviam significativamente do comportamento esperado ao longo do tempo [61]. O primeiro estudo sistemático sobre essa temática foi conduzido por Fox [62], que propôs dois tipos de outliers para séries univariadas: tipo I, que afeta uma única observação pontual, e tipo II, que compromete não apenas a observação atual, mas também as subsequentes, por meio da propagação de seu efeito. Posteriormente, Tsay [63] expandiu essa tipologia para quatro categorias distintas de outliers, abrangendo diferentes padrões de perturbação. Mais adiante, esse arcabouço foi estendido ao contexto multivariado [64], considerando interações entre múltiplas variáveis ao longo do tempo.

Desde então, a literatura tem proposto diversas definições e métodos de detecção, embora ainda não exista consenso quanto à terminologia. Termos como anomalias, observações discordantes, exceções, aberrações, peculiaridades e contaminantes são frequentemente utilizados de forma intercambiável [56]. Do ponto de vista clássico, uma definição amplamente aceita foi proposta por Hawkins [65], segundo a qual um outlier é "*uma observação que se desvia tanto das demais que levanta suspeitas de ter sido gerada por um mecanismo diferente*".

Considerando a estrutura ordenada das séries temporais, os outliers podem ser classificados em diferentes categorias, conforme o contexto e a extensão da anomalia [57, 58]. As categorias mais comuns incluem:

- **Anomalias pontuais:** correspondem a observações que se desviam significativamente do comportamento típico em um instante específico do tempo. Tais desvios podem ser considerados *globais*, quando comparados à distribuição completa da série, ou *locais*, quando avaliados apenas em relação aos valores vizinhos imediatos.
- **Anomalias sequenciais:** consistem em subsequências de pontos consecutivos cujo padrão coletivo é atípico, mesmo que as observações individuais não sejam outliers evidentes. Essas anomalias podem ocorrer

em séries univariadas ou multivariadas, e também podem ser *globais* ou *locais*, dependendo do escopo da análise.

– **Anomalias contextuais:** dizem respeito a valores que são estatisticamente normais em termos absolutos, mas considerados anômalos quando analisados dentro de um determinado contexto temporal, como horário, estação do ano ou ciclo econômico.

3.2 Técnicas de detecção de anomalias em séries temporais

A detecção de anomalias em séries temporais pode ser realizada por diversas técnicas, desde regras estatísticas simples até algoritmos complexos de inteligência artificial. Em geral, o objetivo é modelar o comportamento normal da série e então identificar desvios desse comportamento. A seguir, são descritas as principais abordagens.

3.2.1 Regras estatísticas e limites fixos

As abordagens mais simples utilizam estatísticas descritivas para definir limites além dos quais um ponto é considerado anômalo. Um exemplo clássico é a regra dos 3 sigma: assumindo que os dados sigam aproximadamente uma distribuição normal, qualquer observação que ultrapasse a média em mais de três desvios-padrão (ou fique abaixo em menos de três desvios) é marcada como outliers [66]. De forma semelhante, pode-se utilizar intervalos interquartis, onde valores muito acima do terceiro quartil ou muito abaixo do primeiro quartil são considerados valores atípicos [67].

Testes estatísticos clássicos de detecção de outliers univariados, como o teste de Grubbs [68] ou o método ESD (*Extreme Studentized Deviate*) generalizado [69], detectam a presença de outliers individuais assumindo normalidade dos dados. Em séries temporais estacionárias, essas regras podem ser aplicadas nos resíduos ou na série diferenciada. Embora fáceis de implementar, métodos baseados em regras fixas podem falhar diante de distribuições não normais ou em séries com comportamento não estacionário, resultando em altos índices de falso alarme ou detecções perdidas [62]. Ainda assim, servem como primeiro filtro em muitos sistemas de monitoramento pela sua simplicidade.

3.2.2 Detecção baseada em métodos de previsão

Nessa abordagem, ajusta-se um modelo de previsão nos dados históricos e utiliza-se suas projeções como referência para identificar anomalias. Em essência, o modelo tenta prever o valor "esperado" em cada instante futuro, e então compara-se a diferença (erro de previsão) entre o valor observado e o previsto. Se o erro exceder um limite estabelecido, por exemplo, fora do intervalo de confiança de 95% da previsão, sinaliza-se uma anomalia [70].

Esse conceito pode ser implementado com modelos estatísticos ou de inteligência artificial. Por exemplo, o método de Brutlag [71] integrava a previsão Holt-Winters [72] com bandas de confiança adaptativas, gerando limites superiores e inferiores que se alargam conforme a incerteza da previsão aumenta em horizontes mais longos; pontos fora dessas bandas eram marcados como anômalos.

De maneira geral, qualquer modelo de *forecasting* (previsões) pode ser empregado para auxiliar na detecção de anomalias, uma vez que fornece estimativas pontuais e intervalos de confiança baseados no comportamento histórico da série. Modelos clássicos como o ARIMA (*Autoregressive Integrated Moving Average*) e suas extensões sazonais [73], assim como abordagens modernas baseadas em decomposição aditiva, como o FB-Prophet [74], são amplamente utilizados nesse contexto. A premissa subjacente a essa abordagem é intuitiva: uma observação é considerada anômala quando se desvia significativamente daquilo que seria previsto por um modelo ajustado aos padrões temporais anteriores. Dessa forma, a previsão atua como referência para a identificação de comportamentos atípicos, tornando essa técnica particularmente eficaz na detecção de outliers em uma ampla gama de aplicações práticas.

Por exemplo, no monitoramento de epidemias, modelos de previsão têm sido empregados com sucesso para identificar aumentos inesperados na incidência de doenças [75, 76]. O EARS (*Early Aberration Reporting System*), é um sistema baseado em estatísticas descritivas desenvolvido para detectar precocemente padrões anômalos em dados de saúde pública [77]. Os algoritmos do EARS (C1, C2 e C3) analisam séries temporais e identificam desvios em relação a valores esperados, utilizando médias e desvios padrão de períodos anteriores, sem exigir suposições complexas sobre a distribuição dos dados.

Mais recentemente, algoritmos como o ASMODEE (*Automatic Selection of Models and Outlier Detection for Epidemics*) têm sido utilizados para superar limitações práticas dos métodos tradicionais de vigilância. O ASMODEE realiza a seleção automática do modelo que melhor representa

a tendência recente da série histórica, considerando modelos como regressões lineares, modelos lineares generalizados e regressões bayesianas, com base em critérios como o AIC (*Akaike Information Criterion*) ou validação cruzada. Em seguida, compara os valores observados mais recentes com os intervalos de predição derivados do modelo selecionado, classificando como anomalias os pontos que fogem do padrão esperado. Aplicado a dados simulados da COVID-19, o ASMODEE demonstrou eficácia na detecção precoce de surtos, incorporando tendências e sazonalidades ao processo de modelagem [78].

Em resumo, a detecção baseada em previsão traduz a tarefa de detecção de outlier em um problema de previsão somado à detecção de erro, alavancando décadas de desenvolvimento de modelos preditivos.

3.2.3 Técnicas de aprendizado de máquina: Não supervisionadas e supervisionadas

Técnicas de detecção de outliers baseados em aprendizado de máquina demonstram capacidade para lidar com conjuntos de dados de alta dimensionalidade e correlacionados, identificando anomalias sem a necessidade de assumir distribuições paramétricas, como a distribuição normal [79]. Esses métodos abrangem uma variedade de abordagens que, de forma geral, podem ser agrupadas, em técnicas supervisionadas e não supervisionadas, conforme a presença ou ausência de rótulos nos dados utilizados para o treinamento dos modelos [79].

As abordagens não supervisionadas, como algoritmos de *clustering* (agrupamento), estimação de densidade, aprendizado de uma classe e algoritmos baseados em árvores de decisão são especialmente úteis quando não há rótulos disponíveis, ou seja, quando não se sabe antecipadamente quais observações são anômalas. A ideia central dessas técnicas é que anomalias são observações que não se encaixam bem nos padrões predominantes dos dados [57].

Algoritmos de agrupamento, como o K-means, buscam particionar os dados em grupos homogêneos com base em medidas de similaridade, geralmente utilizando a distância euclidiana [80]. Em séries temporais, isso é realizado convertendo janelas de observações em vetores no espaço vetorial $\mathbb{R}^{k+1}$, permitindo aplicar técnicas de *clustering* como se fossem dados tabulares. No entanto, é importante observar que o agrupamento de subsequências em séries temporais pode gerar agrupamentos espúrios sem significado real, conforme discutido por Keogh e Lin [81].

Já os métodos de estimação de densidade detectam anomalias com base

na concentração local dos dados. O algoritmo DBSCAN [82], por exemplo, agrupa dados em regiões densas e marca como ruído os pontos que não se ajustam a nenhuma dessas regiões densas [82]. Um algoritmo bastante popular é o Local Outlier Factor (LOF) [83], que calcula uma pontuação baseada na densidade de alcançabilidade local em torno de cada ponto, considerando a distância média até seus k vizinhos mais próximos. Pontos com densidade menor que a de seus vizinhos recebem pontuações superiores a 1, sendo interpretados como potenciais outliers.

Dentre as técnicas não supervisionadas, destacam-se também os métodos de aprendizado de uma classe e baseados em árvores de decisão [79], como o One-Class SVM [84] e a Isolation Forest [85, 86]. Esses algoritmos são treinados com dados representando o comportamento normal do sistema e aprendem uma estrutura que engloba esses padrões. O One-Class SVM aprende uma função de decisão que separa os dados normais da origem no espaço de características [87], enquanto o Isolation Forest isola observações em árvores binárias construídas aleatoriamente, pontos anômalos tendem a ser isolados em menos divisões, gerando caminhos mais curtos [85, 86].

Vale notar que, embora esses métodos operem sem rótulos, eles assumem implicitamente que os dados de treinamento representam comportamentos normais [57]. Outro desafio comum dessas abordagens é a escolha dos parâmetros [79].

Quando há dados rotulados, ou seja, quando se conhece previamente quais pontos são normais e quais são anômalos, é possível aplicar métodos supervisionados, tratando a detecção de outliers como um problema de classificação [79]. Técnicas como classificadores binários (por exemplo, Random Forest, SVM e redes neurais) podem ser treinadas para distinguir entre as duas classes. No entanto, por se tratarem de eventos raros e muitas vezes imprevisíveis, as anomalias tendem a estar sub-representadas nos conjuntos de dados, resultando em um forte desbalanceamento ou, em alguns casos, na ausência completa de amostras rotuladas [57, 88]. Essa limitação torna o uso de abordagens supervisionadas menos comum na prática.

3.2.4 Métodos de Deep Learning

Nos últimos anos, abordagens baseadas em redes neurais profundas têm alcançado resultados de ponta em detecção de anomalias em séries temporais [59]. Elas oferecem flexibilidade para modelar relações não lineares complexas e capturar padrões sutis em dados de alta dimensionalidade. Destacam-se duas subcategorias: modelos de reconstrução e modelos de

previsão sequencial, embora várias arquiteturas híbridas existam.

Autoencoders: Um *autoencoder* é uma rede neural treinada para comprimir os dados de entrada em uma representação de menor dimensão (codificação) e, em seguida, reconstruir os dados originais a partir dessa representação. Ao treinar um autoencoder usando apenas dados normais, espera-se que ele reconstrua bem os padrões vistos durante o treinamento, mas falhe em reconstruir padrões anômalos não observados. A métrica de detecção é o erro de reconstrução: se a entrada for reconstituída com erro elevado, é sinal de que não se encaixa nos padrões aprendidos (portanto, possivelmente anômala) [89]. Estudos demonstraram que autoencoders são capazes de detectar anomalias sutis que métodos lineares como PCA (*Principal Component Analysis*) não conseguiam, ao modelar relações não lineares entre variáveis [90].

Variações incluem *denoising autoencoders* (que aprendem a reconstruir dados a partir de versões corrompidas, tornando o modelo mais eficaz a ruído) e VAE (*variational autoencoders*), que impõem uma distribuição probabilística à codificação e permitem avaliar a probabilidade de um novo ponto pertencer ao perfil normal aprendido. Por exemplo, Sakurada e Yairi [90] aplicaram autoencoders para detecção de anomalias em telemetria de satélites, obtendo melhor desempenho que métodos baseados em PCA na identificação de falhas simuladas.

Extensões recentes combinam autoencoders com técnicas adversariais para aprimorar a detecção de anomalias. Um exemplo importante são as Redes Geradoras Adversariais (GANs do inglês, *Generative Adversarial Networks*), que treinam simultaneamente dois componentes: um gerador, responsável por produzir dados que imitam os padrões normais, e um discriminador, que aprende a distinguir entre dados reais e aqueles gerados artificialmente. Nesse contexto, anomalias podem ser identificadas por apresentarem características que dificultam sua geração ou sua correta classificação pelo discriminador. Um exemplo dessa abordagem é o método DAEMON (*Detecting Anomalies with an Energy-based adversarial autoencoder Model*), que utiliza um autoencoder adversarial para para aprender padrões normais e sinaliza anomalias com base no erro de reconstrução e em um discriminador treinado para detectar reconstruções falsas [91].

Em suma, autoencoders fornecem uma forma potente de aprendizado não supervisionado de características relevantes dos dados temporais, e seu erro de reconstrução atua como um indicador de novidade.

Redes Recorrentes: As LSTM (*Long Short-Term Memory*) e outras redes neurais recorrentes foram projetadas para modelar sequências, mantendo um estado interno que pode lembrar informações por longos intervalos. Isso as torna naturalmente adequadas para séries temporais. Para detecção de anomalias, LSTMs podem ser usadas de duas maneiras principais:

1. **Previsão direta:** a rede é treinada para prever o próximo valor (ou próxima janela) da série a partir dos valores anteriores, similar à abordagem preditiva mencionada antes, mas agora aprendendo padrões complexos de forma não linear. Novamente, um grande erro de previsão sugere uma anomalia.

2. **Arquitetura *encoder-decoder:*** refere-se a um modelo composto por dois módulos: um codificador (*encoder*), responsável por comprimir a sequência de entrada em uma representação latente, e um decodificador (*decoder*), que tenta reconstruir a sequência original a partir dessa representação. Essa arquitetura pode ser implementada com diferentes tipos de camadas neurais, como densas, convolucionais, LSTMs, dependendo do domínio e das características dos dados. Quando aplicada a séries temporais, sua estrutura se assemelha a um autoencoder sequencial, onde desvios entre a sequência original e a reconstruída, ou seja, o erro de reconstrução, podem indicar a presença de anomalias.

Malhotra et al. [92] demonstraram sucesso com LSTM *encoder-decoder* para detecção de anomalias em múltiplos sensores, mostrando que a rede podia aprender relações temporais em vários sinais e detectar desvios quando um sensor começava a apresentar comportamento anormal não visto no treinamento. Em aplicações práticas, LSTMs têm detectado anomalias complexas em dados de sistemas de TI e sensores Internet das Coisas (IoT do inglês, *Internet of Things*) onde a correlação temporal é forte [93–95].

Um caso notável foi o trabalho de Hundman et al. [96] aplicado à telemetria de satélites: eles treinaram LSTMs em séries multivariadas de sensores de uma espaçonave e utilizaram um esquema de limiar dinâmico nos resíduos de previsão para detectar anomalias de funcionamento. Esse sistema conseguiu identificar diversos eventos anômalos de forma não supervisionada, servindo como alerta antecipado para engenheiros de missão.

Redes LSTM podem também incorporar múltiplos passos de previsão (seq2seq), detectar anomalias coletivas e, com ajustes, lidar com sazonalidades e tendências. Entretanto, treiná-las requer volume considerável de dados e ajuste cuidadoso de hiperparâmetros para evitar sobreajuste e grande poder computacional [97].

Outra tendência é a integração de modelos baseados em *Graph Neural Networks* (GNNs) com LSTMs, permitindo capturar simultaneamente as dependências espaciais e temporais em séries temporais multivariadas. Essa abordagem tem se mostrado eficaz na detecção de anomalias que se propagam através de estruturas complexas, como redes de sensores ou sistemas de tráfego [98]. Essas arquiteturas híbridas grafo-sequência representam uma fronteira ativa de pesquisa.

3.3 Desafios comuns na detecção de anomalias

Apesar dos avanços metodológicos, a detecção de anomalias em séries temporais ainda enfrenta desafios teóricos e práticos. Um dos principais obstáculos é a escassez de rótulos, o que dificulta a formação de conjuntos de dados representativos. Essa limitação compromete a aplicação de métodos supervisionados e motiva o uso de técnicas não supervisionadas, as quais, por sua vez, enfrentam dificuldades relacionadas à calibração de hiperparâmetros e à escolha de representações temporais adequadas [56, 57, 79].

Outro desafio relevante refere-se à natureza dinâmica e não estacionária dos dados em muitas aplicações, como séries temporais epidemiológicas, transações financeiras e sensores industriais. A presença de ruído e a variabilidade natural nos dados, especialmente em medições realizadas por sensores, frequentemente resultam em valores espúrios e ausentes. Métodos que assumem estabilidade estatística ao longo do tempo podem falhar em contextos nos quais padrões normais mudam com frequência, exigindo a adoção de estratégias adaptativas ou modelos incrementais capazes de aprender continuamente e de forma eficiente em ambientes em constante mudança [99].

A complexidade aumenta em séries multivariadas, nas quais anomalias podem se manifestar não em valores individuais, mas nas relações entre variáveis. Nesse cenário, métodos univariados mostram-se limitados, enquanto modelar diretamente a distribuição conjunta dos dados enfrenta a chamada "maldição da dimensionalidade" [100]. Abordagens como redução de dimensionalidade, modelagem de dependências com redes bayesianas, ou técnicas específicas para dados multivariados, têm sido propostas com o objetivo de capturar padrões relevantes [101].

Outro ponto crítico envolve a sazonalidade, tendências e mudanças de regime. Métodos que não consideram componentes sazonais podem interpretar padrões periódicos como anomalias, enquanto mudanças graduais no comportamento dos dados ao longo do tempo, também conhecidas como *concept drift*, tornam ineficientes modelos previamente calibrados. Técnicas de

decomposição, modelagem adaptativa e detecção de pontos de transição são alternativas utilizadas para mitigar esse tipo de viés [102].

Em síntese, os principais desafios da detecção de anomalias em séries temporais envolvem a escassez de rótulos, a presença de ruído, a alta dimensionalidade, complexas correlações multivariadas e a dinamicidade temporal. Essas dificuldades explicam por que não existe uma solução universal; a seleção e o sucesso de um método dependem intrinsecamente do contexto da aplicação e das características dos dados.

3.4 Direções futuras de pesquisa

Com o avanço de aplicações baseadas em IoT e dados em fluxo contínuo, cresce a demanda por modelos adaptativos e de detecção em tempo real. Técnicas de aprendizado incremental, detecção de concept drift e frameworks de aprendizado contínuo são áreas-chave para viabilizar sistemas que se ajustam dinamicamente a mudanças nos padrões de comportamento sem intervenção humana [103]. Algoritmos meta-cognitivos, capazes de monitorar sua própria performance e decidir quando se reconfigurar, constituem uma linha de pesquisa emergente.

Outro campo em expansão é a integração de conhecimento de domínio e interpretabilidade dos modelos. Métodos híbridos que combinam aprendizado de máquina com modelos dinâmicos, têm o potencial de aumentar a confiabilidade e reduzir sinais de alertas falsos [32]. Simultaneamente, técnicas de *Explainable AI* vêm sendo adaptadas a contextos temporais, com o objetivo de fornecer explicações locais e interpretáveis sobre as detecções em séries temporais, o que é essencial em domínios regulados, como saúde ou finanças [104].

No âmbito da modelagem, novas arquiteturas neurais, como redes baseadas em atenção multiescala e modelos generativos profundos, vêm sendo exploradas para capturar padrões temporais complexos em diferentes escalas. Abordagens de *ensemble learning*, que combinam diferentes modelos com características complementares, têm se mostrado eficazes na construção de sistemas de detecção mais robustos e sensíveis a múltiplos tipos de anomalia [105].

Por fim, o futuro da detecção de anomalias aponta para sistemas cada vez mais autônomos, adaptativos e integrados a contextos reais. À medida que cresce a dependência de sistemas automatizados em domínios críticos, a capacidade de detectar comportamentos anômalos de forma confiável torna-se um requisito essencial para a segurança, eficiência e tomada de decisão.

4
Técnicas da Inteligência Artificial

O aprendizado de máquina (ML do inglês, *Machine Learning*) é um subconjunto de procedimentos e técnicas da Inteligência Artificial (IA) que se utiliza da capacidade dos algoritmos em aprender com dados de treinamento específicos de problemas e melhorar continuamente o desempenho com a experiência acumulada. Desta forma é possível otimizar o processo de construção de modelos analíticos e resolver tarefas associadas, sem a necessidade de ser programado explicitamente, ou seja, os algoritmos são capazes de aprender a partir de dados e fazer previsões ou tomar decisões sem serem programados para cada tarefa. [106].

Os algoritmos de ML têm sido aplicados com sucesso em diversos campos da ciência, fornecendo avanços em áreas como visão computacional [107], finanças [108], astrofísica [109], física [110], biologia [111, 112] e medicina [113]. Na saúde coletiva, particularmente na epidemiologia, o ML tem contribuído na modelagem de doenças infecciosas e na identificação de fatores de risco, que são cruciais para detectar e monitorar surtos de doenças infecciosas [32, 114, 115]

Em geral, dois tipos principais de algoritmos são usados: aprendizado supervisionado e aprendizado não supervisionado. A diferença entre eles é definida pelo processo pelo qual o algoritmo aprende sobre dados.

O aprendizado supervisionado é definido pelo uso de conjuntos de dados rotulados para treinar algoritmos que classificam dados ou predizem resultados com precisão. Conforme os dados de entrada são alimentados no modelo, ele otimiza suas pontuações até que esteja ajustado de maneira adequada. Isso ocorre como parte do processo de validação cruzada para garantir que o modelo evite super ajuste ou subajuste.

O aprendizado não supervisionado usa algoritmos de ML para analisar e agrupar conjuntos de dados não rotulados. Esses algoritmos descobrem padrões ocultos ou agrupamentos de dados sem a necessidade de intervenção humana. A capacidade deste método de descobrir semelhanças e diferenças nas informações o torna ideal para análise exploratória de dados e reconhecimento de padrões.

Neste capítulo, exploramos técnicas do aprendizado de maquina para detecção de outliers. Normalmente os outliers são considerados valores discrepantes, e são comumente entendidos como instâncias, ações ou objetos que estão fora do normal, como já comentado anteriormente. Estatisticamente, isso se refere a pontos de dados ou padrões inesperados que não se conformam a

um comportamento esperado [116, 117]. Essa definição pode ser mais explorada ao se considerar um conjunto abstrato de dados descritos por um determinado número de funções. Nesse caso, outlier designa qualquer ponto que não possa ser ajustado a pelo menos uma dessas funções, originando-se, em vez disso, de uma distribuição desconhecida, estranha aos demais dados. Por outro lado, quaisquer pontos que possam ser ajustados a essas funções descritivas são considerados pontos de dados normais (ou *inliers*) [117].

Existem dois tipos gerais de detecção de outliers: global e local. Os outliers globais ficam fora do intervalo normal para um conjunto de dados inteiro, enquanto os outliers locais podem estar dentro do intervalo normal para todo o conjunto de dados, mas fora do intervalo normal para os pontos de dados circundantes [118]. Dependendo da quantidade, tipo, rotulagem e outras características de um determinado conjunto de dados, a maneira como tais anomalias são identificadas podem variar de maneira significativa.

Ao abordar dados sindrômicos, que são não rotulados, os algoritmos não supervisionados emergem como uma ferramenta imprescindível e adequada, pois não necessitam de rótulos. Dessa forma, utilizaremos neste trabalho quatro técnicas de deteção de outlier (TDO) não supervisionadas conhecidas: Floresta de Isolamento (ISF do inglês, *Isolation Forest*), Fator de Anomalia Local (LOF do inglês, *Local Outlier Factor*), Máquina de Vetor de Suporte de Classe Única (OCSVM do inglês, *One-class Support Vector Machine*) e Detecção de Anomalia Baseado em Cópula (COPOD do inglês, *Copula-Based Outlier Detection*).

Essas técnicas são frequentemente usadas para detecção de anomalias em diversas áreas de pesquisa e disciplinas de aplicação. Em geral, a detecção de anomalia, por algoritmos não supervisionados, envolve a análise dos dados e uma atribuição de pontuação de anomalia (ou *score*) aos dados, com base nas características inerentes ao conjunto de dados [118, 119]. Este score quantifica o grau de anormalidade de cada ponto de dados, permitindo uma avaliação relativa entre eles. Diversos estudos têm evidenciado a eficácia dessas TDOs, como demostrado em [79, 119–121].

4.1 Isolation Forest

O algoritmo *Isolation Forest* (ISF) [85, 86], é uma TDO não supervisionada inspirado no algoritmo de *Random Forest* [122], que faz uso do conceito de isolamento, ou seja, implementa um conjuntos de árvores de isolamentos (ou floresta de isolamento) projetadas para particionar recursivamente o conjunto de dados, conforme indicado na Fig. 4.1. Neste

processo, uma árvore de isolamento é obtida pela divisão aleatória do espaço de características até que cada ponto de dados esteja contida em um nó folha terminal da árvore. Aos pontos de dados são atribuídas pontuações de anomalia que são inversamente relacionadas ao comprimento do ramo entre a raiz e o nó folha da árvore. Na análise final, as anomalias são identificadas com base em suas pontuações médias em todas as árvores na floresta.

Posto isto, o algoritmo opera em duas etapas: i) a etapa de treino, que corresponde essencialmente à construção da floresta, e ii) a chamada etapa de pontuação, onde é gerada a pontuação de anomalia para cada ponto do conjunto de dados.

Seja um conjunto de dados $\mathbf{X} \in \mathbb{R}^{n \times m}$ consistindo de n pontos de dados ($x_i \in \mathbb{R}^n$, $i = 1, 2, ..., n$) e m características ($x_j \in \mathbb{R}^m$, $j = 1, 2, ..., m$). A construção de uma árvore na floresta de isolamento é realizada gerando nós na árvore isolamento $\mathcal{T}$ ao amostrar aleatoriamente um subconjunto dos dados de treinamento de tamanho $n_s \leqslant n$. Após isso, uma característica $j \in m$ é aleatoriamente selecionada, juntamente com uma divisão aleatória no ponto s no intervalo $[x_{j,\min}, x_{j,\max}]$. O subconjunto de dados $x_j \leqslant s$ é atribuído ao ramo esquerdo da árvore e $x_j > s$, ao ramo direito na árvore, ou seja, para cada árvore $\mathcal{T}$, uma característica x_j é escolhida aleatoriamente e, dentro dessa característica, um valor entre seu mínimo e máximo é selecionado para dividir os pontos de dados e expandir a árvore. Vale ressaltar que cada árvore de isolamento é uma árvore binária completa, isto é, cada nó ou é terminal (sem ramificações) ou possui duas ramificações: esquerda e direita. Se uma árvore estiver completamente desenvolvida e os dados forem distinguíveis, haverá n nós externos e $n-1$ nós internos. Portanto, a árvore de isolamento tem $2n-1$ nós.

O processo é repetido para cada um dos ramos descendentes na árvore $\mathcal{T}$ e de todas as árvores na floresta ($\mathcal{T} = 1, 2, ..., N$), terminado quando a profundidade máxima da árvore ou o tamanho mínimo do nó terminal $n_{\min}$ é atingido. Ao final da construção de todas as árvores $\mathcal{T}$, o treinamento é concluído.

Uma vez que uma floresta de isolamento tenha sido construída, o comprimento do caminho esperado, $h(x_i)$, necessário para isolar a amostra x_i dos dados de treinamento é computado a partir da média dos comprimentos dos caminhos em todas as árvores. A pontuação de anomalia $s(x_i)$, para a i-ésima amostra nos dados de treinamento é definida como

$$s(x_i, n) = 2^{-\frac{E[h(x_i)]}{c(n)}} \tag{4-1}$$

onde $E[h(x_i)]$ é a média das profundidades alcançadas por x_i em todas as árvores, e $c(n)$ [123] é um fator de normalização calculado como

$$c(n) = 2H(n-1) - \frac{2(n-1)}{n} \tag{4-2}$$

onde $H(i)$ é o i-ésimo número harmônico, que pode ser aproximado por $\ln(i)+\gamma$, onde γ é a constante de Euler-Mascheroni, aproximadamente 0,57721.

Isso garante que, quando $E[h(x_i)] = c(n)$, a pontuação de anomalia é $s(x_i) = 0,5$. Além disso, para grandes valores de $h(x_i)$ em comparação com $c(n)$, a pontuação de anomalia tende a 0, e para pequenos valores de $h(x_i)$ em comparação com $c(n)$, $s(x_i)$ tende a 1. Assim, utilizando $s(x_i)$, é possível fazer a seguinte avaliação: (i) Se os pontos de dados retornam pontuações muito próximo de 1, então eles são classificados globalmente como anomalias; (ii) Se os pontos de dados apresentam pontuações bem menores que de 0,5, então podem ser considerados seguramente como pontos de dados normais; e (iii) Se todos os pontos de dados retornam pontuações aproximadamente igual a 0,5, então a amostra inteira não possui realmente nenhuma anomalia distinta [85, 86].

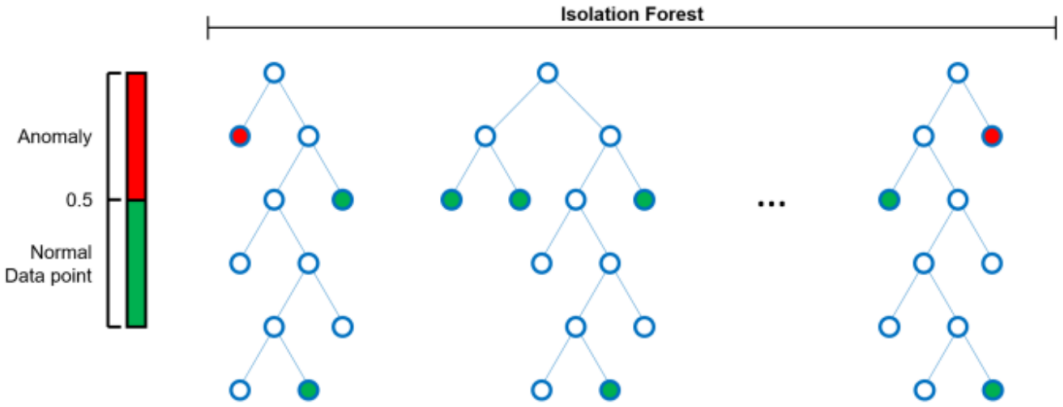


Figura 4.1: Visão geral do algoritmo ISF. Círculos verdes representam amostras normais comuns e os círculos vermelhos representam outliers. Figura retirada do artigo Ref. [124].

O ISF pode ser encontrado implementado nas bibliotecas Scikit-learn [125] e Python Outlier Detection (PyOD) [126], sendo possível experimentar diversos hiperparâmetros. Entre os mais usados se encontram o número de estimadores (quantidade de árvores), número máximo de amostras utilizadas por árvore, número de características utilizadas por cada árvore e contaminação no conjunto de dados (estimativa de outliers) [79].

Em suma o ISF é uma TDO baseado em uma abordagem diferente das demais (estatísticas, clustering, vizinhos mais próximos, etc.), sendo a primeira TDO proposta na categoria baseada em isolamento. Ela possui um requisito de memória baixo e pequeno [85, 86] e um tempo de execução linear, proporcional ao tamanho do conjunto de dados. Possui excelente escalabilidade que torna este método adequado para grandes conjuntos de dados, bem como para processamento em tempo real. A eficácia do algoritmo é influenciada principalmente por dois parâmetros: o número de árvores (N) e o tamanho da subamostra (n).

Quando um grupo de anomalias é grande e denso, o ISF converge rapidamente em sua capacidade de detecção e também minimiza erros de classificação, conhecidos como *swamping* e *masking* [127, 128], uma vez que ele depende de amostras aleatórias e não de todo o conjunto de dados para construir a floresta de isolamento [128]. Entretanto, o ISF tem algumas limitações: por ser sensível à escolha das características, é mais adequado para conjuntos de dados numéricos, e, às vezes, os resultados podem não ser intuitivos, dependendo do contexto dos dados.

4.2 Local Outlier Factor

Outra TDO amplamente utilizada é o *Local Outlier Factor* (LOF). Inspirado pelo conceito de agrupamento baseado em densidade local, o algoritmo LOF foi proposto por Breunig et al. [83]. Esse algoritmo foi desenvolvido como uma técnica não supervisionada para identificação de outlier, considerando individualmente cada ponto no conjunto de dados e atribuindo um *fator* de ser um outlier, ou seja, quantificando o quão outlier o ponto de dados é.

A ideia central do LOF é a não utilização do conceito de anomalia como um atributo binário, o qual é comumente adotado pela maioria das TDOs; em vez disso, propõe-se tratar a anomalia dentro de uma densidade local. Esta abordagem local foca exclusivamente na vizinhança imediata de cada ponto de dados, calculando quão distante um ponto de dado está em relação aos seus vizinhos [83]. Esse algoritmo foi uma inovação na detecção de anomalias de forma localizada e a partir dele, diversos outros foram desenvolvidos, conforme discutido na revisão [118].

A metodologia aplicada pelo LOF pode ser entendida a partir das seguintes definições:

Definição 1. k-distância de um ponto de dados a.

A distância entre dois pontos de dados a e b pode ser calculada usando um espaço n-dimensional Euclidiano, como:

$$d(a,b) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2}. \tag{4-3}$$

Seja um conjunto de dados D e um número inteiro positivo k. Para um ponto de dados a, a k-distância(a) é a distância $d(a,b)$ entre a e o ponto de dados vizinho mais distante b ($b \in D$) nas seguintes condições:

1. Pelo menos, k pontos de dados $b' \in D - \{a\}$ satisfaz a condição $d(a,b') \leqslant d(a,b)$;
2. No máximo, $k-1$ pontos de dados $b' \in D - \{a\}$ satisfaz a condição $d(a,b') < d(a,b)$.

Definição 2. k-vizinhos mais próximos de a.

Aqui, o significado de k-vizinhos mais próximos (kNN do inglês, *k-Nearest Neighbors*) de a é que qualquer ponto de dados b cuja distância até o ponto de dados a não é maior que a k-distância(a). Esses kNN formam a chamada vizinhança de k de distância de a, conforme descrito:

$$N_k(a) = \{b \in D - \{a\} | d(a,b) \leqslant k\text{-distância}(a)\}. \tag{4-4}$$

Definição 3. Distância de Alcançabilidade de a em relação a b.

Seja k um número inteiro positivo. A distância de alcançabilidade (RD do inglês, *Reachability Distance*) de um ponto de dados a em relação ao ponto de dados b é definida como:

$$RD_k(a,b) = \max\{k\text{-distância}(a), d(a,b)\}. \tag{4-5}$$

ou seja, é determinada considerando o máximo entre a k-distância de a e a distância real entre a e b.

Definição 4. Densidade de Alcançabilidade Local de a.

Em algoritmos de agrupamento baseados em densidade, dois parâmetros são usados para definir a noção de densidade: i) um número mínimo de pontos de dados e ii) um volume [118]. Breunig et al. [83] definiram $Dl_{N_k}(a,b)$ para $b \in N_k(a)$ como uma medida de volume. Assim, a densidade de alcançabilidade

local (LRD do inglês, *Local Reachability Density*) do ponto de dados a é definida como:

$$\mathrm{LRD}_{N_k}(a) = \frac{|N_k(a)|}{\sum_{b \in N_k(a)} RD_{N_k}(a,b)}. \tag{4-6}$$

Definição 5. Fator LOF de a.

Uma vez calculada a média das razões da LRD do ponto de dados a e dos vizinhos N_k mais próximos do ponto de dados a, uma pontuação LOF é atribuída a cada ponto de dados, conforme a expressão:

$$LOF(a) = \frac{\sum_{b \in N_k(a)} \frac{\mathrm{LRD}(b)}{\mathrm{LRD}(a)}}{|N_k(a)|}. \tag{4-7}$$

Para determinar se um ponto de dados a é um outlier, utiliza-se a pontuação ou fator LOF. Se para a maioria dos pontos a dentro de um cluster, o LOF de a for aproximadamente 1, isso indica uma densidade local similar à de seus vizinhos [83]. Valores do fator LOF superiores a 1 para um ponto de dados a sugerem que ele é uma anomalia, refletindo uma densidade local inferior à dos pontos adjacentes, conforme ilustrado na Fig. 4.2.

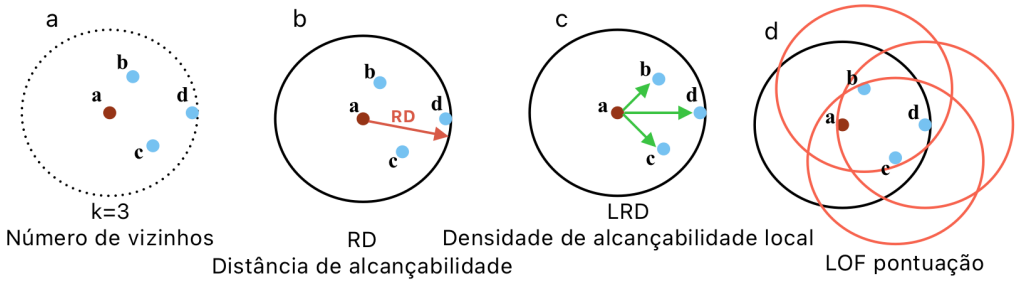


Figura 4.2: Ilustração do processo do algoritmo LOF. (a) Define-se o K-vizinhos mais próximos; (b) Calcula-se a distância de alcançabilidade (RD); (c) Calcula-se a densidade de alcançabilidade local (LRD); (d) Os pontos de dados são considerados outliers se tiverem o "fator" densidade local significativamente maior que os demais vizinhos.

O algoritmo LOF pode ser encontrado já implementado nas bibliotecas *Scikit-learn* [125] e *PyOD* [126] e requer ajuste dos hiperparâmetros. Os principais hiperparâmetros para ajuste são o número de vizinhos, k, a ser considerado para cada ponto de dados e a métrica para medir a distância, sendo que a mais usada é a euclidiana [79].

O LOF destaca-se por sua capacidade de detectar anomalias locais em conjuntos de dados com estruturas complexas, especialmente em ambientes de dados estáticos. Em comparação com outras TDOs, o LOF apresenta menor sensibilidade à quantidade de hiperparâmetros. Adicionalmente, o LOF é severamente afetado pela alta dimensionalidade dos dados, resultando em uma intensa demanda computacional em cenários com grande volume de dados [118, 79]. Embora o LOF ofereça um método para identificar outliers locais, que é sua principal especialidade, sua utilização deve ser cuidadosamente avaliada considerando as características específicas dos dados em análise.

4.3 One-Class Support Vector Machine

O algoritmo *One-Class Support Vector Machine* (OCSVM) é outra TDO não supervisionada, estatisticamente não paramétrica, popularmente aplicada para a classificação de apenas uma classe, denominada classe alvo [87]. Proposto inicialmente por Schölkopf et al. [84], essa abordagem busca identificar uma fronteira de decisão que separa a maioria dos dados, a classe alvo, dos pontos de dados que não pertencem a essa classe, considerados outliers.

Para entender o OCSVM, é essencial começar pelo conceito de Máquina de Vetor de Suporte (SVM do inglês, *Support Vector Machine*), introduzido por Vladimir Vapnik e ajustado por Vapnik e Corinna Cortes [129]. O SVM, originalmente desenvolvido para classificar dados em dois grupos, também conhecido como SVM de classe binária, baseia-se no princípio de mapear os dados do espaço de entrada para uma dimensão superior, chamada espaço de características. Neste espaço, os dados podem ser separados por equações lineares conforme o teorema de Mercer, também conhecido como função kernel [124]. O objetivo é encontrar o hiperplano de separação ideal, utilizando o coeficiente associado à função kernel para separar os dados, em duas classe distintas de dados. Os pontos de dados que se encontram nos limites desta margem são chamados de vetores de suporte. Entre as funções kernel mais utilizadas estão a linear, a polinomial, a gaussiana (função de base radial) e a sigmoide [124, 87].

A ideia central do OCSVM é semelhante à do SVM, mas tem como objetivo identificar uma única classe, que representa o comportamento normal dos dados, enquanto os dados que não seguem esse padrão são classificados como outliers. Esse classificador separa os dados de treinamento da origem (referência para os outliers) por meio de um hiperplano de separação com margem máxima, conforme mostrado na Fig. 4.3. Para garantir a separação

dos dados-alvo da origem, Schölkopf et al. [84] utilizaram o kernel gaussiano, no qual os dados são projetados na superfície de metade de uma hiperesfera de raio unitário, centrada na origem do espaço de projeção [87].

O primeiro passo do processo de treinamento é transformar os dados de entrada usando a função kernel e mapear os dados do espaço de entrada para o espaço de características (alta dimensão). Então, o algoritmo encontra o melhor hiperplano separador dos dados de treinamento para maximizar a margem

$$\mathcal{M} = \frac{b}{\|\mathbf{w}\|}, \tag{4-8}$$

onde $\mathbf{w}$ é o vetor normal do hiperplano e b é o seu viés (que representa a distância do hiperplano à origem). Para separar os dados da origem se propõe resolver o seguinte problema de otimização:

$$\min_{(\mathbf{w},b,\xi)} \left[\frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i - b \right], \tag{4-9}$$

sujeito a desigualdade

$$(\mathbf{w}\cdot\Phi(x_i)) \geqslant b - \xi_i \quad ,$$

e

$$\xi_i \geqslant 0, \quad \forall i = 1,\ldots,n \quad ,$$

onde ν é o coeficiente de regularização ou o parâmetro que controla a separação dos dados ou a proporção de outliers, variando entre 0 a 1 ($\nu \in [0,1]$), ou seja, controla o equilíbrio entre maximizar a distância do hiperplano à origem e o número de pontos de dados contidos pelo hiperplano. n representa o número de pontos no conjunto de treinamento e os ξ são variáveis *slack* (folga) que são usadas para modelar os erros de separação, e têm valores positivos ($\xi > 0$). O problema de otimização na equação 4-9 é geralmente resolvido pela sua forma dual:

$$\min_{\alpha} \left[\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j K(x_i,x_j) \right] \tag{4-10}$$

sujeito a restrições

$$\sum_{i=1}^{n}\alpha_i = 1 \quad \text{e} \quad 0 \leqslant \alpha_i \leqslant \frac{1}{\nu n}, \quad \forall i = 1,\ldots,n$$

onde α_i é um multiplicador de Lagrange (ou "peso") associado ao i-ésimo dado de treinamento x_i. Os dados para os quais $\alpha_i > 0$ são denominados "vetores

de suporte". A função $K(x_i, x_j)$ representa o kernel, que calcula a similaridade entre os dados x_i e x_j.

Ao utilizar a função kernel para projetar os vetores de entrada em um espaço de características, permite-se a definição de fronteiras de decisão não lineares. Dada uma função de mapeamento de características:

$$\phi : X \rightarrow \mathbb{R}^N \tag{4-11}$$

onde ϕ mapeia vetores de treinamento do espaço de entrada X para um espaço de características de alta dimensão, podemos definir a função kernel como:

$$K(x, y) = \langle \phi(x), \phi(y) \rangle \tag{4-12}$$

Os principais kernels comumente utilizados são:

- **Kernel linear:** $K(x, y) = (x \cdot y)$;
- **Kernel polinomial:** $K(x, y) = (x \cdot y + 1)^d$, onde d é o grau do polinômio;
- **Kernel Gaussiano:** $K(x, y) = e^{-\|x-y\|^2/(2\sigma^2)}$, onde σ^2 é a variância;

O processo do OCSVM é ilustrado na Fig. 4.3.

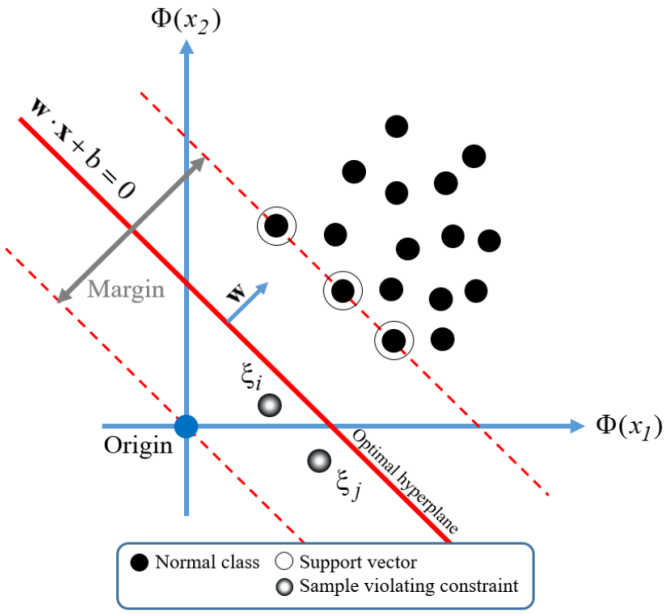


Figura 4.3: Ilustração gráfica do funcionamento do One-Class Support Vector Machine, destacando a separação entre os dados normais (pontos pretos) e os outliers, que são dados que violam a margem de separação (pontos cinzas). Os círculos vazios ao redor dos pontos representam os vetores de suporte, que definem a margem de separação entre os dados normais e a origem. O hiperplano ótimo é representado pela linha vermelha sólida. Figura retirada do artigo Ref. [124].

