CEFET/RJ
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO

UM NOVO MÉTODO HÍBRIDO DE DETECÇÃO DE ANOMALIAS BASEADO EM EMD E ARIMA

Jéssica Barbosa de Souza Costa

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Ciência da Computação, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ, como parte dos requisitos necessários à obtenção do grau de mestre.

Orientadores:
Eduardo Soares Ogasawara, Ronaldo Fernandes Santos Alves

Rio de Janeiro,

5 de Fevereiro de 2025

Um Novo Método Híbrido de Detecção de Anomalias Baseado em EMD e ARIMA

Dissertação de Mestrado em Ciência da Computação, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ.

Jéssica Barbosa de Souza Costa

Aprovada por:

Presidente, Prof. Eduardo Soares Ogasawara, D.Sc.

Ronaldo Fernandes Santos Alves, D.Sc.

Cristiano Siqueira Boccolini, D.Sc.

Fabio Andre Machado Porto, D.Sc.

Kele Teixeira Belloze, D.Sc.

Rio de Janeiro,

5 de Fevereiro de 2025

Ficha catalográfica elaborada pela Biblioteca Central do CEFET/RJ

C837 Costa, Jéssica Barbosa de Souza
Um novo método híbrido de detecção de anomalias baseado em EMD e ARIMA / Jéssica Barbosa de Souza Costa – 2025.
47f : il. color. , enc.

Dissertação (Mestrado). Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, 2025.
Bibliografia : f. 41-47.
Orientador: Eduardo Soares Ogasawara.
Ronaldo Fernandes Santos Alves

1. Análise de séries temporais. 2. Análise de variância. I. Ogasawara, Eduardo Soares. (Orient.). II. Alves, Ronaldo Fernandes Santos (Orient.). III. Título.

CDD 519.55

Elaborada pelo bibliotecário Leandro Mota de Menezes – CRB/7 nº 5281

DEDICATÓRIA

Dedico este trabalho à minha família, que com determinação e amor proporcionou a valorização da educação acima das dificuldades, e aos ancestrais que, com resiliência, delinearam o caminho que hoje sigo.

AGRADECIMENTOS

Agradeço a Deus pela vida e pela força para superar os desafios deste trabalho. Ao meu marido, Gabriel, por seu apoio constante e por sempre estar ao meu lado, aprimorando ideias e oferecendo suporte nos momentos necessários. Minha gratidão à minha família, cujo incentivo e ênfase na educação foram fundamentais, e aos amigos, por sua amizade incondicional ao longo deste projeto. Aos orientadores, pela paciência e ensinamentos valiosos que guiaram meu processo de aprendizagem. Agradeço também às instituições CAPES, FAPERJ e à Fundação Bill e Melinda Gates pelo apoio às pesquisas. Finalmente, obrigado a todos que contribuíram, direta ou indiretamente, para o desenvolvimento desta pesquisa.

RESUMO

Um Novo Método Híbrido de Detecção de Anomalias Baseado em EMD e ARIMA

Jéssica Barbosa de Souza Costa

Orientadores:
Eduardo Soares Ogasawara, Ronaldo Fernandes Santos Alves

Resumo da Dissertação submetida ao Programa de Pós-graduação em Ciência da Computação do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ como parte dos requisitos necessários à obtenção do grau de mestre.

Anomalias são definidas como desvios comportamentais em relação a padrões esperados e conseguir as indentificar é um desafio importante. A detecção de anomalias é uma atividade fundamental na análise de séries temporais. Ela permite a tomada de decisões baseada em dados em muitas atividades de controle e monitoramento, como cuidados de saúde, qualidade da água, análise de reflexão sísmica e exploração de petróleo. Existem muitos métodos de detecção de anomalias, mas escolher os métodos apropriados é complexo devido à natureza intrínseca das séries temporais. Há uma demanda por métodos de detecção de anomalias que sejam robustos, minimizem falsos positivos e adaptáveis a cenários não estacionários. Este trabalho introduz a Decomposição de Modo Empírico Refinada (REMD, do inglês Refined Empirical Mode Decomposition) como uma abordagem híbrida para atender a essa necessidade, integrando os modelos de Decomposição de Modo Empírico (EMD, do inglês Empirical Mode Decomposition) e Autoregressão Integrada de Médias Móveis (ARIMA, do inglês Autoregressive Integrated Moving Average). A concepção do REMD visa otimizar os pontos fortes de ambos os métodos e superar suas limitações. Ele é avaliado em comparação com métodos de ponta em diversos conjuntos de dados. O REMD demonstra desempenho superior, com pontuação até três vezes melhor no F1.

Palavras-chave:
Anomalia, Detecção, ARIMA, EMD, REMD

Rio de Janeiro,

5 de Fevereiro de 2025

ABSTRACT

Um Novo Método Híbrido de Detecção de Anomalias Baseado em EMD e ARIMA

Jéssica Barbosa de Souza Costa

Advisors:
Eduardo Soares Ogasawara, Ronaldo Fernandes Santos Alves

Abstract of dissertation submitted to Programa de Pós-graduação em Ciência da Computação - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ as partial fulfillment of the requirements for the degree of master.

Anomalies are defined as behavioral deviations from expected patterns, and identifying them is a significant challenge. Anomaly detection is a fundamental activity in time series analysis. It enables data-driven decision-making in many control and monitoring activities, such as healthcare, water quality, seismic reflection analysis, and oil exploration. There are many anomaly detection methods, but choosing the appropriate methods is complex due to the intrinsic nature of time series. There is a demand for robust anomaly detection methods that minimize false positives and are adaptable to non-stationary scenarios. This work introduces the Refined Empirical Mode Decomposition (REMD) as a hybrid approach to meet this need, integrating Empirical Mode Decomposition (EMD) and Autoregressive Integrated Moving Average (ARIMA) models. The design of REMD aims to optimize the strengths of both methods and overcome their limitations. It is evaluated against state-of-the-art methods across various datasets. REMD demonstrates superior performance, achieving up to three times better F1 scores.

Key-words:
Anomaly, Detection, ARIMA, EMD

Rio de Janeiro,

5 de Fevereiro de 2025

Sumário

I Introdução **8**

II Referencial teórico **10**

II.1 Série temporal 10

II.2 Anomalias 13

II.3 Detecção de Anomalias Baseada em Modelos 16

II.4 Decomposição de Modo Empírico (EMD) 16

II.5 Modelo Autorregressivo Integrado de Médias Móveis (ARIMA) 18

II.6 Rugosidade média (R_a) 21

III Trabalhos relacionados **23**

IV Método **26**

IV.1 Decomposição de Modo Empírico (EMD) 27

IV.2 Separação de Frequências 27

IV.3 Ajuste do Modelo 28

IV.4 Identificação de Anomalias 28

V Resultados **30**

V.1 Configuração Experimental 30

V.2 Métricas de Desempenho 33

V.3 Análise Detalhada das Séries do Yahoo 35

V.4 Análise Geral de Detecção 36

V.5 Desempenho por tipo de anomalia 39

V.5.1 Série com anomalia pontual 40

V.5.2 Série com anomalias sequenciais 40

V.5.3 Série com falsas anomalias 40

V.5.4 Série com alteração na tendência 41

V.5.5 Série com alteração na média 42

V.5.6 Série com alteração na variança 42

V.5.7 Desempenho sumarizado 43
V.6 Análise geral de processamento 45

VI Conclusão **46**

Referências **48**

Lista de Figuras

II.1 Série temporal: Sequência de pontos ordenados no tempo. Neste exemplo há a ocorrência de uma anomalia. 11

II.2 Série decomposta usando o método STL. Neste método são gerados três componentes: sazonalidade, tendência e ruído. 14

II.3 Série artificial com três perfis de anomalias - Pontuais, coletivas e contextuais. 15

II.4 Processo de extração das IMF's. Neste processo é gerado um envelope, na série original, e calculado o valor médio ao longo da série. 17

II.5 Resultado da decomposição da série original usando o EMD. Formada por sete IMF's e uma série de resíduo. 18

II.6 O cálculo da rugosidade é resultado do grau de oscilação de uma superfície de comprimento L. 21

II.7 Relação entre rugosidade média e frequência - A IMF com maior grau de oscilação, apresenta maior rugosidade e maior frequência dominante. 22

V.1 Séries temporais do dataset Yahoo Labs. Esse dataset apresenta série com anomalias pontuais, predominantemente. 31

V.2 Séries temporais do dataset NAB - Esse dataset apresenta anomalias pontuais e sequenciais. 32

V.3 Séries temporais do dataset *Gecco Challenge* - Esse dataset apresenta anomalias sequênciais, predominantemente. 33

V.4 Série temporal do Yahoo analisada. Formada por doze pontos de anomalia. 36

V.5 Comparação das detecções com REMD, EMD e FBIAD na série temporal do Yahoo. 37

V.6 Série do dataset Yahoo com anomalia pontual 40

V.7 Série do dataset Gecco com anomalias sequênciais 41

V.8 Série do dataset Numenta com falsas anomalias 41

V.9 Série simulada com alteração de tendência ao longo do tempo. 42

V.10 Série simulada com alteração no valor médio ao longo do tempo. 43

V.11 Série do dataset Numenta com alteração de variança ao longo do tempo. 43

Lista de Tabelas

V.1 Parâmetros por método 34

V.2 Matriz de Confusão para a Série do Yahoo 37

V.3 Resultados Resumidos para o Conjunto de Dados do Yahoo 38

V.4 Resultados Resumidos para o Conjunto de Dados da Numenta 38

V.5 Resultados Resumidos para o Conjunto de Dados do GECCO 39

V.6 Análise Geral de Detecção 39

V.7 Resultados para as séries simuladas 44

V.8 Tempo de execução (s) por Método e Dataset 45

Lista de Acrônimos

ARIMA	Modelo Autorregressivo Integrado de Médias Móveis.
ARMA	Modelo Média Móvel Autorregressivo.
EMD	Decomposição de Modo Empírico.
FBIAD	Forward and Backward Inertial Anomaly Detector.
IMF	Função de Modo Intrínseco.
NAB	Numenta Anomaly Benchmark.
REMD	Decomposição de Modo Empírico Refinada.
SOTA	State Of The Art.
STL	Seasonal-Trend decomposition using Loess.

Capítulo I Introdução

A análise de séries temporais desempenha um papel fundamental em diversos campos de pesquisa e aplicação, fornecendo percepções valiosas sobre o comportamento dos dados ao longo do tempo [Gabarda and Cristóbal, 2010]. Por meio dessa análise, é possível identificar componentes sazonais, tendências, padrões e anomalias, permitindo a previsão de cenários futuros e a tomada de decisões baseada em dados [Bisgaard and Kulahci, 2011]. À medida que mais dados são disponibilizados e monitorados, a necessidade de processar e analisar grandes volumes de informações em diferentes aplicações tem se tornado cada vez mais importante, como no mercado financeiro [Christy Jackson et al., 2021] ou na gestão de infraestruturas críticas [Bei An and Liu, 2023].

Entre suas aplicações, a detecção de anomalias se destaca pela relevância em setores como a saúde, qualidade da água, análise de reflexão sísmica e perfuração e exploração de petróleo. Por exemplo, na saúde, séries temporais de sinais vitais são analisadas para detectar padrões anômalos que possam indicar condições médicas críticas, como arritmias cardíacas [Moody and Mark, 1984; Clifford and McSharry, 2011]. Na qualidade da água, a detecção de anomalias em séries temporais de parâmetros de qualidade é crucial para identificar episódios de contaminação que poderiam representar riscos à saúde pública [Perelman et al., 2012; Tonatzin and Rabatel, 2003]. Na análise de reflexão sísmica, anomalias em séries temporais de dados sísmicos ajudam a identificar eventos tectônicos ou a detecção de reservatórios de hidrocarbonetos [Wu et al., 2019; Yingst and Henden, 2006]. Na perfuração e exploração de petróleo, a identificação de anomalias pode indicar problemas nos equipamentos ou novas reservas de petróleo [Bach et al., 2014; Chen and Malli, 2015].

A detecção de anomalias possibilita a tomada de decisões proativas, melhorando consideravelmente a eficiência operacional, garantindo maior segurança e prevenindo potenciais problemas. Especificamente, essa prática ajuda a agilizar a identificação de falhas, reduzindo o tempo de resposta e mitigando riscos antes que se transformem em problemas críticos. Isso não apenas melhora o desempenho geral dos sistemas, mas também minimiza interrupções, resultando em operações mais seguras e confiáveis [Du et al., 2019].

Anomalias referem-se a mudanças de comportamento que desviam do padrão esperado [Chandola et al., 2009; Gupta et al., 2014]. Identificá-las e classificá-las é desafiador, e escolher o método apropriado para tal propósito é complexo. Características como a forma pela qual a mudança no comportamento se manifesta na série temporal original e a duração do fenômeno contribuem para a

especialização de métodos para determinados tipos de problemas, resultando em uma propensão a gerar falsos positivos em situações divergentes [Salles et al., 2023a]. Portanto, a complexidade reside não apenas na natureza intrínseca das anomalias em séries temporais, mas também na necessidade de escolher abordagens de detecção que considerem adequadamente a especificidade de cada cenário.

Neste cenário dinâmico, a necessidade de métodos robustos e altamente adaptáveis é fundamental, independentemente da origem da série temporal analisada [Schmidl et al., 2022]. Além disso, é essencial que esses métodos possam ser aplicados a diferentes séries temporais sem a necessidade de prévio conhecimento de sua natureza intrínseca. Há também uma demanda por métodos que reduzam a ocorrência de falsos positivos, que lidem adequadamente com dados ruidosos e com cenários não estacionários.

A Decomposição de Modo Empírico Refinada (REMD) foi projetada para abordar esse problema. Trata-se de um método híbrido que integra os benefícios da decomposição de séries temporais usando a Decomposição de Modo Empírico (EMD) [Lei et al., 2013] e da modelagem de séries temporais usando Modelo Autorregresivo de Médias Móveis (ARIMA) [Hamilton, 2020]. O REMD foi concebido com a visão estratégica de combinar dois métodos relativamente adaptáveis para otimizar suas forças e superar limitações.

A abordagem REMD aproveita a capacidade do EMD em lidar com séries temporais não lineares e não estacionárias, decompondo-as em componentes oscilatórios chamados Funções de Modos Intrínsecos (IMFs). Ao adicionar o ARIMA no método, que permite a modelagem precisa dos componentes estacionários da série temporal, proporciona-se uma base robusta para previsões e identificação de anomalias.

O REMD foi avaliado através de comparações com métodos estatísticos e de aprendizado de máquina de última geração, conhecidos como *State Of The Art* (SOTA), no contexto da detecção de anomalias. Foram utilizados diferentes conjuntos de dados, que apresentavam propriedades variadas, como volatilidade, tendência e a presença ou ausência de sazonalidade. Os resultados mostram que o REMD superou os métodos SOTA, alcançando valores de F1 até três vezes superiores. Além disso, o REMD demonstrou ser adaptável a diferentes escalas temporais, desde dados de alta frequência, como transações financeiras, até dados de baixa frequência, como indicadores econômicos mensais. Essa versatilidade evidencia o potencial do REMD para aplicações em diversos domínios e contextos operacionais.

Esse trabalho está organizado em mais cinco capítulos. O capítulo II fornece uma visão geral sobre a detecção de anomalias, o capítulo III discute trabalhos relacionados, o capítulo IV apresenta o REMD, o capítulo V discute os resultados da avaliação experimental, e finalmente, o capítulo VI apresenta as considerações finais.