Uma das vantagens do OCSVM é sua capacidade de modelar fronteiras de decisão não lineares. No entanto, a escolha do kernel e dos parâmetros associados é crucial para seu desempenho. Em comparação com outras TDOs, ele pode ser computacionalmente mais caro, especialmente com grandes volumes de dados. O fato do OCSVM tentar separar os dados da origem com hiperplano de margem máxima é considerado seu ponto fraco, pois isso exige que a origem seja um membro da classe atípica [87]. Na literatura, entre os kernels mais comumente utilizados, o kernel Gaussiano — também conhecido como kernel de Função de Base Radial (RBF, do inglês *Radial Basis Function*) — é o único que, de forma consistente, apresenta bom desempenho [87].

Em conclusão, o algoritmo OCSVM oferece uma abordagem robusta e flexível para a detecção de anomalias, especialmente em cenários onde a separação não-linear entre os dados normais e anomalias é esperada. Tem sido aplicado com sucesso a muitos problemas práticos de classificação de uma classe [87]. Contudo, uma calibração cuidadosa dos hiperparâmetros é essencial para obter bons resultados [124] .

4.4 Copula-Based Outlier Detection

O algoritmo *Copula-Based Outlier Detection* (COPOD) foi introduzido por Li et al. [130], sendo uma técnica não paramétrica inovadora para a detecção multivariada de anomalias. Esta TDO distingue-se por sua abordagem estatística e a completa falta de hiperparâmetros. Além disso, é capaz de funcionar com e sem divisões de aprendizagem, pode operar como um algoritmo supervisionado ou não supervisionado e encontra-se implementado ao conjunto de TDOs do PyOD [126].

O algoritmo COPOD é baseado na exploração de propriedades de cópulas, que são derivadas de funções de distribuição cumulativa (FDC). As cópulas são funções que nos permitem separar distribuições marginais da estrutura de dependência de uma distribuição multivariada dada. Formalmente, define-se uma cópula d-variada, $C : [0,1]^d \rightarrow [0,1]$, como uma FDC de um vetor aleatório $(U_1, U_2, ..., U_d)$ com distribuições marginais uniformes no intervalo [0,1]

$$C(u) = P(U_1 \leqslant u_1, ..., U_d \leqslant u_d), \tag{4-13}$$

onde $P(U_j \leqslant u_j) = u_j$ para $j = 1, ..., d$ e $u_j \in [0,1]$. É bem conhecido que, aproveitando o Teorema do Limite Central através da amostragem de uma distribuição dada, ela pode ser transformada em uma distribuição uniforme

[117]. Ademais, qualquer distribuição uniforme pode ser transformada em qualquer função por meio de amostragem inversa [117, 130], conforme:

$$X_j = F_j^{-1}(U_j) \sim F_j. \tag{4-14}$$

Considerando isso, o Teorema de Sklar [131] pode ser aplicado, afirmando que, para quaisquer variáveis aleatórias $(X_1, ..., X_d)$ com função de distribuição conjunta $F(x_1, ..., x_d)$ e distribuições marginais $F_1, ..., F_d$, existe uma cópula tal que:

$$F(x) = C(F_1(x_1), ..., F_d(x_d)). \tag{4-15}$$

Em outras palavras, uma cópula é uma função nos permite descrever a distribuição conjunta de $(X_1, ..., X_d)$ usando apenas suas marginais. Isso proporciona muita flexibilidade ao modelar conjuntos de dados de alta dimensão, pois podemos modelar cada dimensão separadamente.

Sklar [131] também mostra que se F tem distribuições marginais $F_1, ..., F_d$, então existe uma cópula C tal que a equação 4-15 seja válida. Além disso, se as marginais são contínuas, então C pode ser unicamente determinada. Da mesma forma, substituindo a equação 4-15 na equação 4-13, obtemos a equação da cópula expressa em termos das FDCs conjuntas e das FDCs inversas

$$\begin{aligned} C(u) &= P(F_{X_1}(X_1) \leqslant u_1, ..., F_{X_d}(X_d) \leqslant u_d), \\ C(u) &= P(X_1 \leqslant F_{X_1}^{-1}(X_1), ..., X_d \leqslant F_{X_d}^{-1}(X_d)), \\ C(u) &= F_X(F_{X_d}^{-1}(u_1), ..., F_{X_d}^{-1}(u_d)). \end{aligned} \tag{4-16}$$

Assim, o Teorema de Sklar assegura que uma cópula pode ser estabelecida para qualquer função de distribuição cumulativa multivariada que tenha marginais contínuas. O citado teorena, oferece uma formulação explícita para a construção da cópula.

Portanto, uma cópula é um caso especial de FDC multivariado definido pela uniformidade da probabilidade marginal de cada variável no intervalo de [0, 1]. Porém, a natureza contínua destas FDCs impõe requisitos computacionais bastante caros que não se adaptam bem em casos multivariados. Em vez disso, pode ser usada uma função de distribuição cumulativa empírica (FDCE), definida como uma função degrau que aproxima o verdadeiro FDC através de um sistema com pontos de dados espaçados com uma frequência de $1/n$, sendo n a quantidade total de dados [117].

Considere X um conjunto de dados d-dimensional com n pontos de dados.

Define-se $X_{j,i}$ como o i-ésimo ponto de dados da j-ésima dimensão. Assim, a FDCE, $\hat{F}(x)$, é definida como:

$$\hat{F}(x) = P((-\infty, x]) = \frac{1}{n}\sum_{i=1}^{n}\Pi(X_i \leqslant x). \tag{4-17}$$

Ao aproveitar a equação 4-14, podemos obter a cópula empírica $\hat{U}_i$ por:

$$(\hat{U}_{1,i}, ..., \hat{U}_{d,i}) = (\hat{F}_1(X_{1,i}), ..., \hat{F}_d(X_{d,i})). \tag{4-18}$$

Finalmente, substituindo as cópulas empíricas na primeira igualdade da equação 4-16 , temos:

$$\hat{C}(u_1, ..., u_d) = \frac{1}{n}\sum_{i=1}^{n}\Pi(\hat{U}_{1,i} \leqslant u_1, ..., \hat{U}_{d,i} \leqslant u_d). \tag{4-19}$$

Nelsen [132] demonstra que uma cópula empírica, usando dados de n pontos uniformemente espaçados entre $\{1/n, 2/n, ..., 1\}$, cria uma distribuição que inicialmente possui marginais uniformes e discretas. Com o aumento de n, essa cópula empírica tende a se aproximar da cópula teórica, conforme indicado pelo Teorema do Limite Central.

A detecção de outliers com COPOD é um processo de três etapas. Primeiro, calcula-se as FDCEs baseadas no conjunto de dados de interesse. Segundo, usa-se as FDCEs para produzir a cópula empírica, e finalmente, usa-se a cópula empírica para estimar a probabilidade de cauda, que corresponde à pontuação de outlier.

Suponha que o algoritmo recebe uma entrada d-dimensional $X = (X_{1,i}, X_{2,i}, ..., X_{d,i}), i = 1, ..., n$, e produz um vetor de pontuação de outliers $O(X) = [X_1, ..., X_n]$. As pontuações de outliers estão entre $(0, \infty)$ e devem ser usadas comparativamente. Em outras palavras, a pontuação não indica a probabilidade de X_i ser um outlier, mas sim a medida relativa de quão provável X_i é comparado a outros pontos no conjunto de dados. Quanto maior $O(X_i)$, mais provável é que X_i seja um outlier.

Na primeira etapa, o COPOD ajusta d FDCEs da cauda esquerda, $\hat{F}_1(x), ..., \hat{F}_d(x)$, usando a equação 4-17, e d FDCEs da cauda direita $F_1(x), ..., F_d(x)$ substituindo X por $-X$. Também calcula um vetor de assimetria, $b = [b_1, ..., b_d]$, onde

$$b_i = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)^3},$$

é a expressão padrão para estimar a assimetria [130].

Na segunda etapa, calculam-se as cópulas empíricas para cada X_i conforme estabelecido na equação 4-18, obtendo-se $\hat{U}_{d,i} = \hat{F}_d(x_i)$, que representa as transformações para a cauda esquerda, e $\hat{V}_{d,i} = \hat{F}_d(x_i)$, que representa as transformações para a cauda direita, respectivamente. Além disso, a cópula empírica ajustada para assimetria é calculada, utilizando $\hat{W}_{d,i} = \hat{U}_{d,i}$ se a assimetria b_d for negativa; caso contrário, utiliza-se $\hat{V}_{d,i}$. Por fim, calcula-se a probabilidade de se observar um ponto pelo menos tão extremo quanto cada x_i ao longo de cada dimensão d, tomando-se o máximo do logaritmo negativo, definido como:

$$O_d(x_i) = -\max\left\{\log\left(\hat{U}_{d,i}\right), \log\left(\hat{V}_{d,i}\right), \log\left(\hat{W}_{d,i}\right)\right\}. \tag{4-20}$$

Essa medida é gerada a partir das probabilidades fornecidas pela cópula empírica das caudas esquerda e direita, e da cópula corrigida por assimetria, e serve como a pontuação de outlier para a dimensão d. O uso do logaritmo negativo assegura que valores mais altos da função $O_d(x_i)$ indiquem um maior grau de anormalidade da dimensão d, aproveitando a propriedade monótona da função logarítmica [130].

Intuitivamente, quanto menor a probabilidade da cauda, maior é seu logaritmo negativo, e assim associa-se que um ponto é um outlier se ele tem uma pequena probabilidade da cauda esquerda, uma pequena probabilidade da cauda direita ou uma pequena probabilidade da cauda corrigida por assimetria. A Figura 4.4 ilustra a identificação de pontos de dados anômalos com base nas pontuações de outlier calculadas pelo algoritmo COPOD, destacando como as probabilidades das caudas contribuem para a detecção de anomalias.

O algoritmo COPOD é particularmente adequado em cenários de alta dimensionalidade, onde o fenômeno conhecido como *maldição da dimensionalidade* [100] pode obscurecer os padrões nos dados para outras TDOs. O uso de cópulas empíricas, a habilidade de desacoplar e reacoplar as distribuições marginais e a dependência estrutural entre as dimensões tornam o COPOD uma ferramenta poderosa e flexível para a detecção de anomalias em uma vasta gama de aplicações práticas.

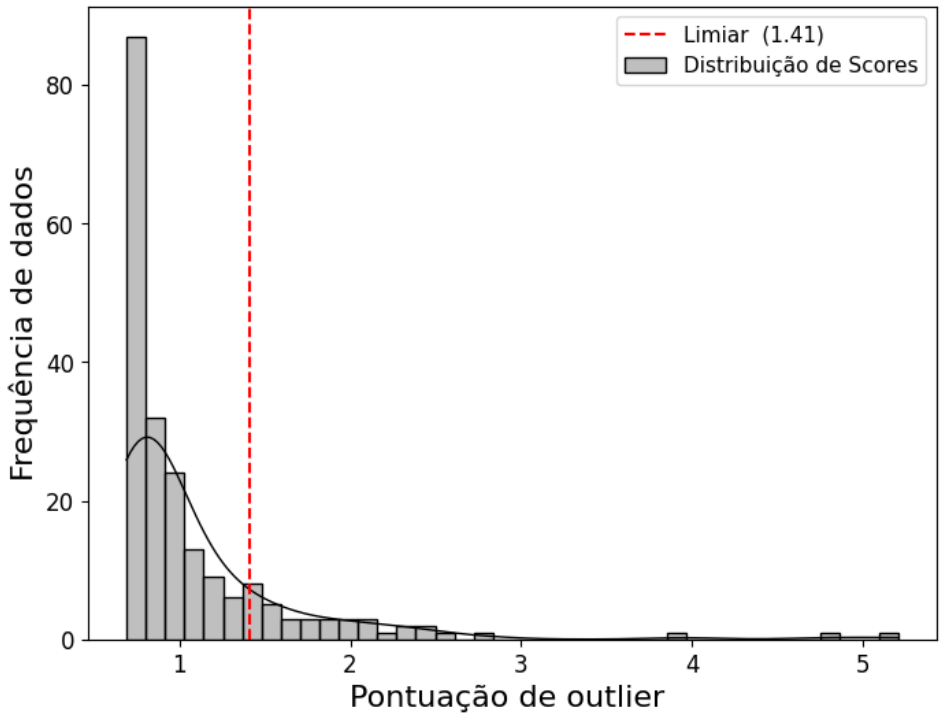


Figura 4.4: Distribuição das pontuações de outliers geradas pelo algoritmo COPOD. As barras cinzas representam a frequência dos dados, enquanto a linha preta corresponde à curva de densidade, ajustada as pontuações. A linha vermelha tracejada indica o limiar de detecção de anomalias. Esse limiar foi definido com base nos 40% de contaminação considerados nos dados da APS para a região imediata de Belo Horizonte. Valores de pontuação superiores ao limiar, situados na região de cauda, são considerados outliers.

5
Conceitos de Ciências de Redes

Nas últimas décadas, assistiu-se ao desenvolvimento de um novo ramo na ciência: a teoria das Redes Complexas, atualmente conhecida como Ciência das Redes [133–135]. Este processo envolveu um esforço interdisciplinar de físicos, matemáticos e outros cientistas, com o objetivo de estabelecer conceitos e métodos precisos aplicáveis à análise de grandes conjuntos de dados e à construção de representações adequadas para as interações entre os componentes de sistemas complexos.

Embora a ciência das redes seja um campo de estudo extremamente antigo, com raízes que remontam a 1736 na aplicação da teoria matemática dos grafos a problemas práticos, como a solução do problema das pontes de Königsberg [136], o estudo sistemático de redes ficou relativamente estagnado até o século XX. Na década de 1950, Erdős e Rényi [137] formalizaram a teoria dos grafos aleatórios, introduzindo um modelo rede onde cada par de vértices é conectado aleatoriamente com a mesma probabilidade, resultando em uma distribuição de grau de Poisson. No entanto, foi somente no final da década de 1990 que a ciência das redes realmente ganhou notoriedade, quando vários cientistas de diferentes áreas começaram a utilizar redes como modelos para fenômenos físicos, biológicos e sociais. Em particular, os trabalhos de Duncan Watts e Steven Strogatz [138], e de Albert-László Barabási e Réka Albert [139], estimularam um renovado interesse na análise matemática de redes aplicadas ao mundo real.

Particularmente, a física estatística desempenhou um papel fundamental no desenvolvimento da ciência das redes, pois seus objetos de estudo são sistemas compostos por um grande número de elementos em interação, com principal interesse na determinação do comportamento macroscópico (ou coletivo) desses sistemas a partir das leis microscópicas que governam sua dinâmica [140]. Por essas razões, o desenvolvimento da teoria de sistemas complexos incorporou muitos conceitos originados na física estatística, e essa observação é igualmente válida para as redes complexas. Conceitos como invariância de escala, dimensão fractal, transporte, expoentes críticos, sincronização e agrupamentos (clustering), todos oriundos da física estatística, foram integrados à ciência de redes.

No estudo da ciência de redes, a identificação de princípios gerais que relacionam a estrutura e dinâmica da rede tem sido objetivo principal

de diversas investigações. [6]. Os princípios estruturais identificados são frequentemente fascinantes e sugestivos [141]. Exemplos proeminentes incluem arquiteturas de pequeno mundo [138] e livres de escala [139], e a estrutura de comunidades [142].

Entre as diferentes abordagens utilizadas para analisar a importância relativa de cada vértice e de cada arestas para a estrutura e dinâmica de uma rede complexa, avaliar a estrutura da comunidade é geralmente o primeiro passo [142]. Comunidades são subgrupos de vértices que estão mais densamente conectados entre si do que com outros nós da mesma rede.

A maioria das redes de mundo real exibe estrutura modular, ou seja, seus vértices estão organizados em grupos, chamados comunidades, clusters ou módulos. Esses grupos de vértices compartilham propriedades comuns e desempenham papéis semelhantes no sistema real [142, 143].

Portanto, a detecção de comunidades é de grande importância em várias áreas da ciência, tendo vasta aplicação. Como por exemplo: i) ao agrupar clientes da Web com interesses semelhantes e que estão geograficamente próximos pode melhorar o desempenho dos serviços na internet [144]; ii) identificar grupos de clientes com interesses comuns em redes de relações de compra ajuda a criar sistemas de recomendação e aumentam as oportunidades de negócios [145]; iii) identificação de grupos de proteobactérias a partir de redes biológicas contribuiram para questões de inferência filogenética [146, 147]; iv) detecção de comunidades em redes espaciais permitem apoiar a tomada de decisões de desenvolvimento e estratégias de regionalização, bem como, pode auxiliar na determinação de novas fronteiras (limites) administrativas, econômicas ou com base no risco epidemiológico [148–151]; v) estratégias de vigilância sentinelas, podem ser baseadas em estruturas modulares de redes de contato humano [152, 153]. A detecção de comunidades também é importante por outras razões. A identificação dos módulos e seus limites permite uma classificação dos vértices, de acordo com sua posição estrutural nos módulos e, possivelmente, demostra a organização hierárquica, apenas usando as informações codificadas na topologia da rede [154].

5.1 Redes monocamadas

Uma rede monocamada R com n vértices é formalmente representada como um par ordenado de conjuntos disjuntos (X, E), onde X é um conjunto não vazio de vértices (ou nós), indexados por $i = 1, 2, 3, \ldots, n$, e E é um subconjunto formado por pares não ordenados de elementos de X, isto é, $E = \{(i, j) \ ; \ i, j \in X\}$ representa o conjunto de arestas da rede. Cada par

$(i, j) \in E$ indica a existência de uma conexão entre os vértices i e j. Nesse caso, dizemos que i e j são adjacentes, isto é, ligados por uma aresta comum.

As conexões entre vértices podem possuir uma direção, caracterizando redes direcionadas. Quando não há direção associada às arestas, a rede é dita não direcionada. Além disso, redes podem ser ponderadas, quando cada conexão possui um peso associado — por exemplo, representando a intensidade do fluxo de pessoas entre municípios. Também é possível que existam múltiplas arestas entre um mesmo par de vértices, bem como auto-laços, nos quais um vértice se conecta a si próprio [7].

Quanto à evolução temporal, uma rede pode ser considerada estática quando sua estrutura permanece inalterada ao longo do tempo, isto é, o número de vértices, arestas e suas conexões não se modificam. Por outro lado, uma rede é dita dinâmica quando esses elementos variam com o tempo. Apesar disso, redes dinâmicas podem ser analisadas como estáticas dentro de determinados intervalos temporais, desde que as mudanças estruturais nesse período sejam inexistentes ou irrelevantes.

Uma maneira comum de representar redes monocamadas não ponderadas é por meio da matriz de adjacência A:

$$A_{ij} = \begin{cases} 1 & \text{se os vértices } i \text{ e } j \text{ estão conectados} \\ 0 & \text{caso contrário.} \end{cases} \tag{5-1}$$

Essa matriz contém todas as informações sobre as conexões entre os vértices. No caso de redes não direcionadas, a matriz é simétrica.

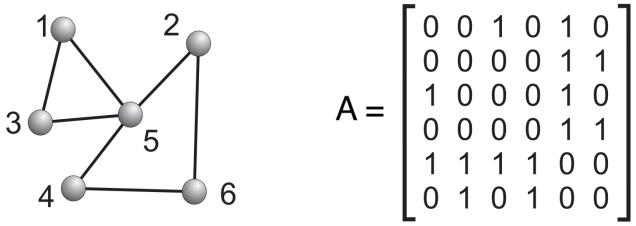


Figura 5.1: Exemplo de rede simples e sua matriz de adjacência.

Métricas de centralidade fornecem medidas fundamentais para caracterizar a estrutura da rede. Por exemplo, a centralidade de grau reflete as propriedades locais da rede subjacente, pois mede o número médio de conexões entre os nós. Outras métricas, como a centralidade de intermediação, fornecem informações sobre a estrutura geral da rede, pois se baseiam no cálculo e na contagem dos caminhos mais curtos [7].

Matematicamente, a centralidade de grau k_i de um vértice i é o número de conexões que ele tem com outros vértices, ou seja, sua quantidade de vizinhos.

Em uma rede não direcionada com n vértices, tem-se:

$$k_i = \sum_{j=1}^{n} A_{ij} \tag{5-2}$$

A partir dessa medida, três grandezas válidas para redes não direcionadas podem ser derivadas. A primeira é o grau médio, que representa a média aritmética dos graus de cada vértice, e é dada pela seguinte fórmula:

$$< k >= \frac{1}{n} \sum_{i=1}^{n} k_i. \tag{5-3}$$

Como o somatório de k_i sobre todos os vértices da rede é igual ao dobro de conexões, assim, defini-se o número de arestas da rede como:

$$m = \frac{1}{2} \sum_{i} k_i = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ji}. \tag{5-4}$$

Por fim, a distribuição de grau $P(k)$ descreve a probabilidade de que um vértice escolhido aleatoriamente possua exatamente k conexões. Essa função caracteriza como os vínculos estão distribuídos entre os nós da rede. Para que $P(k)$ represente uma distribuição de probabilidade válida, ela deve satisfazer a condição de normalização:

$$\sum_{k=1}^{\infty} P(k) = 1, \tag{5-5}$$

essa distribuição pode ser estimada como:

$$P(k) = \frac{n_k}{n}, \tag{5-6}$$

onde n_k representa o número de vértices com grau k, e n é o número total de vértices na rede.

Antes de introduzirmos o conceito de centralidade de intermediação, é importante apresentar algumas métricas fundamentais associadas a caminhos e distâncias em redes.

Um caminho entre dois vértices i e j é definido como uma sequência de vértices conectados por arestas, sem repetições. O comprimento do caminho corresponde ao número de arestas que compõem essa sequência. A distância entre dois vértices, por sua vez, é definida como o comprimento do menor caminho entre eles, chamado de caminho geodésico.

Com base nesses conceitos, pode-se definir o comprimento do caminho

característico, ou distância média da rede, por meio da expressão:

$$L = \frac{1}{n(n-1)} \sum_{i \neq j} d(i,j), \tag{5-7}$$

onde $d(i,j)$ representa a distância entre os vértices i e j, e n é o número total de vértices da rede. Outra métrica intimamente relacionada é a eficiência média da rede, proposta por Latora e Marchiori [155], a qual é definida como:

$$\eta(R) = \frac{1}{n(n-1)} \sum_{i \neq j} \frac{1}{d(i,j)}. \tag{5-8}$$

Essa medida avalia a capacidade de tráfego na rede, ou seja, o quão eficientemente a informação pode ser transmitida entre pares de vértices. Redes com distâncias curtas entre os nós tendem a apresentar maior eficiência.

Outro indicador relacionado à capacidade de tráfego é o diâmetro da rede, denotado por D, que corresponde à maior distância geodésica finita observada entre quaisquer dois vértices da rede. Isto é,

$$D = \max_{i,j \in X} d(i,j). \tag{5-9}$$

O diâmetro é uma métrica importante para avaliar a extensão da rede e sua acessibilidade global.

Fundamentado nesses conceitos, uma outra representação matricial bastante útil de uma rede é a sua matriz de vizinhança V [156, 157]. Para uma rede monocamada, ela é definida como:

$$V_{ij} = \sum_{l=1}^{D} l A_{ij}^{l} \tag{5-10}$$

onde D é o diâmetro da rede e ${A^l}_{ij}$ é definida sobre a distância d entre nós adjacentes. Logo:

$$A_{ij}^{l} = \begin{cases} 1 & d(i,j) = l, \\ 0 & \text{caso contrario .} \end{cases} \tag{5-11}$$

A matriz de vizinhança V sintetiza as informações presentes nos elementos A_{ij}^{l}, incorporando o fator l para indicar explicitamente a ordem de vizinhança entre os vértices. Essa estrutura facilita a análise e a visualização de propriedades relacionadas às distâncias entre pares de vértices na rede, permitindo identificar rapidamente o nível de proximidade entre os vértices.

Como exemplo, considere a Figura 5.1; a matriz de vizinhança correspondente é dada por:

$$V = 1\begin{pmatrix} 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \end{pmatrix} + 2\begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} + 3\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

$$V = \begin{pmatrix} 0 & 2 & 1 & 2 & 1 & 3 \\ 2 & 0 & 2 & 2 & 1 & 1 \\ 1 & 2 & 0 & 2 & 1 & 3 \\ 2 & 2 & 2 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 2 \\ 3 & 1 & 3 & 1 & 2 & 0 \end{pmatrix} \tag{5-12}$$

Vamos agora definir a centralidade de intermediação. É uma métrica que avalia a influência de um vértice no papel de intermediador da comunicação dentro de uma rede. Essa medida quantifica a frequência com que um vértice v atua como ponto de passagem nos caminhos geodésicos entre pares distintos de vértices [158]. Seja $\varphi(i,j)$ o número total de caminhos geodésicos entre os vértices i e j, e $\varphi_v(i,j)$ a quantidade desses caminhos que passam por v. A centralidade de intermediação do vértice v é então definida por:

$$b_v = \sum_{i \neq j}^{n} \frac{\varphi_v(i,j)}{\varphi(i,j)}. \tag{5-13}$$

O conceito de centralidade de intermediação de vértices também foi estendido às arestas [159, 160]. Enquanto aquela mede a influência de um vértice na conexão em pares de outros vértices, a centralidade de intermediação de arestas mede a importância de uma aresta no mesmo contexto, sendo definida por:

$$b_e = \sum_{i \neq j}^{n} \frac{\varphi_e(i,j)}{\varphi(i,j)}, \tag{5-14}$$

onde $\varphi(i,j)$ denota o número de caminhos geodésicos entre os vértices i e j, e $\varphi_e(i,j)$ denota o número de caminhos mais curtos conectando i e j que passam pela aresta e. Este conceito é útil para identificar arestas críticas em uma rede que, se removidas, podem desconectar a rede, determinando possíveis módulos na mesma, como proposto por Girvan e Newman [160], marcando o início de uma nova era no campo da detecção de comunidades [142].

Como mencionado anteriormente, a capacidade de mapear a estrutura de redes complexas tem inúmeras aplicações. Um simples cálculo de distância entre redes torna as estruturas modulares explícitas. A distância euclidiana, ou dissimilaridade $d(\alpha,\xi)$, entre duas redes α e ξ é definida pela soma das diferenças positivas entre os elementos das duas matrizes de vizinhança correspondentes $V(\alpha)$ e $V(\xi)$ [157]. Assim temos

$$d^2(\alpha,\xi) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\frac{V_{ij}(\alpha)}{D_\alpha} - \frac{V_{ij}(\xi)}{D_\xi} \right)^2 , \tag{5-15}$$

onde D_α e D_ξ representam os diâmetros de rede correspondentes. Se $D_\alpha = D_\xi$, um fator comum é removido dos dois denominadores. Para as situações comuns, onde $D_\alpha \neq D_\xi$, a definição mostrou-se bastante adequada, pois normaliza todos os termos na soma. É importante salientar que a definição 5-15 requer que as redes tenham exatamente o mesmo número de vértices n. No entanto, o procedimento pode ser estendido a pares de redes que possuem aproximadamente o mesmo número de vértices, incluindo-se de forma arbitrária vértices desconectados na rede com menor número de vértices, assim igualando o número de vértices nas duas redes. Os elementos da matriz de vizinhança envolvendo estes vértices são todos nulos na rede onde eles foram introduzidos.

Uma aplicação direta desta medida é a caracterização de diferentes redes obtidas a partir de uma matriz de similaridade genética S (dependente de um parâmetro σ) entre grupos de organismos, conforme demonstrado em vários estudos [146, 147, 161–163]. Aqui, α e ξ denotam duas redes obtidas a partir da matriz S ao se escolher dois valores próximos ao limiar de similaridade σ, tais que $\alpha \leqslant \sigma$, $\xi \leqslant \alpha + \delta\sigma$, com $\delta\sigma \ll 1$. Usando esta análise, podem ser identificados os valores de σ para os quais o caráter modular da rede é explicado de forma ótima. Isso permite determinar o conjunto mínimo de arestas que devem ser incluídas na rede para obter as informações relevantes necessárias à elucidação de sua estrutura.

Embora a medida de dissimilaridade entre redes permita identificar valores do parâmetro σ com forte indício de estrutura modular, ela não fornece, por si só, a decomposição explícita dessas comunidades. Para isso, é necessário

um método que explore e classifique a estrutura modular da rede, com base em métricas topológicas relevantes. Um dos métodos mais amplamente utilizados com esse propósito é o algoritmo hierárquico de Girvan-Newman (NG) [160], baseado na centralidade de intermediação de arestas.

O método de NG é amplamente utilizado na identificação de comunidades em de redes monocamadas. O processo inicia com o cálculo da centralidade de intermediação para todas as arestas na rede e, em seguida, remove-se a aresta com a maior centralidade. Este passo é repetido iterativamente, levando a uma desconexão gradual da rede até que todas as arestas tenham sido removidas.

Como resultado, é possível construir um dendrograma que segue o processo de divisão total da rede. Cada etapa de remoção de aresta forma-se um gráfico hierárquico, onde cada divisão indica a formação de uma nova comunidade. Assim, o dendrograma ilustra como as comunidades são separadas ao longo do tempo, com cada nível do gráfico correspondendo a um estado da rede após a remoção de determinadas arestas. Esse resultado é crucial para entender a estrutura comunitária da rede e para identificar em que ponto a rede se divide de forma significativa.

A organização modular de uma rede pode ser quantificada pela métrica de modularidade Q, que avalia a qualidade da divisão dos nós em comunidades [164], ou seja, quantificando se a partição da rede é ou não significativa, em comparação com aquela definida por um modelo nulo. A função Q é definida pela expressão:

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - P_{ij} \right] \delta(c_i, c_j), \tag{5-16}$$

onde $\delta(c_i, c_j) = 1$ se as atribuições da comunidade c_i e c_j dos vértices i e j são iguais e 0 caso contrário. P_{ij} representa o valor esperado da aresta entre os vértices i e j segundo um modelo nulo, sendo dado por $\frac{k_i k_j}{2m}$, onde k_i e k_j são os graus dos vértices e m é o número total de arestas da rede. A chamada métrica de Newman-Girvan é bastante empregada para identificar e aferir a modularidade de uma rede, sendo este o modelo nulo mais popular da literatura. A variedade e a complexidade das propriedades e tipos de redes são extensas, e descrever todas elas está além do escopo desta tese. No entanto, discussões detalhadas podem ser encontradas em [6, 7, 135, 165]

6
Detecção de Surtos Epidêmicos

A importância da vigilância epidemiológica tem crescido nos últimos anos devido ao aumento das ameaças à saúde pública, que são amplificadas pela rápida disseminação de doenças infecciosas. Esse cenário é exacerbado pelo crescimento populacional e pelos crescentes riscos ambientais [166]. Diante disso, as autoridades sanitárias estão em busca de estratégias eficazes que permitam uma detecção rápida de surtos inesperados no número de casos de doenças ou eventos de saúde, de forma a possibilitar a investigação sobre a origem do surto e a implementação de medidas de contenção. O grande desafio para o estabelecimento de estratégias eficazes reside no desenvolvimento de algoritmos que ofereçam um bom equilíbrio entre sensibilidade e especificidade, a fim de detectar a grande maioria dos surtos sem gerar demasiados alarmes falsos positivos [167]. Também, que sejam capazes de lidar com uma vasta gama de dados em saúde, tornando-os práticos e eficazes para uso real [168].

Em geral os algoritmos usam metodologias para detecção temporal de surtos de forma local, enquanto que a integração da distribuição espacial dos dados é menos enfatizado [169]. O princípio geral é identificar um intervalo de tempo em que o número observado nos dados de um evento sob vigilância é significativamente maior do que o esperado. Esta identificação baseia-se principalmente num processo de duas etapas: primeiro, estima-se um número esperado para os dados do evento de interesse para a unidade de tempo (por exemplo uma semana ou um dia) e depois compara-se com o valor observado por um teste estatístico. Um sinal de alerta precoce é acionado se o valor observado for significativamente diferente do valor esperado. A principal diferença entre os algoritmos de detecção reside no método como o valor esperado é estimado [170].

Na literatura, foram propostos diversos algoritmos baseados em diferentes metodologias para detectar surtos a partir de dados de vigilância [77, 78, 171–173]. Dentre eles, os de maior evidência são: i) o algoritmo conhecido como "Sistema de Relatório de Aberração Antecipada"(EARS do inglês, *Early Aberration Reporting System*) [77] baseado em processos de controle de Shewhart [174], desenvolvido e usado como sistema padrão nos Centros de Controle e Prevenção de Doenças (CDC do inglês, *Centers for Disease Control and Prevention*) dos Estados Unidos para conduzir vigilância sindrômica semanal [175]. O EARS também foi adaptado e aplicado ao sistema nacional *EpiSurv* da Nova

Zelândia para monitorar dados semanais de vigilância de doenças notificáveis [176]; ii) o algoritmo aprimorado baseado em regressão quase-Poisson [172], também conhecido como "Farrington Flexível", desenvolvido e atualmente usado na Saúde Pública da Inglaterra para detecção semanal de surtos de doenças infecciosas [172].

O algoritmo EARS é amplamente utilizado como um padrão de referência para avaliar o desempenho de outros algoritmos, devido à sua capacidade de operar com dados históricos de linha de base limitados (menos de cinco anos de dados) [177, 176]. Nesse sentido, vamos nos limitar ao detalhamento do EARS e apresentar informações básicas sobre o projeto ÆSOP [31], no qual este trabalho encontra-se inserido, que tem por objetivo desenvolver um sistema de vigilância sindrômica para o Brasil.

6.1 Early Aberration Reporting System (EARS)

O EARS [77] foi projetado como um sistema de vigilância sindrômica "pronto para uso" para monitorar eventos de grande escala para os quais poucos ou nenhum dado anterior existe. Desde 11 de setembro de 2001, coincidentemente data da explosão das torres gêmeas, o sistema EARS tem sido cada vez mais utilizado como um sistema de vigilância padrão. Está disponível através de suas três variantes usuais: EARS-C1, EARS-C2 e EARS-C3 [175].

Os métodos C1, C3 e uma forma modificada do C2 também são implementados no sistema BioSense do CDC. O BioSense é um programa nacional destinado a melhorar as capacidades de vigilância biossanitária em tempo quase real dos governos federal, estadual e local, incluindo tanto a "consciência situacional de saúde"quanto o "reconhecimento e resposta a eventos" [175].

Estes métodos foram concebidos para serem métodos semelhantes ao CUSUM (*Cumulative Sum*) [178] e, de fato, pelo menos um artigo [179] os referem explicitamente como CUSUMs. No entanto, o C1 e o C2 são, na verdade, variantes do procedimento de Shewhart [174] que usam uma média móvel de amostras e o desvio padrão da amostra para padronizar cada observação. O algoritmo C1 utiliza sete observações anteriores à observação atual para calcular a média da amostra e o desvio padrão da amostra. O C2 é semelhante ao C1, mas usa sete observações anteriores a um atraso de duas observações. O algoritmo C3 combina informações das estatísticas C2 conforme descrito abaixo.

Seja $Y(t)$ a contagem observada para o período semanal t, representando, por exemplo, o número de atendimentos que chegam a uma sala de emergência

de um hospital com uma síndrome específica no tempo t.

O método C1 calcula a estatística $C_1(t)$ como

$$C_1(t) = \frac{Y(t) - \hat{Y}_1(t)}{S_1(t)} \tag{6-1}$$

onde $\hat{Y}_1(t)$ e $S_1(t)$ são, respectivamente a média móvel de amostras e o desvio padrão da amostra

$$\hat{Y}_1(t) = \frac{1}{7}\sum_{i=t-7}^{t-1} Y(i) \quad e \quad S_1^2(t) = \frac{1}{6}\sum_{i=t-7}^{t-1} [Y(i) - \hat{Y}_1(i)]^2.$$

Conforme implementado no sistema EARS, o método C1 emite um sinal no tempo t quando a estatística $C_1(t)$ ultrapassa um limiar h, que é fixado em três desvios padrão acima da média de amostras:

$$C_1(t) > 3.$$

O método C2 é semelhante ao método C1, mas incorpora uma defasagem de duas semanas na média e nos cálculos do desvio padrão. Especificamente, ele calcula

$$C_2(t) = \frac{Y(t) - \hat{Y}_2(t)}{S_2(t)} \tag{6-2}$$

onde

$$\hat{Y}_2(t) = \frac{1}{7}\sum_{i=t-9}^{t-3} Y(i) \quad e \quad S_2^2(t) = \frac{1}{6}\sum_{i=t-9}^{t-3} [Y(i) - \hat{Y}_3(i)]^2,$$

e no EARS ele sinaliza quando $C_2(t) > 3$.

O método C3 utiliza as estatísticas C2 da semana t e das duas semanas anteriores, calculando a estatística $C_3(t)$ como

$$C_3(t) = \sum_{i=t}^{t-2} \max[0, C2(i) - 1]. \tag{6-3}$$

Dentro do método EARS, um sinal é emitido quando $C_3(t) > 2$.

6.2 Alert-Early System of Outbreaks with Pandemic Potential (ÆSOP)

Nesse contexto, o projeto ÆSOP (*Alert-Early System of Outbreaks with Pandemic Potential*) foi proposto como uma iniciativa brasileira de vigilância de última geração [31], desenvolvido pelo Centro de Integração de Dados e Conhecimentos para Saúde (Cidacs) da Fiocruz Bahia, em parceria com a Universidade Federal do Rio de Janeiro (COPPE/UFRJ) e com apoio estratégico e financeiro da Fundação Rockefeller, visando desenvolver estratégias e um sistema de alerta antecipado para surtos com potencial pandêmico.

O principal objetivo do ÆSOP é criar um sistema operacional que permita às autoridades de saúde a identificar precocemente sinais de emergência epidemiológica em nível municipal antes que estes evoluam para grandes crises sanitárias. Para isso, o sistema integra a análise de diversas fontes de dados, oriundas de atendimentos da APS, vendas de medicamentos, menções em redes sociais e mídias locais [31].

Inicialmente, o ÆSOP está focado na detecção de surto associados a síndromes respiratórias agudas, mas tem a perspectiva de expansão para outras síndromes relevantes para a saúde pública brasileira. O sistema busca padrões anômalos, associados à aumentos semanais inesperados nas series temporais de atendimentos da APS associados a síndromes respiratórias fora da sazonalidade típica. Esse tipo de análise permite identificar potenciais sinais de alerta precoce antes do registro oficial de surtos, o que pode ser decisivo para a implementação de medidas preventivas e mitigadoras [31].

A metodologia empregada combina diferentes ferramentas de análise epidemiológica, incluindo os modelos EARS [77], EVI (*Epidemic Volatility Index*) [180] e um novo modelo – MMAING [32] – desenvolvido pela equipe de modelagem do ÆSOP, para analisar esses dados, permitindo a detecção de anomalias que possam indicar o início de um surto. Além disso, são utilizadas abordagens estatísticas, matemáticas e computacionais para modelar cenários de risco correlacionado com variáveis climáticas, mobilidade urbana, e dados de medicamentos de venda livre.

Para além do monitoramento sindrômico, o ÆSOP visa também a modelagem da propagação de doenças a partir da identificação de anomalias geolocalizadas. O projeto ainda investe no aprimoramento e desenvolvimento de técnicas de metagenômica, capazes de identificar rapidamente os possíveis patógenos responsáveis por causar o aumento de casos, inclusive aqueles até então desconhecidos pela ciência.

O desenvolvimento do ÆSOP tem se dado de forma colaborativa,

envolvendo ativamente profissionais da saúde, pesquisadores, gestores públicos e representantes de órgãos governamentais em encontros técnicos e workshops. Essa construção participativa visa garantir que o sistema atenda às necessidades reais do território brasileiro, aumente a adesão institucional e fortaleça o sentimento de pertencimento e responsabilidade entre os usuários.

Resultados preliminares evidenciam a efetividade dos sinais de alerta precoce no âmbito do projeto ÆSOP. Em uma análise retrospectiva envolvendo mais de 589 milhões de atendimentos registrados na APS entre os anos de 2017 e 2020, foi possível antecipar picos de Síndrome Respiratória Aguda Grave (SRAG) com base em aumentos prévios de casos leves detectados nas unidades básicas de saúde [181]. Em outro estudo conduzido no estado da Bahia, a partir do monitoramento de dados sindrômicos da APS, identificou-se sinais de entrada do vírus SARS-CoV-2 em 18 das 21 RGIs analisadas, com pelo menos uma semana de antecedência em relação ao aumento dos casos oficialmente confirmados, evidenciando o potencial do ÆSOP como ferramenta promissora para vigilância epidemiológica proativa [182].

O ÆSOP representa uma inovação na vigilância epidemiológica, ao integrar múltiplas fontes de dados e utilizar tecnologias avançadas para a detecção precoce de surtos. Sua implementação tem o potencial de transformar as estratégias de resposta a emergências de saúde pública, fornecendo subsídios para a formulação de políticas mais eficazes e baseadas em evidências. Ao fortalecer a capacidade de resposta do sistema de saúde, o projeto contribui ainda para a redução dos impactos sociais associados a doenças infecciosas emergentes [31].

7
Materiais e Métodos

Neste capitulo abordamos a descrição dos dados e a construção dos modelos para indicar a emissão de sinais de alerta precoce (EWS do inglês, *Early Warning Signals*), considerando critérios de confiabilidade e avaliação.

7.1
Dados sindrômicos

O Sistema Nacional de Informações em Saúde Primária (SISAB) do Brasil é uma fonte valiosa de dados coletados das unidades de APS, composta por Unidades Básicas de Saúde (UBS) e as Unidades de Saúde da Família (USF). Esses dados fornecem *insights* sobre os padrões de saúde pública, cobrindo uma ampla gama de condições médicas, incluindo as Infecções de Vias Aéreas Superiores (IVAS).

Os dados provenientes do SISAB englobam registros de todos os atendimentos realizados pelo setor público em primeiro nível de atenção à saúde, distribuídos por municípios e codificados conforme a Classificação Internacional de Doenças (CID-10) e a Classificação Internacional de Atenção Primária (CIAP-2). Com uma cobertura que alcança pelo menos 75% da população, o sistema de atendimentos da APS no Brasil oferece uma base de dados extensa e representativa [181].

Este conjunto de dados é processado, tratado e disponibilizado pelo projeto ÆSOP. A base final é um compilado de dados sindrômicos, referentes aos atendimentos primários à saúde relacionados a infecções respiratórias, registrados o período de Janeiro de 2017 até Agosto de 2024.

Foram selecionados 50 códigos do CID-10 e CIAP-2, pelo grupo de clínica do projeto ÆSOP, que correspondem a condições potencialmente correlacionadas ao IVAS. Entre estes, destacam-se códigos diretamente relacionados, tais como J00 (Nasofaringite aguda, ou resfriado comum), J06 (Infecções agudas das vias aéreas superiores de múltiplas localizações e não especificadas), e R74 (Infecção aguda do aparelho respiratório superior), além dos demais códigos detalhados nas tabelas 7.1 e 7.2.

Por fim, para os trabalhos apresentados nesta tese, os dados municipais foram agregados em Regiões Geográficas Imediatas (RGIs), sendo que cada RGI é composta por um conjunto de municípios que têm como referências primárias uma rede urbana e um centro urbano local onde a população próxima

busca bens, serviços, e trabalho, conforme definido pelo Instituto Brasileiro de Geografia e Estatística (IBGE).

Tabela 7.1: Tabela CIAP-2: Códigos usados para definir condições provavelmente relacionadas a IVAS.

Tipo	**Código**	**Descrição**
CIAP-2	A03	Febre
CIAP-2	R01	Dor atribuída ao aparelho respiratório
CIAP-2	R02	Dificuldade respiratória / dispneia
CIAP-2	R03	Respiração ruidosa
CIAP-2	R04	Outros problemas respiratórios
CIAP-2	R05	Tosse
CIAP-2	R07	Espiro e congestão nasal
CIAP-2	R08	Outros sinais / sintomas nasais
CIAP-2	R21	Sinais / Sintoma da garganta
CIAP-2	R23	Sinais / Sintoma da voz
CIAP-2	R25	Expectoração / mucosidade anormal
CIAP-2	R29	Sintoma / queixa respiratória, outros
CIAP-2	R71	Tosse convulsa
CIAP-2	R74	Infecção aguda do aparelho respiratório superior
CIAP-2	R75	Sinusite crônica / aguda
CIAP-2	R76	Amigdalite aguda
CIAP-2	R77	Laringite / traqueíte aguda
CIAP-2	R78	Bronquite crônica
CIAP-2	R80	Gripe sem pneumonia
CIAP-2	R81	Pneumonia
CIAP-2	R83	Outra Infecção respiratória
CIAP-2	R99	Outras doenças respiratórias

Tabela 7.2: Tabela CID-10: Códigos usados para definir condições provavelmente relacionadas a IVAS.