Capítulo II Referencial teórico

Existem vários conceitos essenciais para a compreensão deste trabalho. Entre os mais importantes estão: séries temporais, anomalias, detecção de anomalias baseada em modelos, Decomposição de Modo Empírico (EMD), Modelo Autorregressivo Integrado de Médias Móveis (ARIMA) e rugosidade. Esses conceitos serão detalhados ao longo deste capítulo. Na Seção II.1, serão discutidos os fundamentos das séries temporais, incluindo suas definições, classificações e componentes. Compreender esses fundamentos é essencial para o desenvolvimento de modelos eficazes de análise de dados temporais. Em seguida, na Seção II.2, abordaremos o conceito de anomalias em séries temporais. Serão descritas as diferentes formas de anomalias e como elas podem ser identificadas e classificadas, o que é crucial para a detecção de comportamentos atípicos nos dados.

A Seção II.3 se concentra na detecção de anomalias baseada em modelos. Serão discutidos os métodos tradicionais e métodos modernos para identificar anomalias em séries temporais, enfatizando as técnicas que têm mostrado eficácia em diferentes cenários de aplicação. Na Seção II.4, será explorada a Decomposição de Modo Empírico (EMD). Esta técnica avançada de decomposição de séries temporais será detalhada, ressaltando suas vantagens e seu uso na análise de dados temporais não lineares e não estacionários. A compreensão da EMD é fundamental, pois o método principal deste trabalho, o REMD, é uma variação da EMD que integra a modelagem ARIMA.

Na Seção II.5, discutiremos o Modelo Autorregressivo Integrado de Médias Móveis (ARIMA). Esta seção explicará como o ARIMA pode ser utilizado para modelar e prever séries temporais, bem como para identificar anomalias por meio dos resíduos do modelo. O ARIMA é parte integrante do REMD, contribuindo para a modelagem dos componentes estacionários das séries temporais.

Finalmente, na Seção II.6, será apresentada a medida de rugosidade média (R_a). Esta seção incluirá a definição da rugosidade, como ela é calculada e sua aplicação prática na análise de séries temporais. No contexto do REMD, a rugosidade é utilizada para separar as Funções de Modo Intrínseco (IMFs) de alta frequência das de baixa frequência, aprimorando a detecção de anomalias.

II.1 Série temporal

Uma série temporal é uma coleção de observações, pontos, organizadas cronologicamente, ou seja, sequenciadas ao longo do tempo, e descrevem um fenômeno de interesse. Formalmente, uma

série temporal Y é descrita como uma sequência de observações $\langle y_1, y_2, ..., y_n \rangle$, onde $y_i \in \mathbb{R}$ e n representa o número de observações. Aqui, y_1 representa a observação mais antiga, enquanto y_n denota a mais recente. Esta representação refere-se a uma série temporal com apenas uma variável, conhecida como série temporal univariada [Shumway and Stoffer, 2017; Fu, 2011]. A Figura II.1 apresenta um exemplo de uma série Y, constituída de y_i pontos ordenados no tempo.

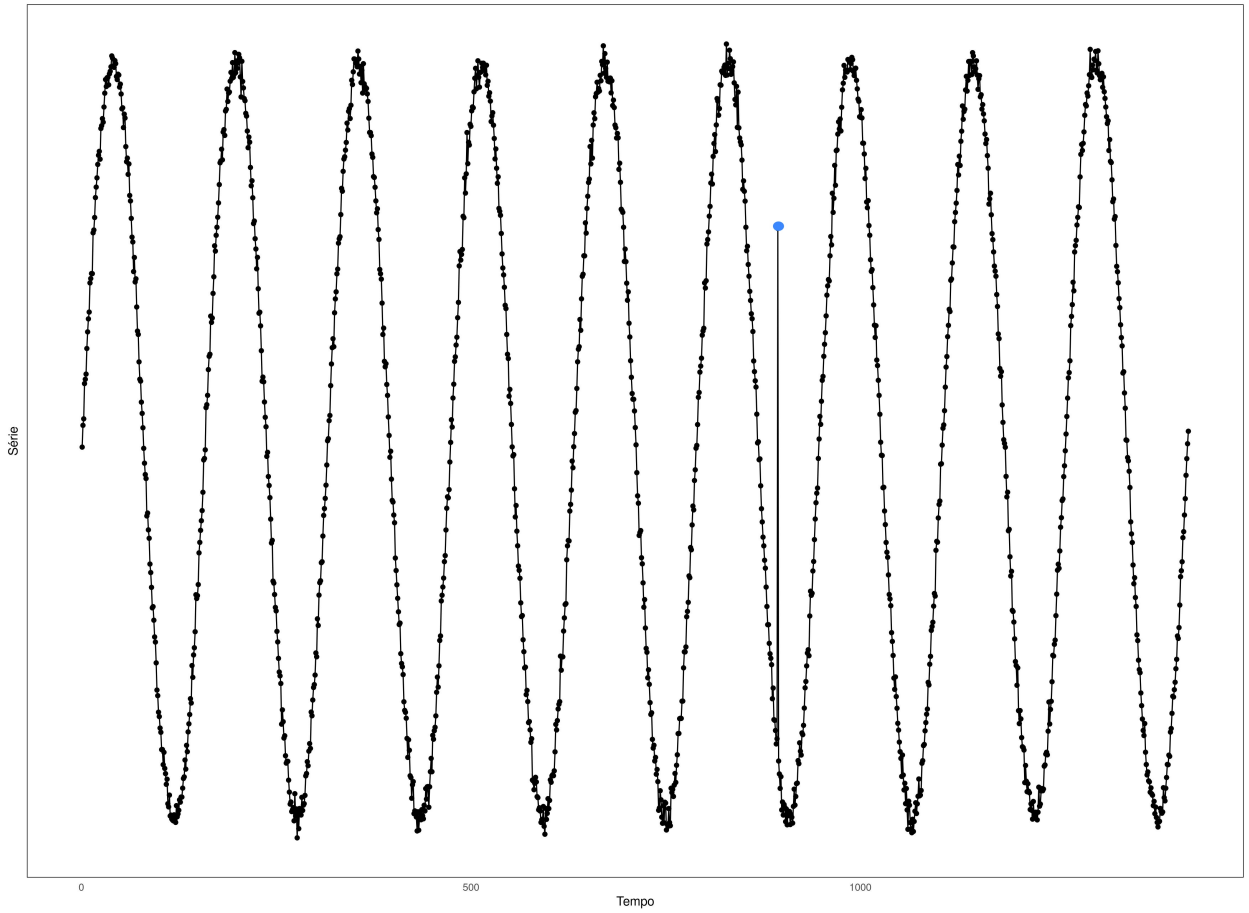


Figura II.1: Série temporal: Sequência de pontos ordenados no tempo. Neste exemplo há a ocorrência de uma anomalia.

As séries temporais podem ser categorizadas em univariadas e multivariadas. Uma série temporal univariada envolve a observação de uma única variável ao longo do tempo, permitindo a análise de um único aspecto do fenômeno observado, como a temperatura diária ou o preço de uma ação. Em contraste, uma série temporal multivariada registra múltiplas variáveis simultaneamente ao longo do tempo, como temperatura, umidade e pressão atmosférica em estudos ambientais [Shumway and Stoffer, 2017; Fu, 2011]. Embora a análise multivariada aumente a complexidade, ela fornece uma visão mais rica e abrangente das inter-relações entre variáveis, permitindo a construção de modelos mais sofisticados e robustos.

Outra classificação fundamental em séries temporais é a distinção entre séries estacionárias e não-estacionárias. Uma série temporal estacionária possui propriedades estatísticas constantes ao

longo do tempo, tais como média, variância e autocovariância. Isso é importante, pois muitos métodos de análise e previsão de séries temporais pressupõem a estacionaridade das séries. Uma série não-estacionária, por outro lado, apresenta mudanças nessas propriedades ao longo do tempo, o que pode complicar a análise e a modelagem. Para determinar a estacionaridade de uma série, são frequentemente utilizados testes estatísticos como o Teste de Dickey-Fuller (ADF), que avalia a presença de uma raiz unitária, e o Teste KPSS (Kwiatkowski-Phillips-Schmidt-Shin), que verifica a hipótese nula de estacionaridade contra a alternativa de uma raiz unitária [Shumway and Stoffer, 2017]. Caso uma série seja identificada como não-estacionária, técnicas de transformação como diferenciação e aplicação de logaritmos são muitas vezes empregadas para estabilizar a média e a variância da série, tornando-a estacionária.

Por fim, mais uma característica importante de uma série temporal é a presença ou ausência de sazonalidade. Sazonalidade refere-se a padrões recorrentes que ocorrem em intervalos regulares, frações do tempo analisado, como semanal, mensal, trimestral ou anual, sendo uma característica frequentemente encontrada em dados econômicos e ambientais. A identificação de sazonalidade pode ser feita por várias técnicas, incluindo análise de componentes principais (PCA), funções de autocorrelação e decomposição de séries temporais, como a Decomposição Clássica ou a *Seasonal-Trend decomposition using Loess* (STL) [Shumway and Stoffer, 2017; Cleveland et al., 1990]. Compreender e ajustar tanto a estacionaridade quanto a sazonalidade são passos essenciais para a análise precisa e eficaz de séries temporais, permitindo a extração de informações substanciais e melhor desempenho na construção de modelos preditivos.

Para entender e modelar melhor as séries temporais, é necessário compreender os componentes que as formam e decompor a série nos elementos tendência, sazonalidade e resíduos. Existem várias técnicas disponíveis para essa decomposição, cada uma com suas próprias vantagens e aplicabilidades. A Decomposição Clássica, por exemplo, separa a série em uma soma ou produto de seus componentes (determinando se a série é aditiva ou multiplicativa), sendo bastante intuitiva e amplamente usada em análises exploratórias [Cleveland et al., 1990]. Outro método popular é a Decomposição STL, que é robusta e adaptável a uma variedade de dados sazonais, permitindo que cada componente seja extraído de forma mais flexível [Cleveland et al., 1990].

A Transformada de Fourier é uma técnica poderosa para a decomposição espectral de séries temporais estacionárias, decompondo a série temporal em suas frequências componentes e proporcionando uma visão clara dos ciclos periódicos presentes nos dados [Brillinger, 1981]. O periodograma, um sumário descritivo da série temporal e análise das frequências existentes, é fundamental nesta abordagem, pois representa os valores observados como uma superposição de ondas senoidais de várias frequências [Shumway and Stoffer, 2017]. Essa metodologia é prática para muitas séries temporais que exibem flutuações cíclicas, cujas frequências, muitas vezes, não são previsíveis antes da

observação dos dados. Essas flutuações cíclicas resultam em um conjunto discreto de frequências, conhecidas como frequências de Fourier.

Para séries temporais não estacionárias, a Decomposição Empírica de Modos (EMD) destaca-se por não pressupor linearidade nem estacionaridade, separando a série em modos intrínsecos de oscilação e capturando dinâmicas complexas e não lineares [Huang et al., 1998a]. Adicionalmente, técnicas multivariadas como a Análise de Componentes Principais (PCA) são extremamente úteis para séries temporais multivariadas, identificando padrões comuns entre múltiplas variáveis ao decompor a matriz de dados em componentes ortogonais [Jolliffe, 2002]. Cada uma dessas metodologias oferece informações distintas e pode ser escolhida com base nas características específicas dos dados e nos objetivos da análise, permitindo a extração de informações mais precisas e relevantes.

Muitos métodos de modelagem de séries temporais são projetados considerando uma série temporal estacionária. Se uma série temporal viola qualquer uma das condições de uma série estacionária, ela é classificada como não-estacionária. Essa falta de estacionaridade indica que as séries temporais mudam suas dinâmicas entre diferentes intervalos, resultando em alterações progressivas na relação de dependência entre as observações de inicio e fim da série [Cao, 2003]. A não-estacionaridade torna a detecção de anomalias mais difícil.

Uma observação formada por uma série temporal y_i é comumente caracterizada pelos componentes de tendência (β_i), sazonalidade (π_i) e resíduos (ω_i) como descrito na Equação II.1 e exibido na Figura II.2. Tendências denotam valores crescentes, decrescentes ou estacionária; sazonalidade refere-se à recorrência periódica de padrões; e os resíduos abrangem as variações imprevisíveis remanescentes dos componentes anteriores [Bell et al., 2018]. Se a tendência e a sazonalidade forem bem caracterizadas, é esperado que o resíduo (ω_i) tenha uma distribuição Gaussiana com média zero e variância finita. Isso é comumente conhecido como ruído branco [Tsay, 2010].

$$y_i = \beta_i + \pi_i + \omega_i \tag{II.1}$$

II.2 Anomalias

Observações atípicas na série temporal que possuem algum significado especial são comumente referidas como anomalias [Chandola et al., 2009]. Essas instâncias se destacam devido à sua diferença notável em relação às demais observações [Atluri et al., 2018]. Nesta situação, as anomalias podem ser modeladas como observações isoladas em relação ao restante da série temporal, utilizando funções de similaridade ou distância [Gupta et al., 2014].

As anomalias podem ser divididas em anomalias pontuais, coletivas ou contextuais. Anomalias pontuais se referem a observações únicas cujo comportamento desvia de forma significativa das

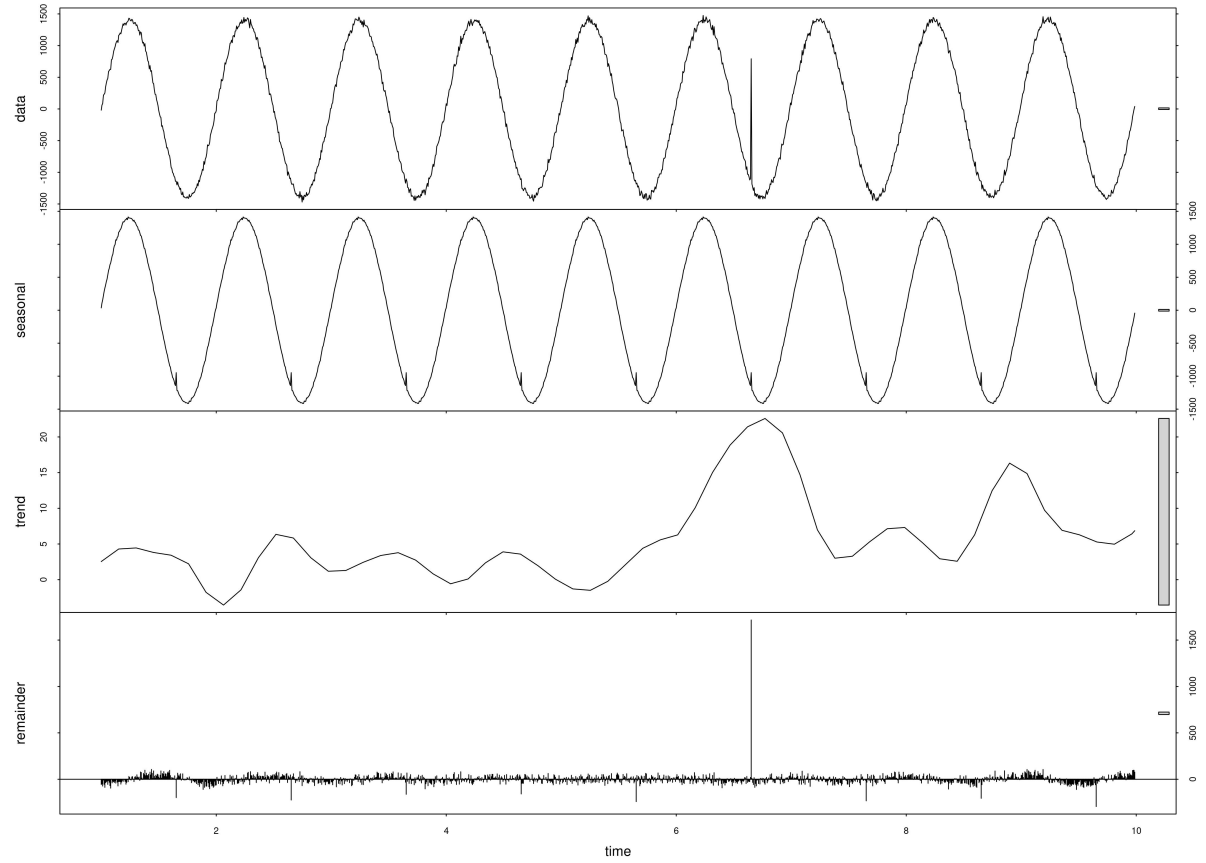


Figura II.2: Série decomposta usando o método STL. Neste método são gerados três componentes: sazonalidade, tendência e ruído.

demais (conforme ilustrado na Figura II.1 na seção II.1, pelo ponto de cor azul), enquanto as anomalias coletivas envolvem um grupo de observações que, em conjunto, não seguem o padrão geral. Já as anomalias contextuais dependem do contexto específico da série temporal, o que torna sua detecção particularmente desafiadora [Gupta et al., 2014]. Um exemplo contendo os três casos é apresentado na Figura II.3. Nessa imagem temos a ocorrência das anomalias pontuais, em azul, das coletivas, em verde, sendo essas oscilantes do padrão por desviarem da média da série, e das contextuais, em vermelho, que apesar de não desviarem da média histórica da série, contextualmente para o perfil do dado, pode representar um problema.

Considerando uma série temporal Y, a análise de discrepâncias é a forma mais simples de identificar anomalias. Suas posições de índice podem ser expressas como $A(Y)$ por meio da Equação II.2, onde $Q_1(Y)$ e $Q_3(Y)$ representam o primeiro e o terceiro quartis, respectivamente, e o $IQR(Y)$ é o intervalo interquartil para Y [Gupta et al., 2014]. Se a série temporal for paramétrica, ou seja, se enquadrar nas características descritas na Seção II.1, as discrepâncias podem ser identificadas usando a Equação II.3 baseada no Teorema do Limite Central [Han et al., 2022], onde $\overline{Y}$ é a média de Y e $\sigma(Y)$ é o desvio padrão de Y.

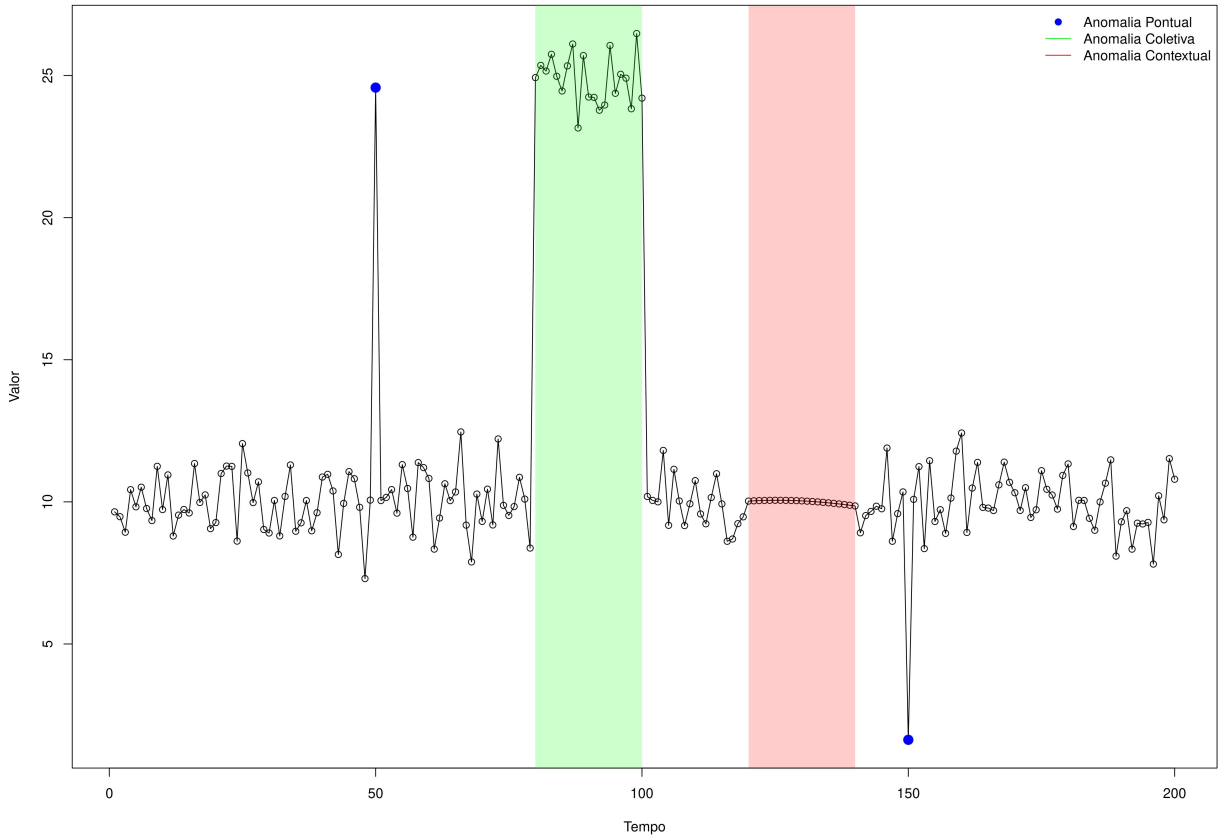


Figura II.3: Série artificial com três perfis de anomalias - Pontuais, coletivas e contextuais.

$$\begin{aligned} &A(Y) = \{i\}, \forall i \in \{1, \cdots, n\}, |Y| = n \mid \\ &y_i \not\subseteq [Q_1(Y) - 1.5 \cdot IQR(Y), Q_3(Y) + 1.5 \cdot IQR(Y)] \end{aligned} \tag{II.2}$$

$$\begin{aligned} &A(Y) = \{i\}, \forall i \in \{1, \cdots, n\}, |Y| = n \mid \\ &y_i \not\subseteq [\overline{Y} - 2.698 \cdot \sigma(Y), \overline{Y} + 2.698 \cdot \sigma(Y)] \end{aligned} \tag{II.3}$$

A detecção de anomalias é crucial em várias aplicações práticas. Por exemplo, na manutenção preditiva de equipamentos industriais, a identificação precoce de um padrão de vibração anômalo em uma máquina pode indicar um problema iminente, permitindo que a manutenção seja realizada antes que ocorra uma falha catastrófica. As anomalias seriam observações que desviam drasticamente do padrão normal de operação.

Outro exemplo é a detecção de fraudes financeiras. Em sistemas de monitoramento de transações bancárias, uma anomalia pode ser uma transação, incomum, em abundância ou em localização, não habitual, indicando possíveis atividades fraudulentas. A rápida detecção dessas anomalias pode

prevenir perdas financeiras significativas para bancos e clientes.

II.3 Detecção de Anomalias Baseada em Modelos

No contexto de séries temporais, as anomalias podem se manifestar em tendências ou volatilidade. Anomalias de tendência ocorrem quando uma observação se desvia da tendência esperada na série temporal. Por outro lado, anomalias de volatilidade ocorrem quando há uma mudança significativa na variação das amostras da série temporal [Gupta et al., 2014; Aminikhanghahi and Cook, 2017].

Há um interesse específico em identificar os pontos fora do padrão, mais conhecidos como *outliers*, que possam indicar uma anomalia no processo de geração da série temporal Y. Essas anomalias são denominadas anomalias de tendência (TA). Se considerarmos $\hat{Y}$ como uma estimativa do processo gerador de Y, obtida ao ajustar um modelo α, onde $\hat{Y} = \alpha(Y)$, podemos observar que os resíduos ω (ruído branco) são obtidos removendo $\hat{Y}$. Portanto, as anomalias de tendência (TA) em Y são comumente identificadas como $TA(Y, \hat{Y})$ através da Equação II.4.

$$TA(Y, \hat{Y}) = A(\omega) \mid \omega_i = y_i - \hat{y}_i \tag{II.4}$$

II.4 Decomposição de Modo Empírico (EMD)

A Decomposição de Modo Empírico (EMD) é uma técnica eficaz de análise temporal para séries não lineares e não estacionárias. Este método permite decompor uma série temporal em componentes oscilatórios intrínsecos chamados Funções de Modo Intrínseco (IMFs) [Lei et al., 2013; Ambrosino et al., 2020; Huang et al., 1998b]. Diferente de outras técnicas, a EMD não pressupõe uma formulação matemática específica para os dados, adaptando-se às variações complexas da série analisada.

Uma IMF é uma função que satisfaz duas condições essenciais: (1) o número de extremos (pontos máximos e mínimos) e o número de cruzamentos por zero devem ser iguais ou diferir no máximo por uma unidade ao longo de toda a série, e (2) em qualquer ponto, o valor médio dos envelopes definidos pelos máximos e mínimos locais deve ser zero. Essas funções representam modos oscilatórios simples embutidos na série temporal, capturando diferentes escalas de oscilação intrínsecas aos dados.

Os envelopes desempenham um papel crucial na definição das IMFs. Um envelope consiste em duas curvas: uma que suaviza os máximos locais e outra que suaviza os mínimos locais da série. Esses envelopes criam uma espécie de "emolduramento"para a série temporal, capturando suas oscilações. O valor médio entre esses dois envoltórios é utilizado para garantir que a IMF tenha um valor médio zero, eliminando tendências de longo prazo e isolando oscilações de curto prazo. O processo de EMD é aplicado a partir da série temporal Y e envolve duas etapas principais: a

extração das IMFs e a determinação do resíduo final ω. Ao final do processo, a série temporal Y pode ser representada como a soma das IMFs C_j $(1 \leq j \leq k)$ e do resíduo ω, ou seja, $Y = \sum_{j=1}^{k} c_j + \omega_i$. As IMFs capturam diversas faixas de frequência, desde as mais altas até as mais baixas, enquanto o resíduo ω simboliza a tendência média da série temporal Y [Lei et al., 2013].

A Figura II.4 e o Algoritmo 1 ilustram o processo de extração das IMF's usando o algoritmo EMD. O algoritmo inicializa os parâmetros com a série temporal Y, incluindo o índice i da IMF, o resíduo inicial r_0 (igual a Y), e um conjunto vazio de IMFs. Durante a iteração principal, o algoritmo extrai sequencialmente cada IMF. A extração de cada IMF envolve um loop interno onde a série atual $h_{i,k}$ é submetida a uma análise para identificar máximos e mínimos locais.

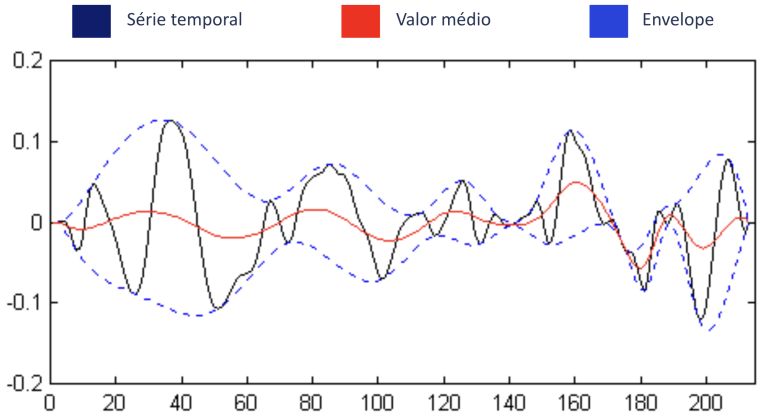


Figura II.4: Processo de extração das IMF's. Neste processo é gerado um envelope, na série original, e calculado o valor médio ao longo da série.

A função *envelop* (em azul na Figura II.4 e descrita como envelope) desempenha um papel importante ao extrair máximos e mínimos locais da série $h_{i,k-1}$, usando interpolação *spline* cúbica para criar os envelopes superior e inferior. A etapa seguinte envolve verificar se a função resultante após a subtração, $h_{i,k}$, é uma IMF. Se for, a IMF é adicionada ao conjunto de IMFs e o loop interno se encerra. Caso contrário, o índice k é incrementado e o processo é repetido [Huang et al., 1998b]. Uma informação adicional relevante para compreender o motivo pelo qual a função de envelope utiliza uma *spline* cúbica é que isso garante um comportamento suave, possibilitando transições naturais entre os pontos sem incrementar desnecessariamente a complexidade do processamento.

Após a extração da IMF, o resíduo subsequente r_{i+1} é calculado subtraindo a última IMF extraída do resíduo atual. A função *hasEndpoints* do Algoritmo 1 verifica se o resíduo ainda possui no mínimo dois pontos extremos. Se possuir, o índice i é incrementado e o processo continua para a extração da próxima IMF. Se não, o algoritmo é concluído, retornando o conjunto de IMFs e o último resíduo ω. Na Figura II.5 temos o EMD aplicado a série apresentada anteriormente na Figura II.1, nela temos todas as IMF's resultantes da aplicação do EMD até restar apenas o resíduo.

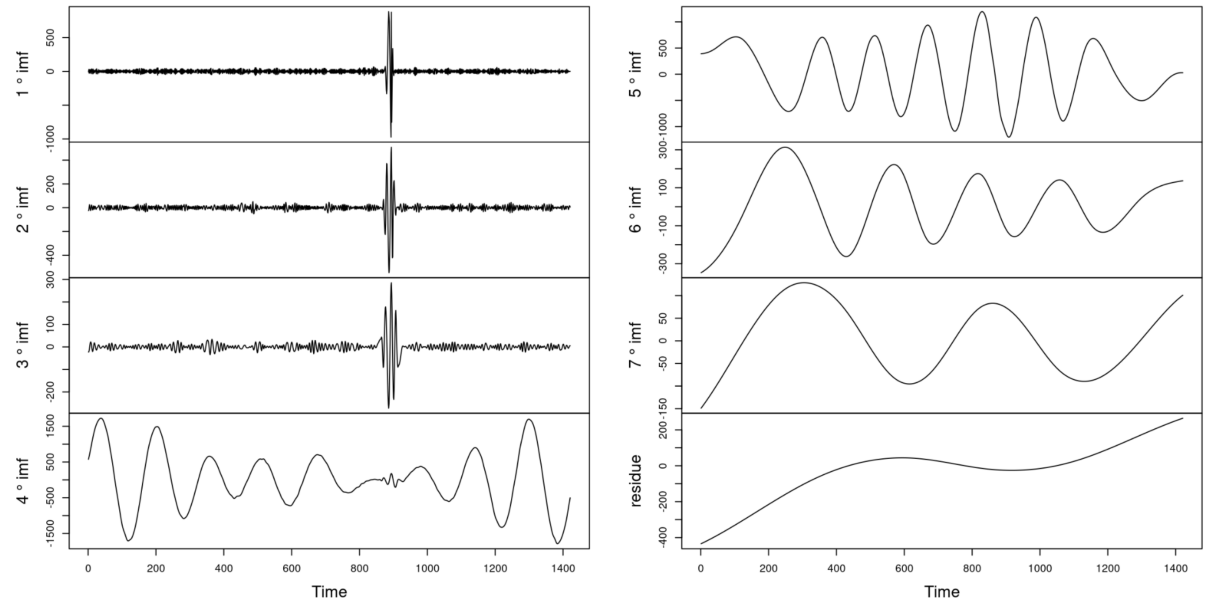


Figura II.5: Resultado da decomposição da série original usando o EMD. Formada por sete IMF's e uma série de resíduo.

A EMD permite a decomposição detalhada de séries em componentes interpretáveis, facilitando a análise de fenômenos complexos em dados temporais. Esta técnica robusta fornece uma ferramenta adaptativa poderosa para a compreensão das dinâmicas temporais inerentes aos dados [Lei et al., 2013; Ambrosino et al., 2020]. Em resumo, a EMD é uma abordagem versátil e eficaz para a análise temporal, especialmente útil na presença de séries temporais não lineares e não estacionárias.