Tipo	**Código**	**Descrição**
CID-10	J00	Nasofaringite aguda
CID-10	J01	Sinusite aguda
CID-10	J02	Faringite aguda
CID-10	J03	Amigdalite aguda
CID-10	J04	Laringite e traqueíte agudas
CID-10	J06	Infecções agudas das vias aéreas superiores de localizações múltiplas e não especificadas
CID-10	J09	Influenza devida a vírus Identificado da Gripe Aviária
CID-10	J10	Influenza devida a outro vírus da influenza identificado
CID-10	J11	Influenza devida a vírus não identificado
CID-10	J12	Pneumonia viral não classificada em outra parte
CID-10	J13	Pneumonia devida a Streptococcus pneumoniae
CID-10	J14	Pneumonia devida a Haemophilus influenzae
CID-10	J15	Pneumonia bacteriana não classificada em outra parte
CID-10	J16	Pneumonia devido a outros organismos infecciosos, não classificada em outra parte
CID-10	J17	Pneumonia em doenças classificadas em outra parte
CID-10	J18	Pneumonia por microorganismo não especificado
CID-10	J20	Bronquite aguda
CID-10	J21	Bronquiolite aguda
CID-10	J22	Infecção aguda não especificada das vias aéreas inferiores
CID-10	J80	Síndrome do desconforto respiratório do adulto
CID-10	R05	Tosse
CID-10	R06	Anormalidades da respiração
CID-10	R07	Dor de garganta e peito
CID-10	R43	Distúrbios do olfato e do paladar
CID-10	R50	Febre de outra origem desconhecida e de outras origens
CID-10	U07	Emergência de uso de U07
CID-10	B34	Doenças por vírus de localização não especificada
CID-10	B97	Vírus como causa de doenças classificadas em outros capítulos

7.2 Dados simulados

A geração de séries sintéticas é uma abordagem amplamente utilizada na avaliação de modelos estatísticos, sendo usada para simular dados que replicam o comportamento esperado de variáveis em situações onde grandes volumes de dados reais são escassos, inexistentes ou incompletos [183].

Primeiramente, descrevemos como os dados de base são simulados e, em seguida, como os surtos são gerados. As simulações foram projetadas para refletir a variabilidade temporal observada nos dados sindrômicos da APS, incorporando explicitamente características como volume, tendência e sazonalidade — um aspecto inovador desta abordagem.

As simulações são um recurso potencial para avaliações semelhantes e podem ser usadas por pesquisadores para testar outros algoritmos. No contexto do MMAING, as séries simuladas desempenham um papel crucial ao fornecer um ambiente controlado para testar a eficácia e a robustez do modelo em diferentes cenários epidemiológicos simulados.

A metodologia proposta para geração da série sintética é projetada para replicar e expandir as características estatísticas do conjunto de dados reais desse trabalho, os quais são registrados em intervalos semanais. O processo é elaborado assegurarando que os dados simulados reflitam não apenas as propriedades fundamentais dos dados originais mas também possibilitem a introdução de novas dinâmicas através dos ruídos. O procedimento é dividido em três etapas principais: i) a geração da serie sintética baseada na distribuição dos dados originais, ii) a introdução de ruído aleatório de forma a simular a presença de surtos, e iii) a formulação final da série sintética junto ao catalogo dos ruídos. Cada uma das etapas incorpora conceitos matemáticos específicos para assegurar a correlação aos dados originais e a utilidade dos dados gerados.

1) Geração da série sintética: Essa etapa envolve a suavização da série de dados originais, $x(t)$, para destacar características fundamentais, particularmente a tendência central. Utilizamos uma média móvel, $\overline{x}$, calculada com uma janela de 8 semanas para alcançar esse objetivo, conforme definido pela equação

$$\overline{x}(t) = \frac{1}{8} \sum_{j=t-3}^{t+4} x(j), \tag{7-1}$$

onde t varre cada semana na série temporal, e a soma é ajustada para manter o tamanho original da série, evitando distorções nas extremidades. Esse ajuste

ocorre nos primeiros e últimos pontos da série, onde a janela de 8 semanas não pode ser completamente aplicada devido à falta de dados anteriores ou posteriores. Para contornar essa limitação, o somatório é modificado para incluir apenas os valores disponíveis, reduzindo dinamicamente a janela nos extremos da série. Dessa forma, a média móvel é calculada em toda a extensão da série, garantindo que o número de observações seja preservado.

Com base na série $x(t)$, procede-se ao ajuste de uma distribuição normal para a geração de dados simulados. Esse ajuste é realizado por meio de dois fatores de escala, ϕ e φ, aplicados respectivamente à média móvel $\overline{x}(t)$ e ao desvio padrão $\sigma_{x(t)}$ da série original. Assim, a média ajustada $\mu'(t)$ e o desvio padrão ajustado $\sigma'(t)$ são definidos como:

$$\mu'(t) = \phi \cdot \overline{x}(t), \qquad \sigma'(t) = \varphi \cdot \sigma_{x(t)}.$$

Cada elemento da série sintética $y(t)$ é então obtido como uma realização da variável aleatória

$$Y(t) \sim \mathcal{N}\big(\mu'(t),\, \sigma'^2(t)\big)\,, \tag{7-2}$$

cuja densidade de probabilidade é dada por:

$$f(x) = \frac{1}{\sigma'(t)\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu'(t)}{\sigma'(t)}\right)^2}. \tag{7-3}$$

Os fatores ϕ e φ variam nos intervalos [0,8, 1,2] e [0,2, 0,8], respectivamente, definidos de forma a permitir ajustes sutis na tendência e na variabilidade da série sintética. O intervalo de ϕ contempla variações moderadas (até $\pm 20\%$) na média dos dados, possibilitando a simulação de séries com pequenas flutuações em relação à tendência observada. Por sua vez, φ permite gerar séries com menor dispersão que a original, ao limitar o desvio padrão a, no máximo, 80% do valor real. Essa escolha favorece a criação de séries com menor ruído, sem descaracterizar a estrutura estatística dos dados, o que é desejável em simulações controladas [184].

Esses ajustes geram pequenas alterações sobrepostas à série original, como ilustrado na Figura 7.1, pois não apenas replicam seus padrões centrais, mas também introduzem flutuações controladas. Essas alterações consistem em variações suaves na tendência e na dispersão dos dados, garantindo que a série sintética mantenha as principais características estatísticas da série real, ao mesmo tempo em que incorpora pequenas mudanças.

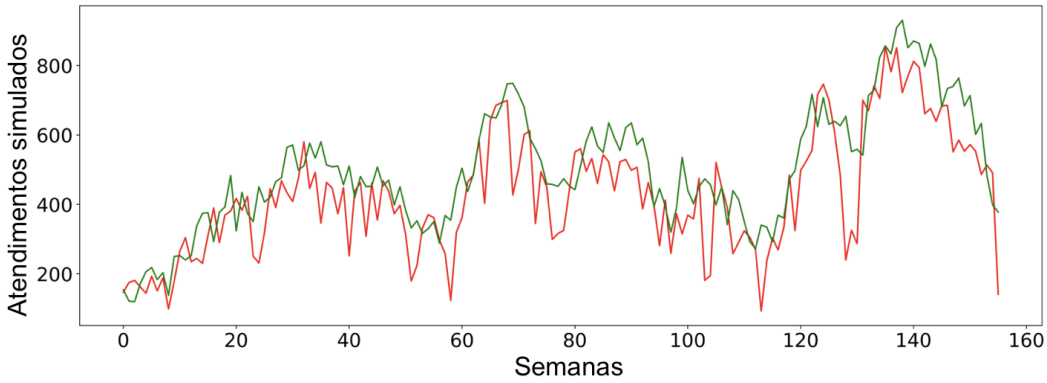


Figura 7.1: Ilustração das etapas para geração dos dados simulados a partir da RGI de Salvador-BA (290001) para o período de 2017 a 2019. Em vermelho, a série temporal original, $x(t)$, e em verde uma realização da série temporal sintética, $y_i(t)$.

2) Simulação de Ruído Aleatório: A introdução de ruído aleatório nas séries sintéticas desempenha um papel fundamental na simulação de eventos imprevisíveis, como surtos, aumentando assim o realismo e a complexidade dos cenários gerados. Esta etapa é projetada para adicionar variações que refletem a volatilidade encontrada em dados do mundo real, enriquecendo significativamente tanto a análise quanto a interpretação dos conjuntos de dados simulados.

O processo de incorporação efetiva do ruído à série sintética inicia-se pela escolha aleatória dos tempos de início t' dos surtos dentro do comprimento total da série l. Esses pontos de início, selecionados aleatoriamente, servem como marcos a partir dos quais o ruído será aplicado. Para cada tempo t' definido, estabelece-se uma duração $\Delta t'$ do intervalo de ruído, que varia aleatoriamente entre 4 e 10 semanas, de acordo com a variação observada na duração dos surtos respiratórios em dois estudos [185, 186].

A seleção dos tempos t' é feita de modo a garantir que o ruído adicionado não ultrapasse o limite final da série, sendo realizada segundo a seguinte distribuição uniforme:

$$t' \sim U(0, l - 10).$$

Conforme o estudo de Utsumi et al. [185], realizado em Instituições de Longa Permanência para Idosos, a mediana da duração dos surtos respiratórios variou entre 18 e 60 dias (∼2,5 a 8,5 semanas), com alguns casos extremos atingindo até 180 dias (∼26 semanas).

Já o estudo de Yan et al. [186] mostrou que a duração dos surtos de influenza simulados varia amplamente, dependendo dos parâmetros

epidemiológicos. A duração mínima observada foi de 43 dias (∼6 semanas) e a máxima de 85 dias (∼12 semanas), sendo influenciada pelo número inicial de casos, taxa de contato, taxa de recuperação e medidas de controle.

Dessa forma, o intervalo escolhido de 4 a 10 semanas reflete de maneira realista a distribuição observada, capturando surtos de curta e média duração, sem extrapolar para extremos menos frequentes.

Esta estratégia assegura que os eventos de surto introduzidos reflitam durações variáveis, mimetizando comportamentos dinâmicos próximos do real. É crucial notar que o ruído será aplicado no intervalo $[t', t' + \Delta t']$, implicando em uma alteração direta nessas semanas específicas e não uma alteração na série como um todo.

Cada intervalo identificado para receber ruído, demarcado por $t' + \Delta t'$, é então sujeito a uma perturbação modelada como uma oscilação. A intensidade dessa oscilação, ou amplitude A_i, juntamente com a sua frequência f_i, são determinadas de maneira aleatória.

A amplitude do ruído é escolhida a partir de um conjunto pré-definido $\mathbb{Z}_A$, definido como

$$\mathbb{Z}_A = \{A_i | A_i \in \mathbb{Z}, 50 \leqslant A_i < \frac{\Delta y}{2}, i = 1, 2, ..., 10\},$$

onde $\Delta y = \max(y) - \min(y)$ indica a amplitude total da série sintética, refletindo a diferença entre os valores máximo e mínimo. Cada valor A_i é um inteiro selecionado aleatoriamente dentro do intervalo $[50, \frac{\Delta y}{2})$, representando a intensidade do ruído a ser adicionado. Esse conjunto contém 10 valores distintos de A_i, permitindo a introdução de variações na intensidade do ruído aplicado aos diferentes segmentos da série sintética .

Adicionalmente, frequências f_i são extraídas aleatoriamente do vetor $\mathbf{f}$, seguindo uma distribuição uniforme $U(0, 1)$, conforme:

$$\mathbf{f} = [f_1, f_2, \ldots, f_8], \quad f_i \sim U(0, 1),$$

garantindo assim uma variedade de comportamentos oscilatórios. Uma vez estabelecidos as semanas de início t' e as durações $\Delta t'$ dos intervalos de ruído, aplicamos o ruído $\xi(t)$ a cada segmento identificado. O ruído é modelado como uma oscilação de amplitudes A_i e frequências f_i aleatórias, é introduzido na série temporal conforme a equação:

$$\xi(t) = A_i \cdot |\sin(2\pi f_i \cdot (t - t'))| + \frac{1}{10^4}, \quad \text{onde } t \in [t', t' + \Delta t'], \tag{7-4}$$

na qual o termo $\frac{1}{10^4}$ é adicionado para assegurar uma perturbação mínima

inicial, simbolizando o início gradual do evento de surto.

3) Composição Final da Série Sintética e Catálogo: Nesta etapa do processo de geração da série temporal sintética, geramos 200 realizações distintas de $y(t)$, seguindo o procedimento demonstrado anteriormente na etapa 1. A partir dessas múltiplas realizações, buscamos consolidar uma representação sintética abrangente e precisa, em uma série única através da média, produzindo a série sintética definitiva $\bar{y}(t)$:

$$\bar{y}(t) = \frac{1}{200} \sum_{i=1}^{200} y_i(t), \tag{7-5}$$

onde $y_i(t)$ indica cada realização individual da série. A figura 7.2 ilustra essa etapa.

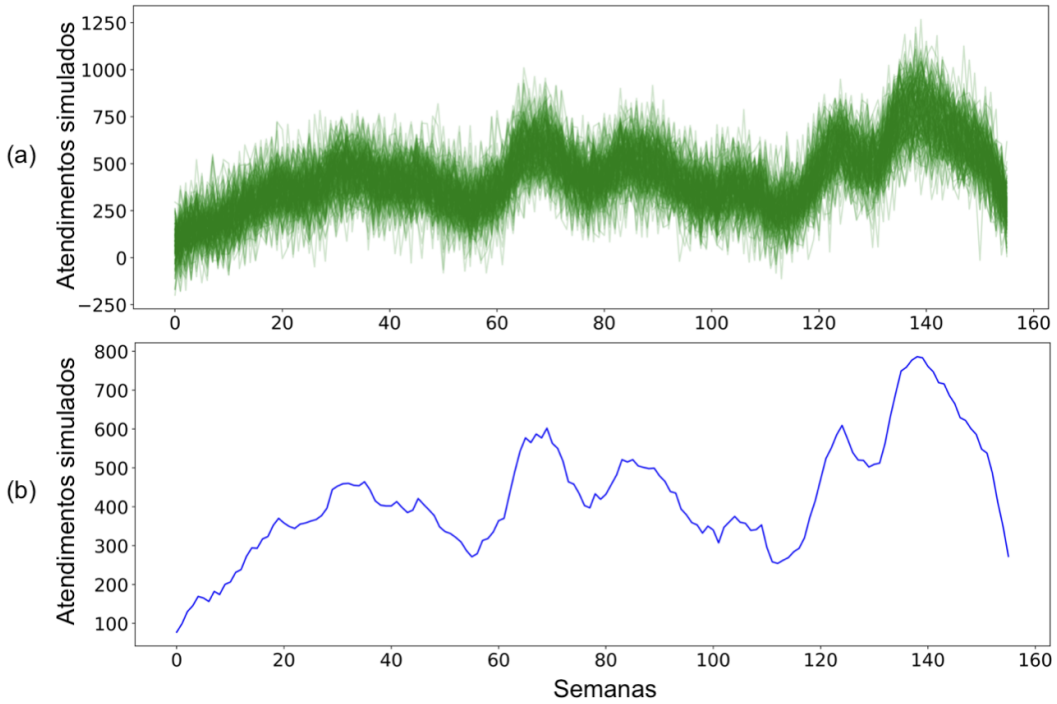


Figura 7.2: Ilustração das etapas para a geração dos dados simulados a partir da RGI de Salvador-BA (290001), no período de 2017 a 2019. a) Conjunto de 200 realizações da série sintética $y_i(t)$, geradas a partir de uma distribuição normal com média $\mu'(t) = \phi \cdot \overline{x}(t)$ e desvio padrão $\sigma'(t) = \varphi \cdot \sigma_{x(t)}$, onde $\phi \in [0{,}8,\ 1{,}2]$ e $\varphi \in [0{,}2,\ 0{,}8]$; b) Série final sem ruído, $\bar{y}(t)$, obtida pela média das realizações simuladas.

Assim, propomos uma representação sintética coesa que suaviza as flutuações individuais, revelando tendências centrais significativas, como por exemplo sazonalidade.

A seguir, adicionamos n ruídos à série $\bar{y}(t)$, para simular eventos variáveis de surto. Conversas com especialistas sugerem a ocorrência de aproximadamente dois surtos respiratórios por ano, o que equivale a aproximadamente 4% do total de semanas em um ano epidemiológico de 52 semanas. Assim, a escolha de 4% do tamanho total l da série reflete essa estimativa empírica e garante que a simulação capture a recorrência esperada desses eventos. A quantidade total m de ruídos possíveis a serem inseridos é proporcional a esse valor, e pode ser descrito pelo conjunto

$$\Xi(t) = \{\xi_i(t) \mid i = 1, 2, \ldots, m\}, \quad m = \lfloor 0.04 \times l \rfloor,$$

onde cada evento de ruído $\xi_i(t)$ é simulado conforme descrito anteriormente na etapa 2, ajustado às características de $\bar{y}(t)$ e definido por amplitudes e frequências variáveis. Esses eventos são aplicados de maneira aleatória, mas sistemática, a semanas selecionadas aleatoriamente da série $\bar{y}(t)$. Embora o conjunto $\Xi(t)$ contenha um máximo de m possíveis eventos de ruído, o número efetivo de eventos n que são inseridos varia aleatoriamente de 1 até m, permitindo a representação de uma frequência variável de surtos na série. Finalmente, a série sintética final, $y'(t)$, é composta segundo a equação:

$$y'(t) = \bar{y}(t) + \sum_{i=1}^{n} \xi_i(t) \tag{7-6}$$

Todo o processo é ilustrado detalhadamente nas quatro subfiguras que compõem a Figura 7.3. Cada subfigura evidencia uma etapa da metodologia, desde a geração inicial das instâncias (b) até a aplicação dos ruídos (d), enriquecendo visualmente a compreensão do método.

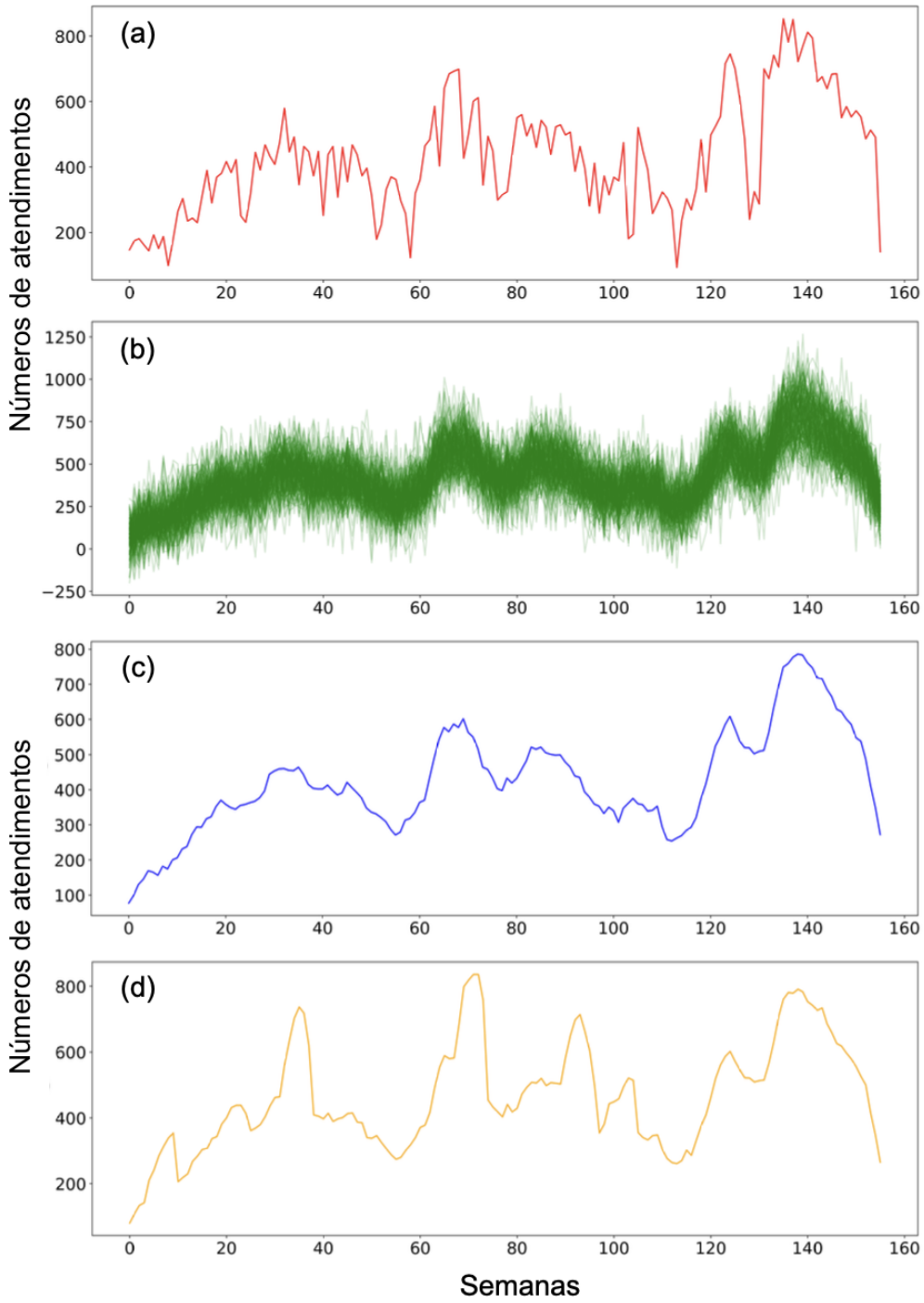


Figura 7.3: Ilustração do processo para geração dos dados simulados a partir da RGI de Salvador - BA (290001). (a) Série temporal original, $x(t)$; (b) Realizações da série temporal sintética $y_i(t)$; (c) Série temporal sintética definitiva sem ruído, $\bar{y}(t)$; (d) Série temporal sintética final com ruído, $y'(t)$.

Além da composição final da série sintética $y'(t)$, elaboramos um catálogo sistemático, que começa pela identificação de sinais anormais presentes na série sintética definitiva $\bar{y}(t)$ antes da adição do ruído. Tais sinais refletem

tendências ou padrões oriundos da série original $x(t)$, importantes para a análise subjacente. Posteriormente, são registrados os intervalos afetados pelos ruídos introduzidos na serie $y'(t)$, marcando as semanas iniciais e finais de cada evento de surto. Esses registros detalhados, são essenciais para as análises, oferecendo uma base para testar e refinar modelos de detecção de anomalias e transições críticas.

Importante ressaltar que a identificação dos sinais anormais na série $\bar{y}(t)$ é conduzida com base no modelo específico em análise. Ao aplicarmos este modelo à série final $y'(t)$, as anormalidades previamente reconhecidas na série $\bar{y}(t)$ são descontadas, isolando efetivamente os ruídos intencionalmente inseridos. Essa estratégia assegura que a análise se concentre nas variações induzidas, permitindo uma avaliação precisa da capacidade do modelo em detectar e interpretar os ruídos simulados.

Essa abordagem é semelhante à abordagem adotada por Neill [187], e ilustra uma nova metodologia para a modelagem de dados temporais simulados. Não só permite a simulação controlada e realista de surtos dentro das séries sintéticas, mas também captura as propriedades essenciais de tendências dos dados originais. Além disso, garantimos que os dados gerados sejam representativos e adequados tanto para avaliar a eficácia de modelos existentes quanto para incentivar o desenvolvimento de novos métodos analíticos.

Por fim, simulamos a partir dos dados do período de 2017 a 2019 de cada RGI, 30 simulações, resultando em $27 \times 30 = 810$ séries temporais simuladas. Isso é feito para explorar uma ampla gama de variações e cenários potenciais refletindo diferentes condições.

7.3 Metodologia de ensemble para detecção de EWS

A análise dos dados sindrômicos tem como objetivo detectar precocemente possíveis emergências de doenças respiratórias, monitorando as tendências de IVAS em nível nacional. Para isso, os dados municipais foram agregados por RGIs, totalizando 510 séries temporais, uma por RGI, conforme definido pelo IBGE.

Dessas, foram analisadas as 27 séries relativas às capitais estaduais (ver Tabela 7.3), abrangendo 41% da população brasileira. Esse recorte oferece uma visão ampla da evolução das síndromes respiratórias no país entre 2017 e 2024, ao longo das semanas epidemiológicas.

Tabela 7.3: Informações individuais das Regiões Geográficas Imediatas

Código	Nome	Estado	Região	Nº de municípios	População
110001	Porto Velho	Rondônia	Norte	05	666953
120001	Rio Branco	Acre	Norte	07	520759
130001	Manaus	Amazonas	Norte	10	2604830
140001	Boa Vista	Roraima	Norte	05	502280
150001	Belém	Pará	Norte	15	2773101
160001	Macapá	Amapá	Norte	04	675363
170001	Palmas	Tocantins	Norte	10	347905
210001	São Luís	Maranhão	Nordeste	13	1656503
220001	Teresina	Piauí	Nordeste	16	1116034
230001	Fortaleza	Ceará	Nordeste	20	4179211
240001	Natal	Rio G. do Norte	Nordeste	24	1734032
250001	João Pessoa	Paraíba	Nordeste	22	1429674
260001	Recife	Pernambuco	Nordeste	16	4107574
270001	Maceió	Alagoas	Nordeste	13	1315995
280001	Aracaju	Sergipe	Nordeste	20	1233495
290001	Salvador	Bahia	Nordeste	16	4064880
310001	Belo Horizonte	Minas Gerais	Sudeste	29	5347854
320001	Vitória	Espírito Santo	Sudeste	10	2100410
330001	Rio de Janeiro	Rio de Janeiro	Sudeste	21	12901184
350001	São Paulo	São Paulo	Sudeste	39	22048504
410001	Curitiba	Paraná	Sul	29	3731769
420001	Florianópolis	Santa Catarina	Sul	17	1180585
430001	Porto Alegre	Rio G. do Sul	Sul	23	3267292
500001	Campo Grande	Mato G. do Sul	Centro-Oeste	13	1131332
510001	Cuiabá	Mato Grosso	Centro-Oeste	14	1104736
520001	Goiânia	Goiás	Centro-Oeste	19	2627637
530001	Distrito Federal	Distrito Federal	Centro-Oeste	01	3094325

Como foi comentado anteriormente, combinamos quatro TDOs não supervisionadas, que são baseados em suposições gerais e treinamento de algoritmos usando grandes volumes de dados, com o número de reprodução dependente do tempo $R(t)$, o que é justificado para levar em conta a existência de uma causa dinâmica para a formação e propagação de surtos. Desta forma demos origem ao modelo final baseado em uma metodologia

de ensemble, denominado Modelo Misto de Inteligência Artificial e Próxima Geração (MMAING).

Procedemos com a avaliação e validação do MMAING, utilizando as seguintes estratégias: i) utilização de dados sindrômicos da APS para realizar uma análise empírica, comparando os resultados de detecção obtidos durante a pandemia de COVID-19 (2020-2022) com os relatórios oficiais do governo brasileiro [188] que detalham a dinâmica de uma pandemia real, ii) comparação de similaridade e desempenho com o algoritmo EARS, utilizando, respectivamente, dados APS e dados simulados, e iii) verificação da relação entre os EWS indicados pelo MMAING com eventuais surtos ocorridos no estado do Amazonas, no segundo quadrimestre de 2024, notificado pelo do Centro de Informações Estratégicas em Vigilância em Saúde (CIEVS) do estado. Nas análises de desempenho foram utilizadas seis métricas estatísticas:

- Métrica 1. Probabilidade de detecção (POD): Se um alarme é gerado pelo menos uma vez entre o início e o fim de um surto, o surto é considerado detectado [166].
- Métrica 2. Sensibilidade (Se): definida como $Se = \frac{VP}{VP+FN}$;
- Métrica 3. Especificidade (Sp): definida como $Sp = \frac{VN}{VN+FP}$;
- Métrica 4. Valor Preditivo Positivo (PPV): definido como $PPV = \frac{VP}{VP+FP}$;
- Métrica 5. F1 definida como a média harmônica da sensibilidade (Se) e do PPV: $F1 = 2 \times \frac{Se \times PPV}{Se+PPV}$;
- Métrica 6. Curva de Característica de Operação do Receptor (ROC): definida como a relação entre a taxa de verdadeiros positivos (Se) e a taxa de falsos positivos (1 - Sp).

onde, semanas com surto e EWS foram consideradas Verdadeiro Positivo (VP); sem surto e sem EWS, Verdadeiro Negativo (VN); com surto, mas sem EWS, Falso Negativo (FN); e sem surto, mas com EWS, Falso Positivo (FP).

7.4 Desenvolvimento do MMAING

Tendo em vista as justificativas para a necessidade de contar com um algoritmos de detecção confiáveis já discutidas em seções anteriores, o MMAING foi proposto e desenvolvido como um modelo de última geração, que facilita a manipulação e implementação em vigilância epidemiológica, oferecendo suporte na tomada de decisões pelas autoridades responsáveis.

7.4.1 Justificativa da proposta

O MMAING [32] se origina da combinação de quatro modelos de aprendizado de máquina — ISF, LOF, OCSVM e COPOD — e de um modelo baseado no tempo de infecção, formulado por meio do método da próxima geração (NGM), que permite estimar o número de reprodução dependente do tempo, $R(t)$.

Na implementação de técnicas de inteligência artificial, a abordagem de ensemble é amplamente utilizada para combinar os resultados e predições de diferentes modelos, com o objetivo de aumentar a precisão, a robustez e a generalização das soluções, além de reduzir a variância e o viés característicos de modelos individuais [189, 190]. Estratégias como *bagging*, *boosting* e esquemas de votação [190] têm sido eficazes em diversos domínios por promoverem consenso entre múltiplas fontes de decisão. Esse tipo de integração tem se mostrado particularmente promissor em diversas aplicações sensíveis, como, por exemplo, na vigilância epidemiológica, ao favorecer tomadas de decisão mais confiáveis e eficazes [190, 191].

A inovação do MMAING reside na combinação, ainda pouco explorada, entre modelos de aprendizado de máquina e um modelo determinístico, integrada a um método de ensemble que seleciona sinais dos modelos base por meio de um sistema de votação, *hard voting* (votação majoritária) [191, 192]. Essa abordagem tem o potencial de facilitar a interpretação dos resultados por meio de um processo de decisão integrativo, além de contribuir para a robustez das detecções ao promover o consenso entre diferentes métodos utilizados.

Para assegurar a confiabilidade das detecções identificadas pelo MMAING, foi adotado um mecanismo de corroboração fundamentado no *princípio da robustez*, conforme descrito por Huppert e Katriel [193]. Esse processo utiliza um esquema de votação majoritária [191], no qual cada modelo contribui com peso igual na decisão final. A detecção considerada corresponde àquela que obtém o maior número de votos entre os modelos selecionados. Especificamente, emprega-se uma estratégia deliberada de seleção de um número ímpar de modelos, neste caso, três entre os cinco modelos base, com o objetivo de maximizar a eficácia e a objetividade do sistema de votação [194].

7.4.2 Modelos baseados em técnicas de ML

As TDOs não supervisionadas, aplicadas à vasta quantidade e diversidade de dados, permitem a identificação de padrões ocultos sem a necessidade de rotulação prévia por especialistas, o que é adequado para o nosso conjunto de

dados da APS. No entanto, a detecção de anomalias pontuais, comum em tais técnicas não supervisionadas, revela-se limitada para a vigilância sindrômica, pois foca em outliers individuais, que raramente indicam um surto [195]. Nossa abordagem, ao contrário, busca padrões coletivos de anomalias, que são mais representativos de um surto em potencial, especialmente quando há um aumento significativo nos atendimentos de saúde em um curto período, delimitado por limite superior aceitável.

No processo de treino e teste, os dados foram divididos em dois subconjuntos baseando-se no período temporal. Para o treinamento dos modelos usamos dados de 2017 a 2019, pressupondo um período de normalidade, devido à estabilidade observada no número de atendimentos ao longo das semanas epidemiológicas para garantir sensibilidade a alterações anormais. Na etapa de teste, usamos os dados de 2020 a 2023 que compreende o período da Covid19 (2020 - 2022) e a normalização no número de atendimentos ao longo das semanas epidemiológicas, bem como a melhoria da coleta de dados (2023). Vale ressaltar que o processo de validação e definição dos hiperparâmetros não é feito de forma individual, mas sim, quando incorporados ao MMAING, como será discutido adiante.

7.4.3 Modelo baseado no NGM

Conforme mencionado na seção 2.2, o parâmetro $R(t)$ é uma medida essencial no monitoramento de epidemias e controle do surgimento de surtos, podendo ser estimado tanto a partir de dados dos casos de infecção confirmados como com a utilização dos modelos compartimentais. Este parâmetro estima o número médio de casos secundários que um caso infectado pode produzir em uma população. O valor $R = 1$ é um limiar de transição crítico; valores acima de 1 indicam um aumento no número de casos (epidemia em expansão), enquanto valores abaixo de 1 indicam uma diminuição (epidemia em declínio).

Entretanto, os dados provenientes da APS, especialmente os dados sindrômicos referentes à IVAS, possuem características diferentes dos dados habitualmente usados. Esses dados não incluem casos confirmados, mas dados sintomáticos gerais de IVAS que podem ou não testar positivo para uma doença específica. Assim, é necessário adaptar meticulosamente os limiares de $R(t)$ que definem o risco de crescimento ou redução do surto. Esta adaptação garante a aplicação precisa do modelo em contextos epidemiológicos que envolvem dados menos específicos e mais variados em termos de diagnóstico clínico.

Portanto, para a contribuição no desenvolvimento do MMAING devido ao modelo NGM, estimamos o parâmetro $\hat{R}(t)$, que é o análogo da grandeza

$R(t)$, utilizando-se do conjunto dados de atendimento APS, $H(t)$, e uma distribuição de intervalo de geração $g(\tau)$. Uma vez que os registros de atendimentos são feitos a cada semana, optamos por adotar a formulação de tempo discreto fazendo o argumento $t \to i$, que leva à discretização da Eq. 2-15, que pode ser escrita como

$$\hat{R}_i = \frac{H_i}{\sum_{j=1}^{n} g_j H_{i-j}}. \tag{7-7}$$

Como já foi demonstrado, a partir da metodologia desenvolvida em [50], podemos obter $g(\tau)$ a partir de um modelo compartimental adequado. No nosso problema, foi escolhido o modelo SEIR, para o qual a distribuição contínua $g(\tau)$ pode ser expressa por:

$$g(\tau) = \frac{\varepsilon e^{-\kappa\tau} + \frac{\kappa}{\gamma-\kappa}\left[e^{-\kappa\tau} - e^{-\gamma\tau}\right]}{\varepsilon + \frac{1}{\kappa} + \frac{1}{\gamma}}.$$

onde os parâmetros ε, κ e γ representam, respectivamente, capacidade do indivíduo exposto em transmitir o patógeno para um indivíduo suscetível, frequência com a na qual os indivíduos saem do compartimento dos expostos e a taxa de recuperação.

Assumimos a priori que a dinâmica da doença a ser detectada precocemente é desconhecida, tomamos então $\kappa = \gamma$ e $\varepsilon \to 0$. Neste caso, a função $g(\tau)$ se reduz a

$$g(\tau) = \gamma^2 \tau e^{-\gamma\tau}. \tag{7-8}$$

Para ser utilizada de forma conveniente, esta forma da distribuição $g(\tau)$ precisa ser convertida para a versão correspondente válida para tempo discreto e em seguida ser normalizada [42]. Para isso faremos a discretização dessa distribuição tomando inicialmente uma progressão geométrica:

$$P(n) = a_1 q^{(n-1)}, \tag{7-9}$$

onde, $q = e^{-\gamma}$ na qual $t = 0$ corresponde $n = 1$.

A partir disso e levando-se em conta que a equação 7-8 pode ser obtida da derivada da função $e^{-\gamma\tau}$ com relação a $e^{-\gamma}$, a expressão para $g(\tau)$ é obtida considerando

$$g(n) = q\frac{\partial}{\partial q}P(n) = a_1(n-1)q^{n-1}, \tag{7-10}$$

e usando o fato que a soma $S(n)$ dos n primeiros termos de $P(n)$ é dada por:

$$S(n) = q\frac{\partial}{\partial q}\sum_{n=1}^{\infty} P(n). \tag{7-11}$$

Em consequência, temos que a soma $S_e(m)$ dos m primeiros termos da função $g(n)$ é dada por:

$$S_e(m) = a_1\ q\frac{\partial}{\partial q}\frac{(1-q^m)}{(1-q)^2}. \tag{7-12}$$

Finalmente, para um intervalo finito m o fator de normalização a_1 é dado por:

$$a_1 = \frac{(1-q)^2}{q\left[1+(m-1)q^m - mq^{m-1}\right]}. \tag{7-13}$$

Assim, conseguimos definir $\hat{R}_i$ com base na PG a partir da equação 7-7 como:

$$\hat{R}_{(i-1)} = \frac{H_{i-1}}{\sum_{j=1}^{\min(i,m)} a_1(j-1)q^{j-1}H_{i-j}} \quad ; \quad q = e^{-\gamma} \tag{7-14}$$

A interpretação do modelo NGM proposto tem por base estabelecer um valor crítico $\bar{R}$ para $\hat{R}(t)$, que desempenhe papel de indicador de sinais de alerte precoce, que também é delimitado por limite superior aceitável.

Estudos recentes têm ressaltado a importância de estabelecer limiares claros de transmissividade de doenças infecciosas baseadas em valores $R(t)$. Por exemplo, um estudo de 2020 utilizou a mediana do $R(t)$ para identificar mudanças nas tendências de transmissibilidade do vírus Ébola, determinando valores de referência para $R(t)$ que prediziam pontos de transições críticas com uma a duas semanas de antecedência em relação às datas dos eventos reais [196]. No ano seguinte, o mesmo grupo de pesquisa aplicou limites a $R(t)$ para construir séries sintéticas de casos de Covid-19, explorando diferentes cenários e limiares [78].

Para o nosso estudo, utilizamos um limiar para $\hat{R}_i > \bar{R} = 1,25$ como indicador no aumento da tendência da série temporal de atendimentos IVAS. Esta escolha se justifica não apenas porque os dados não são de casos confirmados, mas também devido a uma análise que estima o valor médio de $\langle\hat{R}_i\rangle$ em vários contextos no período de 2017 a 2023. Ao analisar os dados em nível nacional, o $\langle\hat{R}_i\rangle$ ficou próximo do valor estabelecido, em torno de 1,24. No entanto, ao dividir os dados em séries temporais para cada estado brasileiro (27 UFs), os valores médios de $\langle\hat{R}_i\rangle$ ficaram na faixa de 1,2 a 1,3,

com o Distrito Federal sendo uma exceção com $\langle \hat{R}_i \rangle$ de 1,5. Subdividindo os dados por Regiões Imediatas (510 RGIs), notou-se uma maior variação nos valores médios de $\langle \hat{R}_i \rangle$, oscilando entre 1,2 e 2,0. Esta variação é significativa, entretanto a adoção de valores muito elevados de $\hat{R}_i$ como limiar pode resultar em alertas emitidos tardiamente em algumas RGIs.

Destacamos que diversos limiares $\bar{R}$ foram testados com o modelo integrado ao MMAING, alcançando a melhor performance com o valor de $\bar{R} = 1,25$. Portanto, foi essencial definir um limiar para $\hat{R}_i$ que refletisse adequadamente a realidade dos dados e transcendesse os parâmetros tradicionais, assegurando a detecção de sinais de alerta precoce com tendências de aumento nos atendimentos por IVAS.

7.4.4 Limite superior para análise dos atendimentos da APS

Conforme mencionado anteriormente, os EWS gerados pelos modelos para uma determinada semana t estão condicionados ao fato de que o número de atendimentos na APS nessa semana ultrapasse o limite superior (ℓ_s), definido como:

$$\ell_s(t) = \bar{x}(t) + z_\alpha \frac{\sigma}{\sqrt{n}} \tag{7-15}$$

Aqui, $\bar{x}$ denota a média amostral para uma determinada semana t, z_α é o valor crítico para um determinado nível de confiança α, σ é desvio padrão da amostra e n representa o tamanho da amostra para a semana t. Este limite permite determinar se os dados da semana desviam da média esperada de atendimentos, categorizando-os como um potencial EWS a ser emitido. O segundo termo, correspondente à multiplicação de z_α pelo erro padrão, ajusta o grau de confiança desejado para detectar desvios, enquanto a divisão por $\sqrt{n}$ reflete a incerteza associada à estimativa da média: quanto maior a amostra, menor a variabilidade esperada e, portanto, mais preciso o limite definido. Nesse trabalho, a escolha da amostra seguiu duas abordagens:

i) Passado recente: Utilização de uma "janela móvel" com tamanho de 5 semanas para refletir os dados de um passado recente, adaptando-se dinamicamente às flutuações no curto prazo. Essa abordagem se mostra útil quando há a necessidade de monitorar variações recentes no conjunto de dados [197].

ii) Passado histórico: O uso de semanas específicas ou correspondentes ao período de 2017 a 2019, com o objetivo de identificar padrões sazonais ou tendências cíclicas. Esses dados podem ser impactados por variáveis externas como alterações climáticas, feriados e outros eventos periódicos.

Portanto, ao identificar uma anormalidade na série temporal, um EWS só será emitido se o número de atendimentos na semana t satisfizer a condição $x(t) > \ell_s(t)$.

7.4.5 Pseudocódigo do algoritmo MMAING

A seguir, apresentamos a lógica computacional que rege o funcionamento do MMAING, estruturada sob a forma de pseudocódigo. Este algoritmo condensa as etapas previamente descritas nas subseções anteriores, abrangendo desde o pré-processamento dos dados até a emissão de sinais de alerta precoce.

O Algoritmo 1 apresenta a lógica principal do MMAING de forma estruturada, destacando os procedimentos centrais do processamento. Por sua vez, o Algoritmo 2 detalha as funções auxiliares que compõem a arquitetura do modelo.

Algorithm 1 Rotina principal

function MMAING(dados D)
    $A \leftarrow \varnothing$
    $D_f \leftarrow$ AgruparDados(D)
    $S \leftarrow$ subconjunto de D_f com RGIs específicas
    for all série $s \in S$ **do**
        $L_h \leftarrow$ LimiteHistorico(s)
        $L_r \leftarrow$ LimiteRecente(s)
        $R_t \leftarrow$ Rt(s)
        $a_5 \leftarrow$ **if** $R_t > 1.2$ **then** -1 **else** 1
        $[a_1, a_2, a_3, a_4] \leftarrow$ DetectarAnomaliasML(s)
        $A_m \leftarrow \{a_1, a_2, a_3, a_4, a_5\}$
        $A_e \leftarrow$ Votacao(A_m)
        if $s_t > \max(L_h[t], L_r[t])$ **and** $A_e = -1$ **then**
            $A \leftarrow A \cup \{t\}$
        end if
    end for
    return A
end function

Algorithm 2 Rotinas auxiliares

function AgruparDados(D)
    Agrupar D por `cod_rgi`, `ano` e `epiweek`
    Calcular a soma de `atend_ivas` para cada grupo
    return D_f
end function

function LimiteHistorico(s)
    Para cada semana t nos respectivos anos de 2017–2019:
        Calcular média μ_t e desvio padrão σ_t das observações nessa semana
        Calcular $L_h[t] \leftarrow \mu_t + z_{0.05} \cdot \frac{\sigma_t}{\sqrt{n}}$
    return vetor L_h
end function

function LimiteRecente(s)
    $l \leftarrow$ tamanho da janela de suavização (e.g., $l = 5$)
    Para cada semana t, calcular média móvel μ_t e desvio padrão σ_t
    Calcular $L_r[t] \leftarrow \mu_t + z_{0.05} \cdot \frac{\sigma_t}{\sqrt{l}}$
    return vetor L_r
end function

function CalcularRt(s)
    Para cada semana i, calcular:
        $q \leftarrow e^{-\gamma}$
        $m \leftarrow$ número máximo de períodos passados
        $a_1 \leftarrow$ constante de normalização
        $\hat{R}_{(i-1)} \leftarrow \dfrac{A_{i-1}}{\sum_{j=1}^{\min(i,m)} a_1(j-1)q^{j-1}A_{i-j}}$
    return vetor $\hat{R}_t$
end function

function DetectarAnomaliasML(s)
    Treinar IFS, LOF, OCSVM e COPOD
    Obter predições a_1 a a_4 com s
    return $[a_1, a_2, a_3, a_4]$
end function

function Votacao(A_m)
    $n \leftarrow$ número de valores iguais a -1 em A_m
    if $n \geqslant 3$ **then**
        return -1
    else
        return 1
    end if
end function

A rotina principal recebe como entrada os dados D, que consistem em séries temporais de atendimentos sindrômicos registrados na APS. Após a etapa de pré-processamento, realizada pela função `AgruparDados`, os dados são organizados em D_f, um conjunto de séries temporais agrupadas por RGI, ano, semana epidemiológica e total de atendimentos por IVAS. A iteração ocorre sobre uma lista específica de séries $s \in S$, extraídas de D_f, nas quais as variáveis de interesse são processadas semana a semana. Para cada série temporal s, são computados dois limites superiores para cada semana t, por meio das funções `LimiteHistorico`, baseada nas semanas t correspondentes aos anos de 2017 a 2019, e `LimiteRecente`, que utiliza uma média móvel suavizada centrada na semana epidemiológica atual. Em seguida, estima-se a série temporal do número de reprodução variável no tempo $\hat{R}(t)$, por meio da função `CalcularRt`. A cada semana t é associada uma classificação binária a_5, assumindo valor -1 (anomalia) quando $\hat{R}(t) > 1{,}25$ e 1 caso contrário. Simultaneamente, quatro modelos de aprendizado de máquina — IFS, LOF, OCSVM e COPOD — geram predições a_1 a a_4. As cinco saídas (a_1 a a_5) compõem o vetor A_m, que é avaliado por meio de uma votação majoritária implementada na função `Votacao`, resultando na decisão A_e. Caso o valor da série na semana t, denotado por s_t, ultrapasse o maior entre os limites $L_h[t]$ e $L_r[t]$, e a votação indique anomalia ($A_e = -1$), é registrado um sinal de alerta correspondente à semana t no conjunto A. Ao final da iteração, A contém todas as semanas com sinais de alerta precoce identificados pelo MMAING.