O processo tradicional de detecção de anomalias usando EMD constitui-se de duas etapas: a primeira é a decomposição da série original em IMFs, e o segundo passo é a detecção de pontos anômalos na primeira IMF, por ser a série que apresenta as oscilações de forma mais intensa. Em seguida, os pontos considerados *outliers* serão classificados como anomalias.

II.5 Modelo Autorregressivo Integrado de Médias Móveis (ARIMA)

O Modelo Autorregressivo de Médias Móveis (ARMA) visa modelar séries temporais estacionárias. ARIMA (Autorregressivo Integrado de Médias Móveis) é uma extensão da ideia de modelar séries para abordar a não estacionaridade nas séries temporais. As propriedades de uma série temporal estacionária são independentes do momento em que são observadas, o que torna os modelos estáveis e previsíveis ao longo do tempo. No entanto, muitas séries temporais do mundo real não são estacionárias devido à presença de tendências ou sazonalidades.

O ARIMA é eficaz porque considera sazonalidade, tendências e outros aspectos não estacionários ao modelar a série temporal. Se uma observação desviar significativamente do modelo ARIMA

Algorithm 1 Algoritmo EMD (adaptado de Lei et al. [2013])

function EMD(Y)
    $i \leftarrow 1$, $r_0 \leftarrow Y$, $IMFs \leftarrow \emptyset$ ▷ Inicialização
    $finish \leftarrow false$
    repeat ▷ Extração da énesima IMF
        $k \leftarrow 1$, $h_{i,0} \leftarrow r_{i-1}$
        $found \leftarrow false$
        repeat
            $m_{i,k-1} \leftarrow \text{envelop}(h_{i,k-1})$
            $h_{i,k} \leftarrow h_{i,k-1} - m_{i,k-1}$
            if isIMF($h_{i,k}$) **then**
                $found \leftarrow true$
                $c_i \leftarrow h_{i,k}$
                $IMFs \leftarrow IMFs \cup c_i$
            end if
            $k \leftarrow k + 1$
        until $found$
        $r_{i+1} \leftarrow r_i - c_i$
        if hasEndpoints(r_{i+1}) **then**
            $i \leftarrow i + 1$
        else
            $finish \leftarrow true$
            $\omega \leftarrow r_{i+1}$
        end if
    until $finish$
    return $IMFs$, ω
end function

ajustado, ele pode ser considerado anômalo [Ray et al., 2018].

Um modelo ARIMA inclui três tipos de parâmetros [Moayedi and Masnadi-Shirazi, 2008]: (p) parâmetro autorregressivo que indica o número de valores passados da série utilizados para prever o valor atual; (d) quantidade de diferenciações aplicadas à série para remover a não-estacionaridade; (q) o parâmetro de média móvel, ele representa o número de termos de erro passados utilizados na construção do modelo.

Um modelo ARIMA pode ser resumido como $ARIMA(p, d, q)$. O processo é generalizado na Equação II.5, onde ω_i é o erro de previsão, ϕ e θ são os coeficientes autorregressivo e de média móvel de ordem finita p e q, respectivamente. Além disso, B representa o operador de defasagem, e d representa a parte integrada do modelo (número de vezes que a série é diferenciada).

$$\left(1 - \sum_{j=1}^{p} \phi_j B^j\right) (1 - B)^d x_i = \left(1 + \sum_{k=1}^{q} \theta_k B^k\right) \omega_i \tag{II.5}$$

A diferenciação (ou integração) transforma uma série temporal não estacionária em uma série estacionária subtraindo o valor atual da série de seu valor anterior. Por exemplo, para um processo de primeira diferenciação (d = 1), a nova série é obtida como $y'_t = y_t - y_{t-1}$.

Uma vantagem significativa dos modelos ARIMA é sua flexibilidade e capacidade de capturar características complexas de séries temporais através da diferenciação e da combinação dos termos autorregressivo e de média móvel. No entanto, a definição precisa dos parâmetros p, d, e q pode ser desafiadora e frequentemente requer uma abordagem iterativa baseada em métodos de seleção de modelos, como o Critério de Informação de Akaike (AIC) ou o Critério Bayesiano de Schwarz (BIC).

Para ajustar esses parâmetros, a análise das funções de autocorrelação (ACF) e autocorrelação parcial (PACF) é frequentemente utilizada. A ACF avalia a correlação entre os valores da série e seus próprios valores defasados, sendo útil para determinar o parâmetro q de média móvel; aqui, procura-se por cortes súbitos nos picos significativos nos primeiros passos do gráfico, o que sugere o número de termos de média móvel necessários. Por outro lado, a PACF mede a correlação entre valores que excluem a influência de passos intermediários, ajudando na identificação do parâmetro p autorregressivo. No gráfico da PACF, a presença de picos significativos, que terminam após um certo atraso, indicam o número de termos autorregressivos apropriados. Em conjunto, a leitura das métricas ACF e PACF permite uma visualização direta da estrutura inerente da série, orientando na seleção precisa dos parâmetros do modelo ARIMA. Após o ajuste inicial do modelo, é importante realizar diagnósticos para assegurar a adequação do modelo, analisando os resíduos e realizando testes de aderência de estacionaridade.

Para garantir que o modelo ajustado é apropriado, recomenda-se observar os resíduos do modelo.

Resíduos ideais devem ser distribuídos aleatoriamente (sem padrões discerníveis) e se assemelham a um ruído branco, indicando que o modelo ARIMA capturou bem as dinâmicas da série temporal analisada.

A detecção eficaz de anomalias em séries temporais usando modelos ARIMA já foi documentada anteriormente [Pena et al., 2013; Blázquez-García et al., 2021]. Ela se baseia basicamente em identificar os pontos anômalos no resíduo de uma série predita pelo ARIMA, ou seja, do que não consegue ser explicado/modelado pelos parâmetros p, d e q.

II.6 Rugosidade média (R_a)

A rugosidade média, denotada como R_a, é uma medida frequentemente usada para quantificar a rugosidade superficial de um material ou superfície. A rugosidade de uma superfície inclui suas irregularidades e ondulações, que podem influenciar significativamente no comportamento em diversas aplicações, como na engenharia de materiais com desgaste, aderência e fricção [Mainsah et al., 2001] . O conceito de R_a permite avaliar de forma objetiva a qualidade da superfície ao examinar a média aritmética das variações absolutas da altura da superfície em relação a uma linha média predefinida ao longo de um comprimento L de avaliação. Na Figura II.6 é exibida uma série sinuosa, com valores que se distanciam dos pontos da linha média ao longo de L.

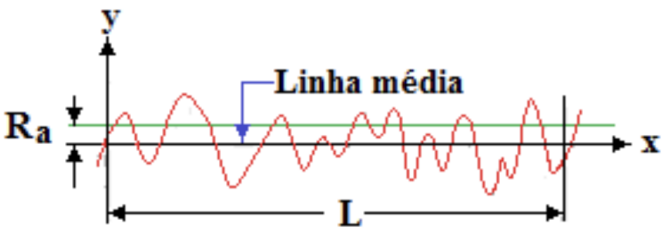


Figura II.6: O cálculo da rugosidade é resultado do grau de oscilação de uma superfície de comprimento L.

O cálculo de R_a é descrito na Equação II.6 e envolve a determinação da média das magnitudes absolutas das variações de altura do perfil da superfície [Whitehouse, 2010] . A linha média é uma referência horizontal que divide o perfil da superfície de forma que a área das protuberâncias acima dessa linha seja igual à área dos vales abaixo dela, onde L representa o comprimento do perfil de avaliação e $y(x)$ é a altura do perfil da superfície em um ponto x.

$$R_a = \frac{1}{L} \int_0^L |y(x)| \, dx \tag{II.6}$$

Esse método de quantificação é amplamente utilizado em diversas indústrias, como na engenharia mecânica para assegurar o funcionamento adequado das peças em movimento, na manufatura para o controle de processos de usinagem e acabamento, e na medicina para a avaliação de implantes e

dispositivos médicos [Stout and Blunt, 2000].

Quando discutimos a rugosidade média no domínio da frequência, consideramos como as variações na altura da superfície se distribuem em termos de frequências. De forma simplificada, quanto maior a rugosidade média de uma superfície, maior será a frequência do sinal quando este é transformado em um espectro de frequência. Isso ocorre porque superfícies com maior rugosidade têm mais irregularidades em distâncias menores, resultando em componentes de alta frequência no espectro [Luo et al., 2008].

A Figura II.7 compara a primeira e a sétima IMFs da série apresentada na Figura II.1. A primeira IMF tem mais oscilações num mesmo espaço de comprimento L; em termos numéricos, isso pode ser medido como uma rugosidade de 0.74 e uma frequência dominante de 0.04 Hertz. Para a sétima IMF, temos uma rugosidade de 0.0002 e uma frequência dominante de 0.001 Hertz. Portanto, ao comparar duas séries temporais e suas rugosidades, se uma delas possui um R_a maior, essa superfície terá uma maior frequência dominante em relação à segunda. Essa análise de frequência é importante para a compreensão detalhada das características da superfície.

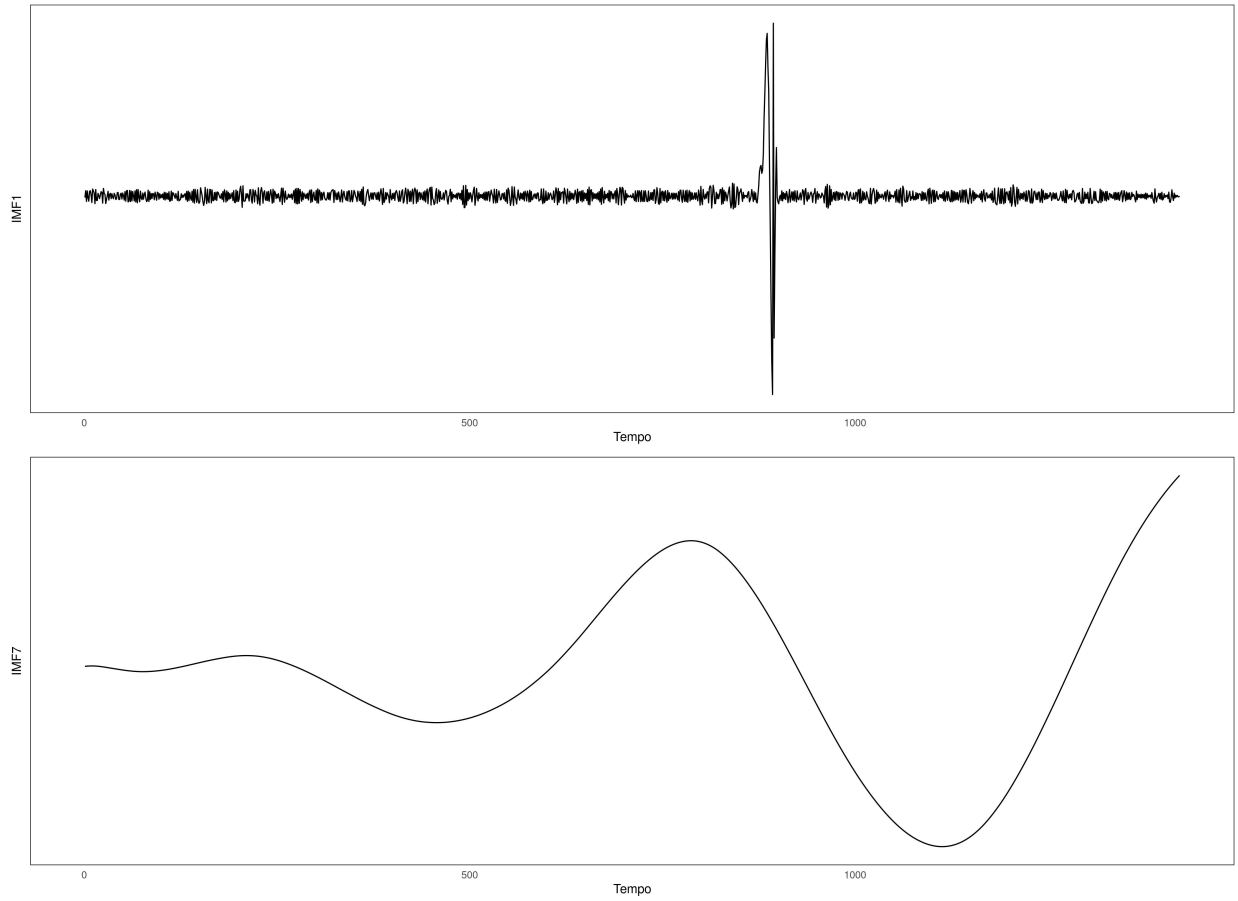


Figura II.7: Relação entre rugosidade média e frequência - A IMF com maior grau de oscilação, apresenta maior rugosidade e maior frequência dominante.

Capítulo III Trabalhos relacionados

Nos últimos anos, muitos pesquisadores desenvolveram métodos para a análise de anomalias em séries temporais [Ambrosino et al., 2020]. De acordo com Chandola et al. [2009], as anomalias podem ser classificadas em três categorias principais: pontuais, coletivas e contextuais. Anomalias pontuais são observações individuais que se destacam como incomuns em relação ao restante da série e geralmente ocorrem em regiões de baixa densidade de valor, indicando que esses valores são isolados e aparecem raramente, destacando-se como *outliers* em um contexto mais amplo.

Por sua vez, as anomalias coletivas envolvem um conjunto de valores que, quando observados em conjunto, formam um padrão anômalo. Embora individualmente possam não parecer incomuns, sua combinação ou sequência foge dos padrões esperados, independentemente da densidade de valor. Já as anomalias contextuais são valores que, apesar de não estarem em regiões de baixa densidade globalmente, são consideradas anômalas em relação aos valores vizinhos ou em um contexto específico. Nesse caso, a densidade de valor é entendida de forma contextual, indicando que um valor pode parecer normal de forma ampla, mas torna-se anômalo em uma análise localizada.

Além dessas categorias, é relevante discutir a importância das anomalias sazonais e de tendência, especialmente em dados de monitoramento contínuo como em sistemas de detecção de fraudes ou de falhas estruturais. A categorização adequada permite a escolha de técnicas mais bem-adaptadas para a detecção de cada tipo específico de anomalia, melhorando a precisão e a eficácia dos métodos implementados.

Essas formulações são inerentemente distintas, exigindo abordagens específicas. Mesmo dentro de uma única formulação de problema, várias técnicas empregando diferentes métodos para detecção de anomalias têm sido propostas. Existem vários tipos de métodos [Chandola et al., 2012], incluindo métodos baseados em agrupamento [Çelik et al., 2011], abordagens de vizinhos mais próximos [Breuniq et al., 2000; Kriegel et al., 2009], estratégias de redução de dimensionalidade [Takeishi and Yairi, 2014], redes neurais [Nanduri and Sherry, 2016; Provotar et al., 2019], Máquinas de Vetores de Suporte (SVM) [Ma and Perkins, 2003], métodos estatísticos como ARIMA [Moayedi and Masnadi-Shirazi, 2008; Thiyagarajan et al., 2017] e Forward and Backward Inertial Anomaly Detector (FBIAD) [Lima et al., 2022], Extreme Learning Machine (ELM) [Huang et al., 2012], Long Short-Term Memory (LSTM) e Redes Neurais Convolucionais 1D (Conv1D) [Salles et al., 2023b], e métodos híbridos [Ambrosino et al., 2020]. No entanto, cada um apresenta desvantagens

relacionadas à especificação de parâmetros, interpretabilidade, generalização e custo computacional [Goldstein and Uchida, 2016; Chandola et al., 2009].

A adaptação e a melhoria desses métodos continuam sendo áreas contínuas de pesquisa, uma vez que se busca aumentar a precisão, reduzir falsos positivos e garantir que os métodos possam ser aplicados de maneira escalável em grandes volumes de dados. A integração de técnicas de aprendizagem profunda com métodos estatísticos tradicionais está emergindo como uma tendência promissora para alavancar o melhor de ambos os mundos [Braei and Wagner, 2020].

Há uma presença significativa de modelos híbridos que utilizam múltiplos métodos. Esses métodos podem combinar várias técnicas distintas, aproveitando as vantagens individuais, juntamente com algoritmos otimizados, proporcionando maior precisão e reduzindo a incerteza geral dos métodos individuais [Ambrosino et al., 2020].

Os modelos híbridos abordam frequentemente a complexidade das séries temporais ao combinar métodos de decomposição com técnicas de previsão e detecção de anomalias. Um exemplo disto é a combinação de decomposição de modos (como EMD) com técnicas preditivas (como ARIMA), explorando o potencial de detecção de características multiescalas em séries temporais complexas.

Cui and Wang [2017] propuseram um sistema de detecção de anomalias e visualização de dados para uma empresa de gestão de instalações, visando detectar anomalias em instalações elétricas de escolas. Esse sistema é baseado em um modelo híbrido que combina regressão polinomial e distribuição gaussiana. Esse estudo exemplifica como modelos híbridos podem ser aplicados para solucionar problemas específicos em ambientes industriais e comerciais, fornecendo previsões mais robustas e detecção de anomalias que podem prevenir falhas sistêmicas.

Lin et al. [2020] propuseram um modelo híbrido não supervisionado para detecção de anomalias em séries temporais. O modelo utiliza um módulo de *autoencoders* variacionais para formar características locais robustas em uma janela curta e um módulo de rede neural recorrente com a arquitetura LSTM para estimar correlações de longo prazo na sequência. Essa combinação permite que o modelo identifique anomalias que podem abranger múltiplas escalas de tempo. Ao explorar os benefícios das técnicas de *autoencoders* variacionais e redes neurais recorrentes, esse método aborda a complexidade de capturar tanto as anomalias locais quanto as globais em séries temporais, destacando seu potencial para aplicações dinâmicas onde os padrões de anomalias podem variar significativamente com o tempo.

Zhang et al. [2019] propuseram um modelo híbrido de três camadas para detecção de *outliers* em dados de sensores. Este modelo incorpora redes neurais, filtro mediano (MF), EMD, árvore de classificação e regressão (CART), assim como métodos autorregressivos (AR) e média móvel exponencialmente ponderada (EWMA). A primeira camada da rede neural é responsável por mitigar a influência de *outliers* evidentes. A segunda camada, composta por módulos EMD-CART-AR, atua

como um modelo para a série temporal. A terceira camada, representada pelo detector EWMA, identifica potenciais *outliers* sutis na série temporal residual e no modelo ajustado.

O modelo de múltiplas camadas exemplifica como diferentes técnicas podem ser integradas para fornecer uma solução abrangente e robusta para a detecção de anomalias, lidando com a complexidade das séries temporais e melhorando a precisão da detecção de *outliers* ao longo de diferentes etapas do processamento. Akoglu and Tong [2019] investigaram técnicas de detecção de anomalias em séries temporais financeiras usando aprendizagem de máquina. Seu trabalho mostra como combinações de modelos estatísticos com técnicas avançadas de aprendizagem de máquina podem ser eficazes na detecção robusta de anomalias em dados altamente voláteis como preços de ações e transações financeiras.

Foi observado que os dados financeiros frequentemente apresentam características que desafiam os modelos tradicionais devido a sua alta variabilidade e presença de *outliers*. As técnicas híbridas que combinam métodos estatísticos com aprendizagem de máquina demonstram uma melhor capacidade de detectar anomalias sem comprometer a robustez dos resultados.

Em sistemas de monitoramento de saúde, Campos and Zimek [2016] utilizaram técnicas de aprendizagem de máquina para detectar anomalias em dados de sinais vitais de pacientes. Este tipo de abordagem é essencial para a detecção precoce de condições críticas de saúde, permitindo intervenções rápidas e precisas.

A aplicação desses métodos em ambientes hospitalares é crucial para o monitoramento contínuo de pacientes, auxiliando na detecção de alterações em padrões vitais que precedem condições críticas, como paradas cardíacas ou complicações pós-cirúrgicas.

No setor energético, [James and Woodall, 2020] propuseram um método híbrido para detecção de anomalias em redes elétricas inteligentes (*smart grids*). O framework combina modelos estatísticos com técnicas de aprendizagem profunda para identificar falhas e prevenir apagões, melhorando significativamente a resiliência e a confiabilidade das redes de distribuição de energia.

A detecção de anomalias em redes elétricas inteligentes é essencial para assegurar a continuidade do fornecimento de energia e a prevenção de apagões. Métodos híbridos permitem a identificação de padrões de falhas que podem passar despercebidos por técnicas tradicionais, reforçando a segurança e a eficiência da distribuição elétrica.

Os trabalhos aqui discutidos destacam a relevância e a diversidade das abordagens para a detecção de anomalias em séries temporais. No entanto, ainda há muitos desafios a serem enfrentados, como a adaptação dos modelos a diferentes tipos de dados e a melhoria da precisão e eficiência das técnicas existentes. A contínua evolução das técnicas de aprendizagem de máquina e estatística promete avanços significativos na detecção de anomalias, proporcionando aplicações cada vez mais eficazes e abrangentes.

Capítulo IV Método

A Decomposição de Modo Empírico Refinada (REMD, do inglês Refined Empirical Mode Decomposition) é uma abordagem inovadora para a detecção de anomalias em séries temporais. Este detector é um método híbrido que combina EMD com o modelo ARIMA. O processo de detecção ocorre em quatro fases: (1) Decomposição por EMD, (2) Separação de Frequências, (3) Ajuste do Modelo e (4) Análise dos Resíduos. REMD é ilustrado no Algoritmo 2 e descrito nas seções seguintes.

Algorithm 2 Algoritmo REMD

function REMD(Y) ▷ Inicialização
    $IMFs, \omega \leftarrow$ EMD(Y) ▷ Algoritmo 1
    $H, L \leftarrow$ FrequencySeparation($IMFs, \omega$)
    $\hat{H} \leftarrow$ ModelAdjustment(H)
    $anomalies \leftarrow$ ResidualAnalysis($H, \hat{H}$)
    return $anomalies$
end function
function FrequencySeparation($IMFs$, ω)
    $RAs \leftarrow$ Roughness($IMFs$)
    $m \leftarrow$ MaxCurvature(RAs)
    $H \leftarrow \{IMFs_1, \ldots, IMFs_m\}$
    $L \leftarrow \{IMFs_{m+1}, \ldots, IMFs_{|IMFs|}, \omega\}$
    return H, L
end function
function ModelAdjustment(H)
    $model \leftarrow$ FitARIMA(H)
    $\hat{H} \leftarrow$ PredictARIMA(H)
    return $\hat{H}$
end function
function ResidualAnalysis($H, \hat{H}$)
    $\omega \leftarrow H - \hat{H}$
    $p_i = \frac{|1-\omega_i|}{\max \omega}$ ▷ Equação 10
    Let P be the set of p_i
    $anomalies \leftarrow A(P)$ ▷ Equação 3
    return $anomalies$
end function

IV.1 Decomposição de Modo Empírico (EMD)

Para a série temporal Y, a Decomposição de Modo Empírico (EMD) decompõe a série em um conjunto de Funções de Modo Intrínsecas (IMFs), conforme mostrado na Equação IV.1. O procedimento passo a passo ao qual a série temporal é submetida está detalhado no Algoritmo 1 no capítulo II seção II.4. Esta decomposição visa obter componentes modais empíricas que capturam diferentes bandas de frequência da série temporal, deixando apenas o resíduo, referido como a tendência principal neste contexto. Este processo é iterativo, e com os mesmos parâmetros definidos para o EMD, diferentes séries temporais podem gerar diferentes Funções Intrínsecas de Modo (IMF) [Huang et al., 1998b].

$$y_i = \omega_i + IMF_{1i} + IMF_{2i} + \cdots + IMF_{ni} \tag{IV.1}$$

Uma vantagem significativa de empregar este método é sua adaptabilidade a diferentes tipos de séries temporais e aplicações, considerando apenas as características próprias das ondas da série temporal.

IV.2 Separação de Frequências

Com as IMFs geradas, é necessária a separação dos componentes considerados para o processo de detecção de anomalias. Para isso, foi criado um método novo para separar as IMF's em frequências altas e baixas usando a rugosidade média da superfície (R_a) [Thomas, 1981]. A técnica utilizada é inspirado no método de separação das componentes apresentado no método RED [Baroni et al., 2024] presente no framework Harbinger. Quanto maior o valor de R_a, mais rugosa é a superfície. Neste contexto, quanto mais rugosa for uma superfície, maior será a amplitude de oscilação da série temporal, menor a distância entre as oscilações, ou seja, mais concentradas no mesmo espaço de análise, e consequentemente, maior será a frequência [Thomas, 1981].

Dentro do processo de separação de frequências, aplica-se o cálculo da curvatura máxima aos R_a calculados. A curvatura máxima mede a taxa de variação da curvatura em relação a uma dada direção. Em um ponto de curvatura máxima, a superfície está curvando mais rapidamente naquela direção específica [Bates and Watts, 1980]. Para as coordenadas x e y no plano, com z = f(x, y), os autovalores λ_1 e λ_2 da matriz Hessiana são calculados [Stewart, 2020]. O valor máximo de curvatura entre λ_1 e λ_2 é, então, utilizado como critério para determinar quais as funções de modo intrínseco (IMFs) serão consideradas de alta frequência.

A série (IMF) que apresentar a rugosidade no ponto de inflexão/ponto de curvatura máxima dividirá as bandas de alta frequência (conjunto de séries de maior frequência) das bandas de baixa frequência (conjunto de séries com menor frequência). Em outras palavras, m é o ponto de otimi-

zação para esse conjunto de séries.

Ao calcular m, duas novas séries temporais podem ser obtidas: a série temporal de alta frequência H (Equação IV.2) e a série temporal de baixa frequência L (Equação IV.3), onde m é o ponto de curvatura máxima, e n é o número máximo de componentes.

$$H = IMF_1 + IMF_2 + \ldots + IMF_m \tag{IV.2}$$

e

$$L = IMF_{m+1} + \ldots + IMF_n + \omega_i \tag{IV.3}$$

Essa abordagem difere significativamente de outras abordagens baseadas em EMD, como descrito na seção II.4, e é uma contribuição importante deste trabalho. Enquanto os métodos baseados em EMD apenas usam a primeira IMF, o REMD inclui as primeiras m IMFs como a componente de alta frequência e as IMFs restantes caracterizam a componente de baixa frequência da série temporal, que incluem componentes de tendência e sazonalidade.

IV.3 Ajuste do Modelo

A terceira etapa foi adicionada para aprimorar a técnica e refinar o método puro de EMD. O modelo ARIMA é aplicado à série temporal H nesta etapa, usando a Equação II.5 como base, e representado de forma simplificada na equação IV.4. O resultado do ajuste do modelo é a série temporal $\hat{H}$.

$$\hat{H} = ARIMA(p, d, q)(H) \tag{IV.4}$$

Esta etapa combina a capacidade de decomposição multiescalar do EMD com o poder preditivo do ARIMA, resultando em uma abordagem robusta para modelagem e análise de séries temporais que capturam tanto os componentes de alta frequência quanto as tendências mais amplas.

IV.4 Identificação de Anomalias

A etapa de identificação de anomalias usa as séries temporais calculadas nas etapas segunda e terceira. Seja ω o resíduo de ambas as séries temporais (H e $\hat{H}$), de modo que $\omega_i = h_i$ - $\hat{h}_i$. O objetivo é identificar a probabilidade p_i de uma observação h_i ser coerente com $\hat{h}_i$, conforme descrito pela Equação IV.5.

$$p_i = 1 - \frac{|\omega_i|}{\max(|\omega|)} \tag{IV.5}$$

Com a probabilidade para cada instante de tempo, as anomalias podem ser calculadas usando o critério de discrepância para distribuição paramétrica (Equação II.3). Todos os pontos no vetor de probabilidade considerados discrepantes são classificados como anomalias.

O valor de saída p_i sempre apresenta valores entre 0 e 1, pois w_i é o resíduo analisado no momento i e ω é a série com todos os resíduos. Logo $\max(|\omega|)$ é o maior valor possível dos resíduos, e a divisão de $|\omega_i|$ por $\max(|\omega|)$ resultará em um valor entre 0 e 1. Para facilitar a compreensão dos dados, é calculado a probabilidade de um dado ponto pertencer à série, e, conforme se deseja determinar se o ponto não pertence à série (anomalia), é calculada a probabilidade complementar de ser uma anomalia.

Capítulo V Resultados

O REMD foi avaliado em três conjuntos de dados e seus resultados foram comparados com métodos de estado da arte (SOTA) de detecção de anomalias, incluindo abordagens estatísticas e de aprendizado de máquina.

V.1 Configuração Experimental

Para a avaliação do método, foram selecionados conjuntos de dados com características diversas: *Yahoo Labs* [Salles et al., 2023a], o Numenta Anomaly Benchmark (NAB)[1], e o *Desafio GECCO 2018*. Esses conjuntos não apenas incorporam anomalias, mas também incluem propriedades importantes como volatilidade, tendência e sazonalidade.

O conjunto de dados denominado *Yahoo Labs* consiste em uma extensa coleção de séries temporais, incluindo tanto séries sintéticas quanto séries em tempo real relacionadas ao tráfego de dados nos serviços do Yahoo[2]. Estas séries temporais são meticulosamente elaboradas para exibir uma variedade de padrões, o que as torna extremamente diversificadas e representativas de situações reais de tráfego de dados. Este conjunto é particularmente valioso porque não apenas reproduz cenários variáveis e multifacetados, mas também oferece informações rotuladas que identificam a presença de anomalias nas séries temporais. Isso é especialmente útil para a avaliação precisa do desempenho de métodos de detecção de anomalias, permitindo comparar diferentes abordagens em um ambiente controlado, mas realístico. O valor deste conjunto de dados é amplamente reconhecido na literatura acadêmica, sendo frequentemente utilizado em pesquisas que envolvem aprendizado de máquina, especialmente em tarefas de detecção de anomalias. A Figura V.1 apresenta quatro(4) séries selecionadas aleatoriamente dentre as disponíveis no dataset *Yahoo Labs*.