7.4.6 Configurações do MMAING

Para avaliar a adaptabilidade e versatilidade do MMAING em cenários que possam exigir rigor moderado ou alto (maior ou menor número de falsos positivos de EWS), adotamos duas distintas configurações dos parâmetros correspondentes a cada um dos algoritmos utilizados. Elas são, denominadas *equilibrada* (EqCf) e *rigorosa* (RiCf), e os valores dos respectivos parâmetros são indicados na Tabela 7.4. As configurações EqCf tendem a diminuir a precisão e aumentar a sensibilidade, resultando em mais EWS e, consequentemente, aumentando o número de falsos positivos. Por outro lado, a RiCf busca aumentar a precisão e reduzir a taxa de falsos positivos, o que pode resultar na falha em emitir alguns EWS que poderiam alertar para o surgimento de surtos efetivamente detectados.

A EqCf usa parâmetros que buscam um compromisso entre sensibilidade, precisão e especificidade. Por exemplo, para os métodos ISF e LOF, o número de estimadores (n_{est}) e vizinhos (n_{nei}) é definido como 500, enquanto a

Tabela 7.4: Configurações de parametrização do MMAING.

Método	Parâmetro	EqCf	RiCf
ISF	n_{est}	500	400
	$\mathcal{C}$	0.4	0.3
LOF	n_{nei}	500	300
	$\mathcal{C}$	0.4	0.3
OCSVM	ν	0.8	0.5
	kernel	RBF	RBF
	γ	0.001	0.001
COPOD	$\mathcal{C}$	0.4	0.3
NGM	$\hat{R}$	1.25	1.30
	γ	0.2	0.2

contaminação $\mathcal{C}$ é estabelecida em 0,4, indicando que esperamos cerca de 40% de pontos anômalos. A configuração rigorosa RiCf implica uma redução no número de árvores ISF e no número de vizinhos LOF para 400 e 300, respectivamente, bem como uma redução na contaminação para 0,3.

O método OCSVM utiliza um valor ν mais alto em EqCf (0,8) comparado com RiCf (0,5), indicando maior flexibilidade na separação de classes. O uso do kernel de Função de Base Radial (RBF) e o valor de γ são mantidos consistentes entre as duas configurações. Como este é o kernel que melhor apresenta resultados na literatura [87], isto sugere que a forma da fronteira de decisão e a complexidade do modelo são consideradas adequadas em ambos os casos. O COPOD e NGM também apresentam diferentes valores de parâmetros entre as duas configurações: no COPOD é mantida uma consistência com ISF e LOF em relação à contaminação $\mathcal{C}$ em ambas as configurações, enquanto que o NGM ajusta o limiar $\hat{R}$ para refletir a rigidez desejada de detecção, assumindo um menor valor na configuração EqCf, e mantém a mesma taxa de recuperação γ para ambas as configurações.

Em contextos de detecção de surtos relacionados a outros tipos de infecção ou em cenários mais graves, essas configurações podem ser adaptadas com objetivos específicos para otimizar a emissão de EWS. Por exemplo, enquanto a medida de sensibilidade é importante para emissão de EWS de forma clara e consistente em cenários onde surtos ocorrem com bastante frequência, a PPV, que mede a probabilidade de um EWS ser um verdadeiro positivo, torna-se especialmente importante quando os surtos são raros ou não ocorrem frequentemente. Para cada uma dessas medidas, a estratégia escolhida pode levar a se definir configurações para os algoritmos e/ou priorizar quais medidas são mais importantes para as necessidades de vigilância [166, 168].

Utilizando as configurações EqCf e RiCf, o MMAING foi testado e analisado em séries simuladas. No capítulo 8, especificamente na seção 8.4, são apresentados os resultados de desempenho para ambas as configurações. Nas demais seções do capítulo, foi adotada exclusivamente a configuração EqCf.

8
MMAING - Resultados e Discussão

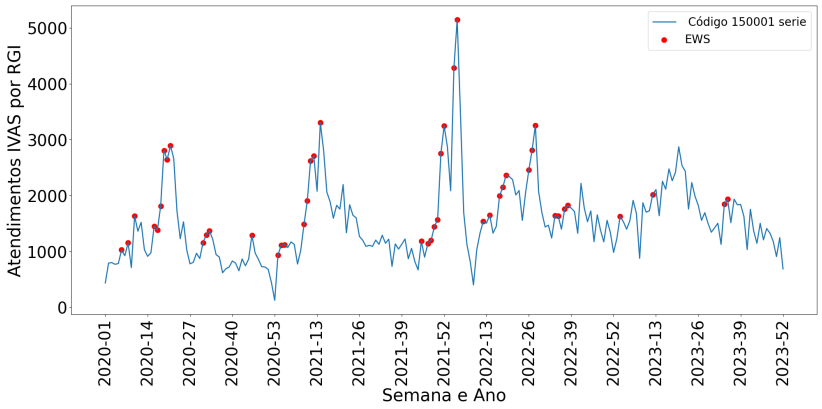
Nesta seção, apresentamos os principais resultados obtidos através da aplicação do MMAING para a detecção de potenciais surtos de IVAS a partir de dados sindrômicos da APS, durante o período de 2020 a 2024 no Brasil.

Para avaliar e validar os resultados obtidos, utilizamos cinco estratégias complementares: (i) aplicação em dados sintéticos rotulados manualmente derivados de dados reais; (ii) aplicação em dados sindrômicos de 27 RGIs correspondentes às capitais brasileiras; (iii) análise empírica do MMAING utilizando como referência o período da pandemia de COVID-19, entre 2020 e 2022; (iv) comparação quantitativa entre os EWS gerados pelo MMAING e os produzidos pelo EARS, entre 2020 e 2023; e (v) comparação com os boletins informativos do Centro de Informações Estratégicas em Vigilância em Saúde (CIEVS) do estado do Amazonas para o segundo quadrimestre (Maio a Agosto) de 2024.

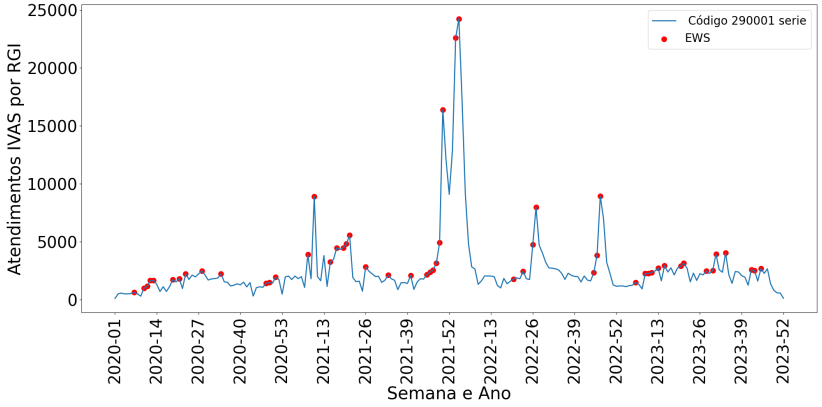
8.1
EWS em dados reais

No intuito de ilustrar a localização dos EWS indicados pelo MMAING nas séries temporais de dados reais, a Figura 8.1 considera 4 séries temporais de atendimentos referentes a IVAS de distintas RGIs, que englobam capitais estrategicamente escolhidas dentro da vasta geografia do Brasil. As RGIs selecionadas são: a) Belém (150001), capital do Pará, localizada na região Norte; b) Salvador (290001) capital da Bahia, localizada na região Nordeste; c) Belo Horizonte (310001), capital de Minas Gerais, localizada na região Sudeste; e d) Porto Alegre (430001), capital do Rio Grande do Sul, localizada no extremo sul.

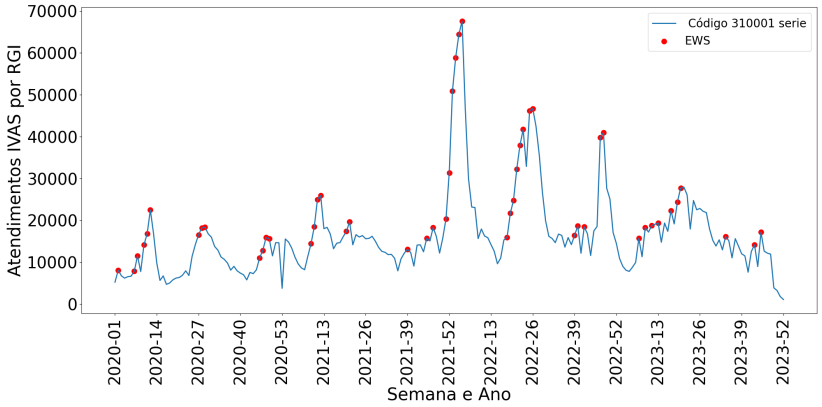
Esta seleção tem como objetivo ilustrar tanto o espectro diversificado de padrões epidemiológicos que ocorreram ao longo do tempo, como também refletir a dinâmica variada das IVAS em diferentes regiões do país.



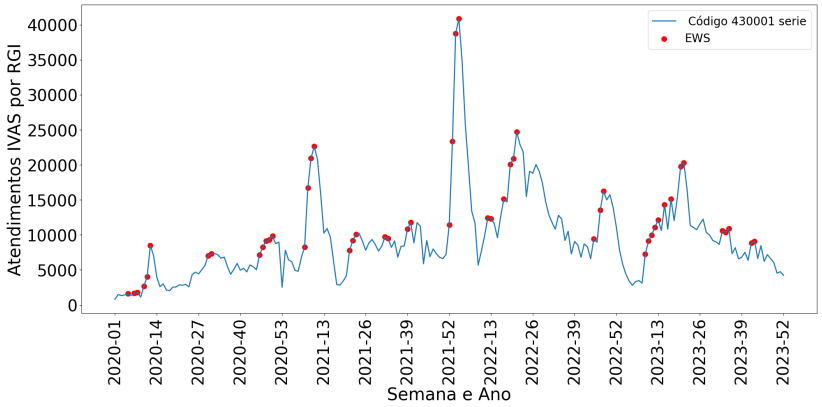
8.1(a): Belém - PA



8.1(b): Salvador - BA



8.1(c): Belo Horizonte - MG



8.1(d): Porto Alegre - RS

Figura 8.1: Detecção de EWS nas séries temporais de atendimentos IVAS entre 2020 e 2023 para RGIs selecionadas.

8.2 Período histórico da Covid-19

Por meio dos EWS indicados nas séries de dados sindrômicos da APS, conduzimos uma análise empírica da recente pandemia de COVID-19 no Brasil, com ênfase nas 27 RGIs (Tabela 7.3). Com base em boletins informativos de órgãos oficiais do governo, detalhamos e definimos aspectos da pandemia em quatro distintas ondas, ocorridas entre 2020 e 2022. As características mais marcantes desta pandemia foram a rápida propagação e evolução do vírus, impulsionadas pela alta transmissibilidade e capacidade de mutação. Isso nos levou a avaliar a capacidade do MMAING em detectar eventos durante uma pandemia real.

Tomamos como referência a análise para Belo Horizonte (310001), conforme apresentado na Figura 8.2, verificando a aplicabilidade do MMAING na identificação das principais tendências e ondas da pandemia, refletidas na série temporal de atendimentos por IVAS. Tais tendências e padrões, detectados tanto nesta quanto em outras RGIs, são indicados na Tabela 8.1.

É importante destacar que os dados da APS refletem moderadamente a linha temporal da COVID-19 no Brasil. O número de atendimentos na APS durante as ondas, conforme definido por boletins oficiais, difere significativamente entre as RGIs. Uma possível explicação para esse fenômeno foi a falta de padronização nas recomendações de saúde pública no Brasil durante a pandemia [182]. Apesar das limitações dos dados, a detecção precoce nas múltiplas RGIs ofereceu evidências da ampla aplicabilidade e utilidade dos dados sindrômicos da APS.

Confome apresentado na Figura 8.2, oficialmente a pandemia no país começou com a confirmação do primeiro caso de COVID-19 no Brasil em 26 de fevereiro de 2020, durante a semana epidemiológica número 9 [188, 182]. O MMAING indicou a emissão de EWS consecutivos nas semanas 7 e 8 (antes dos primeiros registros de casos de COVID-19) e nas semanas 10 a 12 de 2020. A análise indicou que o EWS subsequente, nas semanas 27 a 29, precedeu o pico da primeira onda da doença, registrada entre as semanas 29 e 30, conforme informado em [188].

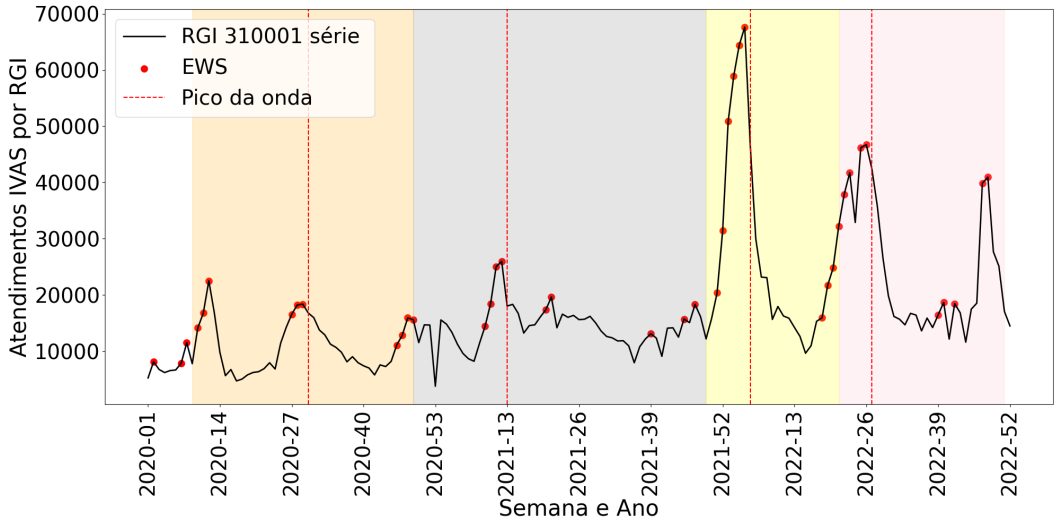


Figura 8.2: Detalhes dos resultados do MMAING para atendimentos IVAS em Belo Horizonte (conforme já exibido na Figura 8.1(c)), restrito ao período COVID-19 (2020-2022) e dividido em quatro ondas distintas de intervalos de tempo sucessivos sombreados por cores diferentes: Onda inicial (laranja); segunda onda, marcada pela chegada da variante Gama (cinza); terceira onda, influenciada pela variante Omicron (amarelo); e quarta onda (rosa); devido a reinfecções de Omicron e suas sublinhagens. As semanas com indicação de emissão EWS são destacadas por pontos em vermelho.

A transição para o intervalo subsequente, denominado de segunda onda (intervalo cinza na Figura 8.2), foi sinalizada por emissão de EWS nas semanas 46, 48 e 49 de 2020, coincidindo com o surgimento da variante Gamma (no início de novembro), que se tornou a principal variante em território brasileiro dois meses depois [188]. No inicio de 2021 foram detectados quatro EWS entre as semanas 9 e 12 antes do pico do segundo onda, que ocorreu entre as semanas 13 e 14. Foi neste contexto que ocorreu o rápido crescimento e predominância da variante Gama, atingindo o ápice em abril de 2021 (semanas 13 a 17). Esta onda foi marcada pelo colapso do sistema de saúde e pela ocorrência de crises sanitárias localizadas, combinando deficiência de equipamentos, de insumos para UTI e esgotamento da força de trabalho da saúde [188].

O MMAING também identificou dois EWS nas semanas 45 e 47 de 2021, as quais anteciparam o início da terceira onda, que começou no verão, durante as semanas epidemiológicas 49 e 50 (intervalo amarelo na Figura 8.2). Essa onda, impulsionada pela variante Omicron, é marcada por um aumento drástico de casos de COVID-19 a partir de dezembro de 2021, com repercussão em janeiro de 2022, culminando com um pico entre as semanas epidemiológicas 5 e 6 de 2022 [188]. Além disso, em novembro de 2021 (semanas 44 a 48), foi identificada uma nova subvariante (BA.1), com crescimento expressivo em

relação a outras sublinhagens Omicron circulantes, levando a um aumento de casos em dezembro. Esses EWS (semanas 45 e 47) antecederam o novo crescimento do número de casos na transição de 2021 para 2022, bem como impacto do surgimento da nova sublinhagem, reforçando assim a eficácia do MMAING na antecipação de tendências epidemiológicas.

O fim oficial da situação de emergência de saúde pública nacional causada pela pandemia da COVID-19, anunciado pelo Ministério da Saúde, entrou em vigor a partir da semana 21 de 2022. No entanto, uma nova onda marcada por reinfecções de Omicron e suas sub-linhagens foi observada, com picos em junho e dezembro. O mês de junho, que se inicia na semana epidemiológica 22 no início da quarta onda (intervalo rosa na Figura 8.2), foi marcado por um aumento elevado de casos e óbitos devido às sub-linhagens BA.4 e BA.5, responsáveis por 79% dos testes positivos à COVID-19 [198]. De acordo com a nossa análise, o MMAING sinalizou corretamente o início da quarta onda com EWS entre as semanas epidemiológicas 18 e 21.

Em suma, com base na série temporal de Belo Horizonte (310001), o MMAING destacou o surto iminente na primeira onda, sinalizando EWS antes mesmo da confirmação oficial do primeiro caso. Em seguida, identificou a chegada da nova onda que correspondem também ao surgimento da variante Gama, caracterizada pelo colapso do sistema de saúde e crises sanitárias locais. De forma análoga, antecipou o início da terceira onda impulsionada pela variante Omicron. Por fim, o MMAING sinalizou corretamente o início da quarta onda, associada às reinfecções causadas pela Ômicron e ao surgimento de novas sublinhagens, como BA.4 e BA.5.

Uma análise abrangente das 27 RGIs (Tabela 8.1) revelou padrões de antecipação durante todo o período da pandemia. Os EWS foram identificados entre 0 e 4 semanas antes do início real de cada onda e evento agravante, evidenciando o potencial do MMAING em fornecer informações oportunas que antecedem mudanças relevantes no cenário analisado. O MMAING conseguiu capturar com sucesso a dinâmica de todas as ondas da COVID-19 nas 27 RGIs. Para a primeira onda, o MMAING sinalizou EWS em 16 (59,3%) das RGIs; na segunda onda, ele indicou EWS em 21 (77,8%) das RGIs. Em relação à terceira onda, os EWS foram identificados em 14 (48,2%) das RGIs, enquanto, na quarta onda, foram identificados em 25 (92,6%) das RGIs. Além disso, em 6 (22%) das RGIs, foi possível sinalizar EWS consistentes em todas as quatro ondas, cobrindo todo o período da pandemia de COVID-19 no Brasil. Esses achados sugerem que o MMAING é eficaz em diferentes cenários e destacam o potencial dos dados da APS para a vigilância epidemiológica e a detecção precoce de surtos.

Tabela 8.1: Detecção precoce (0 a 4 semanas) que antecedem as ondas de COVID-19 usando BLCf do MMAING por RGI.

IGR	Onda 1	Onda 2	Onda 3	Onda 4
110001	-	✓	✓	-
120001	-	✓	✓	✓
130001	-	✓	✓	✓
140001	✓	-	✓	✓
150001	✓	✓	✓	✓
160001	-	-	-	✓
170001	-	-	-	✓
210001	-	-	✓	✓
220001	✓	✓	-	✓
230001	-	-	-	✓
240001	✓	✓	-	✓
250001	✓	✓	-	✓
260001	-	-	-	✓
270001	-	✓	-	-
280001	✓	✓	-	✓
290001	✓	✓	✓	✓
310001	✓	✓	✓	✓
320001	✓	✓	✓	✓
330001	-	✓	✓	✓
350001	✓	✓	✓	✓
410001	✓	✓	-	✓
420001	✓	✓	✓	✓
430001	✓	✓	-	✓
500001	✓	✓	-	✓
510001	-	✓	✓	✓
520001	✓	-	✓	✓
530001	✓	✓	-	✓
Detecção (%)	59,3	77,8	48,2	92,60

8.2.1 Variantes do SARS-CoV-2

No Brasil, no período de 2020 a 2024, foram registradas mudanças na frequência das linhagens dominantes de SARS-CoV-2, segundo dados da Rede Genômica Fiocruz [199], mostrados na Figura 8.3. Esta variação é uma característica notável da rápida movimentação e mutação do vírus.

Inicialmente, a pandemia foi impulsionada principalmente pelas linhagens B.1.1.28 e B.1.1.33, que foram as mais prevalentes até outubro de 2020. Após esse período, destaca-se a circulação de duas variantes de origem nacional, Gama (P.1) e Zeta (P.2), originadas da linhagem B.1.1.28. A partir de setembro de 2021 até janeiro de 2022, sobressaiu a circulação da variante Delta (B.1.617). No ano de 2022, a variante predominante foi a Omicron (BA.1, BA.2, BA.4 e BA.5) [188, 200].

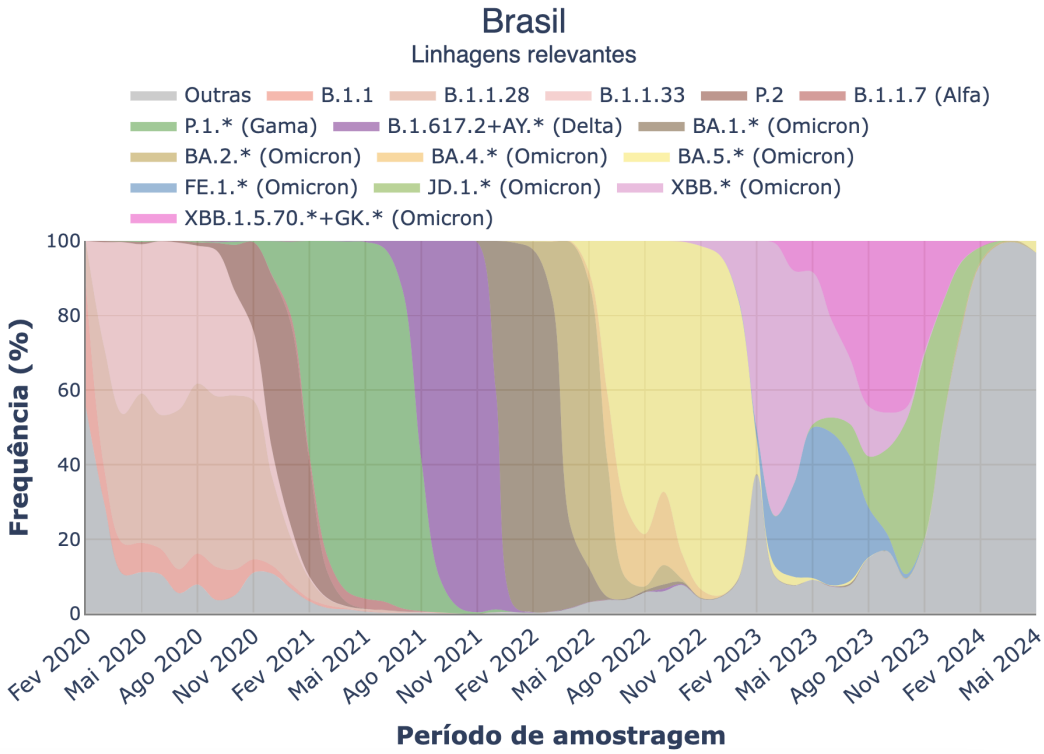


Figura 8.3: Frequência das principais linhagens de SARS-CoV-2 por mês de amostragem no Brasil. Figura retirada da Rede Genômica – Fiocruz (https://www.genomahcov.fiocruz.br/dashboard-pt/ [199]).

A Organização Mundial de Saúde (OMS) divide as variantes do SARS-CoV-2 em dois grupos: variantes de preocupação e variantes de interesse. As principais variantes de preocupação circulantes no período de análise (2020 a 2022), segundo a Rede Genômica Fiocruz [199], são Gama, Delta, Omicron

e suas sublinhagens. A Tabela 8.2 apresenta o período de surgimento dessas variantes no Brasil.

Tabela 8.2: Surgimento das variantes de preocupação do SARS-CoV-2 com relevância no Brasil. De acordo com o levantamento das linhagens pela Fiocruz (Fig. 8.3).

Variante	**Mês / Semanas**	**Ano**
Gama (P.1)	Novembro (44 a 48)	2020
Delta (B.1.617)	Maio (18 a 22)	2021
Omicron (BA.1)	Novembro (44 a 48)	2021
Omicron (BA.2)	Fevereiro (05 a 09)	2022
Omicron (BA.4 e BA.5)	Maio (18 a 22)	2022

A variante Gama, detectada em novembro de 2020, num período 6 meses passou de uma participação, de aproximadamente 10% dos casos, em dezembro de 2020, para mais de 95%, em maio de 2021. A variante Delta, detectada em maio de 2021, também em 6 meses passou de uma participação de 2% dos casos, em junho de 2021, para mais de 99%, em novembro de 2021. Posteriormente, a variante Omicron, detectada em novembro de 2021, variou de 39,8%, em dezembro de 2021, para 99%, a partir de janeiro de 2022. A variante BA.1 prevaleceu sobre as demais variantes, até o surgimento das variantes BA.4 e BA.5, em maio de 2022, cuja participação entre os casos variou de 11%, em maio de 2022, para 92%, em dezembro de 2022 [188, 199, 200]. De forma análoga, ilustramos na Figura 8.4, a análise para Belo Horizonte (310001) a título de referência, que permite ressalta visualmente o marco temporal do surgimento de cada variante e seu tempo de circulação.

Analisando a Figura 8.4, nota-se que os EWS emitidos nas semanas 46 a 48 de 2020 coincidem com o aparecimento da variante Gama, o mesmo acontecendo para as semanas 20 e 21 de 2021 com relação à variante Delta, nas semanas 45 e 47 de 2021 com a Omicron (BA.1), e, finalmente, nas semanas 18 a 23 de 2022 com as variantes Omicrons BA.4 e BA.5.

Portanto, os EWS detectados pelo MMAING a partir de dados de atendimentos de trato respiratório da APS, além de descreverem a dinâmica da pandemia, sinalizaram o aparecimento das variantes e os impactos durante o tempo de circulação das mesmas.

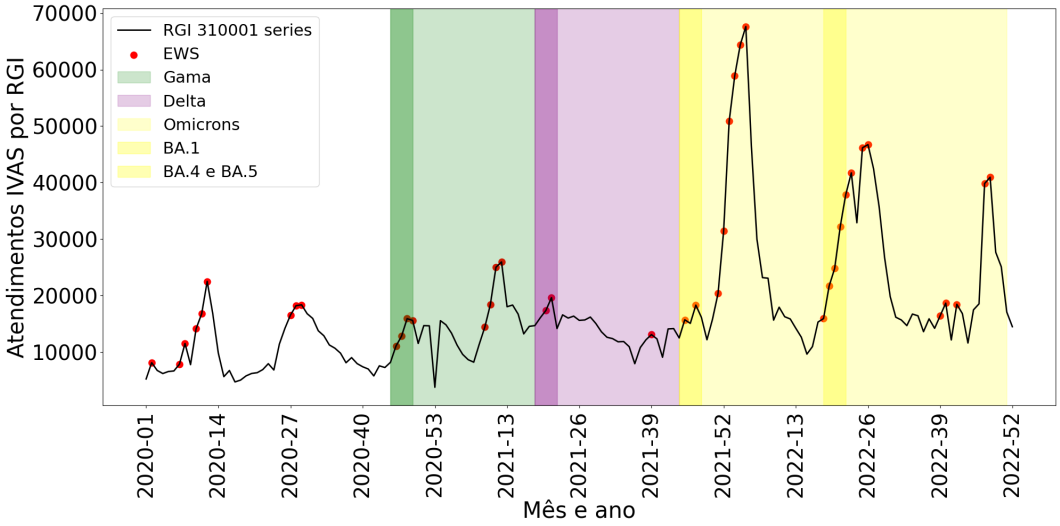


Figura 8.4: Detalhes dos resultados do MMAING para atendimentos IVAS em Belo Horizonte (conforme já exibido na Figura 8.1(c)), restrito ao período COVID-19 (2020-2022), destacando o período temporal de circulação e aparecimento das variantes de preocupação de SARS-CoV-2 em intervalos de tempo sucessivos sombreados por cores diferentes: Gama (verde), Delta (roxo) e Omicron e suas sublinhagens (amarelo), as faixas de coloração mais escura marcam o aparecimento da variante.

8.3 Análise comparativa entre o MMAING e o EARS

Nesta seção, apresentamos uma comparação entre os resultados obtidos pelo MMAING e aqueles gerados pelo método EARS, que, desde 2001, tem sido amplamente utilizado na detecção de surtos e no monitoramento de contagens sindrômicas semanais [177]. Inicialmente, analisamos a concordância entre os métodos na emissão semanal dos EWS, visando, posteriormente, avaliar a eficácia relativa do MMAING frente ao EARS.

8.3.1 Análise em séries temporais reais

Ambos os métodos foram aplicados às series temporais referentes aos atendimentos por IVAS das 27 RGIs, que refletem a atividade sindrômica em diferentes contextos regionais. Inicialmente, realizamos a análise da coincidência na emissão dos EWS semanais entre os dois métodos, métrica que aqui denominamos *compatibilidade*.

A compatibilidade é definida como a proporção de EWS emitidos simultaneamente por ambos os métodos em relação ao total de EWS gerados

pelo MMAING. A compatibilidade percentual é calculada da seguinte forma:

$$\text{Compatibilidade (\%)} = \frac{\text{EWS coincidentes}}{\text{Total de EWS do MMAING}} \times 100 \qquad (8\text{-}1)$$

A Tabela 8.3 resume as médias percentuais de compatibilidade entre o MMAING e as três variações do EARS (C1, C2 e C3), destacando que a variante C2 apresentou a melhor compatibilidade na detecção de EWS, considerando os dados analisados.

Tabela 8.3: Compatibilidade média entre MMAING e variações do EARS (2020 a 2023, considerando todas as 27 RGIs).

Compatibilidade (%)			
EARS	**C1**	**C2**	**C3**
MMAING	63,41	66,23	39,65

Conforme apresentado na Tabela 8.4, também comparamos o número total de EWS emitidos pelo MMAING e pelas três variações do EARS longo do período analisado, considerando as 27 RGIs. Observa-se que o método EARS-C1 apresentou o menor número total de EWS emitidos no período (904), seguido pelo EARS-C3 (1174). Por outro lado, o MMAING e o EARS-C2 geraram o maior volume de detecções, ambos com um total de 1208 EWS no mesmo intervalo.

Na sequência, detalhamos a compatibilidade anual entre os métodos, conforme apresentado na Tabela 8.5. Observa-se que os maiores percentuais de coincidência ocorreram entre o MMAING e o EARS-C2: 74,48% em 2020, 58,65% em 2021, 73,91% em 2022 e 52,67% em 2023. Esses resultados reforçam a maior convergência entre esses dois métodos na detecção de EWS utilizando os dados da APS.

Tabela 8.4: Número de EWS por ano (dados da APS), para EARS (C1, C2 e C3) e MMAING.

Ano	**C1**	**C2**	**C3**	**MMAING**
2020	290	389	370	337
2021	192	246	207	283
2022	271	372	419	345
2023	151	201	178	243
Total	904	1208	1174	1208

Tabela 8.5: Número de EWS coincidentes por ano e percentual de compatibilidade (%) entre MMAING e EARS (C1, C2 e C3).

Ano	**MMAING e C1**	**MMAING e C2**	**MMAING e C3**
2020	229 (67,95%)	251 (74,48%)	159 (47,18%)
2021	166 (58,65%)	166 (58,65%)	75 (26,50%)
2022	246 (71,30%)	255 (73,91%)	176 (51,02%)
2023	125 (51,44%)	128 (52,67%)	69 (28,40%)
Total	766 (63,41%)	800 (66,23%)	479 (39,65%)

Na etapa seguinte, conduzimos uma análise semelhante, considerando individualmente o número de EWS coincidentes para cada uma das 27 RGIs.

Todos os resultados dessa etapa estão apresentados nas Tabelas 8.6, 8.7 e 8.8. Para facilitar a interpretação, analisamos e ilustramos graficamente a compatibilidade entre o MMAING e as diferentes variações do EARS (C1, C2 e C3), utilizando como referência a RGI de Belo Horizonte (310001), conforme mostrado na Figura 8.5.

Nessa representação gráfica (Figura 8.5), são detalhadas as distribuições dos EWS indicados por todos os métodos para a RGI de Belo Horizonte (310001), destacando comparações individuais entre o MMAING e as variações do EARS: C1 (Figura (a) 8.5), C2 (Figura (b) 8.5) e C3 (Figura (c) 8.5). Em cada caso, dois gráficos consecutivos representam as detecções realizadas pelo MMAING (superior) e pelo respectivo método EARS (inferior). Os marcadores azuis destacam os *EWS coincidentes*, alinhados verticalmente por linhas tracejadas, e a combinação com os marcadores vermelhos indica o total de EWS detectados.

Analisando a relação entre os EWS coincidentes (azul) detectados pelo EARS e o total de EWS gerados pelo MMAING (azul + vermelho), observamos correspondências percentuais de 68,63% para o EARS C1, 76,47% para o EARS C2 e 43,14% para o EARS C3, conforme detalhado nas Tabelas 8.6, 8.7 e 8.8. Esses resultados reforçam que a melhor convergência ocorre entre o MMAING e o EARS C2.

Tabela 8.6: Número de EWS coincidentes por RGI (dados da APS), por MMAING e EARS C1.

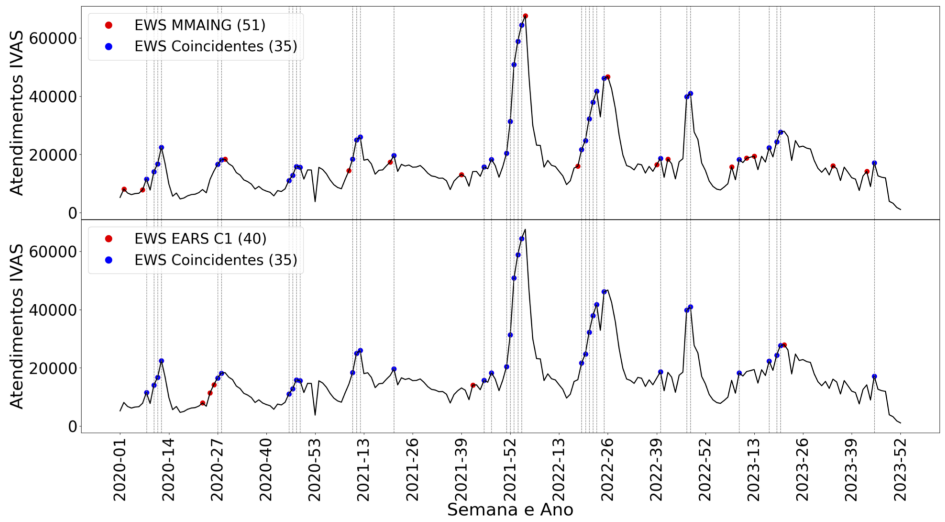
RGI	MMAING	EARS C1	Coincidentes	Compatibilidade (%)
110001	52	35	34	65.38
120001	36	32	29	80.56
130001	42	27	25	59.52
140001	36	36	27	61.11
150001	46	31	27	58.70
160001	36	29	20	55.56
170001	41	38	25	60.98
210001	40	27	23	57.50
220001	56	36	35	62.50
230001	40	39	26	65.00
240001	47	24	24	51.06
250001	42	28	25	59.52
260001	31	27	21	67.74
270001	38	25	21	55.26
280001	51	34	32	62.75
290001	53	27	27	50.94
310001	51	40	35	68.63
320001	57	35	34	59.65
330001	49	36	34	69.39
350001	37	38	23	62.16
410001	41	39	30	73.17
420001	50	30	28	56.00
430001	51	40	33	64.71
500001	49	36	34	69.39
510001	50	34	31	62.00
520001	49	40	39	75.51
530001	37	41	31	83.78
Total	1208	904	766	63.41

Tabela 8.7: Número de EWS coincidentes por RGI (dados da APS), por MMAING e EARS C2.

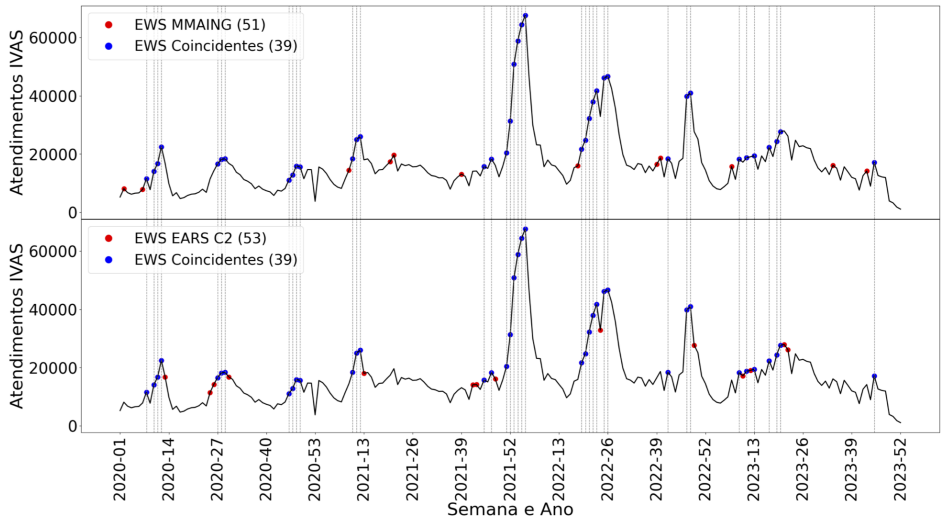
RGI	MMAING	EARS C2	Coincidentes	Compatibilidade (%)
110001	52	42	34	65.38
120001	36	41	26	72.22
130001	42	39	27	64.29
140001	36	47	21	58.33
150001	46	39	22	47.83
160001	36	33	21	58.33
170001	41	49	26	63.41
210001	40	42	26	65.00
220001	56	47	37	66.07
230001	40	46	26	65.00
240001	47	36	27	57.45
250001	42	38	24	57.14
260001	31	40	25	80.65
270001	38	35	21	55.26
280001	51	55	40	78.43
290001	53	44	32	60.38
310001	51	53	39	76.47
320001	57	44	36	63.16
330001	49	53	38	77.55
350001	37	49	27	72.97
410001	41	53	32	78.05
420001	50	45	30	60.00
430001	51	51	31	60.78
500001	49	46	33	67.35
510001	50	45	31	62.00
520001	49	52	37	75.51
530001	37	44	31	83.78
Total	1208	1208	800	66.23

Tabela 8.8: Número de EWS coincidentes por RGI (dados da APS), por MMAING e EARS C3.

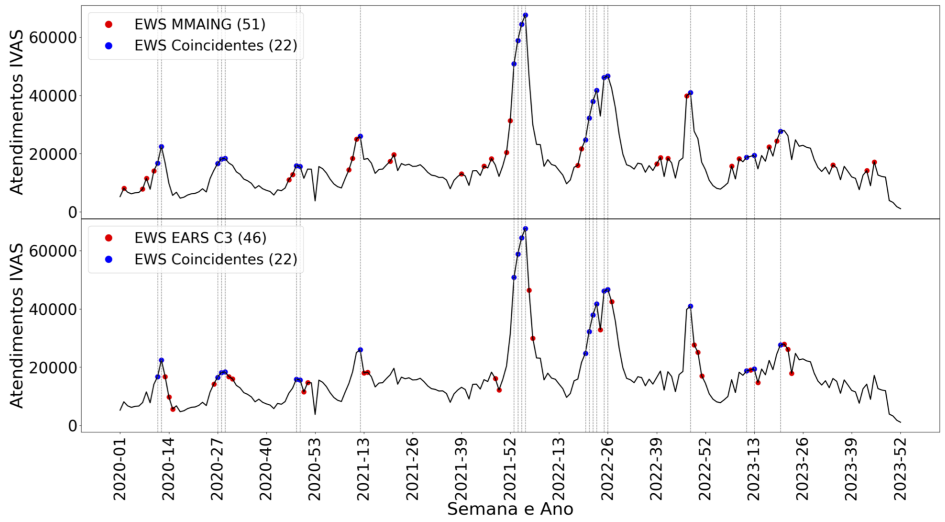
RGI	MMAING	EARS C3	Coincidentes	Compatibilidade (%)
110001	52	42	21	40.38
120001	36	36	11	30.56
130001	42	40	15	35.71
140001	36	43	16	44.44
150001	46	37	11	23.91
160001	36	26	7	19.44
170001	41	48	18	43.90
210001	40	42	17	42.50
220001	56	47	24	48.21
230001	40	43	13	32.50
240001	47	39	14	29.79
250001	42	39	12	28.57
260001	31	37	17	54.84
270001	38	32	10	26.32
280001	51	57	27	52.94
290001	53	37	15	28.30
310001	51	46	22	43.14
320001	57	50	24	42.11
330001	49	54	20	40.82
350001	37	48	20	48.65
410001	41	48	20	48.78
420001	50	48	19	38.00
430001	51	51	19	37.25
500001	49	46	20	40.82
510001	50	45	22	44.00
520001	49	51	23	46.94
530001	37	42	21	56.76
Total	1208	1174	479	39.65



(a): MMAING vs C1



(b): MMAING vs C2



(c): MMAING vs C3

Figura 8.5: Os gráficos ilustram a região imediata de Belo Horizonte (310001). a) Equivalência entre MMAING e C1; b) Equivalência entre MMAING e C2 e c) Equivalência entre MMAING e C3.

Também realizamos uma comparação mais detalhada entre o MMAING e o EARS C2, especificamente para as RGIs selecionadas: Belém (150001), Salvador (290001), Belo Horizonte (310001) e Porto Alegre (430001), conforme ilustrado na Figura 8.6. Em cada subfigura (a-d): a) Belém; b) Salvador; c) Belo Horizonte; e d) Porto Alegre, dois gráficos consecutivos apresentam as detecções realizadas pelo MMAING (superior) e pelo EARS C2 (inferior). Os marcadores seguem o mesmo padrão adotado anteriormente, com marcadores azuis indicando os EWS coincidentes, alinhados por linhas tracejadas verticais, e a combinação dos marcadores azuis e vermelhos indicando o total de EWS detectados.

Considerando a relação entre os EWS coincidentes (azul) detectados pelo EARS C2 e o total de EWS gerados pelo MMAING (azul + vermelho), identificamos percentuais de correspondência variados entre as RGIs: Belém apresentou 47,83%, Salvador 60,38%, Belo Horizonte 76,47% e Porto Alegre 60,78%. Esses resultados destacam não somente a precisão geral do método, mas também evidenciam a variabilidade regional na detecção de EWS entre o MMAING e o EARS C2, conforme detalhado na Tabela 8.7.