O conjunto de dados Numenta Anomaly Benchmark (NAB) é outra fonte crucial de séries temporais rotuladas, projetado especificamente para a avaliação de métodos de detecção de anomalias. O NAB é considerado um padrão de referência na comunidade de detecção de anomalias devido à sua ampla seleção de séries temporais com rótulos apurados. As séries incluídas no NAB são variadas e abrangem tanto dados reais, como os coletados do monitoramento de serviços em nuvem da *Amazon Web Services* (AWS), quanto dados sintéticos, cuidadosamente gerados para representar diferentes

[1]Disponível no pacote OTSAD da linguagem R
[2]https://yahooresearch.tumblr.com/post/114590420346/

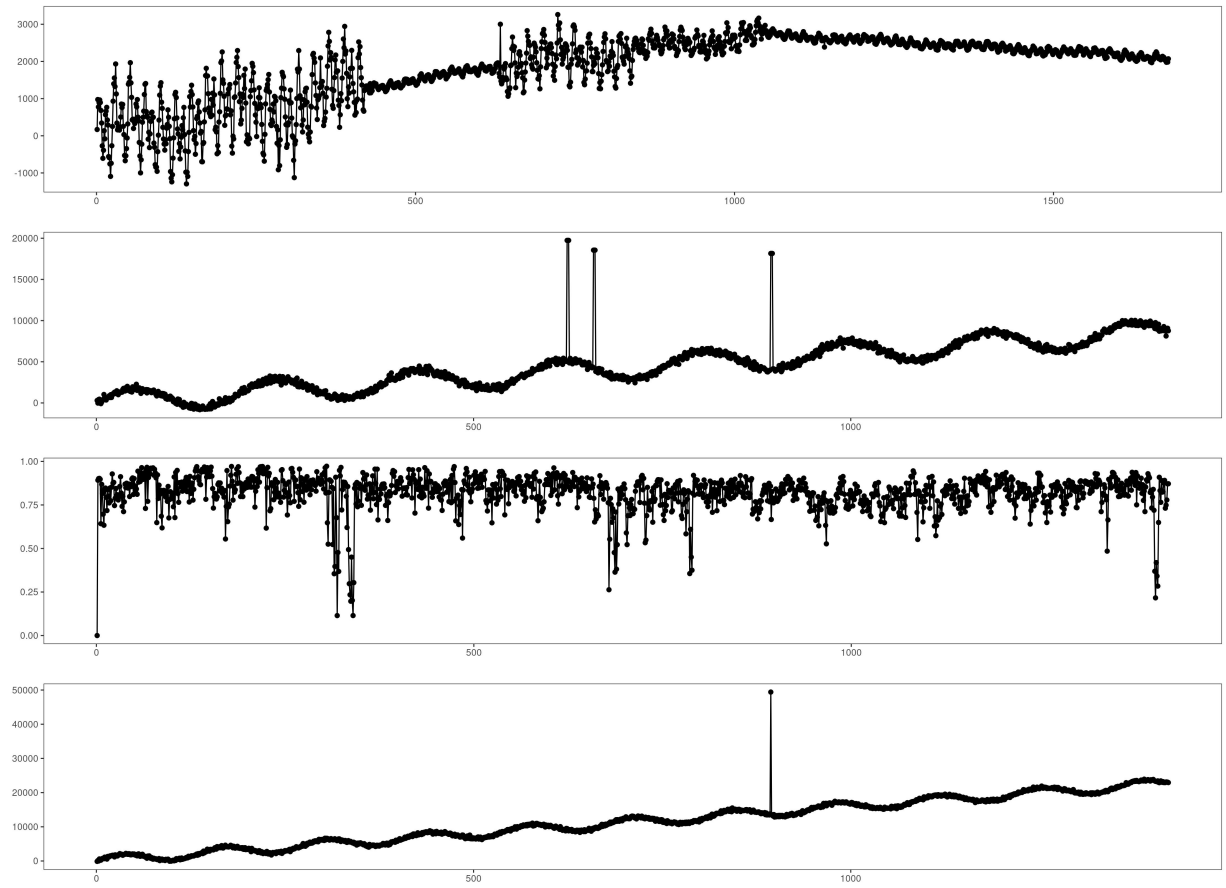


Figura V.1: Séries temporais do dataset Yahoo Labs. Esse dataset apresenta série com anomalias pontuais, predominantemente.

tipos de eventos anômalos. Esta diversidade permite que os pesquisadores testem a robustez e a flexibilidade de seus métodos em diferentes contextos e tipos de dados. A escolha do NAB para este estudo visa avaliar especificamente a eficácia dos algoritmos de detecção de anomalias no contexto do tráfego de dados em ambientes de computação em nuvem, onde a detecção precoce de anomalias pode ser crítica para a manutenção de serviços contínuos e confiáveis. Semelhante ao dado anterior, a Figura V.2 apresenta quatro séries escolhidas aleatoriamente dentre as disponíveis no dataset.

Por fim, o conjunto de dados denominado *Desafio GECCO 2018* [Moritz et al., 2018][3] consiste em uma coleção detalhada de nove variáveis que descrevem a composição da água potável, com observações coletadas a cada minuto. Este conjunto de dados foi desenvolvido para desafiar algoritmos de detecção de anomalias em um domínio complexo e vital, onde a qualidade da água é um fator crítico de saúde pública. A granularidade dos dados, coletados a cada minuto, proporciona uma resolução temporal fina ideal para identificar anomalias que possam afetar a qualidade da água. Um subconjunto deste conjunto de dados, contendo aproximadamente 1.500 observações (equivalente a um dia completo), foi selecionado para análise detalhada. Este subconjunto é particularmente

[3]http://www.spotseven.de/gecco-challenge/gecco-challenge-2018

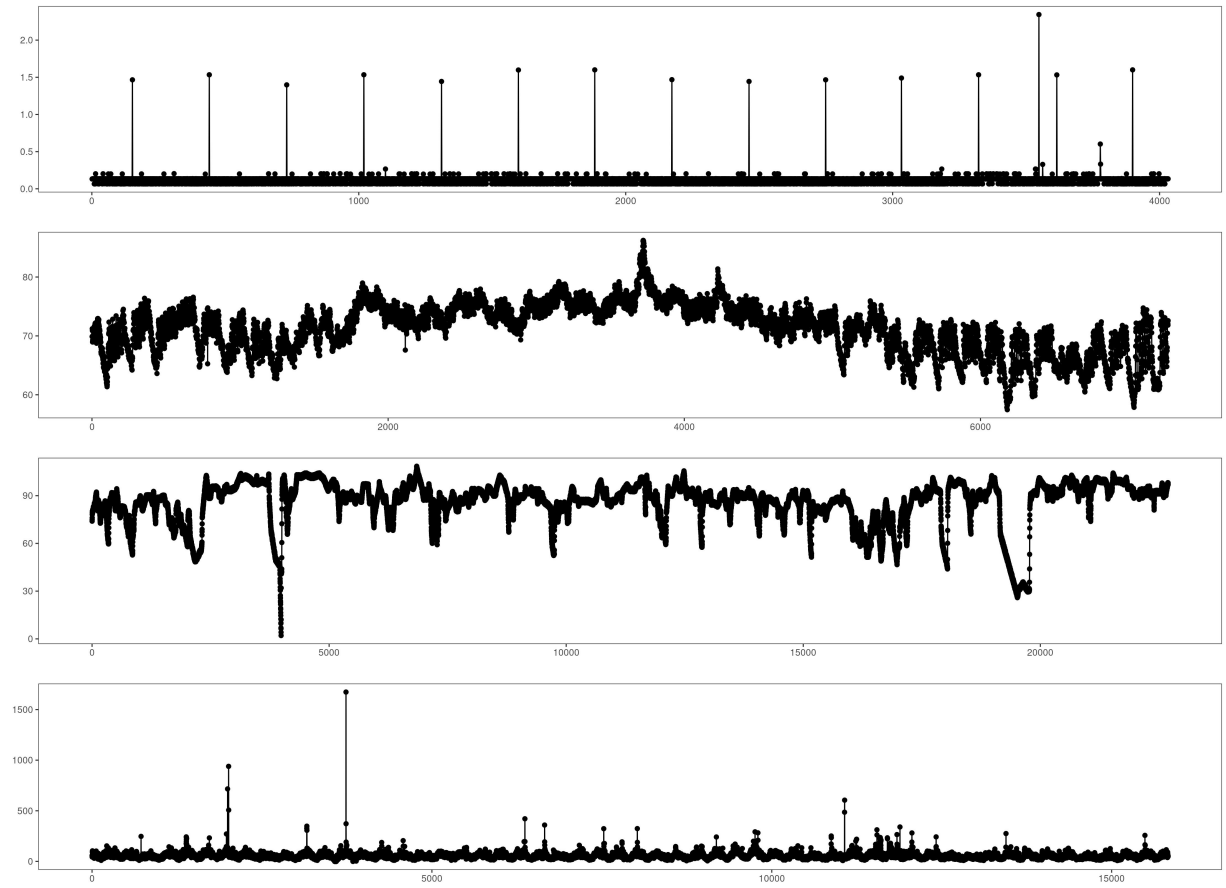


Figura V.2: Séries temporais do dataset NAB - Esse dataset apresenta anomalias pontuais e sequenciais.

importante porque inclui 72 anomalias rotuladas, fornecendo um ponto de referência claro para a avaliação dos métodos de detecção de anomalias. O uso deste conjunto de dados não apenas oferece uma aplicação prática significativa, mas também contribui para a compreensão de como detectar e reagir a anomalias em sistemas críticos de monitoramento ambiental. A Figura V.3 apresenta quatro imagens selecionadas de forma aleatória no dataset original. Entender como as amostras divergem entre si, ajuda a justificar e entender o desempenho do modelo quando submetido aos dados apresentados nas próximas seções.

O método REMD foi comparado com outros métodos estatísticos, como EMD, FBIAD [Lima et al., 2022] e ARIMA, além de métodos de aprendizado de máquina como LSTM, ELM, Conv1D e SVM. Esses métodos foram selecionados por sua relevância e por empregarem diferentes abordagens no processo de detecção. Todos os métodos estão disponíveis no pacote Harbinger [Salles et al., 2023a].

Os parâmetros utilizados para cada método estão descritos na Tabela V.1. Muitos parâmetros são específicos de cada método e, para esses casos, foram considerados os parâmetros padrão.

A análise foi realizada em um processador Intel Core i7-7820X, com 128 GB de RAM, 16 núcleos

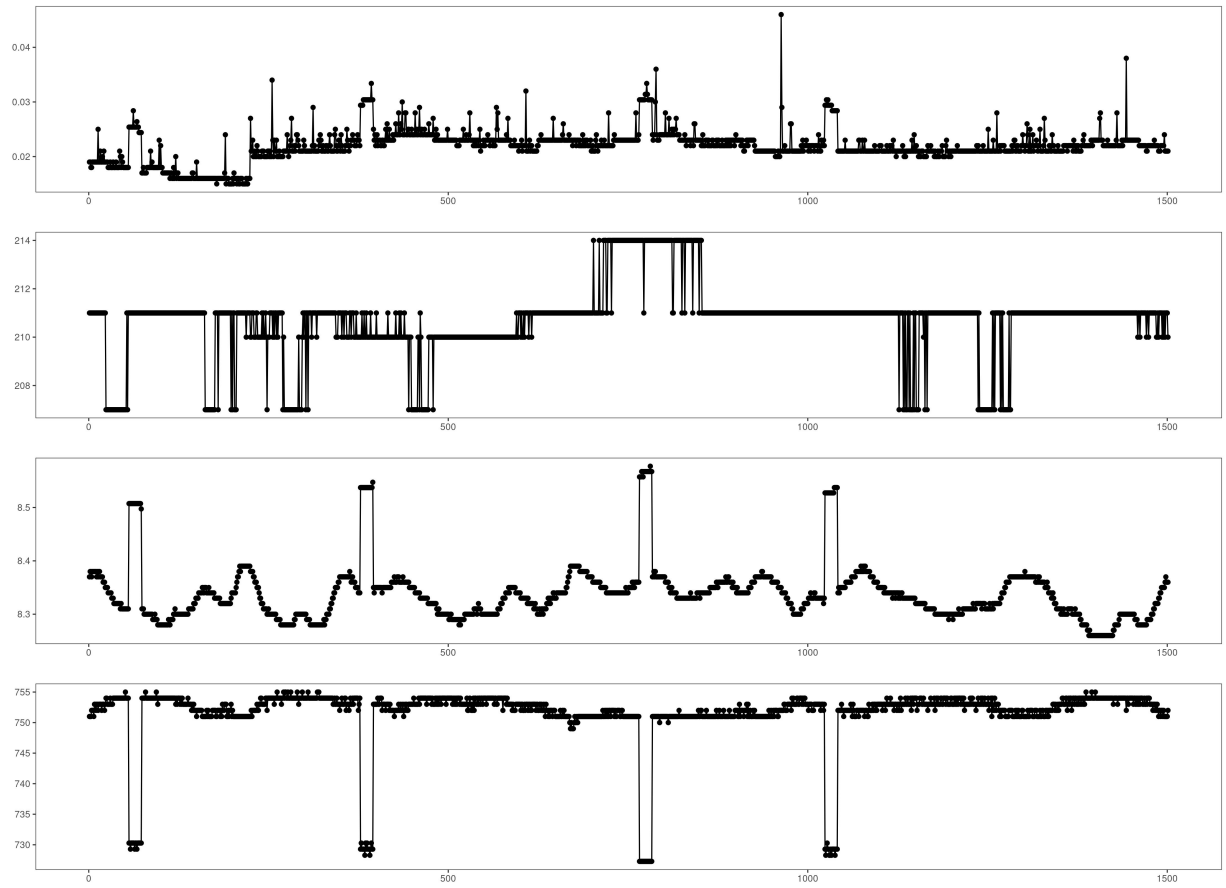


Figura V.3: Séries temporais do dataset *Gecco Challenge* - Esse dataset apresenta anomalias sequênciais, predominantemente.

e sistema operacional Ubuntu 22.04 LTS. O desempenho das técnicas de detecção de anomalias segue princípios semelhantes às métricas de classificação de aprendizado de máquina, utilizando a matriz de confusão para apresentar verdadeiros positivos (TP), verdadeiros negativos (TN), falsos positivos (FP) e falsos negativos (FN). Essa matriz fornece uma visão quantitativa do desempenho da detecção [Han et al., 2022].

V.2 Métricas de Desempenho

Após a construção da matriz de confusão, composta por verdadeiros positivos (TP), verdadeiros negativos (TN), falsos positivos (FP) e falsos negativos (FN), um conjunto de métricas pode ser derivado. Destacam-se cinco medidas que oferecem uma compreensão aprofundada do problema de detecção de anomalias: precisão, revocação, F1, acurácia e acurácia balanceada.

A precisão, mensurada pela proporção de detecções corretas sobre o total de detecções realizadas, é apresentada na Equação V.1. Esta métrica fornece uma indicação clara da exatidão das detecções efetuadas:

Tabela V.1: Parâmetros por método

Método	**Parâmetro**	**Valor**
EMD, REMD	Trial	5
EMD, REMD	Noise	0.1
EMD, REMD	Critério de Parada	0.00001
FBIAD	w (sliding windows size)	30
FBIAD	Tipo de Janela	Deslizante
FBIAD	Parâmetro de Sobreposição	50%
LSTM	Epochs	10000
LSTM	Learning Rate	0.001
LSTM	Tamanho do Batch	32
ELM	Actfun	Purelin
ELM	N de Neurônios na Camada Oculta	100
Conv1D	Epochs	10000
Conv1D	Learning Rate	0.001
Conv1D	Tamanho do Filtro	3
Conv1D	Strides	1
Conv1D	Padding	Valid
SVM	Kernel	Radial
SVM	C	1.0
SVM	Gamma	Scale

$$\text{Precisão} = \frac{TP}{TP + FP} \tag{V.1}$$

A revocação, também conhecida como sensibilidade, avalia a integridade da detecção, representando a proporção de anomalias corretamente identificadas em relação ao total de anomalias existentes, conforme descrito na Equação V.2:

$$\text{Revocação} = \frac{TP}{TP + FN} \tag{V.2}$$

A medida F1, ilustrada na Equação V.3, combina de forma ponderada a precisão e a revocação, fornecendo, assim, uma avaliação equilibrada do desempenho geral:

$$\text{F1} = 2 \times \frac{\text{Precisão} \times \text{Revocação}}{\text{Precisão} + \text{Revocação}} \tag{V.3}$$

A acurácia quantifica tradicionalmente a taxa geral de previsões corretas na amostra, conforme a Equação V.4. Entretanto, em eventos raros, ela não fornece poder de diferenciação significativo, já que todos os resultados tendem a apresentar valores próximos de 1. Isso a torna uma escolha menos adequada para este trabalho [Han et al., 2022]:

$$\text{Acurácia} = \frac{TP + TN}{TP + TN + FP + FN} \tag{V.4}$$

A especificidade mede a capacidade do método de identificar corretamente os casos negativos, sendo calculada pela proporção de verdadeiros negativos em relação ao total de casos realmente negativos, como demonstrado na Equação V.5:

$$\text{Especificidade} = \frac{TN}{TN + FP} \tag{V.5}$$

Essa métrica é crucial em contextos onde é importante minimizar a ocorrência de falsos positivos, garantindo maior confiabilidade na detecção de eventos negativos.