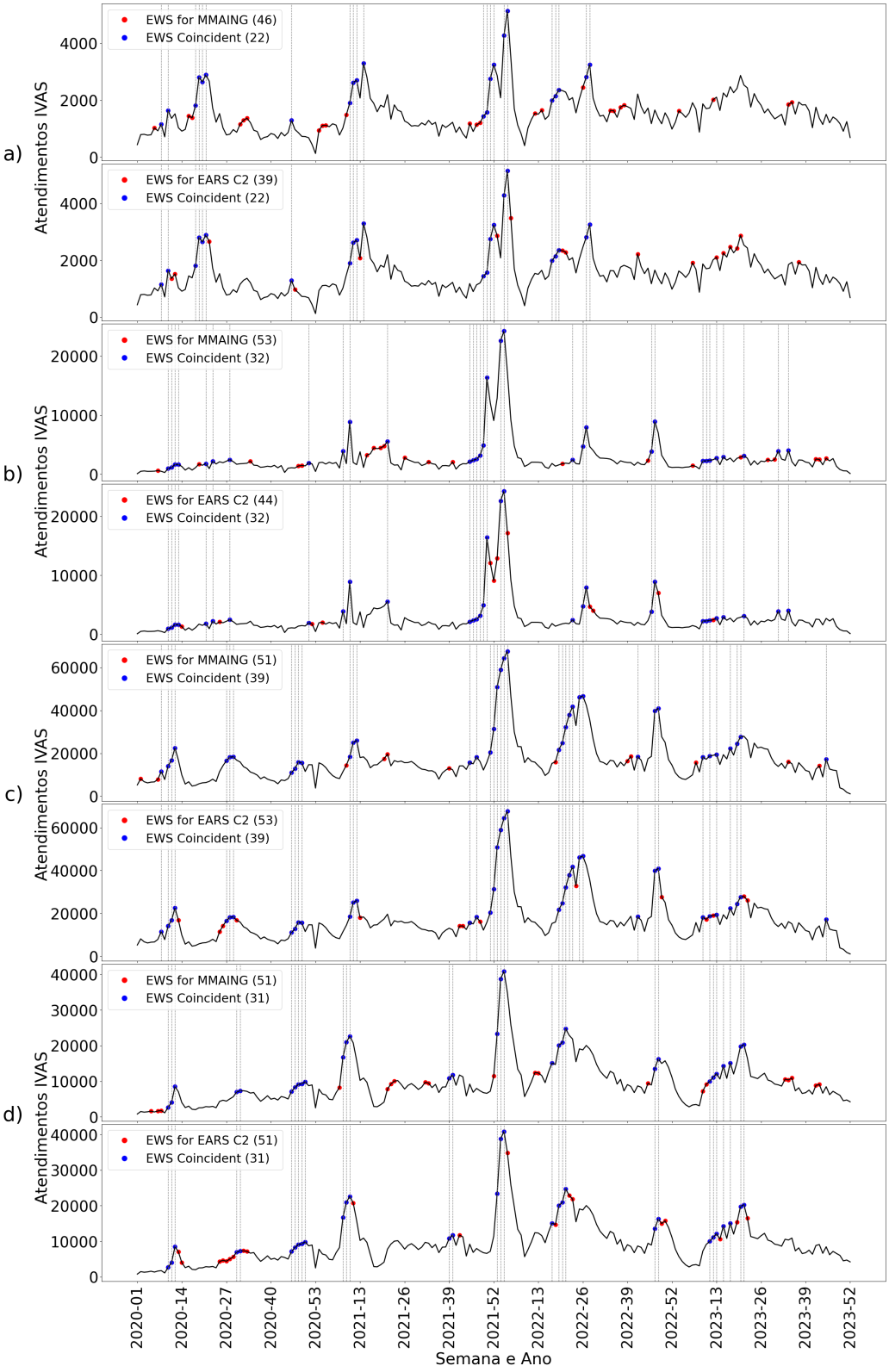


Figura 8.6: EWS para quatro RGIs: a) Belém (150001); b) Salvador (290001); c) Belo Horizonte (310001) e d) Porto Alegre (IGR 430001) de 2020 a 2023. Os gráficos superiores e inferiores indicam detecções pelo MMAING e EARS, respectivamente. Marcadores azuis indicam EWS coincidentes para ambos os métodos. A soma dos marcadores vermelhos e azuis corresponde ao total de EWS.

8.3.2 Análise em séries temporais sintéticas

Além disso, realizamos uma avaliação de performance utilizando dados sintéticos e obtivemos o desempenho dos métodos MMAING e EARS (C1, C2 e C3), nas 810 séries simuladas de cenários únicos. Detalhamos os resultados para as mesmas RGIs selecionadas, conforme apresentado na Tabela 8.9. Observamos que todos os métodos, exceto o EARS C3, tiveram desempenhos semelhantes.

O MMAING demonstrou um melhor POD e sensibilidade em comparação com as variações do EARS; no entanto, apresentou um PPV inferior ao EARS C1 e manteve um PPV próximo ao EARS C2. Todos os métodos exibiram pontuações F1 e especificidades semelhantes. O MMAING superou o EARS C3 em todas as métricas.

A Figura 8.7 apresenta a distribuição das métricas de desempenho – POD, PPV, sensibilidade, pontuação F1 e especificidade – para os métodos avaliados. O MMAING demonstrou consistência em quase todas as métricas, mantendo pontuações altas e variações mínimas, além de uma boa probabilidade de detecção, indicando um equilíbrio robusto entre verdadeiros positivos e negativos de EWS. Em contraste, as variantes do EARS apresentaram maiores variações em suas métricas. O EARS C1 e C2 possivelmente equilibram a identificação correta dos EWS positivos (PPV) e a captura do máximo de EWS positivos possíveis (sensibilidade), respectivamente. O EARS C3 destacou-se por métricas abaixo da média, exceto por sua especificidade altamente concentrada, o que pode ser vantajoso quando os custos de falsos positivos são elevados. A análise da pontuação F1, que equilibra PPV e sensibilidade, sugere que todos os modelos mantiveram um nível moderadamente alto de desempenho equilibrado, com o MMAING oferecendo uma leve vantagem em consistência.

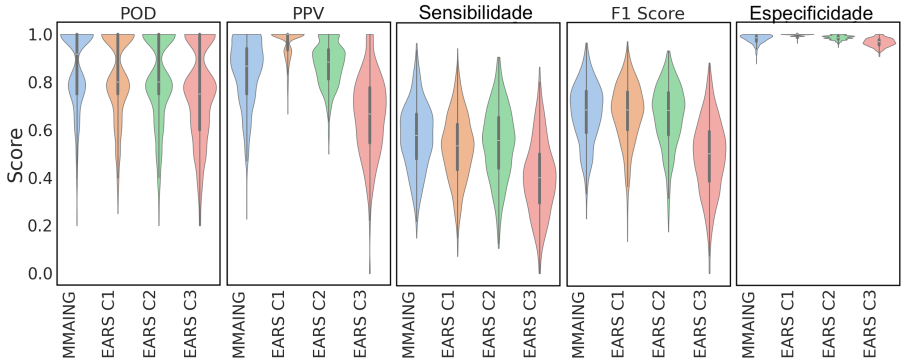


Figura 8.7: Distribuição das pontuações avaliadas entre os diferentes modelos considerando 30 simulações para 27 IGRs dos estados capitais brasileiros.

Tabela 8.9: Desempenho médio do MMAING e EARS (C1, C2 e C3) para as IGRs de Belém (150001), Belo Horizonte (310001) e Porto Alegre (430001).

MMAING					
IGR	POD	PPV	Sensibilidade	F1	Especificidade
150001	0.75	0.85	0.50	0.62	0.99
290001	0.86	0.92	0.56	0.69	0.99
310001	0.81	0.88	0.57	0.68	0.99
430001	0.87	0.83	0.59	0.68	0.98
27 IGRs	0.86	0.85	0.59	0.68	0.98
EARS C1					
IGR	POD	PPV	Sensibilidade	F1	Especificidade
150001	0.80	0.97	0.50	0.64	1.00
290001	0.88	0.98	0.53	0.68	1.00
310001	0.80	0.97	0.53	0.68	1.00
430001	0.73	0.99	0.49	0.64	1.00
27 IGRs	0.83	0.97	0.53	0.68	1.00
EARS C2					
IGR	POD	PPV	Sensibilidade	F1	Especificidade
150001	0.80	0.86	0.54	0.66	0.99
290001	0.85	0.90	0.54	0.66	0.99
310001	0.80	0.90	0.56	0.67	0.99
430001	0.73	0.93	0.52	0.65	0.99
27 IGRs	0.80	0.86	0.55	0.66	0.99
EARS C3					
IGR	POD	PPV	Sensibilidade	F1	Especificidade
150001	0.72	0.64	0.38	0.46	0.96
290001	0.76	0.66	0.38	0.47	0.96
310001	0.71	0.67	0.41	0.50	0.97
430001	0.65	0.79	0.43	0.54	0.98
27 IGRs	0.75	0.66	0.40	0.49	0.97

A Figura 8.8 ilustra as curvas ROC (Receiver Operating Characteristic) para os resultados obtidos das 27 RGIs pelos quatro modelos de detecção —

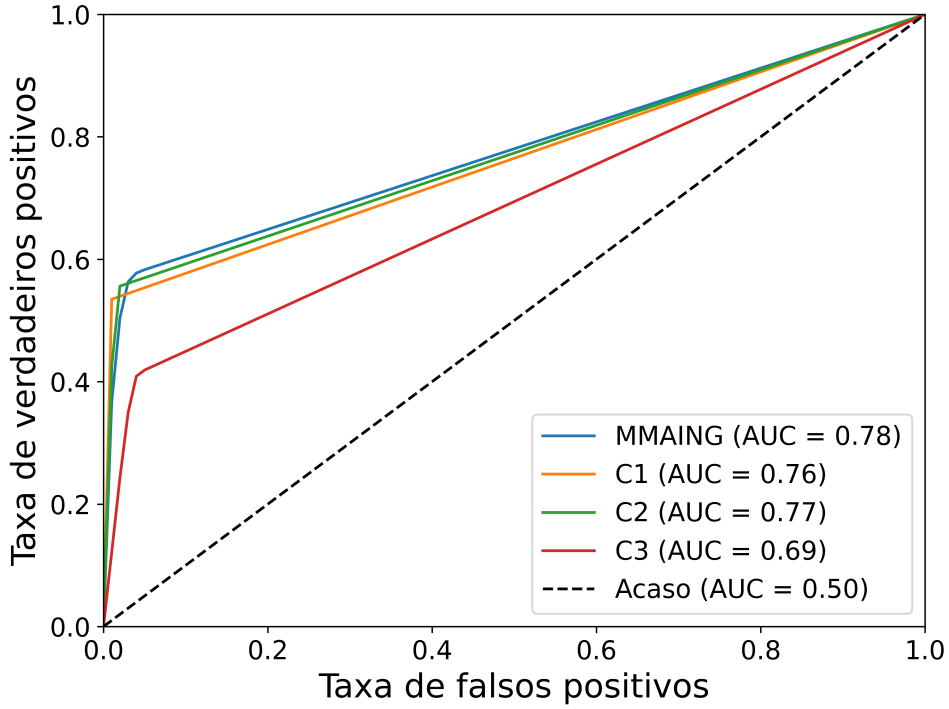


Figura 8.8: Capacidade discriminatória do MMAING e variações do EARS (C1, C2 e C3) ilustrada por suas curvas ROC.

MMAING, EARS C1, C2 e C3. O MMAING mostrou-se robusto, com uma Área Sob a Curva (AUC) média de 0.78, demonstrando um equilíbrio eficaz entre sensibilidade e especificidade, sugerindo uma capacidade consistente de discriminar entre classes positivas e negativas. Os modelos EARS C1 e C2, com AUCs de 0.76 e 0.77, respectivamente, apresentaram um desempenho semelhante ao do MMAING, indicando também uma boa precisão de classificação. No entanto, a ligeira diferença na AUC sugere que o MMAING teve um desempenho geral ligeiramente superior. O EARS C3, com uma AUC de 0.69, mostrou uma clara diminuição no desempenho em comparação com os outros modelos, sugerindo menor precisão de classificação e uma tendência a apresentar uma taxa mais alta de falsos positivos para EWS.

No geral, a análise da curva ROC sugere que o MMAING pode ser a escolha preferida para aplicações que exigem um equilíbrio entre a identificação correta de EWS positivos e a prevenção de falsos positivos.

Embora nenhum modelo tenha se destacado consistentemente em todas as métricas de desempenho, o MMAING emergiu como uma opção promissora para um sistema de alerta precoce. Este modelo não só se mostrou mais oportuno, mas também registrou um POD ligeiramente superior aos demais,

um aspecto importante se a prioridade do sistema de vigilância for a detecção abrangente de surtos. Além disso, o MMAING alcançou a maior sensibilidade entre os modelos avaliados, uma métrica essencial para a clareza e consistência de EWS, além de um bom PPV, que indica a probabilidade de um alerta ser verdadeiro. Sua alta especificidade é essencial para minimizar os alertas falsos, e sua taxa de pontuação F1, que equilibra PPV e sensibilidade, reforça a confiabilidade na detecção de surtos reais, reduzindo o número de alertas falsos. Também exibiu a melhor curva ROC, reforçando a avaliação de sua eficácia.

8.4 Performance das configurações do MMAING

Para avaliar a eficácia das configurações EqCf e RiCf do MMAING, cujos parâmetros estão detalhados na Tabela 7.4, e verificar se essas configurações atendem aos seus objetivos específicos, EqCf visando equilíbrio entre precisão e sensibilidade e RiCf focando em alta precisão, examinamos sua adaptabilidade em diversos cenários simulados.

Empregamos uma abordagem de teste cego, que é caracterizada pela divisão de conjuntos de dados para treinamento e validação. O modelo é treinado exclusivamente com dados reais, enquanto a validação é realizada em uma série temporal simulada completamente desconhecida do modelo durante sua fase de treinamento. Esta abordagem garante que a avaliação do desempenho do modelo seja realizada sob condições imparciais, refletindo a sua capacidade de generalização para novos dados. As 810 séries temporais simuladas com cenários únicos, foram aplicadas para cada configuração.

Apresentamos os resultados de desempenho do MMAING nas diferentes configurações para os 27 RGIs. As Tabelas 8.10 e 8.11 exibem, respectivamente, o desempenho das configurações EqCf e RiCf do MMAING. Com base nos resultados médios para as duas configurações (Tabelas 8.10 e 8.11), observamos que o MMAING apresentou ligeira variação entre PPV e especificidade, enquanto as principais diferenças foram para POD, sensibilidade e escore F1. Nota-se uma melhora na precisão e uma diminuição na sensibilidade da RiCf em relação a configuração EqCf, como era esperado. Entretanto, esperava-se uma melhora no valor do PPV.

Tabela 8.10: Métricas de Desempenho obtidas pelo EqCf do MMAING para dados sintéticos baseados em RGI

RGI	PPV	POD	Sensibilidade	F1	Especificidade
110001	0.78	0.85	0.56	0.64	0.97
120001	0.78	0.84	0.50	0.60	0.98
130001	0.93	0.88	0.60	0.72	0.99
140001	0.83	0.74	0.50	0.62	0.98
150001	0.85	0.75	0.50	0.62	0.99
160001	0.91	0.97	0.62	0.73	0.99
170001	0.89	0.94	0.64	0.74	0.99
210001	0.75	0.88	0.58	0.63	0.96
220001	0.88	0.92	0.61	0.71	0.99
230001	0.95	0.89	0.64	0.75	0.99
240001	0.83	0.83	0.51	0.62	0.98
250001	0.70	0.78	0.47	0.55	0.97
260001	0.94	0.94	0.66	0.77	0.99
270001	0.92	0.84	0.56	0.69	0.99
280001	0.88	0.85	0.56	0.67	0.99
290001	0.92	0.86	0.56	0.69	0.99
310001	0.88	0.81	0.57	0.68	0.99
320001	0.75	0.85	0.53	0.61	0.96
330001	0.92	0.87	0.56	0.69	0.99
350001	0.95	0.88	0.65	0.76	0.99
410001	0.69	0.84	0.57	0.62	0.96
420001	0.79	0.78	0.54	0.63	0.97
430001	0.83	0.87	0.59	0.68	0.98
500001	0.78	0.90	0.61	0.68	0.97
510001	0.83	0.88	0.59	0.68	0.98
520001	0.92	0.95	0.69	0.78	0.99
530001	0.88	0.95	0.64	0.73	0.99
Média Total	0.85	0.86	0.59	0.68	0.98

Tabela 8.11: Métricas de Desempenho obtidas pelo RiCf do MMAING para dados sintéticos baseados em RGI

RGI	PPV	POD	Sensibilidade	F1	Especificidade
110001	0.79	0.78	0.43	0.54	0.98
120001	0.78	0.62	0.28	0.40	0.99
130001	0.94	0.81	0.56	0.69	0.99
140001	0.84	0.73	0.39	0.52	0.99
150001	0.93	0.64	0.37	0.52	1.00
160001	0.92	0.87	0.56	0.68	0.99
170001	0.90	0.83	0.54	0.66	0.99
210001	0.81	0.75	0.42	0.53	0.98
220001	0.89	0.80	0.43	0.57	0.99
230001	0.93	0.79	0.50	0.64	0.99
240001	0.89	0.58	0.31	0.45	0.99
250001	0.71	0.57	0.26	0.37	0.98
260001	0.90	0.88	0.54	0.67	0.99
270001	0.94	0.69	0.42	0.57	1.00
280001	0.85	0.68	0.36	0.49	0.99
290001	0.93	0.67	0.37	0.52	0.99
310001	0.87	0.77	0.41	0.55	0.99
320001	0.71	0.64	0.34	0.45	0.98
330001	0.92	0.70	0.44	0.58	1.00
350001	0.93	0.82	0.54	0.68	0.99
410001	0.72	0.69	0.40	0.50	0.97
420001	0.72	0.54	0.28	0.39	0.98
430001	0.87	0.68	0.39	0.52	0.99
500001	0.91	0.73	0.46	0.60	0.99
510001	0.88	0.79	0.48	0.61	0.99
520001	0.94	0.86	0.58	0.70	0.99
530001	0.94	0.67	0.42	0.57	0.99
Média Total	0.87	0.73	0.44	0.57	0.99

Os resultados médios para EqCf foram 0,86 para POD, 0,85 para PPV, 0,59 para sensibilidade, 0,68 para pontuação F1 e 0,98 para especificidade, indicando que o MMAING tem uma alta probabilidade de detecção de surto, sensibilidade razoável e uma boa taxa de PPV. Estes valores também sugerem

que a maioria dos EWS tem uma probabilidade considerável de ser verdadeiro e foram emitidos corretamente, implicando em uma confiabilidade média de 79%, útil para vigilância de surtos frequentes. Quanto ao RiCf, percebe-se que o resultado médio entre as regiões foi de 0,73 para POD, 0,87 para PPV, 0,44 para sensibilidade, 0,57 para escore F1 e 0,99 para especificidade. Estes resultados indicam que o RiCf apresentou um comportamento mais rigoroso, mantendo maior precisão, mas reduzindo significativamente a probabilidade de detecção e a sensibilidade. Esse comportamento garante que qualquer EWS emitido provavelmente será verdadeiro. Em consequência, nem todos os EWS verdadeiros existentes na série seriam detectados, mas proporciona um sistema de detecção confiável, extremamente útil para a vigilância de surtos não frequentes. Esses resultados são comparáveis aos de outros modelos usados no monitoramento da saúde pública [24, 78, 167, 166, 191].

Essa validação do MMAING, realizada através da análise das séries sintéticas geradas a partir de séries reais, garantiu a escolha de configurações adequadas para o modelo, especialmente na parametrização EqCf, proporcionando assim uma adaptação adequada ao seu grau de aplicabilidade.

Em uma outra análise, os resultados obtidos para as duas configurações do MMAING, aplicadas às 810 séries sintéticas foram agrupados em cinco regiões geográficas brasileiras: 7 para o Norte (N); 9 para o Nordeste (NE); 4 para Sudeste (SE); 3 para o Sul (S) e por fim, 4 para o Centro-Oeste (C). A Tabela 8.12 apresenta a média dos resultados para a EqCf na parte superior, enquanto na parte inferior apresenta a média dos resultados para a RiCf.

O fato de que a entrada aleatória de surtos na série sintética ainda depende dos dados reais explica uma pequena, mas perceptível, variabilidade de pontuação entre as regiões geográficas do Brasil (Tabela 8.12). As pontuações mais altas e mais baixas foram obtidas para as regiões Centro-Oeste e Sul respectivamente, como mostrado na Tabela 8.12; não está claro como explicar esse comportamento, exceto pelo pequeno número de estados em cada região geográfica (4 e 3 respectivamente), o que causa um aumento nas flutuações. De fato, dois RGIs nas regiões Centro-Oeste (520001 e 530001) e uma na regiões Sul (430001) levaram a resultados acima da média nacional.

Tabela 8.12: Métricas de desempenho médio para MMAING – EqCf (superior) e RiCf (inferior).

MMAING – EqCf					
Região	POD	PPV	Sensibilidade	F1	Especificidade
Norte	0.83	0.85	0.56	0.67	0.98
Nordeste	0.87	0.86	0.57	0.68	0.98
Sudeste	0.86	0.87	0.58	0.69	0.98
Sul	0.83	0.77	0.57	0.64	0.97
Centro-Oeste	0.91	0.85	0.63	0.72	0.98
27 IGRs	0.86	0.85	0.59	0.68	0.98
MMAING – RiCf					
Região	POD	PPV	Sensibilidade	F1	Especificidade
Norte	0.76	0.86	0.44	0.56	0.99
Nordeste	0.72	0.87	0.43	0.56	0.99
Sudeste	0.77	0.86	0.47	0.60	0.99
Sul	0.67	0.75	0.36	0.47	0.98
Centro-Oeste	0.75	0.89	0.47	0.60	0.99
27 IGRs	0.73	0.87	0.44	0.57	0.99

8.5 Validação do MMAING com dados do CIEVS-AM

Nesta seção apresentamos uma avaliação mais precisa dos protocolos desenvolvidos pelo projeto ÆSOP para a identificação e emissão de alertas de surtos epidêmicos.

Em cooperação com o Centro de Informações Estratégicas em Vigilância em Saúde (CIEVS) do estado do Amazonas, juntamente com a supervisão do CIEVS Nacional, foram disponibilizados ao projeto ÆSOP relatórios semanais devidamente preenchidos pelo CIEVS-AM para a análise dos EWS emitidos para o estado do Amazonas.

Para cada relatório, foram analisadas três colunas importantes: i) *Vigilância local confirma o sinal?*, ii) *Houve aumento de atendimentos não detectados?* e iii) *Observações da vigilância local*. A partir da interpretação dessas três colunas, foram rotuladas as semanas em que surtos reais foram identificados. Com base nesses rótulos, realizamos uma análise de performance

para o MMAING durante o segundo quadrimestre de 2024, abrangendo as semanas epidemiológicas de 18 a 28 para 62 municípios.

A Figura 8.9 mostra a quantidade de semanas rotuladas com surtos reais para diferentes municípios do estado do Amazonas, contabilizando um total de 92 semanas com surto. Observa-se que há diversos municípios que não indicaram qualquer semana com surto.

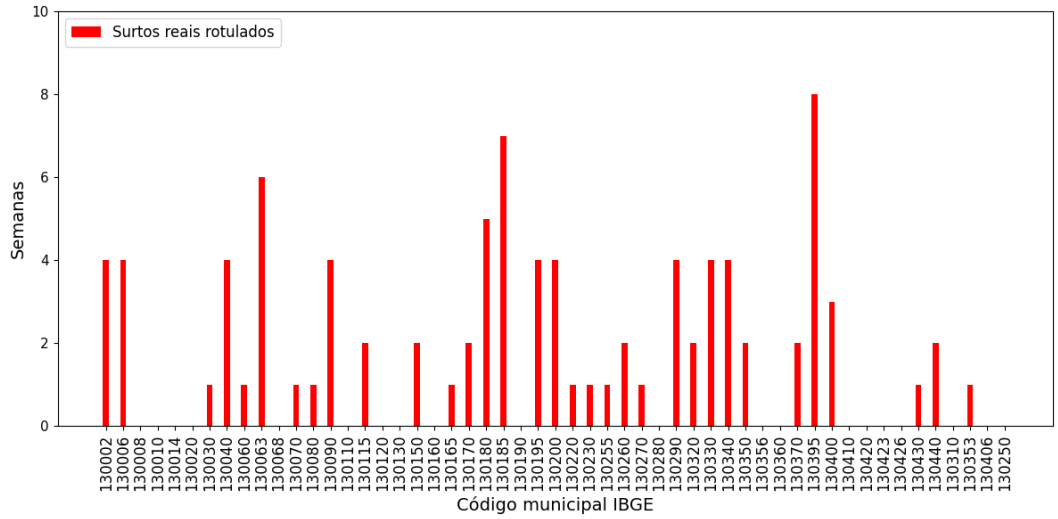


Figura 8.9: Distribuição municipal da quantidade de surtos rotulados para o estado do Amazonas.

É importante mencionar que o MMAING analisa os dados de APS apenas para séries que passaram por um controle de qualidade, denominado de "Indice de Qualidade de Dado" (DQI do inglês, *Data Quality Index*), que foi desenvolvido por uma equipe do projeto ÆSOP. O DQI permite classificar como "aptos" para análise apenas os dados que atendem aos critérios de completude, tempestividade e consistência. Durante o período total considerado (11 semanas), foram realizadas 437 análises, correspondendo às instâncias em que os dados semanais dos municípios do estado do Amazonas foram considerados aptos segundo o DQI. As demais 245 instâncias foram excluidas da avaliação devido à baixa qualidade dos dados.

As métricas calculadas nessa análise incluem, o Valor Preditivo Positivo (PPV), Sensibilidade, Especificidade e F1-Score, considerando um intervalo de 0 a 3 semanas a partir da emissão do EWS pelo MMAING. Isso significa que foi assumido que um EWS emitido na semana t é válido por quatro semanas consecutivas $(t, t+1, t+2, t+3)$, contando a semana de emissão. Os resultados dessas análises estão apresentados na Tabela 8.13.

O MMAING apresentou resultados satisfatórios, com um PPV de 0.74, Sensibilidade de 0.75, Especificidade de 0.93 e F1-Score de 0.75. A matriz de

Tabela 8.13: Desempenho do MMAING para surtos reais relatados pelos CIEVS para o estado do Amazonas.

	MMAING			
UF	**PPV**	**Sensibilidade**	**Especificidade**	**F1-Score**
AM	0.74	0.75	0.93	0.75

confusão é apresentada na Figura 8.10 para ilustrar a relação entre os EWS verdadeiros negativos (1º quadrante, superior esquerdo), falsos positivos (2º quadrante, superior direito), falsos negativos (3º quadrante, inferior esquerdo) e verdadeiros positivos (4º quadrante, inferior direito).

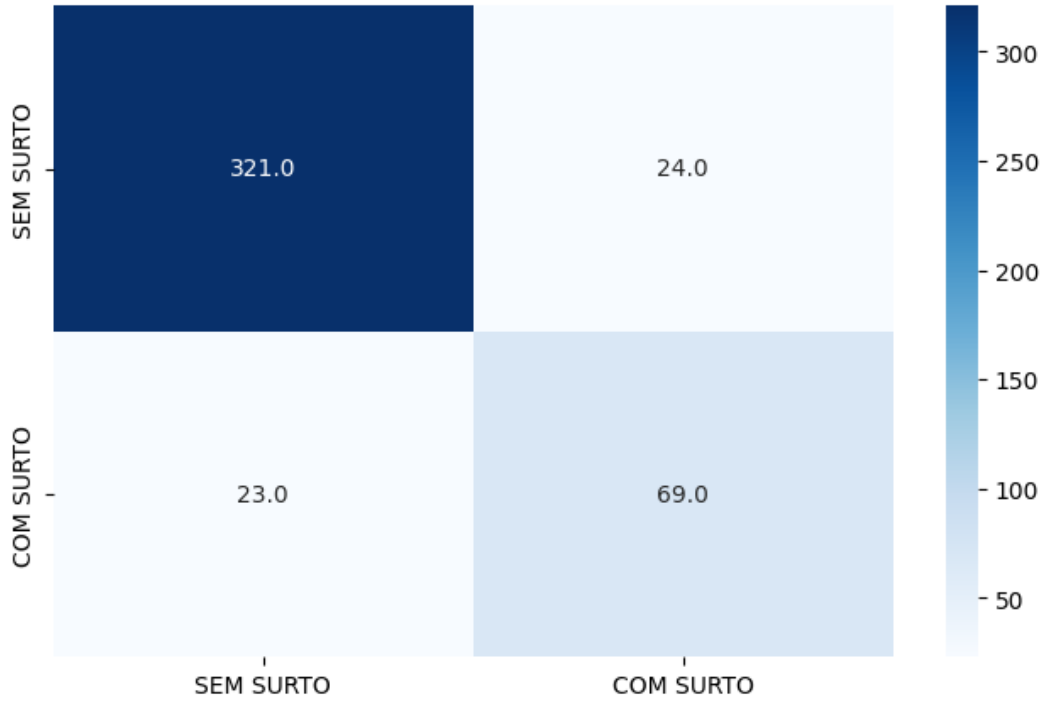


Figura 8.10: Distribuição dos acertos (VP = 69 e VN = 321) e erros (FP = 24 e FN = 23) do MMAING, conforme ilustrado na matriz em tonalidades distintas de azul.

Em 69 dessas instâncias, o MMAING emitiu EWS corretamente a ocorrência do surto, mostrando que o modelo foi eficaz em identificar esses eventos precocemente. Em 321 instâncias, o modelo também acertou ao indicar que não houve surto, confirmando que as condições estavam normais. No entanto, em 24 instâncias, o modelo emitiu um falso EWS para surto, ou seja, identificou um surto que, na realidade, não aconteceu. Já em 23 instâncias, o modelo não sinalizou os surtos que ocorreram, deixando de emitir um EWS quando deveria. Esses resultados mostram-se promissores e ressaltam a importância de continuar aprimorando o MMAING.

A matriz de confusão apresentada na Figura 8.10 reforça os resultados das métricas, demonstrando que o MMAING foi eficaz tanto na detecção precoce de surtos reais, com taxa de acerto de 75%, quanto na exclusão de instâncias sem surtos. A alta sensibilidade do modelo reflete sua capacidade de identificar a maioria dos surtos reais, com poucos falsos negativos (23 instâncias). Por outro lado, a elevada especificidade evidencia sua eficiência em minimizar falsos positivos, pois apenas 24 instâncias foram erroneamente sinalizadas com surtos. Esses resultados se traduzem em um F1-Score robusto, que equilibra bem a PPV e a sensibilidade, indicando que o MMAING mostra-se uma ferramenta eficiente e confiável para contribuir na vigilância epidemiológica do estado do Amazonas.

9
Transmissibilidade Espacial

O aumento da mobilidade humana ampliou consideravelmente os riscos epidemiológicos [201]. O mundo contemporâneo é extremamente interconectado, como evidenciado pelas extensas e diversificadas redes de transporte (terrestre, aéreo e marítimo), que continuam a se expandir em alcance, velocidade e volume de passageiros e mercadorias [201]. Pesquisas recentes indicam que a movimentação das pessoas desempenha um papel importante na disseminação de doenças infecciosas [50, 202–204]. Um exemplo notável foi a recente pandemia da COVID-19, marcada pela rápida disseminação global e pelas diversas mutações do patógeno. No Brasil o vírus emergente se disseminou pelo país muito rapidamente, alcançando regiões distantes dos grandes centros urbanos em poucas semanas [205]. Este cenário ressalta e reforça ainda mais a importância da vigilância de síndromes respiratórias para detecção precoce de circulação de novos vírus. Assim, o desenvolvimento de sistemas focados em ferramentas de alerta precoce é um requisito fundamental para aprimorar a preparação e a resposta a pandemias [31, 166].

Assim, apresentamos neste capítulo um estudo fundamentado nos princípios da vigilância sentinela, utilizando dados de fluxo de pessoas em conjunto com dados de atendimentos para IVAS na APS. O objetivo deste estudo é analisar a dinâmica de disseminação de doenças respiratórias e identificar as principais regiões que atuam como focos de propagação. Assim, aplicamos abordagens da ciência de redes e da modelagem metapopulacional de transmissão de patógenos já apresentadas nas seções 2.5 e 5.1, para investigar as interações entre RGIs selecionadas no território brasileiro.

9.1
Vigilância sentinela

A vigilância sentinela é uma abordagem comumente usada na saúde pública, concentrando-se em monitorar dados clínicos e epidemiológicos provenientes de unidades selecionadas em regiões estratégicas, para detectar precocemente a emergência local de novas doenças. O objetivo central é identificar as unidades que são mais suscetíveis e estratégicas para o controle da disseminação, enfrentando o desafio epidemiológico de detectar indivíduos ou grupos que seriam infectadas precocemente e com alta probabilidade em

deflagrar um surto infeccioso. [206–208]. Normalmente ela é implementada a partir da seleção de alguns hospitais em regiões específicas para rastrear, ou testar com mais frequência, uma infecção específica [206].

Com os avanços tecnológicos na coleta e disponibilidade de dados de saúde, os sistemas de vigilância sentinela têm ganhado destaque no mundo contemporâneo, devido ao rápido desenvolvimento desses sistemas nas grandes cidades da Europa e da América do Norte [209]. Esse movimento foi especialmente impulsionado após a pandemia de H1N1 em 2009, quando diversos países implementaram redes sentinelas de monitoramento para doenças infecciosas agudas, com atenção especial às doenças respiratórias [210]. No contexto europeu, desde 2015, todos os estados membros do Centro Europeu de Prevenção e Controle de Doenças (European Centre for Disease Prevention and Control – ECDC) reportam dados de vigilância de influenza sazonal a partir de casos de síndrome gripal na atenção primária [211]. Além das unidades sentinela, muitos países utilizam também dados de notificação voluntária de unidades não sentinela para monitoramento situacional e identificação viral. Nos Estados Unidos, o CDC adota uma abordagem semelhante, usando vigilância sentinela de síndrome gripal para monitorar infecções respiratórias que requerem atendimento ambulatorial [212].

No Brasil, foi instaurada a Rede Sentinela de Síndrome Gripal que é parte de um sistema de vigilância epidemiológica criado em 2000 para monitorar a circulação dos vírus respiratórios no país. Ela tem foco especial na vigilância do vírus influenza, por meio da identificação da circulação dos vírus, de acordo com a patogenicidade e virulência em cada período sazonal, a existência de situações inusitadas ou o surgimento de novo subtipo viral. Por conseguinte, a rede sentinela fornece informações para a detecção precoce de novos vírus, guia a adequação da vacina da influenza sazonal, bem como o monitoramento dos vírus respiratórios já circulantes [213].

Essa rede é composta atualmente por 314 unidades sentinelas de saúde distribuídas em todas as regiões do Brasil, que coletam e analisam dados sobre atendimentos de casos de síndrome gripal (SG) e síndrome respiratória aguda grave (SRAG). A vigilância se baseia na amostragem de pacientes com sintomas gripais, cujas amostras são enviadas para laboratórios de referência para identificação viral. Esses dados são registrados no Sistema de Vigilância Epidemiológica da Gripe (SIVEP-Gripe), integrado ao Departamento de Informática do SUS, denominado DATASUS.

Atualmente, além das atividades de rotina para vigilância de influenza e outros vírus respiratórios, as unidades sentinelas incorporaram atividades para

a notificação do vírus SARS-CoV-2 na sua rotina [213].

No entanto, essa abordagem ainda enfrenta desafios, como a cobertura geográfica insuficiente das unidades sentinelas e a seleção de suas localizações muitas vezes baseada em fatores como infraestrutura preexistente, disponibilidade de orçamento ou conveniência política, o que limita sua representatividade [202, 206, 214].

9.1.1 Rede sentinela na Bahia

Na Bahia, a primeira unidade sentinela foi implantada em 2013, no município de Salvador, e desde então a rede estadual, que é uma sub-rede da rede nacional de vigilância sentinela da síndrome gripal, tem sido continuamente ampliada. Em 2022, o número de unidades sentinelas aumentou de 5 para 12, distribuídas entre 7 Núcleos Regionais de Saúde (NRS), conforme demonstrado na Tabela 9.1. Essa expansão promove um maior conhecimento epidemiológico sobre a circulação viral em diferentes regiões do território baiano.

Tabela 9.1: Distribuição de unidades sentinelas na Bahia

Núcleo Regional de Saúde	**Município Sentinela**	**Unidades**
Nordeste	Alagoinhas	1
Oeste	Barreiras	1
Centro - Leste	Feira de Santana	1
Sul	Ilhéus	1
Norte	Juazeiro	1
Extremo - Sul	Porto Seguro	1
Leste	Salvador	5
Leste	Santo Antônio de Jesus	1

Os critérios usados pelas autoridades responsáveis para a seleção das unidades sentinelas vão desde parâmetros populacionais, como locais com maior concentração e fluxo de pessoas, estratégicos para a vigilância de eventos locais como a introdução de novos agentes infecciosos ou subtipos de influenza, até características específicas dos serviços de saúde. Esses critérios incluem:

- Locais com maior concentração e fluxo de pessoas;
- Serviços de saúde com demanda espontânea e com atendimento 24 horas (exemplo: pronto-atendimento, emergência e ambulatório);
- Serviços de saúde que preferencialmente atendam a todas as faixas etárias, sem priorizar determinadas especialidades;

- Locais voltados para a vigilância por motivos de população de trabalhadores de granjas e frigoríficos que produzem ou abatem aves e suínos;
- Número de atendimentos por síndrome gripal com importância epidemiológica;
- Unidades de saúde públicas e privadas;
- Hospitais com Núcleos de Epidemiologia.

No entanto, vale ressaltar que esses critérios não estão fundamentados em estudos científicos específicos e aprofundados, mas sim em parâmetros básicos e operacionais [215]. Essa abordagem, embora prática, pode limitar a eficácia da vigilância ao desconsiderar uma análise mais detalhada dos fatores locais de natureza epidemiológica e social que influenciam a propagação de doenças.

O Governo da Bahia, em parceria com os municípios, tem buscado continuamente expandir a rede sentinela, conforme regulamentado pela Portaria GM/MS 183/2014. Esta portaria estabelece o incentivo financeiro para a implantação e manutenção de ações e serviços públicos estratégicos de vigilância em saúde, conforme previsto no art. 18, inciso I, da Portaria nº 1.378/GM/MS, de 9 de julho de 2013. Além disso, ela define os critérios para financiamento, monitoramento e avaliação dessas ações. O governo planeja instalar novas unidades sentinelas nas regiões sudoeste e centro-norte do estado, fortalecendo a cobertura e capacidade de vigilância em todo o estado [215].

9.2 Dados e Métodos

Nesta seção, detalhamos os conjuntos de dados utilizados e a metodologia aplicada neste estudo, com base em critérios de avaliação específicos.

9.2.1 Dados de mobilidade: Rodoviários e Aquaviários

O estudo utilizou dados de mobilidade de pessoas em larga escala, obtidos de fontes oficiais do governo, com o objetivo de analisar as dinâmicas de mobilidade por meio das redes rodoviária e fluvial.

Os dados de mobilidade rodoviária e fluvial, que abrangem todo o território brasileiro, referem-se ao ano de 2016 e incluem as frequências de tráfego de veículos de transporte de passageiros, coletados e disponibilizados pelo Instituto Brasileiro de Geografia e Estatística (IBGE) [216].

A metodologia adotada pelo IBGE para a construção desse banco de dados baseou-se em diversas fontes, incluindo pesquisas em terminais rodoviários e hidroviários, bilheterias, paradas de ônibus municipais, cooperativas de transporte aquaviário, operadores autônomos e comunicações diretas com empresas de transporte.

Com o objetivo de garantir uma cobertura mais abrangente, foram considerados também os modos de transporte alternativos e informais, como vans, peruas e micro-ônibus. Para padronizar os dados e viabilizar comparações entre diferentes modais de transporte, as frequências foram convertidas para uma métrica comum. A frequência de tráfego foi obtida por meio da soma do número de partidas semanais entre cada par de municípios. Quando os dados estavam disponíveis apenas em bases quinzenais ou mensais, os valores foram ajustados por meio de multiplicadores (0,5 para dados quinzenais e 0,25 para mensais), de modo a estimar a frequência semanal [216].

O ônibus foi adotado como unidade de referência (valor 1). As frequências de veículos menores, como vans e carros, foram ajustadas por um fator de 0,25. Para embarcações fluviais, aplicaram-se multiplicadores específicos, de forma a equivaler sua frequência à dos ônibus [216].

Para integrar esses dados ao presente estudo, estimou-se o número de passageiros multiplicando-se a frequência total de tráfego por um fator de 40, correspondente ao número médio de ocupantes de um ônibus. Como os dados disponibilizados representam o acumulado anual, os valores foram divididos por 52 para refletir o fluxo médio semanal. Por fim, os dados foram agregados segundo as RGIs.

Portanto, os dados utilizados contêm informações sobre a frequência semanal estimada de passageiros entre pares de regiões imediatas, calculadas com base no acúmulo de partidas ao longo do ano de 2016, o que possibilita uma análise detalhada da mobilidade nesses modos de transporte.

A base de dados original pode ser acessado em: `https://tinyurl.com/ligacoes-rodoviarias-2016`.

Além dos dados de mobilidade, este estudo também incorpora dados sindrômicos da APS, os quais foram discutidos em detalhes na Seção 7.1.

9.2.2 Desenho do estudo

A intensificação da globalização e as mudanças nos padrões de circulação humana, acompanhadas pelo aumento do risco de introdução e disseminação de novos vírus e variantes, tornam oportuno o estudo da rede sentinela atualmente proposta no Brasil, a fim de identificar suas robustezes e fragilidades, que possam subsidiar novos desenhos.

Conforme discutido na Seção 2.5, a abordagem metapopulacional permite modelar a propagação de doenças infecciosas em um contexto espacial, considerando o fluxo de indivíduos entre populações interconectadas por redes de mobilidade. A aplicação desse formalismo requer dados sobre a estrutura espacial das populações, os padrões de movimentação entre elas, bem como dados epidemiológicos.

Nesse contexto, conduzimos um estudo de caso focado no estado da Bahia, que é politicamente dividido em 417 municípios. Esses municípios, conforme critérios estabelecidos pelo IBGE, estão agrupados em 34 RGIs conforme explicitado na Tabela 9.2. A Bahia é o maior estado da região Nordeste e o quinto maior do Brasil, com uma extensão territorial de aproximadamente 564.733 km^2, representando cerca de 7% do território nacional. Além disso, é o quarto estado mais populoso do país, com aproximadamente 14.141.626 habitantes, segundo o Censo IBGE de 2022 [217].

A análise aqui apresentada se baseia em dados que foram organizados por RGIs. Assim, a partir dos dados de mobilidade organizados, foi gerado uma rede ponderada a partir de dados de transporte rodoviário-aquaviário. A rede é composta por 34 nós, representando as RGIs, e as arestas representam a existência de fluxo de pessoas entre as RGIs, sendo o peso delas o número total de pessoas deslocadas. Para garantir a simetria das redes, consideramos o fluxo total entre duas RGIs como a soma dos fluxos de ida de A para B e de B para A, presumindo que o fluxo é pendular [50], ou seja, não envolve deslocamentos permanentes.

Tabela 9.2: Detalhamento das RGIs do estado da BA.

Código	Nome	Sigla	Nº Municípios	População
290001	Salvador	SA	16	4064880
290002	Alagoinhas	Al	17	507494
290003	Santo Antônio de Jesus	SJ	14	295278
290004	Cruz das Almas	CA	12	288664
290005	Valença	Vç	8	257665
290006	Nazaré – Maragogipe	N-M	7	184245
290007	Ilhéus – Itabuna	I-I	22	650652
290008	Teixeira de Freitas	TF	13	458167
290009	Eunápolis - Porto Seguro	E-P	8	387910
290010	Camacan	Cn	8	132940
290011	Vitória da Conquista	VC	30	815516
290012	Jequié	Jq	16	344134
290013	Brumado	Bd	12	228008
290014	Ipiaú	Ip	13	218247
290015	Itapetinga	Ig	6	153712
290016	Guanambi	Gb	24	480683
290017	Bom Jesus da Lapa	BJ	7	233847
290018	Barreiras	Br	17	515348
290019	Santa Maria da Vitória	SM	7	138306
290020	Irecê	Ic	19	409863
290021	Xique-Xique – Barra	X-B	10	222932
290022	Juazeiro	Jz	9	521703
290023	Senhor do Bonfim	SB	9	299084
290024	Paulo Afonso	PA	7	197492
290025	Ribeira do Pombal	RP	7	205704
290026	Euclides da Cunha	EC	5	189690
290027	Cícero Dantas	CD	6	129375
290028	Jeremoabo	Jm	5	97955
290029	Feira de Santana	FS	33	1241854
290030	Jacobina	Jb	16	325532
290031	Itaberaba	Ib	12	223359
290032	Conceição do Coité	CC	7	196507
290033	Serrinha	Sr	5	191400
290034	Seabra	Sb	10	177138

Em seguida, é executado um estudo na rede de mobilidade para explorar a estrutura modular por meio de métodos de detecção de comunidades, aplicando ao final do processo o algoritmo conhecido de Newman-Girvan (NG) [160].