Para a presente análise, a acurácia balanceada, apresentada na Equação V.6, é adotada para ajustar seu enfoque ao lidar com classes desbalanceadas, assegurando uma avaliação mais justa e precisa. Essa medida é particularmente relevante na análise de fenômenos raros, como anomalias, ao garantir que as classes incomuns recebam a mesma consideração que as classes frequentes:

$$\text{Acurácia Balanceada} = \frac{\text{Sensibilidade} + \text{Especificidade}}{2} \tag{V.6}$$

Com base nos resultados de detecção dos algoritmos aplicados a diferentes conjuntos de dados, é possível comparar a eficácia na identificação de anomalias em relação aos dados rotulados. O desempenho do REMD é analisado de forma abrangente e comparado a outros métodos, destacando-se na subseção V.3 para o dataset Yahoo. Resultados agregados de desempenho são descritos na subseção V.4, enquanto a análise sob diferentes características de anomalia, para avaliar cenários de bom ou mau desempenho, está na subseção V.5. Adicionalmente, a subseção V.6 examina o desempenho considerando o tempo de processamento.

V.3 Análise Detalhada das Séries do Yahoo

Nesta subseção, são apresentados os resultados dos algoritmos aplicados ao conjunto de dados Yahoo de forma detalhada. O estudo de caso considerou a amostra A3Benchmark-TS39, uma série temporal que exibe características relevantes para análise, como sazonalidade, tendência e anomalias em momentos distintos da amostra, conforme ilustrado na Figura V.4. A série compreende 1.680 observações, das quais 12 são identificadas como anomalias (pontos em verde), já previamente identificadas. Além disso, trata-se de uma série não estacionária com presença de sazonalidade, aspectos essenciais para evidenciar o diferencial do uso do REMD em diversas aplicações. Os resultados referentes a esta amostra específica são abordados nesta seção e estão apresentados na Tabela V.2.

Para todo o conjunto de dados Yahoo, o REMD mostrou o melhor desempenho em F1. O desempenho satisfatório para o conjunto de dados geral é apresentado na Tabela V.3.

A Figura V.5 mostra a série temporal A3Benchmark-TS39 e seus resultados para os três melhores

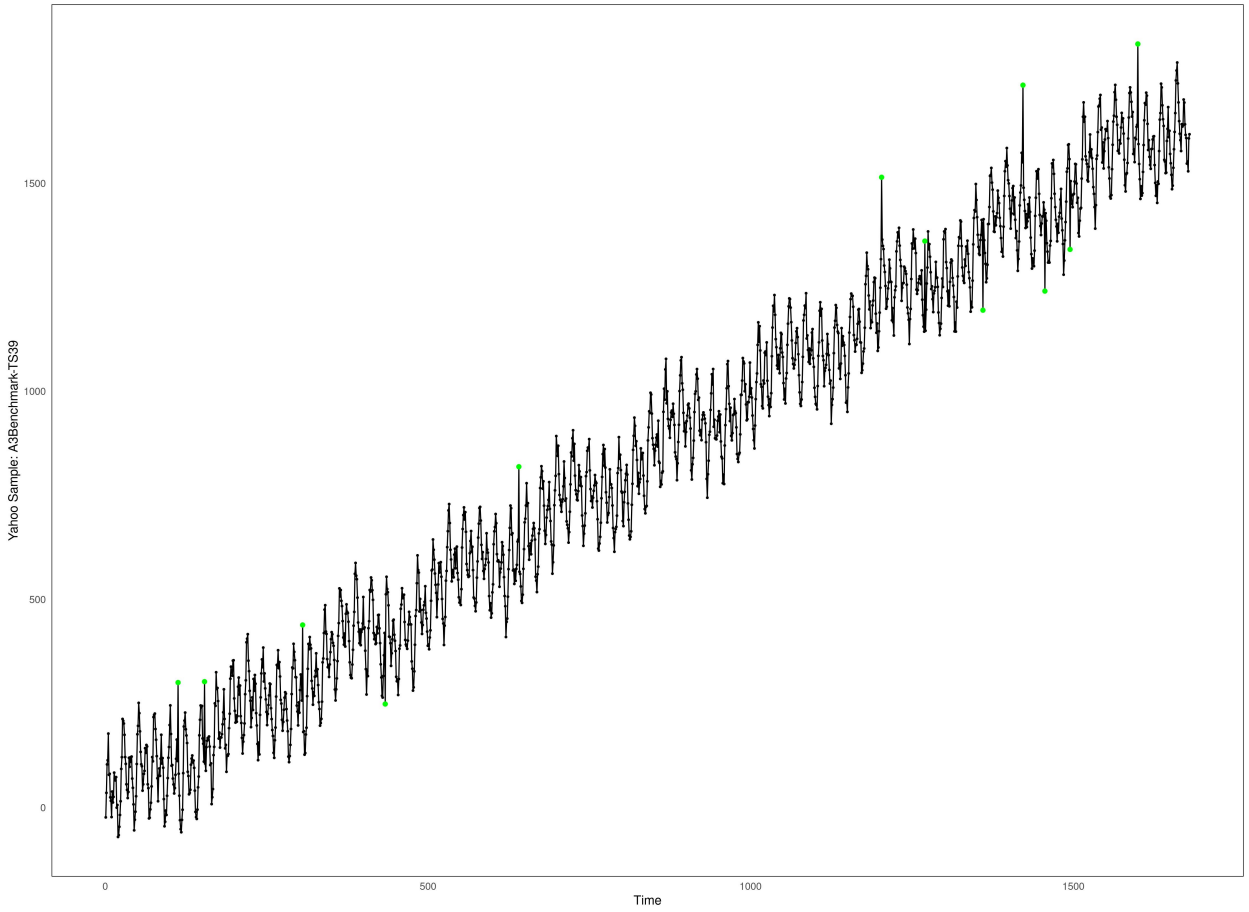


Figura V.4: Série temporal do Yahoo analisada. Formada por doze pontos de anomalia.

métodos: REMD, EMD e FBIAD, respectivamente. Na Figura, pontos azuis são falsos negativos, pontos verdes são verdadeiros positivos e pontos vermelhos são falsos positivos. A Tabela V.2 apresenta a matriz de confusão para esta série temporal em estudo aplicada a diferentes modelos de detecção de anomalia.

O método REMD mostra resultados superiores em comparação com seus melhores concorrentes. Isso pode ser observado analisando a Figura V.5 e a matriz de confusão (Tabela V.2). Das 12 anomalias para esta série temporal, o REMD identificou corretamente nove anomalias e não gerou nenhum falso positivo. O EMD apresentou o segundo melhor desempenho de acordo com F1, estando entre os três métodos que identificaram mais anomalias, com um total de oito anomalias e 13 falsos positivos. O método FBIAD ficou em terceiro lugar de acordo com F1, identificando 11 anomalias, porém, por outro lado, com 89 falsos positivos.

V.4 Análise Geral de Detecção

Para os três conjuntos de dados estudados, o desempenho médio segundo as quatro principais métricas (precisão, revocação, F1 e acurácia balanceada) são apresentados nas tabelas V.3, V.4 e V.5. A Tabela V.6 resume todos os resultados.

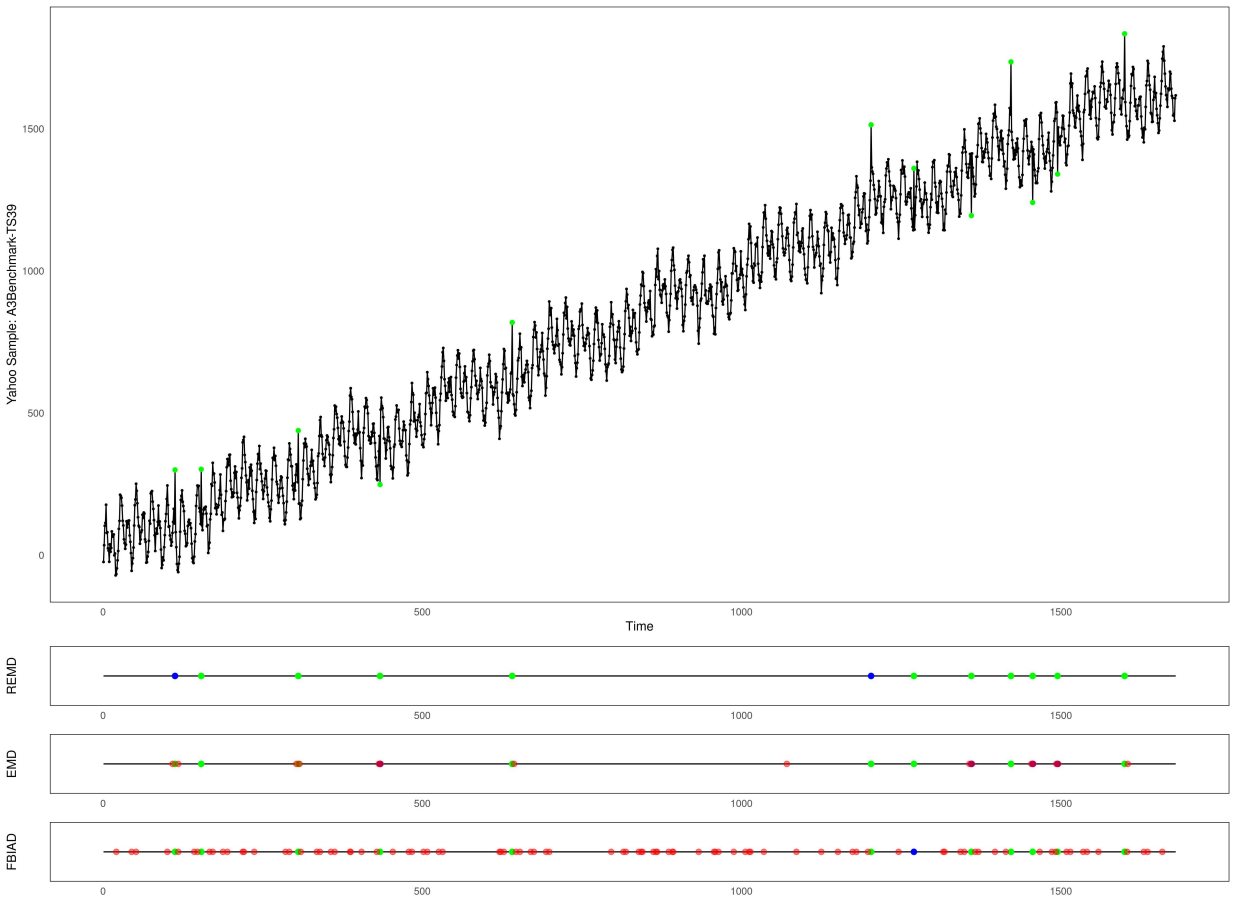


Figura V.5: Comparação das detecções com REMD, EMD e FBIAD na série temporal do Yahoo.

Tabela V.2: Matriz de Confusão para a Série do Yahoo

Método	TP	TN	FP	FN
REMD	9	1668	0	3
EMD	8	1655	13	4
FBIAD	11	1579	89	1
ARIMA	10	1560	108	2
LSTM	8	1634	34	4
ELM	12	1590	78	0
Conv1D	5	1597	71	7
SVM	12	1562	106	0

As tabelas contendo a comparação das métricas destacam, em negrito, os melhores resultados de cada métrica em suas respectivas colunas. Os valores estão sublinhados quando não há diferença significativa nos valores das métricas entre os melhores métodos, com um nível de significância estatística de 5%.

Para o conjunto de dados Yahoo, o método REMD mostra desempenho superior, em média, em duas das quatro métricas analisadas. Especialmente em F1, ele se mostra quase duas vezes melhor do que o segundo melhor método, EMD. A precisão também é notável, pois quando executada para várias amostras desse conjunto de dados, é maior do que o segundo lugar, em média, mais do

que três vezes, demonstrando como o método foi capaz de capturar pontos de anomalia nas séries temporais. Os resultados para todo o conjunto de dados do Yahoo esta disponível na Tabela V.2.

Tabela V.3: Resultados Resumidos para o Conjunto de Dados do Yahoo

Método	Precisão	Revocação	F1	Acurácia balanceada
REMD	**0.896**	0.539	**0.620**	0.769
EMD	0.263	0.639	0.347	0.813
FBIAD	0.115	0.669	0.158	0.807
ARIMA	0.049	**0.781**	0.087	**0.859**
LSTM	0.059	0.725	0.102	0.837
ELM	0.045	0.709	0.080	0.823
Conv1D	0.045	0.681	0.081	0.810
SVM	0.035	0.752	0.065	0.838

No segundo conjunto de dados de estudo (Tabela V.4), o benchmark Numenta, o método REMD continua mostrando desempenho superior em F1 e precisão. Novamente, seu resultado é significativamente melhor do que o do segundo lugar. O REMD não alcançou posição de destaque em revocação e acurácia balanceada.

Tabela V.4: Resultados Resumidos para o Conjunto de Dados da Numenta

Método	Precisão	Revocação	F1	Acurácia balanceada
REMD	**0.726**	0.405	**0.468**	0.702
EMD	0.002	0.503	0.012	0.729
FBIAD	0.008	**0.632**	0.015	**0.776**
ARIMA	0.016	0.583	0.030	0.760
LSTM	0.007	0.595	0.013	0.760
ELM	0.016	0.565	0.031	0.748
Conv1D	0.007	0.591	0.013	0.766
SVM	0.012	0.586	0.023	0.752

Os resultados para o terceiro conjunto de dados apresentam uma ordem de grandeza distinta (Tabela V.5). Apesar de o ARIMA mostrar o melhor desempenho, nenhum método alcançou um nível de significância estatística. Diante disso, as métricas de revocação, F1 e acurácia balanceada indicam um empate entre os métodos. Para desempate, é possível considerar a precisão como critério, onde o EMD destaca-se com o melhor resultado, seguido pelo REMD em segundo lugar.

A Tabela V.6 apresenta uma consolidação dos resultados obtidos em todos os conjuntos de dados avaliados. A partir dela, é possível confirmar o desempenho superior do REMD, seguido pelo EMD e FBIAD, utilizando a métrica F1 como critério de escolha. Esta ordem permanece inalterada mesmo ao adotar a métrica de precisão como critério de desempate, registrando uma diferença significativa em relação ao segundo colocado em ambos os casos. No que se refere à acurácia e revocação, o

Tabela V.5: Resultados Resumidos para o Conjunto de Dados do GECCO

Método	Precisão	Revocação	F1	Acurácia balanceada
REMD	0.220	0.041	0.063	0.511
EMD	**0.424**	0.069	0.079	0.519
FBIAD	0.078	0.075	0.070	0.507
ARIMA	0.078	**0.092**	**0.080**	0.512
LSTM	0.060	0.078	0.065	0.506
ELM	0.069	0.089	0.076	**0.512**
Conv1D	0.200	0.035	0.051	0.507
SVM	0.050	0.078	0.060	0.500

método baseado em ARIMA ocupa a posição de destaque. Vale ressaltar que o método REMD se configura como uma abordagem híbrida, incorporando elementos de EMD e ARIMA, métodos que merecem atenção especial na análise devido aos seus resultados positivos.

Tabela V.6: Análise Geral de Detecção

Método	Precisão	Revocação	F1	Acurácia balanceada
REMD	**0.684**	0.386	**0.448**	0.690
EMD	0.243	0.408	0.207	0.696
FBIAD	0.066	0.528	0.085	0.731
ARIMA	0.045	**0.556**	0.067	**0.746**
LSTM	0.041	0.534	0.062	0.735
ELM	0.041	0.517	0.063	0.726
Conv1D	0.036	0.519	0.055	0.724
SVM	0.030	0.542	0.049	0.732

Ressalta-se que, conforme as tabelas V.3, V.4, V.5, o REMD é geralmente superado por outros métodos na revocação. Esse resultado está relacionado ao foco na precisão devido à probabilidade de anomalias descritas na Equação IV.5.