Além disso, realizamos estudos de centralidade de intermediação para identificar RGIs críticas no espalhamento. Esses métodos nos permitiram entender a organização estrutural das redes e destacar as RGIs que exercem um papel central na difusão de doenças entre regiões.

Paralelamente, aplicamos aos dados das 34 RGIs um modelo SIR metapopulacional [50], que combina os dados de atendimentos IVAS da APS com os dados de fluxos de maneira integrada. Essa abordagem permitiu analisar como uma RGI exporta potenciais infecções respiratórias para outras, utilizando o fluxo de pessoas pelas rodovias entre RGIs como vetor de transmissão. O modelo SIR metapopulacional, detalhado na seção 2.5, fornece uma visão intrínseca de como a doença se espalha espacialmente, favorecendo a identificação de possíveis locais que podem iniciar um surto.

Este formalismo permite obter resultados para $\mathcal{T}_{ij}(t)$, que representa a taxa de novas infecções em i causadas por indivíduos previamente infectados em j, em sua forma discreta:

$$\mathcal{T}_{ij}(t) = R_{ij}(t) \sum_{\tau=0}^{t} g_{ij}(\tau) B_j(t-\tau)\Delta t, \tag{9-1}$$

onde $B_j(t)$ representa o número de novas infecções (casos) na região j entre t e $t + \Delta t$; $R_{ij}(t)$ é o número médio de novas infecções, durante o período infeccioso, causadas por um indivíduo infectado em j a suscetíveis em i; e $g_{ij}(t, \tau)$ representa a distribuição do intervalo de geração entre a infecção de um indivíduo e a infecção secundária subsequente.

Somando a expressão 9-1 sobre todas as regiões j vizinhas ao nó i, obtém-se o número total de novas infecções em i no tempo t:

$$B_i(t) = \sum_{j}^{n} \mathcal{T}_{ij}(t). \tag{9-2}$$

A expressão 9-1 representa uma forma geral da versão discreta da equação de renovação, cujo desenvolvimento é detalhado no trabalho de Jorge et al. [50].

Neste trabalho, para poder usar dados sindrômicos, é necessário estender a interpretação da Eq. 9-1, definindo $\widehat{\mathcal{T}}_{ij}(t)$ de forma análoga a $\mathcal{T}_{ij}(t)$ para os dados de sindrômicos APS, $H(t)$, com o objetivo de identificar hubs, que funcionam como pontos centrais de transmissão. Nesta formulação estendida, $\widehat{\mathcal{T}}_{ij}(t)$ representa o número estimada de potenciais infecções respiratórias entre i e j, usando dados sindrômicos $H(t)$ como um proxy para $B(t)$.

9.2.3
Índice Sentinela e Análise Integrada da Vigilância

Para encerrar esta seção, voltamos nossa atenção para outra contribuição deste trabalho: dado um conjunto de centros populacionais interconectados (municípios, RGIs, estados, etc.), introduzimos uma nova métrica, denominada *Índice Sentinela* (IS), para avaliar a adequação de cada elemento desse conjunto para sediar uma unidade sentinela com o objetivo de otimizar a vigilância em saúde. Isso é feito por meio de uma combinação de resultados quantitativos apresentados nas subseções anteriores, os quais se baseiam na detecção de comunidades em redes, na centralidade de intermediação e em um modelo metapopulacional. Esses três conceitos e as ferramentas derivadas deles não apenas permitem o desenho de uma rede sentinela, mas também possibilitam avaliar o desenho atual de redes sentinelas existentes, potencialmente sugerindo novas diretrizes para sua expansão e/ou reformulação.

Mais especificamente, para um determinado centro populacional j, IS_j é definido como a média geométrica entre três medidas: i) a média temporal do número total normalizado de potenciais infecções exportadas $\langle \sum_i \widehat{\mathcal{T}}_{ij}(t) \rangle$, ii) o valor da centralidade de intermediação normalizado (b_j) e iii) a população normalizada (p_j) de j, ou seja,

$$IS_j = \left((b_j)(p_j) \left\langle \sum_i \widehat{\mathcal{T}}_{ij}(t) \right\rangle \right)^{1/3} . \tag{9-3}$$

A média geométrica é a única medida de tendência central que preserva a proporcionalidade entre variáveis, sendo especialmente útil para combinar grandezas de escalas distintas sem exigir padronização prévia [218]. Sua estrutura multiplicativa a torna apropriada para sintetizar métricas que, embora distintas em natureza, como centralidade de rede, população e exportação de potenciais infecções, possuem relevância equivalente no fenômeno avaliado.

A adequação de cada centro populacional como unidade sentinela pode ser avaliada levando-se em conta o ranking resultante e um limite pré-estabelecido de acordo com as necessidades locais de vigilância.

Como critério de avaliação, consideramos o conjunto de RGIs do estado da Bahia e comparamos as unidades sentinelas propostas com aquelas já estabelecidas na Rede Sentinela para Síndrome Gripal no mesmo estado. Além disso, discutimos possíveis formas de contribuição dessas unidades para aumentar a efetividade da vigilância.

9.3
Resultados e Discussão

Nesta seção, apresentamos os principais resultados obtidos através da abordagem proposta, e investigamos a robustez e fraquezas da rede sentinela implantada no estado da Bahia.

9.3.1
Estudo da Rede Sentinela da Bahia

A rede de mobilidade rodoviária-aquaviária aqui construída para o estado da Bahia é formada por 34 vértices, cada um correspondendo a uma. Suas conexões refletem o número total de pessoas que se deslocam entre cada par de RGI, sendo representada por uma matriz ponderada e simétrica. A partir de métodos de estudo estrutural, podemos inicialmente explorar a configuração modular da rede, empregando limiares ótimos (σ) para definir a mínima conexão entre os vértices.

Para determinar o valor ótimo do limiar σ na rede, e proceder à determinação de comunidades, gera-se um gráfico que relaciona a dissimilaridade $d(\sigma, \sigma + \delta\sigma)$ entre as matrizes de vizinhança de duas redes obtidas para valores próximos de σ, de acordo com a equação 5-15 na seção 5.1. Ao traçarmos $d(\sigma, \sigma + \delta\sigma)$ como função de σ, verifica-se que o gráfico é caracterizado pela presença de picos agudos, apresentado na figura 9.1. Esses picos indicam uma mudança na estrutura modular da rede.

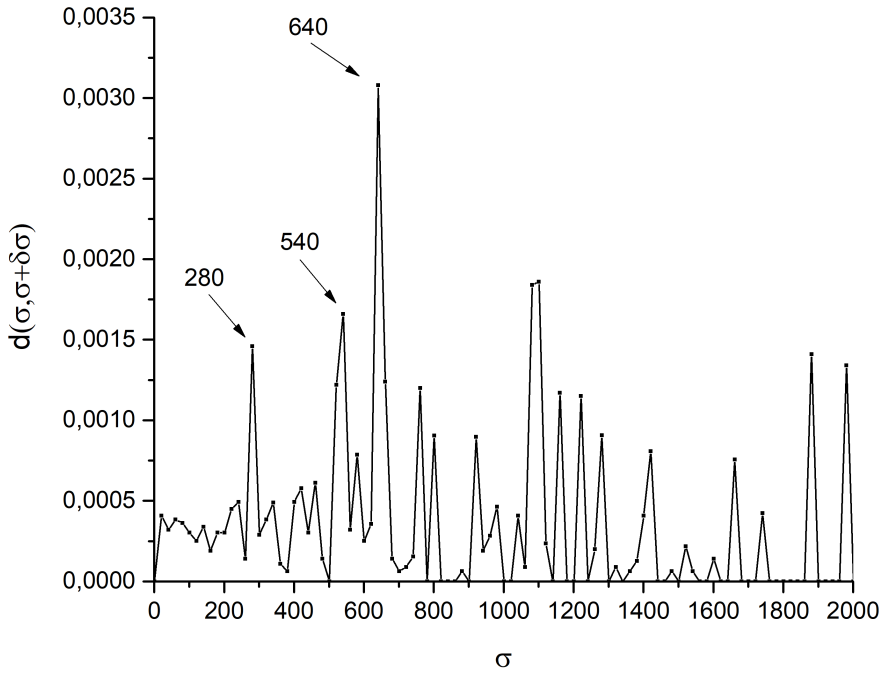


Figura 9.1: Exemplo para ilustrar a determinação do valor ótimo do limiar.

Foram avaliados três valores para σ (280,540,640), escolhidos por representarem pontos relevantes de mudança na estrutura da rede. A partir desses valores, foram geradas três matrizes de adjacência, conforme descrito na Seção 5.1. A matriz de adjacência $M(\sigma)$ é definida pela seguinte relação:

$$M(\sigma)_{ij} = \begin{cases} 1 & \text{se peso entre } i \text{ e } j \text{ exceder } \sigma \\ 0 & \text{caso contrário} \end{cases} \tag{9-4}$$

Observou-se que, nos limiares de 540 e 640, a rede já apresentava um nível alto de desconexão, com a presença de diversos nós isolados. Esse comportamento evidencia uma redução na conectividade da rede, comprometendo a interação entre seus componentes. Diante desse efeito, valores de σ superiores não foram analisados, pois tais configurações inviabilizariam uma avaliação da dinâmica da rede. Por outro lado, para $\sigma = 280$ e valores ligeiramente inferiores, verificou-se que todos os nós, com exceção de um, permaneciam conectados por aproximadamente 226 arestas, proporcionando uma condição adequada para a continuidade da análise. A Figura 9.2 ilustra a configuração da rede utilizada no estudo.

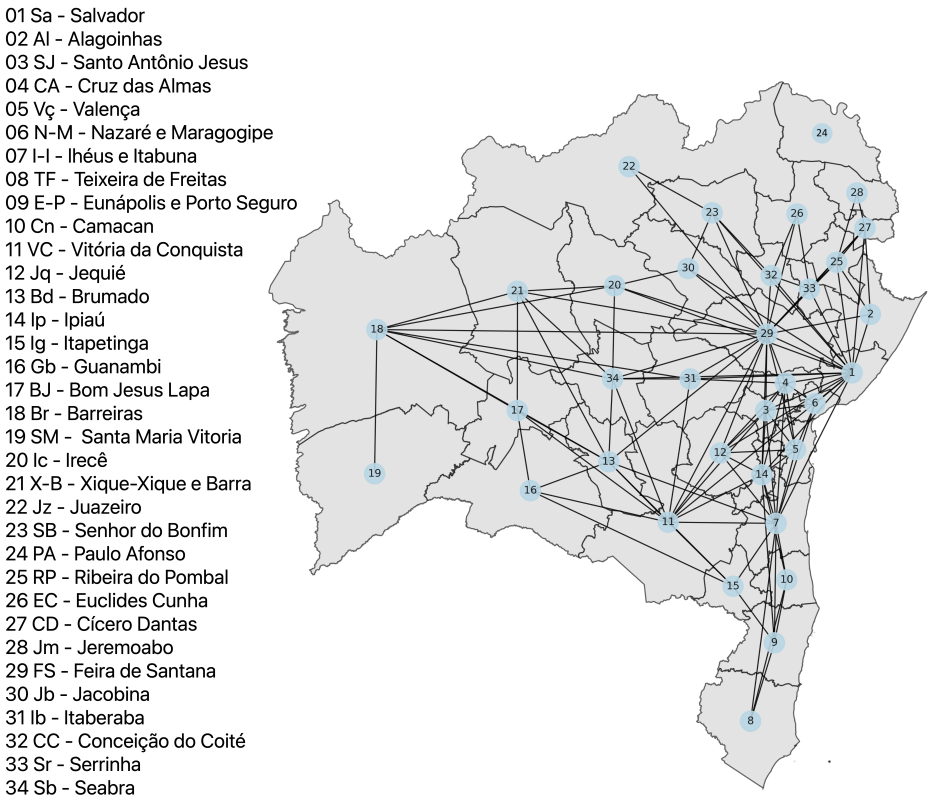


Figura 9.2: Rede de mobilidade centrada no mapa geográfico da Bahia para $\sigma = 280$, composta de 34 nós e 226 arestas.

A partir da rede construída e de sua respectiva matriz de adjacência, $M(280)$, aplicou-se o método de detecção de comunidades de Newman-Girvan (NG) [160], que identifica comunidades com base na centralidade de intermediação das arestas. O algoritmo remove iterativamente as arestas mais centrais, aquelas que mais frequentemente atuam como pontes nos caminhos mais curtos entre pares de nós, promovendo a fragmentação progressiva da rede. Esse processo gera uma divisão hierárquica representada em um dendrograma, no qual cada divisão corresponde à formação de uma nova comunidade. Como resultado, foram identificadas quatro comunidades, com características espaciais coerentes com a geografia do estado e com a esperada concentração de fluxo entre cidades mais próximas, conforme ilustrado na Figura 9.3.

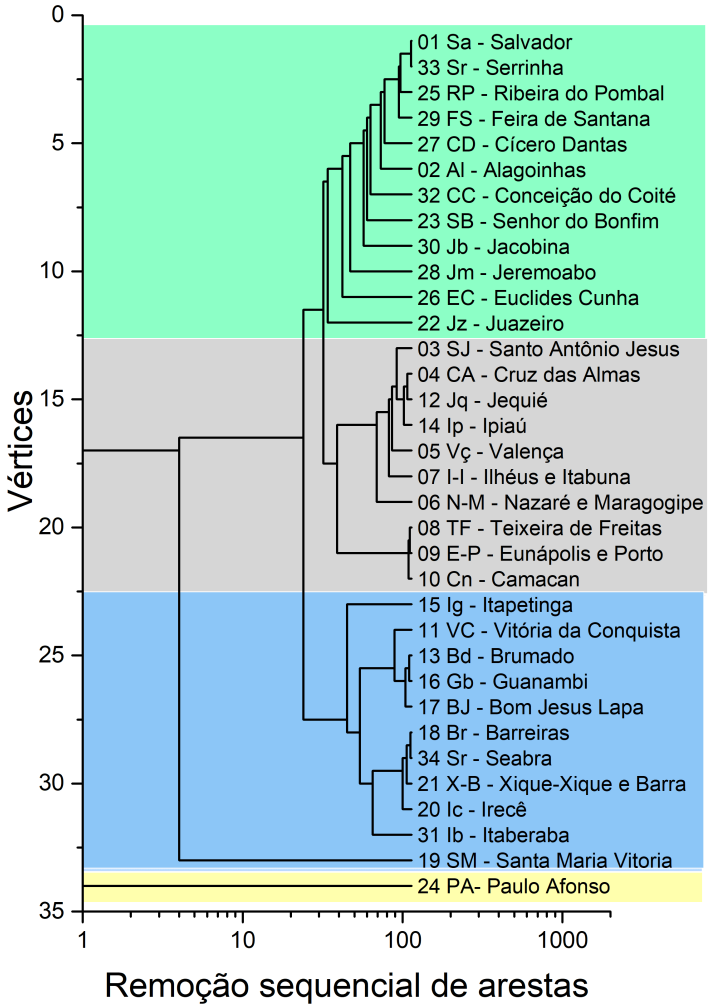


Figura 9.3: Dendrograma da rede de mobilidade rodoviária da Bahia, demonstrando as comunidades: C1 em verde, C2 em cinza, C3 em azul e C4 em amarelo.

A comunidade C1 destacada em cor verde, englobando as RGIs de Salvador e Feira de Santana, abrange grande parte das regiões norte e nordeste do estado. A comunidade C2 cobre a região sul e parte do leste do estado, destacada em tons cinza, e pode ser subdividida em duas áreas distintas: C2.1 que abrange a região leste-sul e a C2.2 que se estende até o extremo sul, conforme ilustrado na Figura 9.4. Já a comunidade C3 ocupa as regiões oeste e sudoeste do estado, destacada em tons azul, podendo também ser subdividida em duas áreas: C3.1 em torno da RGI de Vitória da Conquista (sudoeste), e C3.2 em torno da RGI de Barreiras (oeste), conforme ilustrado na Figura 9.4. Por fim, a comunidade C4 que isola a RGI de Paulo Afonso, indicando ter baixa conectividade com o estado da Bahia.

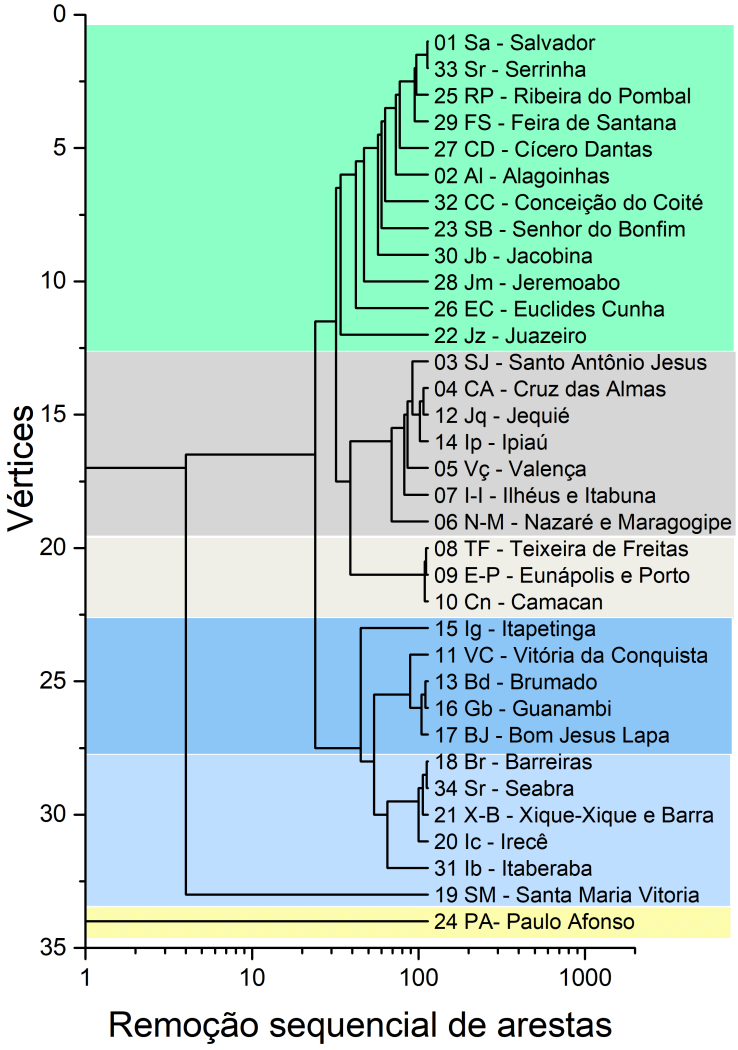


Figura 9.4: Dendrograma da rede de mobilidade rodoviária da Bahia, demonstrando as 4 comunidades, com o detalhamento de subcomunidades em C2 e C3.

Na figura 9.5, ilustramos o detalhamento das seis comunidades obtidas no dendrograma (Figura 9.4) apresentadas no mapa geográfico da Bahia, dividido em 34 RGIs. Essas comunidades estão destacadas por diferentes cores, que traduzem a conectividade, refletindo a proximidade geográfica entre as RGIs dentro da mesma comunidade.

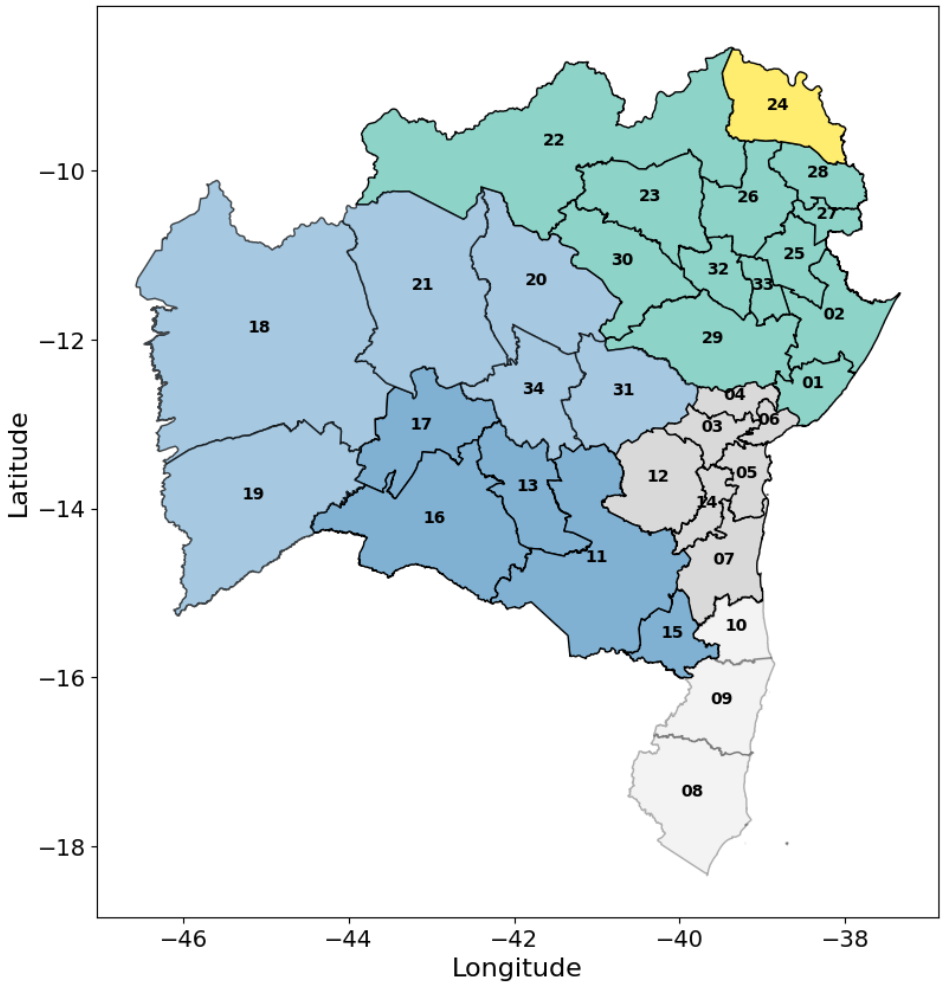


Figura 9.5: Mapa da Bahia dividido pelas 34 RGIs, com comunidades destacadas em diferentes cores: C1 em verde, C2.1 em cinza escuro, C2.2 em cinza claro, C3.1 em azul escuro, C3.2 em azul claro e C4 em amarelo.

Após concluir o estudo das comunidades na rede de mobilidade rodoviária da Bahia, procedemos à análise das medidas de centralidade de intermediação.

Os dois nós com maior medida de centralidade de intermediação foram 29 e 1, correspondendo, respectivamente, a Feira de Santana e Salvador. Esses nós se destacam por apresentarem os maiores valores de centralidade: 0,35 para Feira de Santana e 0,19 para Salvador, refletindo a importância dessas RGIs como hubs centrais de tráfego e interação de pessoas no estado.

Além dos nós 29 e 01, os nós destacados em vermelho (07, 11, 13, 14 e 18) também demonstram sua importância estratégica dentro da rede, apresentando valores de centralidade superiores a 0.05, conforme ilustrado na figura 9.6.

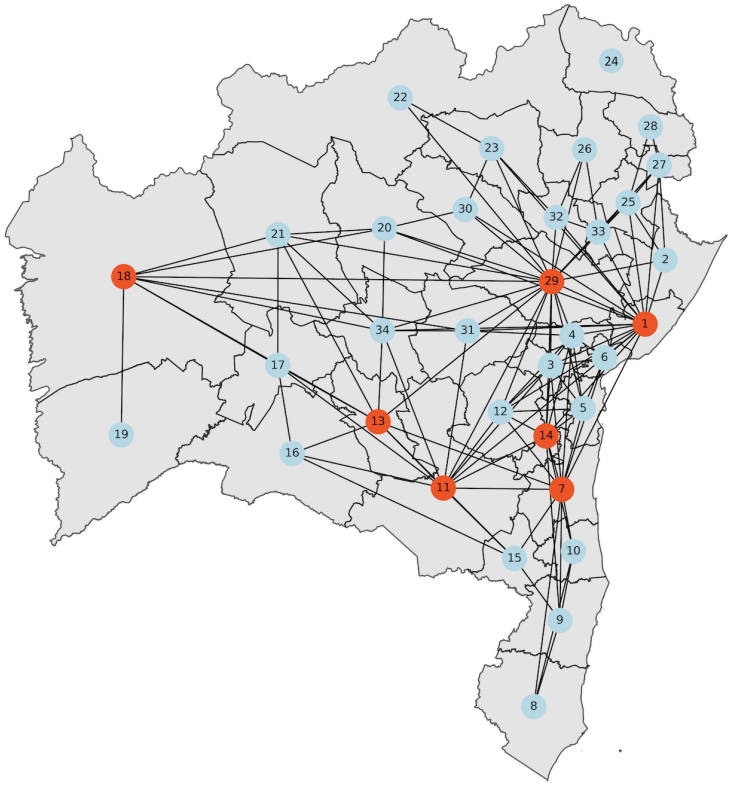


Figura 9.6: Representação da rede de mobilidade rodoviária da Bahia, indicando o nós e destacando em vermelho os nós com centralidade de intermediação maior que $0,05$.

A definição desse limiar foi fundamentada em critérios estatísticos e funcionais. Observa-se que a distribuição da centralidade de intermediação é fortemente assimétrica e dominada por nós com baixa centralidade, conforme ilustrado na Figura 9.7. Dado esse comportamento, o uso de um ponto de corte empírico em 0,05, valor situado acima da mediana e da média, permite isolar um subconjunto de cerca de 20% do total de vértices com maior influência estrutural na rede.

Do ponto de vista funcional, os vértices com centralidade de intermediação superior a 0,05 englobam as capitais regionais do estado [219], caracterizadas por elevada conectividade e importância socioeconômica em suas respectivas regiões, o que demonstra coerência com a realidade territorial da Bahia, conforme ilustrado na Figura 9.6. Em contraste, os demais nós exibiram valores de centralidade menores, reforçando a relevância dos nós destacados, conforme apresentado na Tabela 9.3).

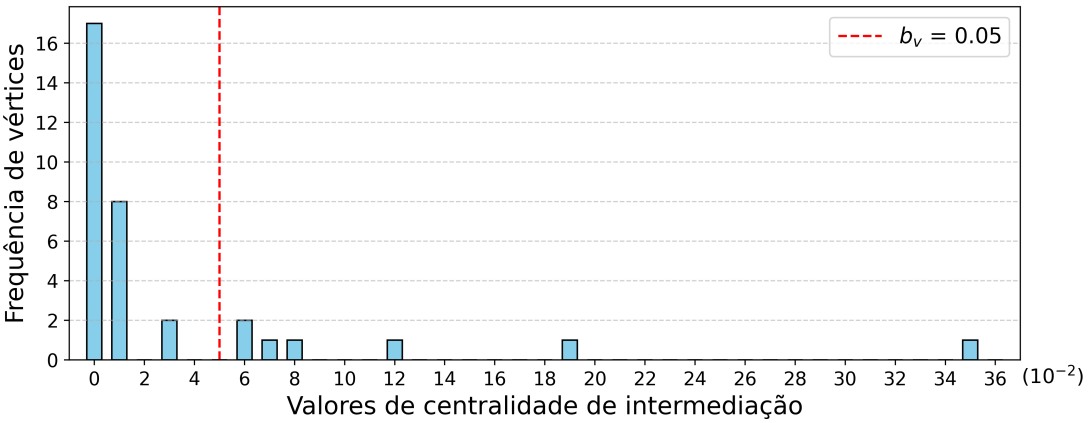


Figura 9.7: Distribuição da centralidade de intermediação dos vértices da rede. O traço vermelho pontilhado indica o limiar adotado de 0,05.

Tabela 9.3: Valores de centralidade de intermediação (b_v) para os vértices da rede rodoviária da Bahia.

Nó	RGI	b_v	Nó	RGI	b_v
01	Salvador	0.19	18	Barreiras	0.07
02	Alagoinhas	0.00	19	Santa Maria da Vitória	0.00
03	Santo Antônio de Jesus	0.01	20	Irecê	0.01
04	Cruz das Almas	0.01	21	Xique-Xique e Barra	0.01
05	Valença	0.00	22	Juazeiro	0.00
06	Nazaré e Maragogipe	0.00	23	Senhor do Bonfim	0.00
07	Ilhéus e Itabuna	0.12	24	Paulo Afonso	0.00
08	Teixeira de Freitas	0.00	25	Ribeira do Pombal	0.03
09	Eunápolis e Porto Seguro	0.01	26	Euclides da Cunha	0.00
10	Camacan	0.00	27	Cícero Dantas	0.03
11	Vitória da Conquista	0.08	28	Jeremoabo	0.00
12	Jequié	0.00	29	Feira de Santana	0.35
13	Brumado	0.06	30	Jacobina	0.00
14	Ipiaú	0.06	31	Itaberaba	0.01
15	Itapetinga	0.01	32	Conceição do Coité	0.00
16	Guanambi	0.00	33	Serrinha	0.01
17	Bom Jesus da Lapa	0.00	34	Seabra	0.01

As unidades sentinelas podem ser selecionadas com base em propriedades específicas da rede [153]. Os métodos de análise de redes aplicados neste estudo

têm demonstrado potencial para detecatar riscos epidemiológicos e identificar unidades sentinelas ideais para intervenções [220].

Além disso, Colman et al. [152] destacam que uma estratégia de vigilância que distribui sentinelas em diferentes regiões é mais eficaz em redes com alta modularidade ou quando há uma estrutura espacial bem definida. Já em redes com alta heterogeneidade de grau, a escolha de nós altamente conectados em relação aos demais, tende a ser mais adequada, pois esses nós podem facilitar a detecção precoce e a disseminação de informações de vigilância.

Verificamos que a rede de mobilidade apresenta uma estrutura comunitária alinhada à realidade geográfica da Bahia, além de ter alta heterogeneidade de grau, conforme evidenciado na Figura 9.8, características que são fundamentais para estratégias eficazes de vigilância sentinela. A modularidade permite a seleção de unidades sentinelas por RGI, dentro das comunidades, garantindo uma cobertura representativa no estado. Já a alta heterogeneidade de grau indica que nós altamente conectados são bons candidatos para detecção precoce e disseminação eficiente de informações epidemiológicas.

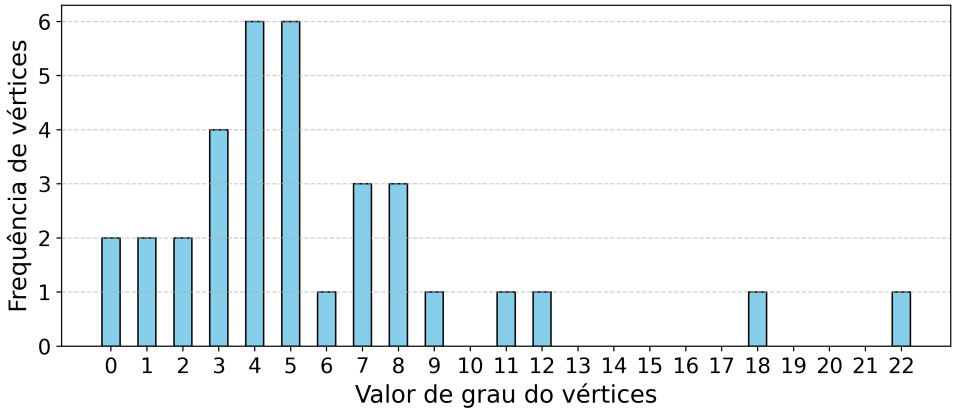


Figura 9.8: Distribuição de graus dos vértices da rede.

Para aprofundar a compreensão das dinâmicas de disseminação, aplicamos o modelo SIR metapopulacional de forma independente a cada uma das comunidades detectadas com base nos dados de mobilidade e síndromes respiratórias. Usando o análogo a equação 9-1,

$$\widehat{\mathcal{T}}_{ij}(t) = \widehat{R}_{ij}(t) \sum_{\tau=0}^{t} g_{ij}(\tau) H_j(t-\tau) \Delta t, \tag{9-5}$$

podemos acessar a contribuição de cada RGI para a disseminação das síndromes respiratórias que ocorrem em cada comunidade individualmente. Assim, ao dividir $\widehat{\mathcal{T}}_{ij}(t)$ por $H_i(t)$ em cada etapa de tempo e calcular a média

temporal, obtém-se a contribuição média de j sobre o número de infecções em i, expressa como fração do total de potenciais infecções. Essa exportação média está representada na subfigura (a) das Figuras 9.9, 9.10 e 9.11. Observa-se um comportamento autoctóctone elevado na transmissão de doenças respiratórias para todas as comunidades, indicando que a maior influência na geração de doenças respiratórias de uma RGI é sobre ela mesma, ou seja, a maioria das potenciais infecções geradas em uma RGI é causada por seus próprios habitantes, o que já era esperado. No entanto, também identificamos RGIs onde as potenciais infecções geradas em outras RGIs contribuíram para a disseminação dentro de uma mesma comunidade.

A seguir, detalhamos os resultados obtidos para cada comunidade:

A Figura 9.9, referente à comunidade C1, apresenta a proporção semanal de transmissão de potenciais infecções respiratórias entre as 12 RGIs que compõem essa comunidade. A matriz de espalhamento das infecções respiratórias (Fig. 9.9 a), mostra as proporções de exportações entre 12 RGIs. Os valores diagonais representam a proporção de potenciais infecções gerados na própria RGI, enquanto os valores fora da diagonal indicam a exportação para outras RGIs. A escala de cores reflete a intensidade da exportação, destacando as maiores taxas em azul. O diagrama de cordas (Fig. 9.9 b) ilustra como ocorre as exportações com taxa superior a 0,7%. Nota-se que a RGI de Feira de Santana (29) é o principal gerador de infecções respiratórias dentro da comunidade C1 (Fig. 9.9 c), com altas taxas de exportação para Serrinha (33) e Conceição do Coité (32), o que é coerente com a proximidade geográfica dessas RGIs.

Da mesma forma, a Figura 9.10, referente à comunidade C2, mostra a taxa semanal de potenciais infecções respiratórias entre as 10 RGIs que formam essa comunidade. A matriz de espalhamento (Fig. 9.10 a) apresenta todas as taxas de potenciais infecções respiratórias exportadas, enquanto o diagrama de cordas (Fig. 9.10 b) destaca exportações com taxas superiores a 0,7%. Cinco RGIs se destacam como principais geradores em C2: Ilhéus-Itabuna (7) com 31 gerações, Jequié (12) com 24, Cruz das Almas (4) com 22 e Santo Antônio de Jesus (3), Valença (5) e Ipiaú (14) todos com 20 gerações (Fig. 9.10 c).

Por fim, a Figura 9.11, referente à comunidade C3, apresenta a taxa semanal de exportação de potenciais infecções respiratórias dessa comunidade. A matriz (Fig. 9.11 a) detalha todas as taxas de potenciais infecções respiratórias exportadas entre as 11 RGIs, e o diagrama de cordas 9.11 b) destaca exportações acima de 0,7%. A RGI de Vitoria da Conquista (11) e Barreiras (13) se destacam como as principais geradoras de potenciais infecções respiratórias dentro da comunidade, reforçando seus papeis estratégicos.

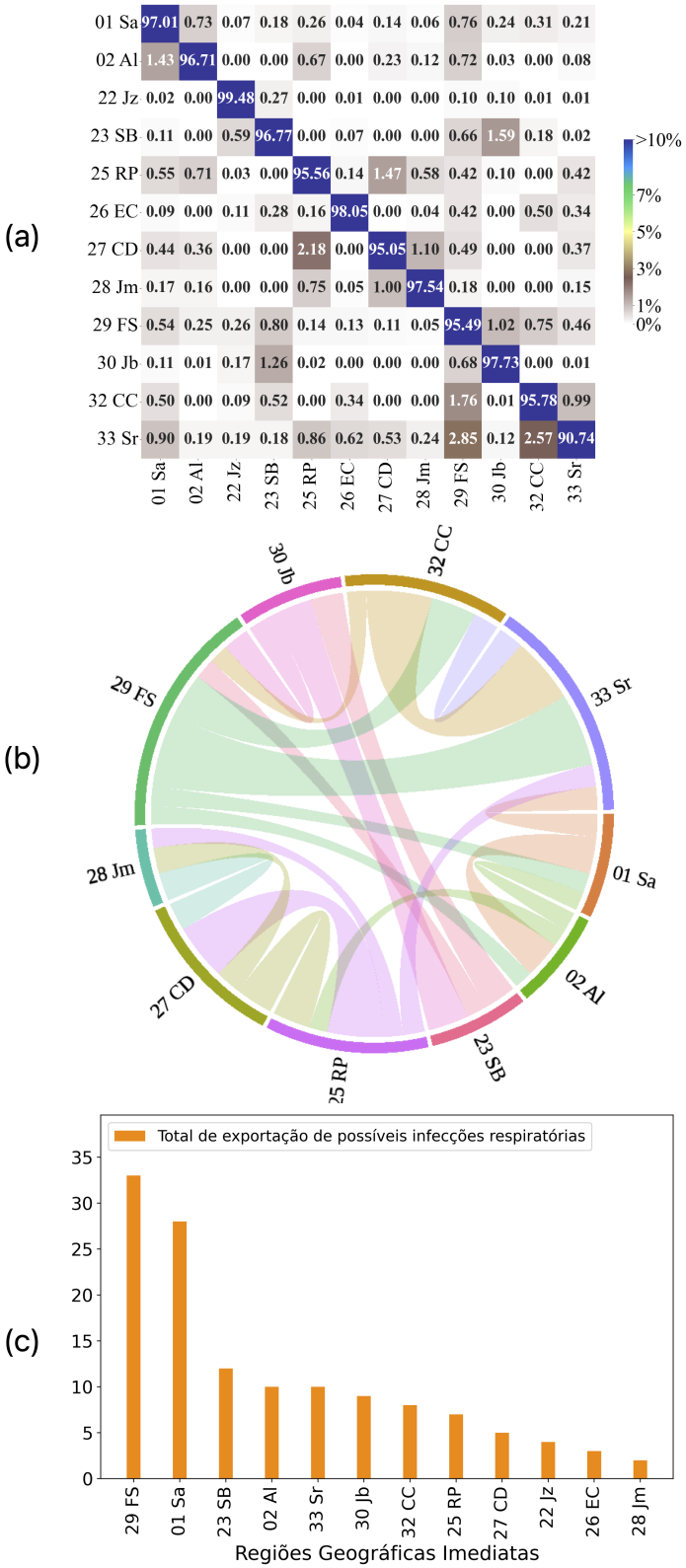


Figura 9.9: Espalhamento de potenciais infecções respiratórias entre RGIs dentro da comunidade C1. **(a)** Matriz de contribuição cruzada de potenciais infecções respiratórias entre as 12 RGIs. Os valores indicam, a proporção média de potenciais infecções em cada RGI (linha) atribuídas à disseminação oriunda de outras RGIs (colunas). **(b)** Diagrama de cordas das exportações de potenciais infecções superiores a 0,7% entre as RGIs, com a coloração dos arcos indicando a RGI de origem da exportação. **(c)** Frequência semanal dos valores totais de exportação de potenciais infecções respiratórias por RGI (representados pelas barras laranja). As RGIs são identificadas pelas seus números e suas siglas.

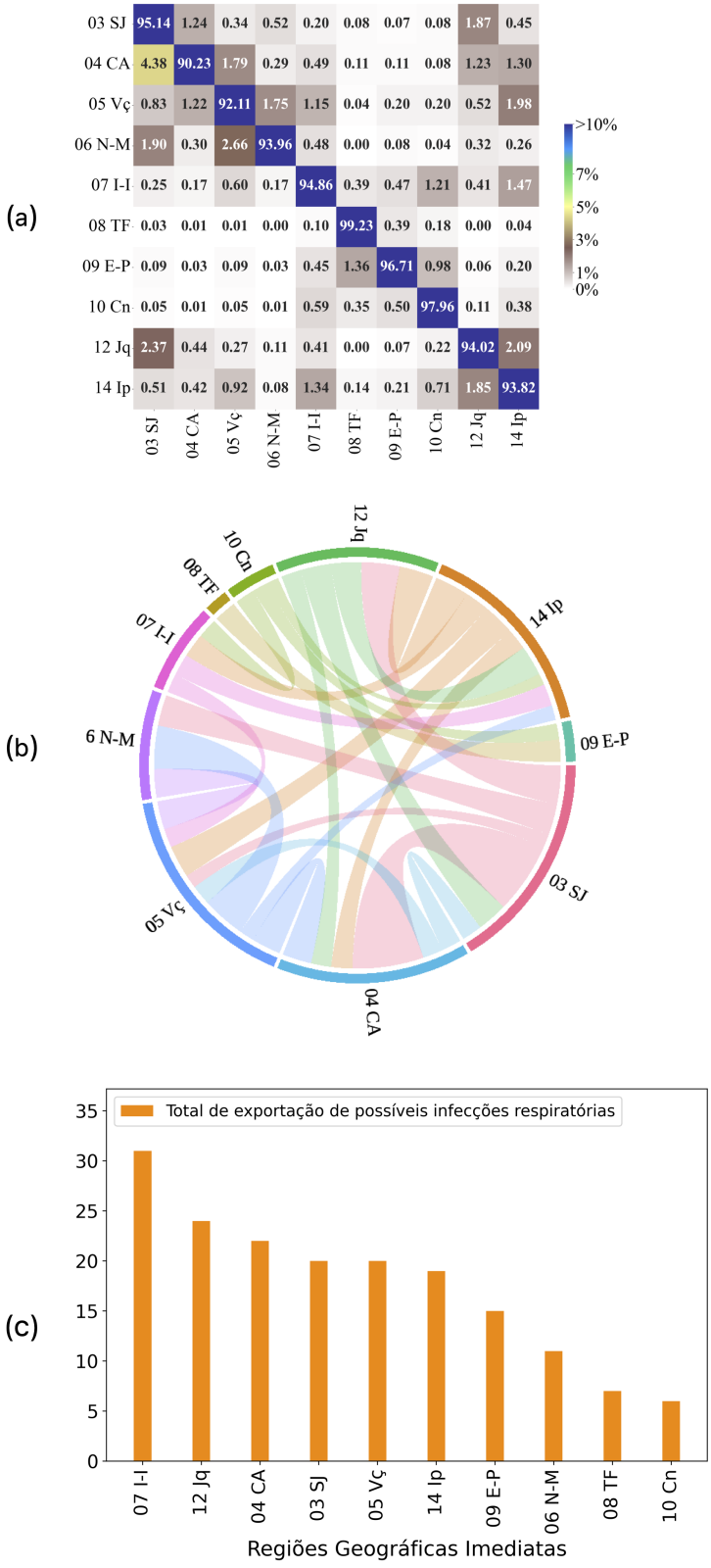


Figura 9.10: Espalhamento de potenciais infecções respiratórias entre RGIs dentro da comunidade C2. **(a)** Matriz de contribuição cruzada de potenciais infecções respiratórias entre as 10 RGIs. Os valores indicam, a proporção média de potenciais infecções em cada RGI (linha) atribuídas à disseminação oriunda de outras RGIs (colunas). **(b)** Diagrama de cordas das exportações de potenciais infecções superiores a 0,7% entre as RGIs, com a coloração dos arcos indicando a RGI de origem da exportação. **(c)** Frequência semanal dos valores totais de exportação de potenciais infecções respiratórias por RGI (representados pelas barras laranja). As RGIs são identificadas pelas seus números e suas siglas.

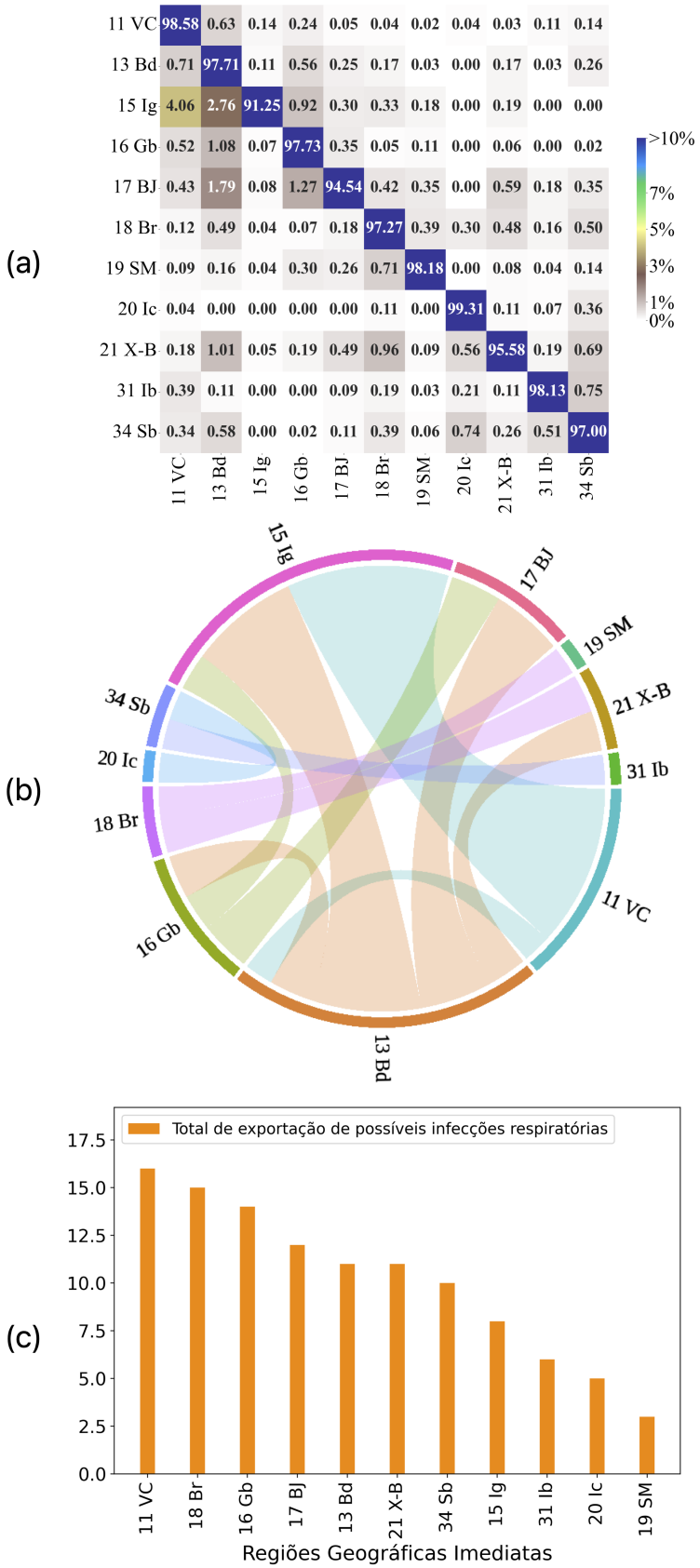


Figura 9.11: Espalhamento de potenciais infecções respiratórias entre RGIs dentro da comunidade C3. **(a)** Matriz de contribuição cruzada de potenciais infecções respiratórias entre as 11 RGIs. **(b)** Diagrama de cordas das exportações de potenciais infecções superiores a 0,7% entre as RGIs, com a coloração dos arcos indicando a RGI de origem da exportação. **(c)** Frequência semanal dos valores totais de exportação de potenciais infecções respiratórias por RGI (representados pelas barras laranja). As RGIs são identificadas pelas seus números e suas siglas.

Para obter uma visão global da dinâmica de exportação de infecções, analisamos simultaneamente as 34 RGIs, organizadas em comunidades previamente identificadas. Verificou-se uma baixa exportação de potenciais infecções entre comunidades, com a maior parte da disseminação ocorrendo no interior de cada comunidade conforme ilustrado na figura 9.12. Esse padrão reforça a adequação do recorte por comunidades como unidade funcional para a modelagem, destacando a relevância das conexões intracomunitárias em relação às intercomunitárias.

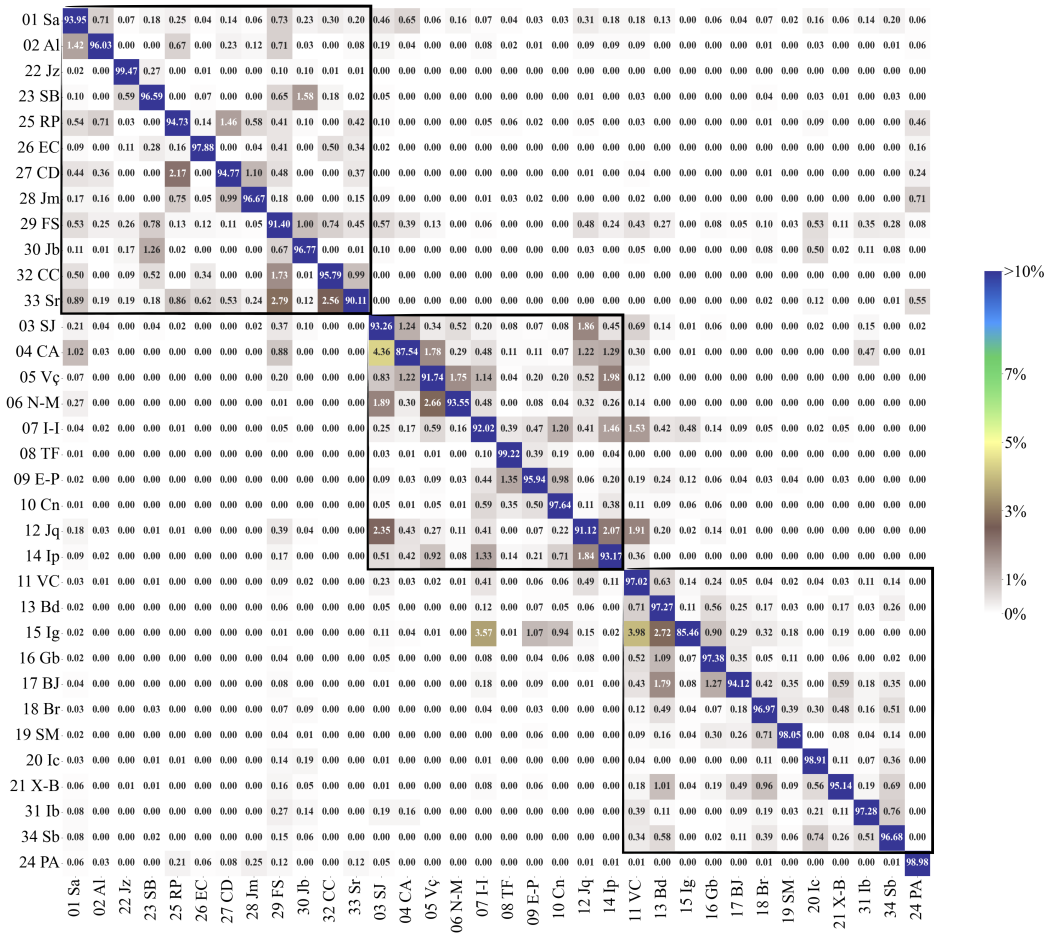


Figura 9.12: Matriz de espalhamento das potenciais infecções respiratórias entre as 34 RGIs. Cada célula representa a fração média de infecções em uma RGI de destino (linha) atribuídas à RGI de origem (coluna), indicando o percentual de potenciais infecçõesexportadas. A presença de blocos destacados evidencia que a maior parte das infecções exportadas ocorrem no interior de cada comunidade detectada. As RGIs são identificadas pelas seus números e suas siglas.

A Figura 9.13 ilustra espacialmente os fluxos médios de exportação entre as 34 RGIs. Nota-se que a maioria das setas permanece concentrada

dentro dos limites das comunidades previamente identificadas, evidenciando uma dinâmica de propagação predominantemente intracomunitária.

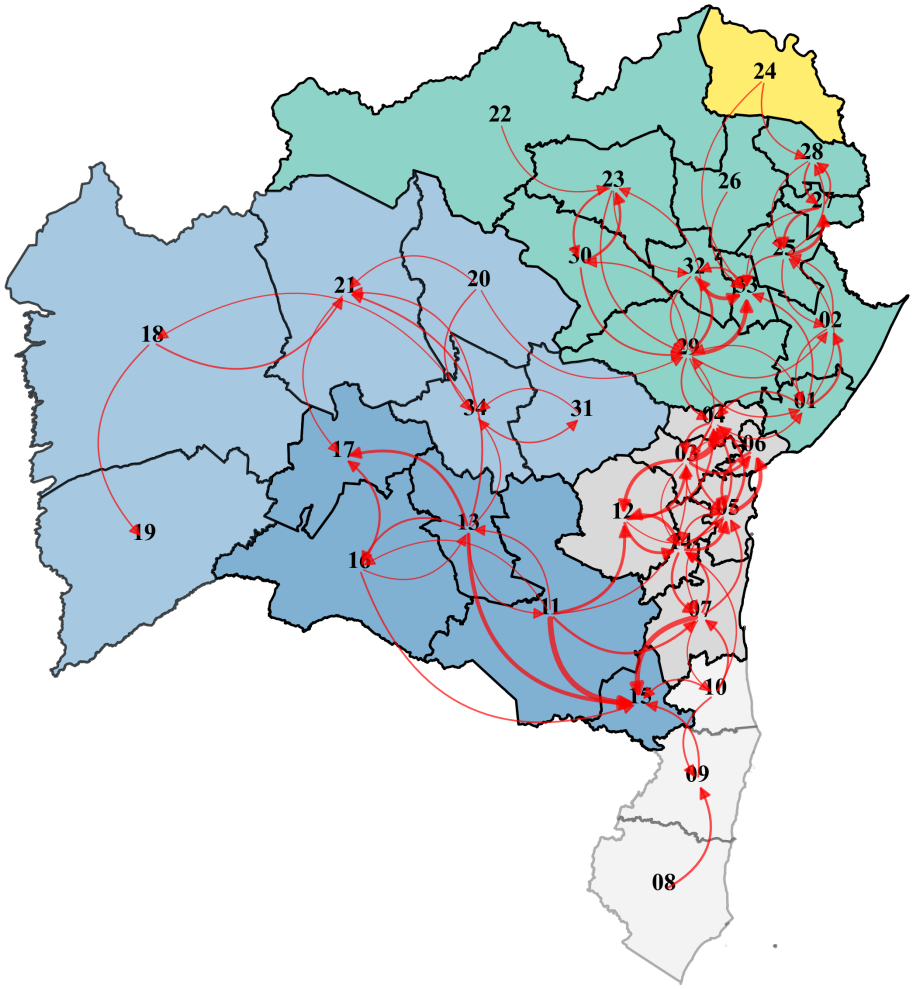


Figura 9.13: Representação espacial dos fluxos médios de exportação de potenciais infecções respiratórias entre as 34 RGIs da Bahia. As setas apresentam espessura proporcional à intensidade da exportação.

Além disso, para cada RGI, comparamos as estimativas do número total de potenciais infecções respiratórias exportadas com o movimento total de pessoas, conforme ilustrado na Figura 9.14. Para investigar a relação entre os dados de mobilidade e a exportação de potenciais infecções respiratórias, utilizou-se a correlação de Spearman. Essa medida estatística não paramétrica (ρ) avalia a relação monotônica entre duas variáveis. Diferentemente da correlação de Pearson, que mensura a linearidade entre variáveis contínuas, o método de Spearman baseia-se em valores de classificação, tornando-o mais adequado para dados que não apresentam distribuição normal [221].

Matematicamente, a correlação de Spearman é definida pela seguinte

equação:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}, \tag{9-6}$$

onde d_i representa a diferença entre as classificações das duas variáveis para a $i-$ésima observação, e n representa o número total de observações.

Encontramos uma forte correlação de Spearman ($\rho = 0,89$) entre as duas variáveis, apesar da presença de desvios observados em alguns RGIs, onde fluxos totais semelhantes resultaram em números diferentes de potenciais infecções respiratórias exportadas (ver Figura 9.14). Os resultados sugerem que outros fatores também influenciam a dinâmica observada.

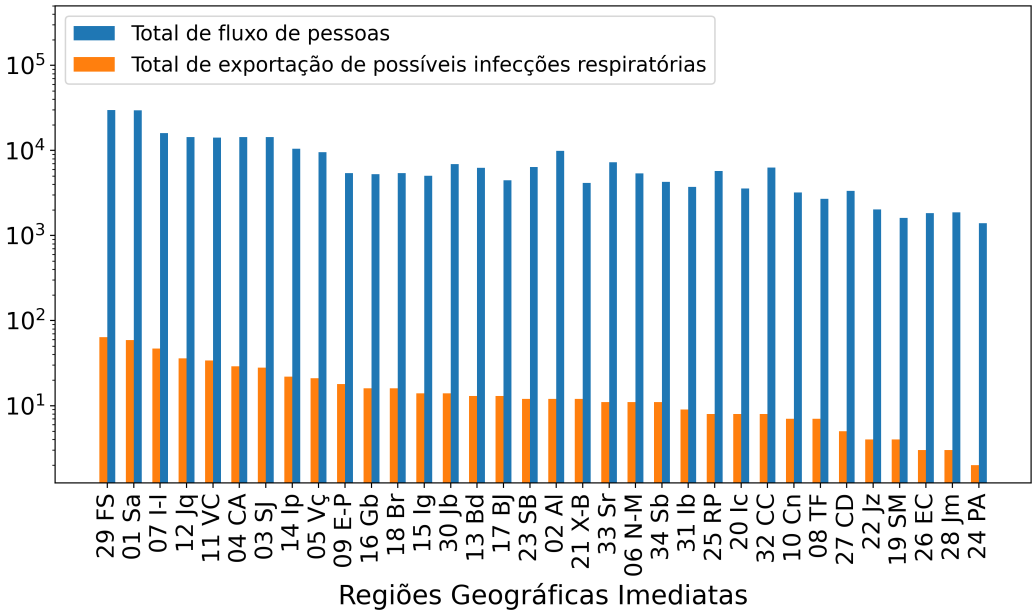


Figura 9.14: Comparação entre estimativas semanais do número total de potenciais infecções respiratórias exportadas e o movimento total de pessoas para cada IGR. As RGIs são identificadas pelas seus números e suas siglas.

Iniciamos a comparação de nossos resultados observando que, atualmente, as 12 unidades da rede sentinela operacional da Bahia estão distribuídas em 8 RGIs, alocadas em três comunidades, conforme segue: na C1 são 4 RGIs – correspondentes a Salvador (5 unidades), Feira de Santana, Alagoinhas e Juazeiro; na C2 encontram-se 3 RCIs – referentes a Ilhéus-Itabuna, Santo Antônio de Jesus e Eunápolis-Porto Seguro; e na C3 há 1 RCI em Barreiras. Na C4 não há nenhuma unidade, conforme indicado em marrom na Figura 9.16.

Dentre essas 12 unidades, 9 são compatíveis com os resultados de nossa análise: na C1, 5 unidades em Salvador (nó 01) e 1 unidade em Feira de Santana

(nó 29); na C2, as unidades de Ilhéus-Itabuna (nó 07) e Santo Antônio de Jesus (nó 03); e na C3, a unidade de Barreiras (nó 18), conforme ilustrado na Figura 9.15.

Entendemos que a falha de detecção das outras 3 RGIs (Juazeiro - nó 22, Alagoinhas - nó 02 e Eunápolis-Porto Seguro - nó 09) deve estar relacionada a um efeito de borda que restringe nossa análise às RGIs ao estado da Bahia. É natural que elas apresentem alta mobilidade com RGIs em estados vizinhos, e ao mesmo tempo elas intermediam um número mais reduzido de caminhos geodésicos entre as RGIs no estado da Bahia. Assim, duas das três medidas utilizadas para avaliar o SI ficam artificialmente reduzidas, o que pode impactar na exclusão delas de nossa análise.

No entanto, nossa análise sugere que oito outras RGIs podem ser candidatas a novas unidades sentinelas. As RGIs assim identificadas são: Senhor do Bonfim (nó 23) e Ribeira do Pombal (nó 25) em C1; Cruz das Almas (nó 4) e Ipiaú (nó 14) em C2; Vitória da Conquista (nó 11), Brumado (nó 13) e Irecê (nó 20) em C3, conforme indicado na Figura 9.15.

Por fim, a comunidade C4, caracterizada por sua posição isolada na rede, necessita de pelo menos uma unidade sentinela, que foi atribuída a RGI de Paulo Afonso (nó 24).

Portanto, os resultados integrados da análise estrutural da rede e do modelo metapopulacional, que culminaram na definição de um índice sentinela, sugerem um possível redesenho para a rede sentinela instalada na Bahia. A Figura 9.16 ilustra a proposta dessa nova sub-rede, na qual foram incorporadas novas regiões sentinelas (pontos marrons), estrategicamente distribuídas entre as diferentes comunidades. Ademais, os resultados reforçam a manutenção das atuais unidades sentinelas em funcionamento (pontos vermelhos).

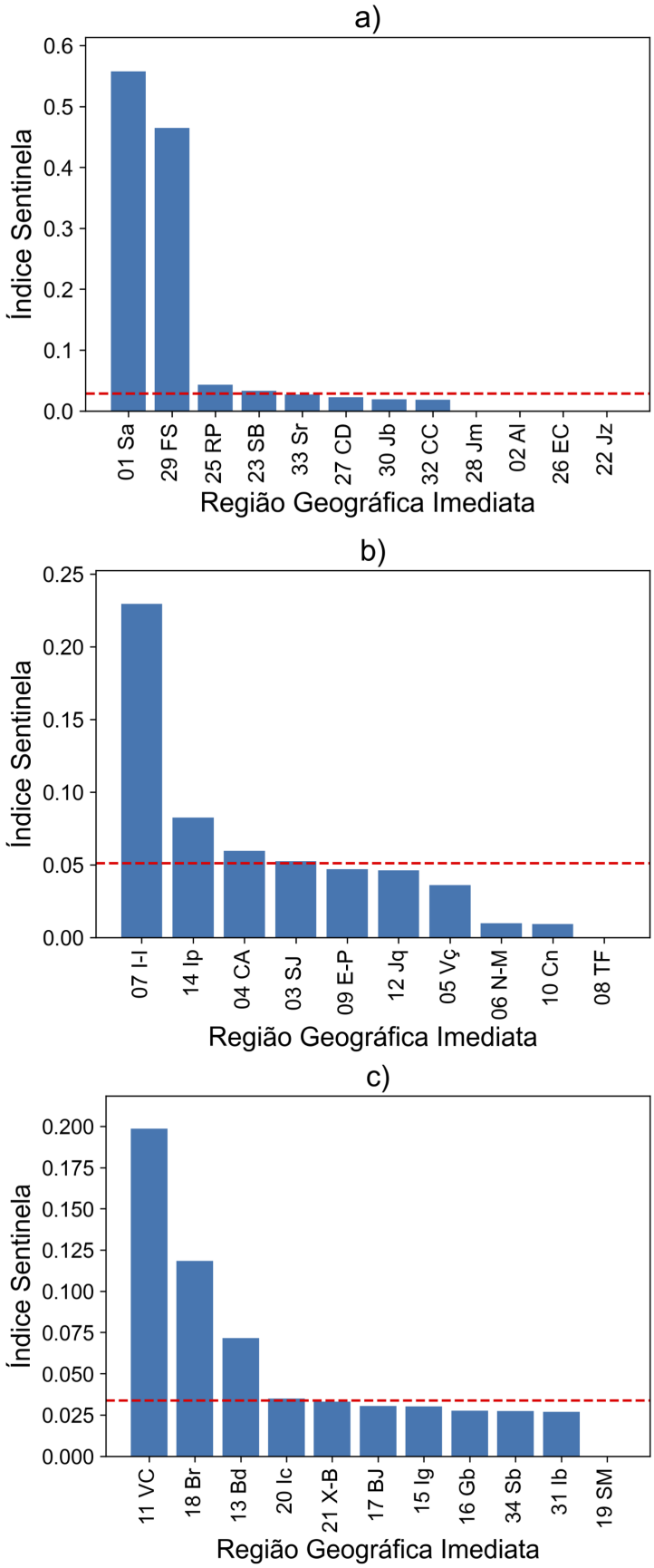


Figura 9.15: Valores do índice sentinela (SI) para cada IGR nas três comunidades distintas C1 (a), C2 (b) e C3 (c). As barras representam o índice obtido para cada RGI. A linha tracejada vermelha marca o limiar de adequação do 60º percentil.

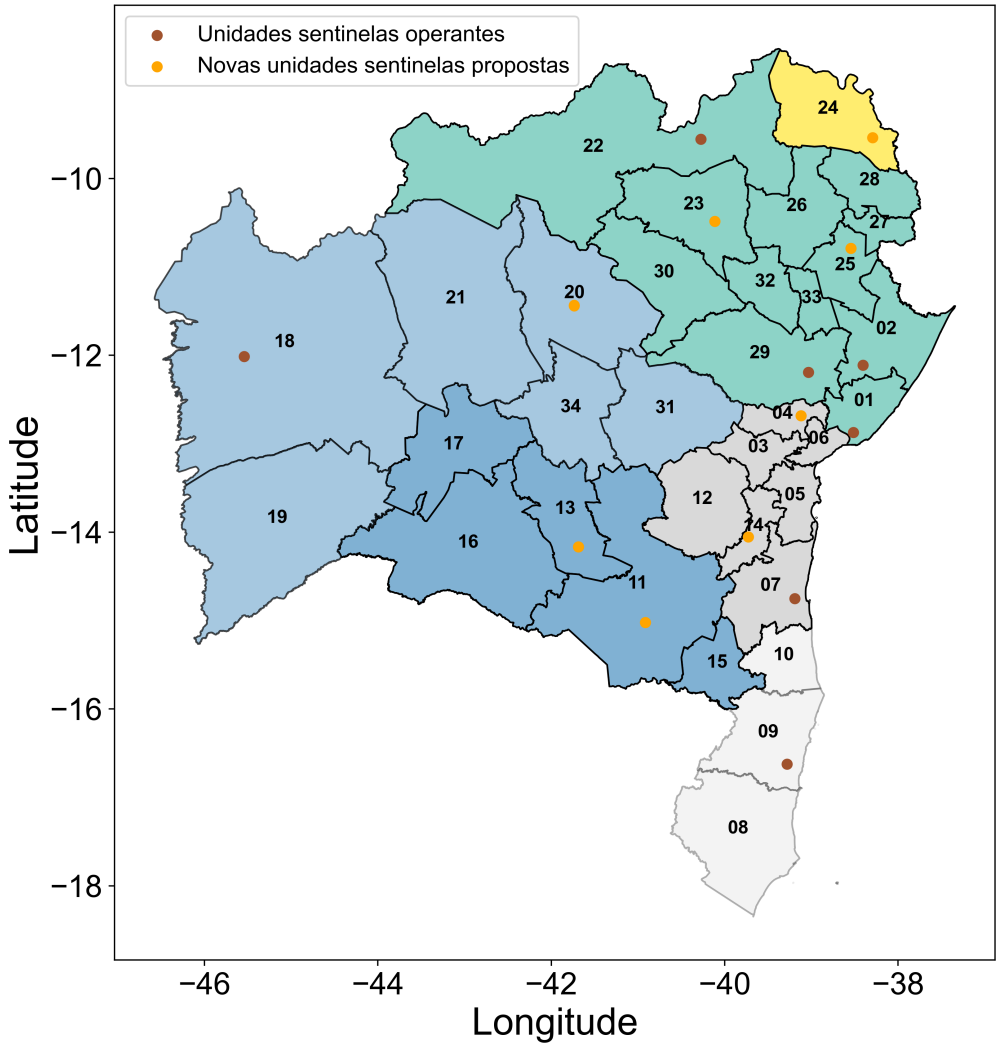


Figura 9.16: Mapa do estado da Bahia, caracterizado pelas comunidades C1 em verde, C2 em colorações em cinza, C3 em colorações em azul e C4 em amarelo, mostrando ainda a localização das unidades sentinelas no estado em funcionamento (pontos vermelhos escuros), e das unidades recomendadas (pontos laranjas).

9.3.2
Detalhamento da proposta de redesenho da Rede Sentinela da Bahia

A análise integrada de vigilância revela que C1 possui uma cobertura sentinela eficaz, baseada em unidades distribuídas em seis regiões imediatas: quatro atualmente operacionais e duas novas propostas. O IS ressalta a importância estratégica de Salvador (nó 01) e Feira de Santana (nó 29) , que atuam como principais polos de trânsito e interação social no estado, sendo os principais responsáveis pela exportação de potenciais infecções dentro de C1.

Embora as outras duas unidades sentinelas operacionais em C1 nas

RGIs de Alagoinhas (nó 02) e Juazeiro (nó 22), não apresentem índices sentinelas elevados (devido às limitações impostas pela rede analisada), elas desempenham um papel crucial na garantia de uma cobertura abrangente, funcionando como polos locais para o monitoramento.

As duas novas unidades sentinelas propostas, Ribeira do Pombal (nó 25) e Senhor do Bonfim (nó 23), apresentam índices sentinelas acima do percentil 60% adotado, e seriam essenciais para complementar os esforços de vigilância em C1, fortalecendo a infraestrutura de monitoramento dentro da comunidade.

A Comunidade C2 possui uma cobertura moderada com suas três unidades sentinelas operacionais. Contudo, a análise integrada de vigilância revelou que uma cobertura mais eficiente seria alcançada com seis unidades sentinelas. Dentro dessa comunidade, Ilhéus-Itabuna (nó 7) tem uma das unidades operacionais, e destaca-se por apresentar o maior índice sentinela de C2, refletindo sua posição estratégica tanto na comunidade como na rede. Trata-se do maior exportador de potenciais infecções respiratórias em C2 e mantém uma forte conexão com Salvador (nó 1 em C1), reforçando sua centralidade de intermediação na rede.

Santo Antônio de Jesus, outra RGI com unidade sentinela operacional, apresenta um índice sentinela acima do percentil 60% e é a quarta RGI em volume de exportação de potenciais infecções respiratórias. Por sua vez, Eunápolis-Porto Seguro, contém uma unidade operacional, não obteve um índice sentinela adequado, mas desempenha um papel essencial na garantia de uma cobertura abrangente em C2. Localizada na subcomunidade C2.2, esta RGI exerce uma função relevante na disseminação local de potenciais infecções respiratórias, inclusive em interações diretas com as RGIs Teixeira de Freitas (nó 8) e Camacan (nó 10).

Além das unidades existentes, Ipiaú (nó 14) surge como um excelente candidato para sediar a quarta unidade sentinela em C2. Com o segundo maior IS da comunidade, Ipiaú desempenha um papel relevante na disseminação de infecções dentro de C2, possui um alto grau de centralidade de intermediação e mantém uma forte conexão com Feira de Santana (nó 29 em C1). Por fim, Cruz das Almas (nó 4), tem o terceiro maior IS em C2, é recomendada em ter uma unidade sentinela adicional, consolidando a infraestrutura de vigilância dentro da comunidade.

Por outro lado, a análise revelou uma carência de unidades sentinelas na Comunidade C3. A rede atual inclui apenas a RGI de Barreiras (nó 18) como sede de uma unidade sentinela operacional. A análise integrada de vigilância confirma Barreiras como região para ter unidade sentinela, ocupando o segundo lugar no índice sentinela em C3. Localizada na subcomunidade C3.1,

Barreiras destaca-se como a segunda maior RGI em termos de centralidade de intermediação, com fortes conexões com Feira de Santana (nó 29 em C1). Ademais, é o segundo maior exportador de potenciais infecções respiratórias dentro da comunidade.

Além de Barreiras, a análise recomenda o estabelecimento de três novas RGIs para sediar unidades sentinelas: a RGI de Vitória da Conquista (nó 11), a RGI de Brumado (nó 13), ambas localizadas na subcomunidade C3.2 e a RGI de Irecê (nó 20), situada na subcomunidade C3.1.

Vitória da Conquista apresentou o maior índice sentinela em C3, destacando-se por sua relevância regional (sudoeste) como um importante polo de circulação social e convergência no estado. Mantém fortes conexões com Salvador e Feira de Santana (ambas em C1) e é o maior exportador de potenciais infecções respiratórias dentro de C3.

Brumado registrou o terceiro maior índice sentinela em C3, com fortes conexões com Feira de Santana (nó 29 em C1) e Ilhéus-Itabuna (nó 7 em C2). Atua como elo direto entre as comunidades C1 e C2 e é um ponto crítico de transmissão local de potenciais infecções respiratórias.

Irecê apresentou um índice sentinela acima do percentil 60%, e é essencial para complementar os esforços de vigilância em C3, particularmente em C3.1. A expansão proposta garantiria uma cobertura abrangente em toda a comunidade, distribuindo as unidades sentinelas entre as subcomunidades C3.1 e C3.2.

A Comunidade C4, devido à sua baixa conectividade com o restante da Bahia e, provável alta conexão com os estados vizinhos do Nordeste, necessita de uma vigilância local em sua única RGI (Paulo Afonso nó 24), integrando-a à rede sentinela para uma cobertura adequada de toda a comunidade.

9.3.3 Considerações

Neste estudo, analisamos as rotas de mobilidade rodoviária de um dos maiores estados do país e demonstramos que o fenômeno de espalhamento aparentemente não está exclusivamente ligado à mobilidade, sendo assim necessário analisar a dinâmica de modelos epidêmicos metapopulacional. Nossa abordagem combinou resultados da composição estrutural da rede, modelagem de dados sindrômicos e padrões de mobilidade, identificando regiões que podem ampliar a cobertura da rede e integrar qualquer sistema de rede sentinela.

Dada as dificuldades inerentes à instalação de sentinelas, como a disponibilidade de recursos e infraestrutura, este estudo propõe uma estrutura que auxilia na priorização de regiões fundamentado cientificamente e orientado

por dados. A seleção baseia-se em duas métricas e conceitos-chave da ciência de redes, a centralidade de intermediação e a estrutura comunitária, e também considera as taxas de potenciais infecções respiratórias geradas entre metapopulações, obtidas a partir do modelo SIR de metapopulação. Os resultados concordam parcialmente com o desenho atual da sub-rede de vigilância sentinela para síndromes gripais na Bahia, mas indicam que ela pode ser aprimorada e expandida, resultando em uma configuração mais eficaz e estratégica da rede sentinela no estado.

10
Conclusões

Sumarizamos aqui as principais conclusões referentes às duas partes do trabalho concernentes à detecção e à propagação de síndromes de infecções respiratórias: o desenvolvimento de um modelo para identificação de EWS (MMAING), e combinação de técnicas de redes complexas, e modelos de metapopulação para examinar a influência da mobilidade humana na exportação infecções respiratórias entre regiões imediatas e para o redesenho da rede sentinela da Bahia. Além disso, abordamos os principais desdobramentos e perspectivas futuras decorrentes deste estudo.

MMAING: modelo misto para detecção de EWS

O MMAING, proposto aqui como um método para detectar EWS baseado em dados de atenção primária à saúde, provou ser eficaz de acordo com as medidas comumente usadas para avaliar sistemas de vigilância em saúde. Ao se basear tanto nas informações dos dados de atendimento quanto no processo dinâmico de transmissão subjacente e empregar diferentes métodos de aprendizado de máquina não supervisionado integrados com um método de próxima geração, ele oferece uma nova perspectiva metodológica que enriquece o conjunto de ferramentas disponíveis para detecção precoce de surtos.

As possíveis limitações enfrentadas no uso do MMAING incluem, em primeiro lugar, a dificuldade em identificar surtos de magnitude extremamente baixa e surtos subjacentes. Outra limitação possível da abordagem é que, na versão desenvolvida e aqui apresentada, o MMAING não considera a propagação espacial dos surtos, embora vários locais possam ser analisados separadamente, como realizado neste trabalho. No entanto, acreditamos ser possível estender a formulação atual para incluir modelos de metapopulação, de forma similar à estratégia desenvolvida no segundo estudo da tese.

Como uma inovação na área de estudo, até onde sabemos, esta é a primeira vez que métodos tão variados foram combinados, contribuindo para o desenvolvimento de mecanismos de tomada de decisão na vigilância epidemiológica. Assim, esperamos que o MMAING seja um complemento valioso aos métodos existentes de detecção de surtos aplicados em séries temporais de dados em saúde, como o EARS [77], Farrington Flexivél [172], ASMODEE [78] e RAMMIE [173].

Desgin da Rede Sentinela da Bahia: uma abordagem metodológica mista

Este estudo não apenas reforça a utilidade das técnicas de análise de rede na vigilância sentinela, mas também apresenta uma metodologia inovadora para identificar regiões ideais para a implantação de unidades sentinelas. A integração dos resultados dos métodos de rede com o modelo metapopulacional criou um mecanismo adaptável a diferentes redes, que promove um sistema de vigilância sentinela adequado. Esta abordagem assegura que decisões baseadas em dados epidemiológicos sejam tomadas com maior eficácia, fortalecendo as estratégias de saúde pública.

Identificamos algumas limitações em nosso estudo. Por exemplo, os dados sobre mobilidade em rodovias não são atuais, limitando-se a 2016. Assim, uma nova análise com dados atualizados poderia fornecer uma representação mais precisa dos padrões de mobilidade rodoviária no país. Também não foram explorados os dados relacionados à mobilidade aérea, devido a malha aérea na Bahia ser restrita.

Apesar das limitações, espera-se que as estratégias desenvolvidas neste trabalho possam ser aplicadas para aprimorar e expandir a rede sentinela de vigilância de doenças respiratórias na Bahia. Assim, as autoridades de saúde podem configurar estrategicamente a rede sentinela, levando em consideração as especificidades regionais de mobilidade e a dinâmica de propagação de doenças, ao selecionar os locais para a implantação das unidades sentinelas. Dessa forma, é possível otimizar a cobertura da vigilância e aprimorar a capacidade de resposta a surtos.

Aplicações futuras e perspectivas de desenvolvimento

Os resultados alcançados nesta tese indicam o potencial do MMAING e das abordagens de análise de redes complexas e modelos metapopulacionais para aprimorar a vigilância epidemiológica. No entanto, o escopo desta pesquisa abre diversas possibilidades de continuidade e aprofundamento.

Uma perspectiva promissora consiste na adaptação do MMAING para o monitoramento de outras doenças de interesse em saúde pública, como as arboviroses (dengue, chikungunya e zika). Essa adaptação demandará ajustes específicos na função de geração do número de reprodução, de modo a incorporar explicitamente a dinâmica populacional do vetor biológico, principalmente o *Aedes aegypti* [222]. Fatores entomológicos como a densidade vetorial, o comportamento das picadas do mosquito, as taxas de oviposição, a mortalidade em diferentes fases do ciclo de vida desempenham um papel determinante na dinâmica de propagação dessas doenças, exigindo

modificações específicas no modelo [222, 223].

Além disso, destaca-se a possibilidade de aprimorar os sinais de alerta precoce gerados pelo MMAING por meio da integração com o volume crescente de dados provenientes de vigilância digital. Fontes como postagens em redes sociais, registros de buscas por sintomas em ferramentas online e informações sobre a venda de medicamentos em farmácias podem fornecer sinais adicionais sobre alterações no comportamento populacional relacionado à saúde. A incorporação dessas fontes de dados, que refletem tendências emergentes de morbidade antes mesmo da confirmação clínica, poderá aumentar a sensibilidade e a antecipação dos alertas, fortalecendo a capacidade do modelo em identificar precocemente potenciais surtos [224].

Outra linha de desenvolvimento futuro envolve a aplicação de modelos de aprendizado profundo (*deep learning*), com destaque para arquiteturas híbridas que combinem diferentes tipos de redes neurais, como as LSTMs (*Long Short-Term Memory*), Redes Neurais Convolucionais (CNNs, do inglês *Convolutional Neural Networks*), Redes Neurais de Grafos (GNNs, do inglês *Graph Neural Networks*) e as Unidades Recorrentes Fechadas (GRUs, do inglês *Gated Recurrent Units*), voltadas para a previsão de séries temporais de atendimentos sindrômicos na APS. Tais modelos têm se mostrado promissores na previsão de tendências epidêmicas e na identificação de padrões complexos e não lineares em dados de saúde pública [225, 226].

Referências Bibliográficas

[1] Luis AN Amaral and Julio M Ottino, "Complex networks: Augmenting the framework for the study of complex systems", The European physical journal B, vol. 38, pp. 147–162, 2004.

[2] Jarosław Kwapień and Stanisław Drożdż, "Physical approach to complex systems", Physics Reports, vol. 515, no. 3, pp. 115–226, 2012.

[3] Alexander N. Pisarchik and Ulrike Feudel, "Control of multistability", Physics Reports, vol. 540, no. 4, pp. 167–218, 2014.

[4] Marten Scheffer, Jordi Bascompte, William A. Brock, Victor Brovkin, Stephen R. Carpenter, Vasilis Dakos, Hermann Held, Egbert H. van Nes, Max Rietkerk, and George Sugihara, "Early-warning signals for critical transitions", Nature, vol. 461, no. 7260, pp. 53–59, 2009.

[5] Sandip V George, Sneha Kachhara, and G Ambika, "Early warning signals for critical transitions in complex systems", Physica Scripta, vol. 98, no. 7, pp. 072002, 2023.

[6] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. Hwang, "Complex networks: Structure and dynamics", Physics Reports, vol. 424, no. 4-5, pp. 175–308, 02 2006.

[7] Mark Newman, Networks: An Introduction, Oxford University Press, 03 2010.

[8] Albert-László Barabási and Márton Pósfai, Network Science, Cambridge University Press, Cambridge, UK, 2016.

[9] Cristina de Albuquerque Possas, "Social ecosystem health: confronting the complexity and emergence of infectious diseases", Cadernos de Saúde Pública, vol. 17, pp. 31–41, 2001.

[10] Oriol Artime and Manlio De Domenico, "From the origin of life to pandemics: Emergent phenomena in complex systems", Philosophical Transactions of the Royal Society A, vol. 380, no. 2227, pp. 20200410, 2022.

[11] Emma Southall, Tobias S. Brett, Michael J. Tildesley, and Louise Dyson, "Early warning signals of infectious disease transitions: a review",

Journal of The Royal Society Interface, vol. 18, no. 182, pp. 20210555, 2025/03/12 2021.

[12] Walter Ullon and Eric Forgoston, "Controlling epidemic extinction using early warning signals", International Journal of Dynamics and Control, vol. 11, no. 2, pp. 851–861, 2023.

[13] Daniele Proverbio, Françoise Kemp, Stefano Magni, and Jorge Gonçalves, "Performance of early warning signals for disease re-emergence: A case study on COVID-19 data", PLOS Computational Biology, vol. 18, no. 3, pp. e1009958–, 03 2022.

[14] Nina Fefferman and Elena Naumova, "Innovation in observation: a vision for early outbreak detection", Emerging Health Threats Journal, vol. 3, no. 1, pp. 7103, 04 2010.

[15] Rehab Meckawy, David Stuckler, Adityavarman Mehta, Tareq Al-Ahdal, and Bradley N. Doebbeling, "Effectiveness of early warning systems in the detection of infectious diseases outbreaks: a systematic review", BMC Public Health, vol. 22, no. 1, pp. 2216, 2022.

[16] Ziqi Li, Fancun Meng, Bing Wu, Dekun Kong, Mengying Geng, Xintong Qiu, Zicheng Cao, Tiancheng Li, Yaqian Su, and Suyang Liu, "Reviewing the progress of infectious disease early warning systems and planning for the future", BMC Public Health, vol. 24, no. 1, pp. 3080, 2024.

[17] C. Raina MacIntyre, Samsung Lim, Deepti Gurdasani, Miguel Miranda, David Metcalf, Ashley Quigley, Danielle Hutchinson, Allan Burr, and David J. Heslop, "Early detection of emerging infectious diseases - implications for vaccine development", Vaccine, vol. 42, no. 7, pp. 1826–1830, 2024.

[18] Roy M Anderson and Robert M May, Infectious Diseases of Humans: Dynamics and Control, Oxford University Press, 05 1991.

[19] Jack Feehan and Vasso Apostolopoulos, "Is COVID-19 the worst pandemic?", Maturitas, vol. 149, pp. 56, 2021.

[20] Philip A Mackowiak, "Prior pandemics. looking to the past for insight into the COVID-19 pandemic", Journal of Community Hospital Internal Medicine Perspectives, vol. 11, no. 2, pp. 163–170, 2021.

[21] Eskild Petersen, Marion Koopmans, Unyeong Go, Davidson H Hamer, Nicola Petrosillo, Francesco Castelli, Merete Storgaard, Sulien Al Khalili,

and Lone Simonsen, "Comparing SARS-CoV-2 with SARS-CoV and influenza pandemics", The Lancet infectious diseases, vol. 20, no. 9, pp. e238–e244, 2020.

[22] Nicholas A Hakes, Jeff Choi, David A Spain, and Joseph D Forrester, "Lessons from epidemics, pandemics, and surgery", Journal of the American College of Surgeons, vol. 231, no. 6, pp. 770, 2020.

[23] James W Buehler, Richard S Hopkins, J Marc Overhage, Daniel M Sosin, and Van Tong, "Framework for evaluating public health surveillance systems for early detection of outbreaks: recommendations from the CDC working group.", MMWR Recomm Rep, vol. 53, no. RR-5, pp. 1–11, May 2004.

[24] Adam T Craig, Robert Neil F Leong, Mark W Donoghoe, David Muscatello, Vio Jianu C Mojica, and Christine Joy M Octavo, "Comparison of statistical methods for the early detection of disease outbreaks in small population settings", IJID regions, vol. 8, pp. 157–163, 2023.

[25] Arnaud Chiolero and David Buckeridge, "Glossary for public health surveillance in the age of data science", J Epidemiol Community Health, vol. 74, no. 7, pp. 612–616, 2020.

[26] Wullianallur Raghupathi and Viju Raghupathi, "Big data analytics in healthcare: promise and potential", Health information science and systems, vol. 2, pp. 1–10, 2014.

[27] Larissa May, Jean-Paul Chretien, and Julie A. Pavlin, "Beyond traditional surveillance: applying syndromic surveillance to developing settings — opportunities and challenges", BMC Public Health, vol. 9, no. 1, pp. 242, 2009.

[28] Stephen S. Morse, "Public health surveillance and infectious disease detection", Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science, vol. 10, no. 1, pp. 6–16, 2024/02/23 2012.

[29] Lerina Aversano, Martina Iammarino, Ilaria Mancino, Debora Montano, and Giovanni Angiulli, "A systematic review on artificial intelligence approaches for smart health devices", PeerJ Computer Science, vol. 10, pp. e2232, 2024.

[30] Ismael Villanueva-Miranda, Guanghua Xiao, and Yang Xie, "Artificial intelligence in early warning systems for infectious disease surveillance: a systematic review", Frontiers in Public Health, vol. Volume 13 - 2025, 2025.

[31] Pablo Ivan P Ramos, Izabel Marcilio, Ana I Bento, Gerson O Penna, Juliane F de Oliveira, Ricardo Khouri, Roberto FS Andrade, Roberto P Carreiro, Vinicius de A Oliveira, and Luiz Augusto C Galvão, "Combining digital and molecular approaches using health and alternate data sources in a next-generation surveillance system for anticipating outbreaks of pandemic potential", JMIR public health and surveillance, vol. 10, pp. e47673, 2024.

[32] Dérick G. F. Borges, Eluã R. Coutinho, Thiago Cerqueira-Silva, Malú Grave, Adriano O. Vasconcelos, Luiz Landau, Alvaro L. G. A. Coutinho, Pablo Ivan P. Ramos, Manoel Barral-Netto, Suani T. R. Pinho, Marcos E. Barreto, and Roberto F. S. Andrade, "Combining machine learning and dynamic system techniques to early detection of respiratory outbreaks in routinely collected primary healthcare records", BMC Medical Research Methodology, vol. 25, no. 1, pp. 99, 2025.

[33] William Ogilvy Kermack and Anderson G McKendrick, "A contribution to the mathematical theory of epidemics", Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character, vol. 115, no. 772, pp. 700–721, 1927.

[34] Pedro Teles, "A time-dependent SEIR model to analyse the evolution of the sars-cov-2 epidemic outbreak in Portugal", arXiv preprint arXiv:2004.04735, 2020.

[35] Juliane F. Oliveira, Daniel C. P. Jorge, Rafael V. Veiga, Moreno S. Rodrigues, Matheus F. Torquato, Nivea B. da Silva, Rosemeire L. Fiaccone, Luciana L. Cardim, Felipe A. C. Pereira, Caio P. de Castro, Aureliano S. S. Paiva, Alan A. S. Amad, Ernesto A. B. F. Lima, Diego S. Souza, Suani T. R. Pinho, Pablo Ivan P. Ramos, and Roberto F. S. Andrade, "Mathematical modeling of covid-19 in 14.8 million individuals in Bahia, Brazil", Nature Communications, vol. 12, no. 1, pp. 333, 2021.

[36] Wuyue Yang, Dongyan Zhang, Liangrong Peng, Changjing Zhuge, and Liu Hong, "Rational evaluation of various epidemic models based on the COVID-19 data of China", Epidemics, vol. 37, pp. 100501

[37] Lingcai Kong, Mengwei Duan, Jin Shi, Jie Hong, Zhaorui Chang, and Zhijie Zhang, "Compartmental structures used in modeling covid-19: a scoping review", Infectious Diseases of Poverty, vol. 11, no. 1, pp. 72, 2022.

[38] Fred Brauer, "Age-of-infection and the final size relation", Math. Biosci. Eng, vol. 5, no. 4, pp. 681–690, 2008.

[39] Odo Diekmann and Johan A. P. Heesterbeek, Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis and Interpretation, vol. 5 of Wiley Series in Mathematical and Computational Biology, John Wiley & Sons, Chichester, UK, 2000.

[40] JA Simpson, L Aarons, WE Collins, GM Jeffery, and NJ White, "Population dynamics of untreated plasmodium falciparum malaria within the adult human host during the expansion phase of the infection", Parasitology, vol. 124, no. 3, pp. 247–263 2002.

[41] Susan Cassels, Cynthia R Pearson, Ann E Kurth, Diane P Martin, Jane M Simoni, Eduardo Matediana, and Stephen Gloyd, "Discussion and revision of the mathematical modeling tool described in the previously published article "modeling hiv transmission risk among mozambicans prior to their initiating highly active antiretroviral therapy"", AIDS care, vol. 21, no. 7, pp. 858–862

[42] Christophe Fraser, "Estimating individual and household reproduction numbers in an emerging epidemic", PloS one, vol. 2, no. 8, pp. e758

[43] Anne Cori, Neil M Ferguson, Christophe Fraser, and Simon Cauchemez, "A new framework and software to estimate time-varying reproduction numbers during epidemics", American journal of epidemiology, vol. 178, no. 9, pp. 1505–1512, 2013.

[44] Kangguo Li, Jiayi Wang, Jiayuan Xie, Jia Rui, Buasiyamu Abudunaibi, Hongjie Wei, Hong Liu, Shuo Zhang, Qun Li, and Yan Niu, "Advancements in defining and estimating the reproduction number in infectious disease epidemiology", China CDC Weekly, vol. 5, no. 37, pp. 829, 2023.

[45] Hiroshi Nishiura and Gerardo Chowell, "The effective reproduction number as a prelude to statistical estimation of time-dependent epidemic trends", Mathematical and statistical estimation approaches in epidemiology, pp. 103–121, 2009.

[46] Hiroshi Nishiura, Natalie M Linton, and Andrei R Akhmetzhanov, "Serial interval of novel coronavirus (COVID-19) infections", International journal of infectious diseases, vol. 93, pp. 284–286, 2020.

[47] Sang Woo Park, David Champredon, and Jonathan Dushoff, "Inferring generation-interval distributions from contact-tracing data", Journal of the Royal Society Interface, vol. 17, no. 167, pp. 20190719, 2020.

[48] Jacco Wallinga and Marc Lipsitch, "How generation intervals shape the relationship between growth rates and reproductive numbers", Proceedings of the Royal Society B: Biological Sciences, vol. 274, no. 1609, pp. 599–604, 2007.

[49] David Champredon and Jonathan Dushoff, "Intrinsic and realized generation intervals in infectious-disease transmission", Proceedings of the Royal Society B: Biological Sciences, vol. 282, no. 1821, pp. 20152026, 2015.

[50] DCP Jorge, JF Oliveira, JGV Miranda, RFS Andrade, and STR Pinho, "Estimating the effective reproduction number for heterogeneous models using incidence data", Royal Society Open Science, vol. 9, no. 9, pp. 220005, 2022.

[51] Odo Diekmann, Johan Andre Peter Heesterbeek, and Johan AJ Metz, "On the definition and the computation of the basic reproduction ratio r 0 in models for infectious diseases in heterogeneous populations", Journal of mathematical biology, vol. 28, pp. 365–382, 1990.

[52] Pauline Van den Driessche and James Watmough, "Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission", Mathematical biosciences, vol. 180, no. 1-2, pp. 29–48, 2002.

[53] David Champredon, Jonathan Dushoff, and David JD Earn, "Equivalence of the erlang-distributed SEIR epidemic model and the renewal equation", SIAM Journal on Applied Mathematics, vol. 78, no. 6, pp. 3258–3278, 2018.

[54] Matt J Keeling and Ken TD Eames, "Networks and epidemic models", Journal of the royal society interface, vol. 2, no. 4, pp. 295–307, 2005.

[55] José Garcia Vivas Miranda, Mateus Souza Silva, José Gabriel Bertolino, Rodrigo Nogueira Vasconcelos, Elaine Cristina Barbosa Cambui, Marcio Luis Valença Araújo, Hugo Saba, Diego Pereira Costa, Soltan Galano Duverger, and Matheus Teles de Oliveira, "Scaling effect in covid-19 spreading: The role of heterogeneity in a hybrid ode-network model with restrictions on the inter-cities flow", Physica D: Nonlinear Phenomena, vol. 415, pp. 132792, 2021.

[56] Ane Blázquez-García, Angel Conde, Usue Mori, and Jose A. Lozano, "A review on outlier/anomaly detection in time series data", ACM Comput. Surv., vol. 54, no. 3, April 2021.

[57] Varun Chandola, Arindam Banerjee, and Vipin Kumar, "Anomaly detection: A survey", ACM Comput. Surv., vol. 41, no. 3, July 2009.

[58] Sebastian Schmidl, Phillip Wenig, and Thorsten Papenbrock, "Anomaly detection in time series: a comprehensive evaluation", Proc. VLDB Endow., vol. 15, no. 9, pp. 1779–1797, May 2022.

[59] Zahra Zamanzadeh Darban, Geoffrey I. Webb, Shirui Pan, Charu Aggarwal, and Mahsa Salehi, "Deep learning for time series anomaly detection: A survey", ACM Computing Surveys, vol. 57, no. 1, 2024.

[60] Said E. Said and David A. Dickey, "Testing for unit roots in autoregressive-moving average models of unknown order", Biometrika, vol. 71, no. 3, pp. 599–607, 12 1984.

[61] Manish Gupta, Jing Gao, Charu C. Aggarwal, and Jiawei Han, "Outlier detection for temporal data: A survey", IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 9, pp. 2250–2267, 2014.

[62] A. J. Fox, "Outliers in time series", Journal of the Royal Statistical Society: Series B (Methodological), vol. 34, no. 3, pp. 350–363, 1972.

[63] Ruey S. Tsay, "Outliers, level shifts, and variance changes in time series", Journal of Forecasting, vol. 7, no. 1, pp. 1–20, 1988.

[64] Ruey S. Tsay, Daniel Peña, and Alan E. Pankratz, "Outliers in multivariate time series", Biometrika, vol. 87, no. 4, pp. 789–804, 2000.

[65] Douglas M. Hawkins, Identification of Outliers, Monographs on Applied Probability and Statistics. Chapman and Hall, London, 1980.

[66] Friedrich Pukelsheim, "The three sigma rule", The American Statistician, vol. 48, no. 2, pp. 88–91, 1994.

[67] Robert Mcgill, John W. , Tukey, , and Wayne A. Larsen, "Variations of box plots", The American Statistician, vol. 32, no. 1, pp. 12–16, 02 1978.

[68] Matthieu Urvoy and Florent Autrusseau, "Application of grubbs' test for outliers to the detection of watermarks", in Proceedings of the 2nd ACM Workshop on Information Hiding and Multimedia Security, New York, NY, USA, 2014, IH&MMSec '14, pp. 49–60, Association for Computing Machinery.

[69] Bernard Rosner, "Percentage points for a generalized esd many-outlier procedure", Technometrics, vol. 25, no. 2, pp. 165–172, 1983.

[70] Chung Chen and Lon-Mu Liu, "Forecasting time series with outliers", Journal of Forecasting, vol. 12, no. 1, pp. 13–35, 2025/07/29 1993.

[71] Joseph D. Brutlag, "Aberrant behavior detection in time series for network monitoring", in 14th USENIX Systems Administration Conference (LISA). 2000, pp. 139–146, USENIX Association.

[72] Peter J. Brockwell and Richard A. Davis, Introduction to Time Series and Forecasting, Springer Texts in Statistics. Springer, New York, 1996.

[73] Rob J. Hyndman and Yeasmin Khandakar, "Automatic time series forecasting: The forecast package for R", Journal of Statistical Software, vol. 27, no. 3, pp. 1–22, 2008.

[74] K. Thiyagarajan, S. Kodagoda, N. Ulapane, and M. Prasad, "A temporal forecasting driven approach using facebook's prophet method for anomaly detection in sewer air temperature sensor system", in 2020 15th IEEE Conference on Industrial Electronics and Applications (ICIEA), 2020, pp. 25–30.

[75] Ghazaleh Babanejaddehaki, Aijun An, and Manos Papagelis, "Disease outbreak detection and forecasting: A review of methods and data sources", 2024.

[76] Apollinaire Batoure Bamana, Mahdi Shafiee Kamalabad, and Daniel L. Oberski, "A systematic literature review of time series methods applied to epidemic prediction", Informatics in Medicine Unlocked, vol. 50, pp. 101571, 2024.

[77] Lori Hutwagner, William Thompson, G Matthew Seeman, and Tracee Treadwell, "The bioterrorism preparedness and response early aberration

reporting system (EARS)", Journal of Urban Health, vol. 80, pp. i89–i96, 2003.

[78] Thibaut Jombart, Stéphane Ghozzi, Dirk Schumacher, Timothy J Taylor, Quentin J Leclerc, Mark Jit, Stefan Flasche, Felix Greaves, Tom Ward, and Rosalind M Eggo, "Real-time monitoring of COVID-19 dynamics using automated trend fitting and anomaly detection", Philosophical Transactions of the Royal Society B, vol. 376, no. 1829, pp. 20200266, 2021.

[79] Siddharth Misra, Hao Li, and Jiabo He, Machine learning for subsurface characterization, Gulf Professional Publishing, 2019.

[80] S. P. Lloyd, "Least squares quantization in pcm", IEEE Transactions on Information Theory, vol. 28, no. 2, pp. 129–137, 1982.

[81] Eamonn Keogh and Jessica Lin, "Clustering of time series subsequence is meaningless: Implications for previous and future research", in Proceedings of the IEEE International Conference on Data Mining, 2005, pp. 115–122.

[82] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise", in Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. 1996, KDD'96, pp. 226–231, AAAI Press.

[83] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander, "Lof: identifying density-based local outliers", Proceedings of the 2000 ACM SIGMOD international conference on Management of data, pp. 93–104, 2000.

[84] Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson, "Estimating the support of a high-dimensional distribution", Neural computation, vol. 13, no. 7, pp. 1443–1471, 2001.

[85] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou, "Isolation forest", 2008 eighth ieee international conference on data mining, pp. 413–422, 2008.

[86] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou, "Isolation-based anomaly detection", ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 6, no. 1, pp. 1–39, 2012.

[87] Abdenour Bounsiar and Michael G Madden, "One-class support vector machines revisited", 2014 International Conference on Information Science & Applications (ICISA), pp. 1–4, 2014.

[88] Haibo He and Edwardo A. Garcia, "Learning from imbalanced data", IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263–1284, 2009.

[89] Lyudmyla Kirichenko, Yulia Koval, Sergiy Yakovlev, and Dmytro Chumachenko, "Anomaly detection in fractal time series with LSTM autoencoders", Mathematics, vol. 12, no. 19, 2024.

[90] Mayu Sakurada and Takehisa Yairi, "Anomaly detection using autoencoders with nonlinear dimensionality reduction", in Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis, New York, NY, USA, 2014, MLSDA'14, pp. 4–11, Association for Computing Machinery.

[91] Xuanhao Chen, Liwei Deng, Yan Zhao, and Kai Zheng, "Adversarial autoencoder for unsupervised time series anomaly detection and interpretation", in Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining, New York, NY, USA, 2023, WSDM '23, pp. 267–275, Association for Computing Machinery.

[92] Pankaj Malhotra, Anusha Ramakrishnan, Gaurangi Anand, Lovekesh Vig, Puneet Agarwal, and Gautam Shroff, "Lstm-based encoder-decoder for multi-sensor anomaly detection", 2016.

[93] Sourav Das, Debasis Samanta, Sayan Chatterjee, and Debasis Giri, "Lstm based neural network model for anomaly event detection in smart homes", Journal of Ambient Intelligence and Humanized Computing, vol. 12, pp. 1–14, 2021.

[94] Zhen Chen, ZhenWan Li, Jia Huang, ShengZheng Liu, and HaiXia Long, "An effective method for anomaly detection in industrial internet of things using xgboost and lstm", Scientific Reports, vol. 14, no. 1, pp. 74822, 2024.

[95] Priyanshu Sinha, Dinesh Sahu, Shiv Prakash, Tiansheng Yang, Rajkumar Singh Rathore, and Vivek Kumar Pandey, "A high performance hybrid lstm cnn secure architecture for iot environments using deep learning", Scientific Reports, vol. 15, pp. 9684, 2025.

[96] Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Soderstrom, "Detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding", in Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, New York, NY, USA, 2018, KDD '18, pp. 387–395, Association for Computing Machinery.

[97] Mingmin Zhang, Dihua Wu, and Rongna Xue, "Hourly prediction of pm2.5 concentration in Beijing based on Bi-LSTM neural network", Multimedia Tools and Applications, vol. 80, no. 16, pp. 24455–24468, 2021.

[98] Zhilei Zhao, Zhao Xiao, and Jie Tao, "Msdg: Multi-scale dynamic graph neural network for industrial time series anomaly detection", Sensors, vol. 24, no. 22, 2024.

[99] Philip Wenig and Ying Zhu, "Anomaly detection in time series: A comprehensive evaluation", Proceedings of the VLDB Endowment, vol. 15, no. 9, pp. 1779–1797, 2022.

[100] Arthur Zimek, Erich Schubert, and Hans-Peter Kriegel, "A survey on unsupervised outlier detection in high-dimensional numerical data", Statistical Analysis and Data Mining: The ASA Data Science Journal, vol. 5, no. 5, pp. 363–387, 2024/04/30 2012.

[101] Zhihan Li, Youjian Zhao, Jiaqi Han, Zhiqiang He, and Jianfeng Lu, "Multivariate time series anomaly detection and interpretation using hierarchical inter-metric and temporal embedding", in Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021, pp. 3220–3230.

[102] Subutai Ahmad, Alexander Lavin, Scott Purdy, and Zuha Agha, "Unsupervised real-time anomaly detection for streaming data", Neurocomputing, vol. 262, pp. 134–147, 2017.

[103] João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia, "A survey on concept drift adaptation", ACM Computing Surveys, vol. 46, no. 4, pp. 44:1–44:37, 2014.

[104] Jannes Vielhaben, Sebastian Lapuschkin, Grégoire Montavon, and Wojciech Samek, "Explainable ai for time series via virtual inspection layers", arXiv preprint, vol. arXiv:2303.06365, 2023.

[105] Bartosz Krawczyk, Leandro L. Minku, João Gama, Jerzy Stefanowski, and Michał Woźniak, "Ensemble learning for data stream analysis: A survey", Information Fusion, vol. 37, pp. 132–156, 2017.

[106] Christopher Bishop, "Pattern recognition and machine learning", Springer google schola, vol. 2, pp. 531–537, 2006.

[107] Victor Wiley and Thomas Lucas, "Computer vision and image processing: a paper review", International Journal of Artificial Intelligence Research, vol. 2, no. 1, pp. 29–36, 2018.

[108] John W Goodell, Satish Kumar, Weng Marc Lim, and Debidutta Pattnaik, "Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis", Journal of Behavioral and Experimental Finance, vol. 32, pp. 100577, 2021.

[109] Snigdha Sen, Sonali Agarwal, Pavan Chakraborty, and Krishna Pratap Singh, "Astronomical big data processing using machine learning: A comprehensive review", Experimental Astronomy, vol. 53, no. 1, pp. 1–43, 2022.

[110] George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang, "Physics-informed machine learning", Nature Reviews Physics, vol. 3, no. 6, pp. 422–440, 2021.

[111] Adi L Tarca, Vincent J Carey, Xue-wen Chen, Roberto Romero, and Sorin Drăghici, "Machine learning and its applications to biology", PLoS computational biology, vol. 3, no. 6, pp. e116, 2007.

[112] Maxwell W Libbrecht and William Stafford Noble, "Machine learning applications in genetics and genomics", Nature Reviews Genetics, vol. 16, no. 6, pp. 321–332, 2015.

[113] Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane, "Machine learning in medicine", New England Journal of Medicine, vol. 380, no. 14, pp. 1347–1358, 2019.

[114] Jenna Wiens and Erica S Shenoy, "Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology", Clinical infectious diseases, vol. 66, no. 1, pp. 149–153, 2018.

[115] Chandini Raina MacIntyre, Xin Chen, Mohana Kunasekaran, Ashley Quigley, Samsung Lim, Haley Stone, Hye-young Paik, Lina Yao, David

Heslop, and Wenzhao Wei, "Artificial intelligence in public health: the potential of epidemic early warning systems", Journal of International Medical Research, vol. 51, no. 3, pp. 03000605231159335, 2023.

[116] Gábor Horváth, Edith Kovács, Roland Molontay, and Szabolcs Nováczki, "Copula-based anomaly scoring and localization for large-scale, high-dimensional continuous data", ACM Transactions on Intelligent Systems and Technology (TIST), vol. 11, no. 3, pp. 1–26, 2020.

[117] Nuno Reis, José Machado da Silva, and Miguel Velhote Correia, "An introduction to the evaluation of perception algorithms and lidar point clouds using a copula-based outlier detector", Remote Sensing, vol. 15, no. 18, pp. 4570, 2023.

[118] Omar Alghushairy, Raed Alsini, Terence Soule, and Xiaogang Ma, "A review of local outlier factor algorithms for outlier detection in big data streams", Big Data and Cognitive Computing, vol. 5, no. 1, pp. 1, 2020.

[119] Markus Goldstein and Seiichi Uchida, "A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data", PloS one, vol. 11, no. 4, pp. e0152173, 2016.

[120] Rémi Domingues, Maurizio Filippone, Pietro Michiardi, and Jihane Zouaoui, "A comparative evaluation of outlier detection algorithms: Experiments and analyses", Pattern recognition, vol. 74, pp. 406–421, 2018.

[121] Pınar Ersoy, "Evolution of outlier algorithms for anomaly detection", Manchester Journal of Artificial Intelligence and Applied Sciences, vol. 2, no. 1, 2021.

[122] Vladimir Svetnik, Andy Liaw, Christopher Tong, J Christopher Culberson, Robert P Sheridan, and Bradley P Feuston, "Random forest: a classification and regression tool for compound classification and qsar modeling", Journal of chemical information and computer sciences, vol. 43, no. 6, pp. 1947–1958, 2003.

[123] Bruno R Preiss, Data structures and algorithms with object-oriented design patterns in C++, John Wiley & Sons, 2008.

[124] Kittikun Kittidachanan, Watha Minsan, Donlapark Pornnopparath, and Phimphaka Taninpong, "Anomaly detection based on gs-ocsvm

classification", 2020 12th international conference on knowledge and smart technology (KST), pp. 64–69, 2020.

[125] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, and Vincent Dubourg, "Scikit-learn: Machine learning in Python", the Journal of machine Learning research, vol. 12, pp. 2825–2830, 2011.

[126] Yue Zhao, Zain Nasrullah, and Zheng Li, "Pyod: A python toolbox for scalable outlier detection", Journal of machine learning research, vol. 20, no. 96, pp. 1–7, 2019.

[127] Ray Bradford Murphy, On tests for outlying observations, Princeton University, 1951.

[128] Yousra Chabchoub, Maurras Ulbricht Togbe, Aliou Boly, and Raja Chiky, "An in-depth study and improvement of Isolation Forest", IEEE Access, vol. 10, pp. 10219–10237, 2022.

[129] Corinna Cortes and Vladimir Vapnik, "Support-vector networks", Machine learning, vol. 20, pp. 273–297, 1995.

[130] Zheng Li, Yue Zhao, Nicola Botta, Cezar Ionescu, and Xiyang Hu, "Copod: copula-based outlier detection", 2020 IEEE international conference on data mining (ICDM), pp. 1118–1123, 2020.

[131] M Sklar, "Fonctions de répartition àn dimensions et leurs marges", Annales de l'ISUP, vol. 8, no. 3, pp. 229–231, 1959.

[132] Roger B Nelsen, An introduction to copulas, Springer, 2006.

[133] Ted G. Lewis, Network Science: Theory and Applications, John Wiley & Sons, Hoboken, NJ, 2011.

[134] Albert-László Barabási, "Network science", Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 371, no. 1987, pp. 20120375, 2024/07/19 2013.

[135] Paulo F. Gomes, "Uma introdução à Ciência de Redes e Teoria de Grafos", Revista Brasileira de Ensino de Física, vol. 46, pp. e20240190, 2024.

[136] Leonard Euler, "Solutio problematis ad geometriam situs pertinentis", Commentarii academiae scientiarum Petropolitanae, vol. 8, pp. 128–140, 1741.

[137] Paul Erdős and Alfréd Rényi, "On the evolution of random graphs", Publications of the Mathematical Institute of the Hungarian Academy of Sciences, vol. 5, pp. 17–61, 1960.

[138] Duncan J. Watts and Steven H. Strogatz, "Collective dynamics of 'small-world'networks", Nature, vol. 393, pp. 440 EP –, 06 1998.

[139] Albert-László Barabási and Réka Albert, "Emergence of scaling in random networks", Science, vol. 286, no. 5439, pp. 509–512, 10 1999.

[140] Eric Bertin, Statistical Physics of Complex Systems: A Concise Introduction, Springer International Publishing, 2 edition, 2016.

[141] Evan A. Variano, Jonathan H. McCoy, and Hod Lipson, "Networks, dynamics, and modularity", Physical Review Letters, vol. 92, no. 18, pp. 188701–, 05 2004.

[142] Santo Fortunato, "Community detection in graphs", Physics Reports, vol. 486, no. 3, pp. 75–174, 2010.

[143] Andrea Lancichinetti, Mikko Kivelä, Jari Saramäki, and Santo Fortunato, "Characterizing the community structure of complex networks", PLOS ONE, vol. 5, no. 8, pp. e11976–, 08 2010.

[144] Balachander Krishnamurthy and Jia Wang, "On network-aware clustering of web clients", SIGCOMM Comput. Commun. Rev., vol. 30, no. 4, pp. 97–110, aug 2000.

[145] P. Krishna Reddy, Masaru Kitsuregawa, P. Sreekanth, and S. Srinivasa Rao, "A graph based approach to extract a neighborhood customer community for collaborative filtering", in Databases in Networked Information Systems, Subhash Bhalla, Ed., Berlin, Heidelberg, 2002, pp. 188–200, Springer Berlin Heidelberg.

[146] Daniel S. Carvalho, Roberto F. S. Andrade, Suani T. R. Pinho, Aristóteles Góes-Neto, Thierry C. P. Lobão, Gilberto C. Bomfim, and Charbel N. El-Hani, "What are the evolutionary origins of mitochondria? a complex network approach", PLOS ONE, vol. 10, no. 9, pp. e0134988–, 09 2015.

[147] Dérick Gabriel F. Borges, Daniel S. Carvalho, Gilberto C. Bomfim, Pablo Ivan P. Ramos, Jerzy Brzozowski, Aristóteles Góes-Neto, Roberto F. S. Andrade, Charbel El-Hani, and Joseph Gillespie, "On the origin of mitochondria: a multilayer network approach", PeerJ, vol. 11, pp. e14571, 2023.

[148] Grant E. Rosensteel, Elizabeth C. Lee, Vittoria Colizza, and Shweta Bansal, "Characterizing an epidemiological geography of the united states: influenza as a case study", medRxiv, p. 2021.02.24.21252361, 01 2021.

[149] Pablo Kaluza, Andrea Kölzsch, Michael T. Gastner, and Bernd Blasius, "The complex network of global cargo ship movements", Journal of the Royal Society Interface, vol. 7, pp. 1093–1103, 2010.

[150] Carlo Ratti, Stanislav Sobolevsky, Francesco Calabrese, Clio Andris, Jonathan Reades, Mauro Martino, Rob Claxton, and Steven H. Strogatz, "Redrawing the map of great britain from a network of human interactions", PLOS ONE, vol. 5, no. 12, pp. e14248–, 12 2010.

[151] Yanfang Yang, Jiandong Cao, Yong Qin, Limin Jia, Honghui Dong, and Aomuhan Zhang, "Spatial correlation analysis of urban traffic state under a perspective of community detection", International Journal of Modern Physics B, vol. 32, no. 12, pp. 1850150, 2024/07/23 2018.

[152] Ewan Colman, Petter Holme, Hiroki Sayama, and Carlos Gershenson, "Efficient sentinel surveillance strategies for preventing epidemics on networks", PLOS Computational Biology, vol. 15, no. 11, pp. e1007517–, 11 2019.

[153] Axel Browne, David Butts, Edgar Jaramillo-Rodriguez, Nidhi Parikh, Geoffrey Fairchild, Zach Needell, Cristian Poliziani, Tom Wenzel, Timothy C. Germann, and Sara Del Valle, "Evaluating disease surveillance strategies for early outbreak detection in contact networks with varying community structure", Social Networks, vol. 79, pp. 122–132, 2024.

[154] Andrea Lancichinetti, Santo Fortunato, and János Kertész, "Detecting the overlapping and hierarchical community structure in complex networks", New Journal of Physics, vol. 11, no. 3, pp. 033015, 2009.

[155] Vito Latora and Massimo Marchiori, "Efficient behavior of small-world networks", Physical Review Letters, vol. 87, no. 19, pp. 198701–, 10 2001.

[156] Roberto F. S. Andrade, José G. V. Miranda, and Thierry Petit Lobão, "Neighborhood properties of complex networks", Physical Review E, vol. 73, no. 4, pp. 046101–, 04 2006.

[157] R. F. S. Andrade, J. G. V. Miranda, S. T. R. Pinho, and T. P. Lobão, "Characterization of complex networks by higher order neighborhood properties", The European Physical Journal B, vol. 61, no. 2, pp. 247–256, 2008.

[158] Linton Freeman, "A set of measures of centrality based on betweenness", Sociometry, vol. 40, no. 1, pp. 35–41, 03 1977.

[159] J. M. Anthonisse, "The rush in a directed graph", Tech. Rep., Centre for Mathematics and Computer Science (CWI), Amsterdam, 10 1971.

[160] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks", Proceedings of the National Academy of Sciences, vol. 99, no. 12, pp. 7821–7826, 2024/07/22 2002.

[161] Roberto F. S. Andrade, Ivan C. Rocha-Neto, Leonardo B. L. Santos, Charles N. de Santana, Marcelo V. C. Diniz, Thierry Petit Lobão, Aristóteles Goés-Neto, Suani T. R. Pinho, and Charbel N. El-Hani, "Detecting network communities: An application to phylogenetic analysis", PLOS Computational Biology, vol. 7, no. 5, pp. e1001131–, 05 2011.

[162] Aristóteles Góes-Neto, Marcelo V. C. Diniz, Daniel S. Carvalho, Gilberto C. Bomfim, Angelo A. Duarte, Jerzy A. Brzozowski, Thierry C. Petit Lobão, Suani T. R. Pinho, Charbel N. El-Hani, Roberto F. S. Andrade, and Michael Hallett, "Comparison of complex networks and tree-based methods of phylogenetic analysis and proposal of a bootstrap method", PeerJ, vol. 6, pp. e4349, 2018.

[163] Daniel S Carvalho, James C Schnable, and Ana Maria R Almeida, "Integrating phylogenetic and network approaches to study gene family evolution: The case of the agamous family of floral genes", Evolutionary Bioinformatics, vol. 14, pp. 1176934318764683, 2024/07/24 2018.

[164] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks", Physical Review E, vol. 69, no. 2, pp. 026113, Feb 2004.

[165] Réka Albert and Albert-László Barabási, "Statistical mechanics of complex networks", Reviews of Modern Physics, vol. 74, no. 1, pp. 47–97, 01 2002.

[166] Angela Noufaily, Roger A Morbey, Felipe J Colón-González, Alex J Elliot, Gillian E Smith, Iain R Lake, and Noel McCarthy, "Comparison of statistical algorithms for daily syndromic surveillance aberration detection", Bioinformatics, vol. 35, no. 17, pp. 3110–3118, 2019.

[167] Gabriel Bédubourg and Yann Le Strat, "Evaluation and comparison of statistical methods for early temporal detection of outbreaks: A simulation-based study", PloS one, vol. 12, no. 7, pp. e0181227, 2017.

[168] Roger Morbey, Gillian Smith, Isabel Oliver, Obaghe Edeghere, Iain Lake, Richard Pebody, Dan Todkill, Noel McCarthy, and Alex J Elliot, "Evaluating multi-purpose syndromic surveillance systems–a complex problem", Online Journal of Public Health Informatics, vol. 13, no. 3, 2021.

[169] Robert W Mathes, Ramona Lall, Alison Levin-Rector, Jessica Sell, Marc Paladini, Kevin J Konty, Don Olson, and Don Weiss, "Evaluating and implementing temporal, spatial, and spatio-temporal methods for outbreak detection in a local syndromic surveillance system", PloS one, vol. 12, no. 9, pp. e0184419, 2017.

[170] Steffen Unkel, C Paddy Farrington, Paul H Garthwaite, Chris Robertson, and Nick Andrews, "Statistical methods for the prospective detection of infectious disease outbreaks: a review", Journal of the Royal Statistical Society Series A: Statistics in Society, vol. 175, no. 1, pp. 49–82, 2012.

[171] CP Farrington, Nick J Andrews, AD Beale, and MA Catchpole, "A statistical algorithm for the early detection of outbreaks of infectious disease", Journal of the Royal Statistical Society: Series A (Statistics in Society), vol. 159, no. 3, pp. 547–563, 1996.

[172] Angela Noufaily, Doyo G Enki, Paddy Farrington, Paul Garthwaite, Nick Andrews, and Andre Charlett, "An improved algorithm for outbreak detection in multiple surveillance systems", Statistics in medicine, vol. 32, no. 7, pp. 1206–1222, 2013.

[173] Roger A Morbey, Alex J Elliot, Andre Charlett, Neville Q Verlander, Nick Andrews, and Gillian E Smith, "The application of a novel 'rising activity, multi-level mixed effects, indicator emphasis'(RAMMIE) method for syndromic surveillance in england", Bioinformatics, vol. 31, no. 22, pp. 3660–3665, 2015.

[174] Walter Andrew Shewhart, Economic Control of Quality of Manufactured Product, Van Nostrand, 1931.

[175] Ronald D Fricker Jr, Benjamin L Hegler, and David A Dunfee, "Comparing syndromic surveillance detection methods: EARS'versus a CUSUM-based methodology", Statistics in Medicine, vol. 27, no. 17, pp. 3407–3429, 2008.

[176] Rochelle E. Watkins, Serryn Eagleson, Bert Veenendaal, Graeme Wright, and Aileen J. Plant, "Disease surveillance using a hidden Markov model", BMC Medical Informatics and Decision Making, vol. 9, no. 1, pp. 39, 2009.

[177] Lori Hutwagner, Timothy Browne, G Matthew Seeman, and Aaron T Fleischauer, "Comparing aberration detection methods with simulated data.", Emerg Infect Dis, vol. 11, no. 2, pp. 314–316, Feb 2005.

[178] G Rossi, L Lampugnani, and M Marchi, "An approximate CUSUM procedure for surveillance of health events", Statistics in medicine, vol. 18, no. 16, pp. 2111–2122, 1999.

[179] Yiliang Zhu, W Wang, D Atrubin, and Y Wu, "Initial evaluation of the early aberration reporting system — Florida", Morbidity and Mortality Weekly Report, vol. 54, no. 123, pp. 1, 2005.

[180] Polychronis Kostoulas, Eletherios Meletis, Konstantinos Pateras, Paolo Eusebi, Theodoros Kostoulas, Luis Furuya-Kanamori, Niko Speybroeck, Matthew Denwood, Suhail A. R. Doi, Christian L. Althaus, Carsten Kirkeby, Pejman Rohani, Navneet K. Dhand, JoséL. Peñalvo, Lehana Thabane, Slimane BenMiled, Hamid Sharifi, and Stephen D. Walter, "The epidemic volatility index, a novel early warning tool for identifying new waves in an epidemic", Scientific Reports, vol. 11, no. 1, pp. 23775, 2021.

[181] Thiago Cerqueira-Silva, Juliane F. Oliveira, Vinicius de Araújo Oliveira, Pilar Tavares Veras Florentino, Alberto Sironi, Gerson O. Penna, Pablo Ivan Pereira Ramos, Viviane S. Boaventura, Manoel Barral-Netto, and

Izabel Marcilio, "Early warning system using primary health care data in the post-covid-19 pandemic era: Brazil nationwide case-study", Cadernos de Saúde Pública, vol. 40, no. 11, pp. e00010024, 2024.

[182] Thiago Cerqueira-Silva, Izabel Marcilio, Vinicius de Araújo Oliveira, Pilar Tavares Veras Florentino, Gerson O Penna, Pablo I Pereira Ramos, Viviane S Boaventura, and Manoel Barral-Netto, "Early detection of respiratory disease outbreaks through primary healthcare data", Journal of Global Health, vol. 13, 2023.

[183] Fouad Bahrpeyma, Mark Roantree, Paolo Cappellari, Michael Scriney, and Andrew McCarren, "A methodology for validating diversity in synthetic time series generation", MethodsX, vol. 8, pp. 101459, 2021.

[184] Ofek Aloni, Gal Perelman, and Barak Fishbain, "Synthetic random environmental time series generation with similarity control, preserving original signal's statistical characteristics", Environmental Modelling & Software, vol. 185, pp. 106283, 2025.

[185] Momoe Utsumi, Kiyoko Makimoto, Nahid Quroshi, and Nobuyuki Ashida, "Types of infectious outbreaks and their impact in elderly care facilities: a review of the literature", Age and ageing, vol. 39, no. 3, pp. 299–305, 2010.

[186] Jing Yan, Suvajyoti Guha, Prasanna Hariharan, and Matthew Myers, "Modeling the effectiveness of respiratory protective devices in reducing influenza outbreak", Risk Analysis, vol. 39, no. 3, pp. 647–661, 2019.

[187] Daniel B Neill, "An empirical comparison of spatial scan statistics for outbreak detection", International journal of health geographics, vol. 8, pp. 1–16, 2009.

[188] Carlos Machado de Freitas, Christovam Barcellos, Daniel Antunes Maciel Villela, Gustavo Corrêa Matta, Lenice Costa Reis, Margareth Crisóstomo Portela, Diego Ricardo Xavier, Raphael Guimarães, Raphael de Freitas Saldanha, and Isadora Vida Mefano, "BOLETIM Observatório Fiocruz COVID-19: Boletim especial: balanço de dois anos da pandemia Covid-19: janeiro de 2020 a janeiro de 2022", Periodical, Fundação Oswaldo Cruz, Rio de Janeiro, 2022, 29 pages.

[189] Thomas G. Dietterich, "Ensemble methods in machine learning", in Multiple Classifier Systems, Berlin, Heidelberg, 2000, pp. 1–15, Springer Berlin Heidelberg.

[190] Palak Mahajan, Shahadat Uddin, Farshid Hajati, and Mohammad A. Moni, "Ensemble learning for disease prediction: A review", Healthcare, vol. 11, no. 12, 2023.

[191] Gaëtan Texier, Rodrigue S. Allodji, Loty Diop, Jean-Baptiste Meynard, Liliane Pellegrin, and Hervé Chaudet, "Using decision fusion methods to improve outbreak detection in disease surveillance", BMC Medical Informatics and Decision Making, vol. 19, no. 1, pp. 38, 2019.

[192] Sami Hadhri, Mondher , Hadiji, , and Walid Labidi, "A voting ensemble classifier for stress detection", Journal of Information and Telecommunication, vol. 8, no. 3, pp. 399–416, 07 2024.

[193] A. Huppert and G. Katriel, "Mathematical modelling and prediction in infectious disease epidemiology", Clinical Microbiology and Infection, vol. 19, no. 11, pp. 999–1005, 2013.

[194] L. Lam and S. Y. Suen, "Application of majority voting to pattern recognition: an analysis of its behavior and performance", IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, vol. 27, no. 5, pp. 553–568, 1997.

[195] Weng-Keen Wong, Andrew Moore, Gregory Cooper, and Michael Wagner, "What's strange about recent events (WSARE): an algorithm for the early detection of disease outbreaks", The Journal of Machine Learning Research, vol. 6, pp. 1961–1998, 2005.

[196] Thibaut Jombart, Christopher I Jarvis, Samuel Mesfin, Nabil Tabal, Mathias Mossoko, Luigino Minikulu Mpia, Aaron Aruna Abedi, Sonia Chene, Ekokobe Elias Forbin, and Marie Roseline D Belizaire, "The cost of insecurity: from flare-up to control of a major Ebola virus disease hotspot during the outbreak in the Democratic Republic of the Congo, 2019", Eurosurveillance, vol. 25, no. 2, pp. 1900735, 2020.

[197] Krishnan Bhaskaran, Antonio Gasparrini, Shakoor Hajat, Liam Smeeth, and Ben Armstrong, "Time series regression studies in environmental epidemiology.", Int J Epidemiol, vol. 42, no. 4, pp. 1187–1195, Aug 2013.

[198] Instituto Todos Pela Saúde, "Em duas semanas, identificação de ba.4 e ba.5 passa de 44% para 79,3% das amostras positivas de sars-cov-2", June 2022, Acesso em: 10 de janeiro de 2024.

[199] Fiocruz, "Genomahcov dashboard", 2024, Acesso em: 11 de julho de 2024.

[200] Cleonice Maria Michelon, "Principais variantes do SARS-CoV-2 notificadas no Brasil", Revista RBAC, vol. 53, no. 2, pp. 1–10, 2021.

[201] Andrew J. Tatem, Simon I. Hay, and David J. Rogers, "Global traffic and disease vector dispersal", Proceedings of the National Academy of Sciences, vol. 103, no. 16, pp. 6242–6247, 2024/10/10 2006.

[202] Juliane F Oliveira, Andrêza L Alencar, Maria Célia L S Cunha, Adriano O Vasconcelos, Gerson G Cunha, Ray B Miranda, Fábio M H S Filho, Corbiniano Silva, Emanuele Gustani-Buss, Ricardo Khouri, Thiago Cerqueira-Silva, Luiz Landau, Manoel Barral-Netto, and Pablo Ivan P Ramos, "Human mobility patterns in Brazil to inform sampling sites for early pathogen detection and routes of spread: a network modelling and validation study", The Lancet Digital Health, vol. 6, no. 8, pp. e570–e579, 2024.

[203] Pedro S. Peixoto, Diego Marcondes, Cláudia Peixoto, and Sérgio M. Oliva, "Modeling future spread of infections via mobile geolocation data and population dynamics. an application to covid-19 in Brazil", PLOS ONE, vol. 15, no. 7, pp. e0235732–, 07 2020.

[204] Chenfeng Xiong, Songhua Hu, Mofeng Yang, Weiyu Luo, and Lei Zhang, "Mobile device data reveal the dynamics in a positive relationship between human mobility and covid-19 infections", Proceedings of the National Academy of Sciences, vol. 117, no. 44, pp. 27087–27089, 2024/10/10 2020.

[205] Flávio C. Coelho, Raquel M. Lana, Oswaldo G. Cruz, Daniel A. M. Villela, Leonardo S. Bastos, Ana Pastore y Piontti, Jessica T. Davis, Alessandro Vespignani, Claudia T. Codeço, and Marcelo F. C. Gomes, "Assessing the spread of covid-19 in Brazil: Mobility, morbidity and social vulnerability", PLOS ONE, vol. 15, no. 9, pp. e0238214–, 09 2020.

[206] Steven M. Teutsch and R. Elliott Churchill, Principles and Practice of Public Health Surveillance, Oxford University Press, USA, New York, 2nd edition, 2000.

[207] Yuan Bai, Bo Yang, Lijuan Lin, Jose L. Herrera, Zhanwei Du, and Petter Holme, "Optimizing sentinel surveillance in temporal network epidemiology", Scientific Reports, vol. 7, no. 1, pp. 4804, 2017.

[208] Petter Holme, "Objective measures for sentinel surveillance in network epidemiology", Physical Review E, vol. 98, no. 2, pp. 022313–, 08 2018.

[209] Lone Simonsen, Julia R Gog, Don Olson, and Cécile Viboud, "Infectious disease surveillance in the big data era: towards faster and locally relevant systems", The Journal of infectious diseases, vol. 214, no. suppl4, pp. S380–S385, 2016.

[210] S. Briand, A. Mounts, and M. Chamberland, "Challenges of global surveillance during an influenza pandemic", Public Health, vol. 125, no. 5, pp. 247–256, 2011.

[211] R Snacken and C Brown, "New developments of influenza surveillance in Europe", Eurosurveillance, vol. 20, no. 4, 2015.

[212] Benjamin J. Cowling, Shuo Feng, Lyn Finelli, Andrea Steffens, and Ashley Fowlkes, "Assessment of influenza vaccine effectiveness in a sentinel surveillance network 2010–13, United States", Vaccine, vol. 34, no. 1, pp. 61–66, 2016.

[213] Laís Picinini Freitas, Cláudia Torres Codeço, Leonardo Soares Bastos, Daniel Antunes Maciel Villela, Oswaldo Gonçalves Cruz, Antonio Guilherme Pacheco, Flavio Codeço Coelho, Raquel Martins Lana, Luiz Max Fagundes de Carvalho, and Roberta Pereira Niquini, "Avaliação do desenho da vigilância sentinela de síndrome gripal no Brasil", Cadernos de Saúde Pública, vol. 40, pp. e00028823 0102–311X, 2024.

[214] Philip M Polgreen, Zunqui Chen, Alberto M Segre, Meghan L Harris, Michael A Pentella, and Gerard Rushton, "Optimizing influenza sentinel surveillance at the state level.", Am J Epidemiol, vol. 170, no. 10, pp. 1300–1306, Nov 2009.

[215] Telessaúde Bahia, "Webpalestra – vigilância sentinela das síndromes gripais", `https://www.youtube.com/watch?v=vVk5Ank6Ly4`, 3 2024, Acessado em 10 de Outubro de 2024.

[216] IBGE, Ligações rodoviárias e hidroviárias: 2016 - Redes e fluxos do território, Instituto Brasileiro de Geografia e Estatística - IBGE, Rio de Janeiro, 2017, 79p.

[217] Instituto Brasileiro de Geografia e Estatística (IBGE), "Panorama do Estado da Bahia", `https://cidades.ibge.gov.br/brasil/ba/panorama`, 2024, Acesso em: 18 Outubro de 2024.

[218] Richard M. Vogel, "The geometric mean?", Communications in Statistics - Theory and Methods, vol. 51, no. 1, pp. 82–94, 2022.

[219] Leonardo Rodrigues Porto, Gildásio Santana Júnior, and Humberto Miranda Nascimento, "Rede urbana do estado da Bahia: o caso de Vitória da Conquista (BA)", RDE-Revista de Desenvolvimento Econômico, vol. 2, no. 37, 2017.

[220] Gail E. Potter, Timo Smieszek, and Kerstin Sailer, "Modeling workplace contact networks: The effects of organizational structure, architecture, and reporting errors on epidemic predictions", Network Science, vol. 3, no. 3, pp. 298–325, 2015.

[221] Joost CF De Winter, Samuel D Gosling, and Jeff Potter, "Comparing the pearson and spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data.", Psychological methods, vol. 21, no. 3, pp. 273, 2016.

[222] Suani T. R. Pinho, C. P. Ferreira, L. Esteva, F. R. Barreto, V. C. Morato e Silva, and M. G. L. Teixeira, "Modelling the dynamics of dengue real epidemics", Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 368, no. 1933, pp. 5679–5693, 2010.

[223] Michael A. Johansson, Katie M. Apfeldorf, Stewart Dobson, and et al., "Nowcasting the spread of chikungunya virus in the Americas", PLOS Neglected Tropical Diseases, vol. 13, no. 1, pp. e0007065, 2019.

[224] Qian Yuan, Elaine O. Nsoesie, Bin Lv, Gong Peng, and Yu Hu, "Social media–based surveillance systems for early warning of infectious disease outbreaks: A systematic review", Journal of Medical Internet Research, vol. 25, pp. e45232, 2023.

[225] Bijaya Adhikari, Xuan Xu, Naren Ramakrishnan, and B. Aditya Prakash, "Epidemiological forecasting with graph neural networks", ACM Transactions on Spatial Algorithms and Systems (TSAS), vol. 5, no. 3, pp. 1–29, 2019.

[226] Xi Jin, Cheng Wang, Yuyang Wang, and et al., "Interpretable and generalizable epidemic forecasting with graph neural networks", Nature Communications, vol. 12, no. 1, pp. 4720, 2021.