V.5 Desempenho por tipo de anomalia

Na criação de uma série sintética, onde são inseridas características específicas a serem mensuradas, ou na escolha de quais séries dos conjuntos de dados serão testadas e analisadas, é possível descrever com maior precisão o desempenho do método. Em todas as séries, o método proposto, REMD, é aplicado, gerando pontos verdes para verdadeiros positivos (detecção correta), vermelhos para falsos positivos (detecção equivocada) e azuis para falsos negativos (não detecção). Os resultados resumidos são apresentados na Tabela V.7

V.5.1 Série com anomalia pontual

A série apresentada na Figura V.6 é uma série do conjunto Yahoo Labs contendo uma anomalia pontual isolada. Este dado foi escolhido aleatoriamente a partir da seleção mostrada na Figura V.1.

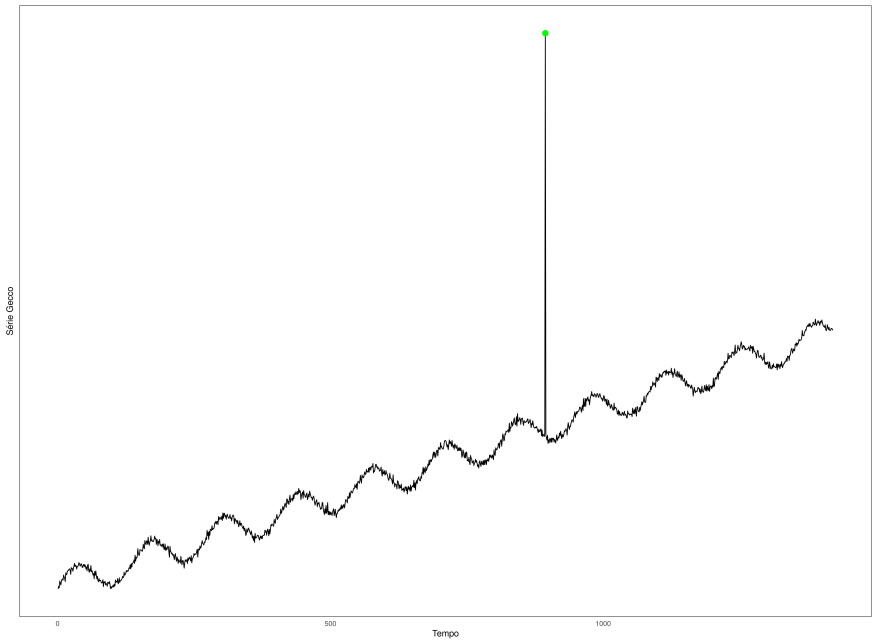


Figura V.6: Série do dataset Yahoo com anomalia pontual

O método REMD demonstrou um excelente desempenho na detecção dessa anomalia pontual e isolada.

V.5.2 Série com anomalias sequenciais

A série apresentada na Figura V.7 foi retirada do conjunto de dados Gecco. Nela, há uma sequência de anomalias identificadas.

Para todos os grupos de anomalias apresentados, o método REMD foi capaz de identificar corretamente a primeira anomalia da sequência. No entanto, não conseguiu identificar as anomalias subsequentes (falso negativo) e gerou um falso positivo quando a série retornou à normalidade. Este comportamento sugere que, embora seja eficaz na identificação do início de uma anomalia consecutiva, o método não é eficaz para a de detecção em cenários de anomalias contínuas.

V.5.3 Série com falsas anomalias

No dataset Numenta (ou NAB), conforme ilustrado na Figura V.8, algumas séries podem induzir uma pré-classificação equivocada de anomalias, que, na verdade, não são eventos anômalos reais. Identificar a verdadeira anomalia, quando presente, pode ser um desafio.

Neste conjunto de dados, há apenas duas anomalias reais. O método REMD conseguiu identificar corretamente uma dessas anomalias, enquanto a outra não foi detectada. Contudo, a vantagem é

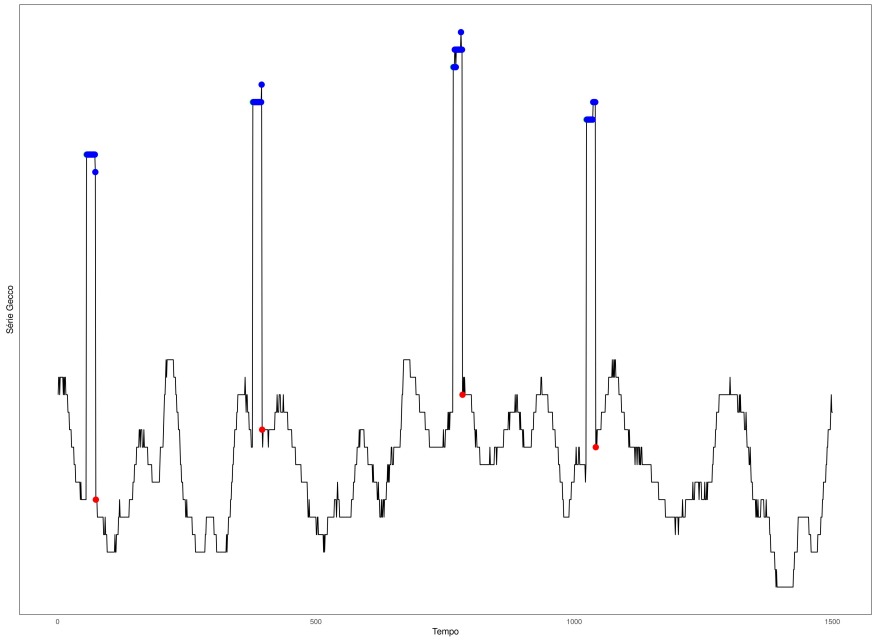


Figura V.7: Série do dataset Gecco com anomalias sequênciais

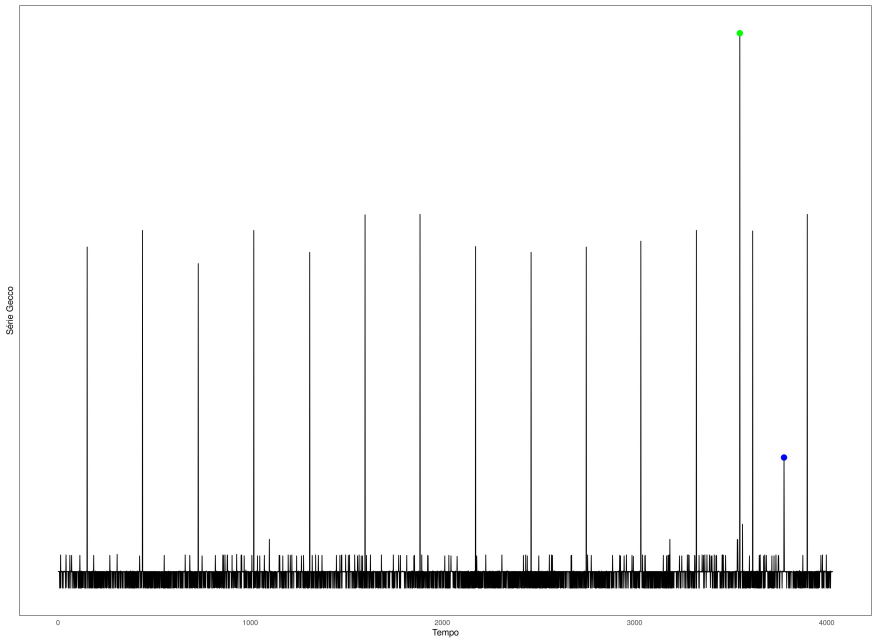


Figura V.8: Série do dataset Numenta com falsas anomalias

que não houve a geração de falsos positivos. Isso indica que o método é eficaz em evitar a acusação errônea de anomalias inexistentes, embora ainda haja espaço para melhorar a sensibilidade e garantir que todas as anomalias reais sejam detectadas.

V.5.4 Série com alteração na tendência

A série apresentada na Figura V.9 simula sete anomalias pontuais, com pontos escolhidos de forma aleatória, cada um gerando uma leve alteração na tendência do dado inicial.

Nas duas primeiras anomalias, o método já evidencia suas limitações na detecção de anomalias

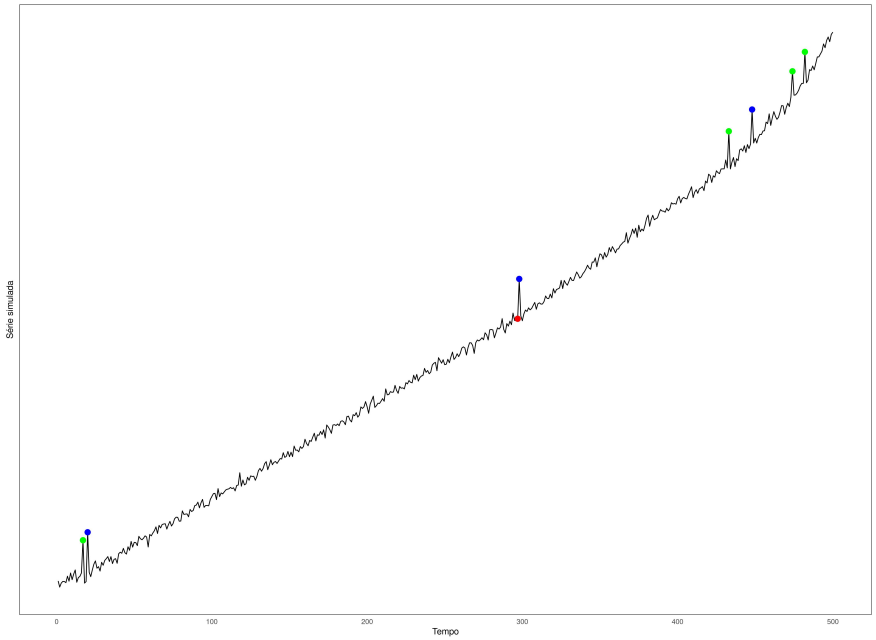


Figura V.9: Série simulada com alteração de tendência ao longo do tempo.

em sequência. Embora o primeiro ponto seja identificado corretamente, o segundo gera um falso negativo. Na terceira marcação, ocorre um falso positivo – o alerta foi gerado no local incorreto, mas bastante próximo do ponto real de anomalia, que não foi identificado devido à proximidade com ao anterior. Nos demais pontos de anomalia, o desempenho do método é bom.

Essa análise sugere que, embora o método REMD seja eficaz para detectar anomalias isoladas, ele pode necessitar de ajustes adicionais para melhorar a precisão na identificação de anomalias próximas e em sequência, caso se escolha esse objetivo.

V.5.5 Série com alteração na média

Na Figura V.10, existem oito anomalias pontuais que, cumulativamente, geram mudanças na média da série, variando entre modificações suaves e intensas.

As anomalias que resultaram em alterações sutis na média não foram identificadas, enquanto aquelas que produziram mudanças intensas foram corretamente detectadas. Um aspecto positivo a se destacar é que o método não gerou nenhum falso positivo durante a análise.

Isso indica que o método REMD se destaca na identificação de anomalias que provocam mudanças abruptas, porém, ainda apresenta limitações quando se trata de detectar alterações sutis.

V.5.6 Série com alteração na variança

Outra característica a ser validada é o desempenho do método quando há alterações na variância da série. Na Figura V.11, são apresentadas três anomalias, todas ocorrendo após alterações na série. Este conjunto de dados pertence ao dataset Yahoo.

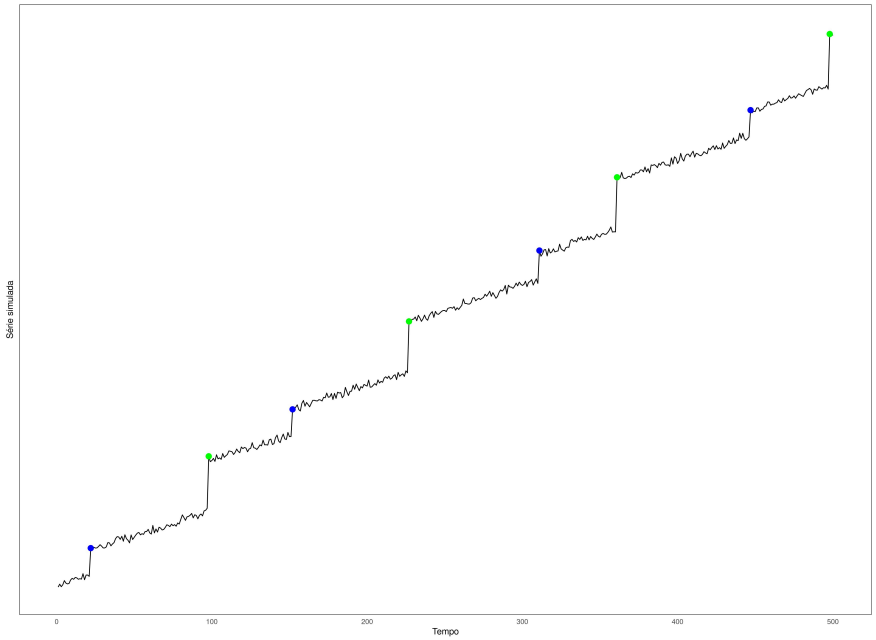


Figura V.10: Série simulada com alteração no valor médio ao longo do tempo.

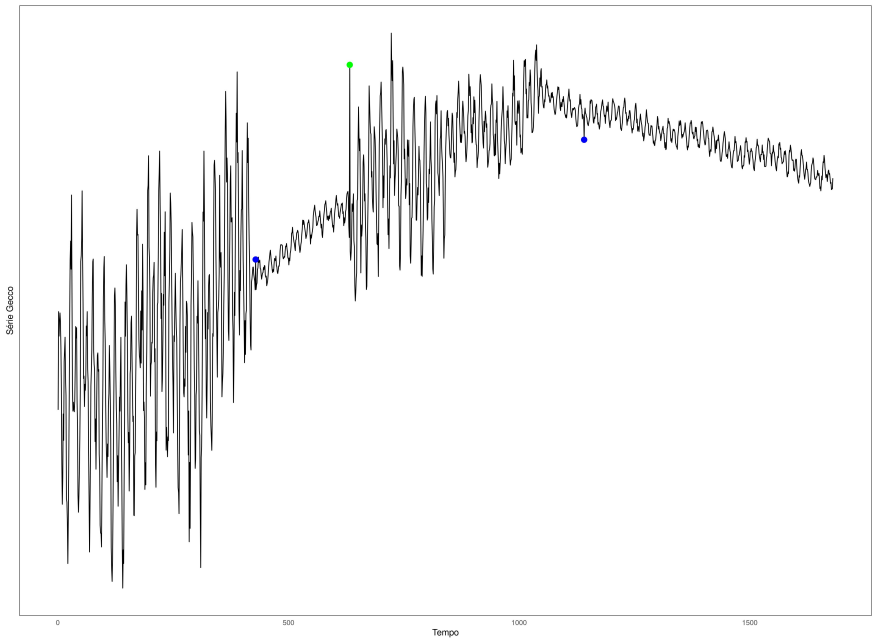


Figura V.11: Série do dataset Numenta com alteração de variança ao longo do tempo.

Das três anomalias, apenas uma foi identificada corretamente, enquanto as outras duas não foram detectadas. No entanto, um ponto positivo é que não houve a geração de falsos positivos, mesmo com a série aparentemente apresentando um alto nível de ruído. Esse resultado sugere que o método REMD é robusto na prevenção de falsos alertas em ambientes ruidosos.

V.5.7 Desempenho sumarizado

Na Tabela V.7, são resumidas as métricas de desempenho para cada um dos casos de estudo.

Após ser submetido a análises específicas, o método demonstrou ser eficiente em evitar alertas

Tabela V.7: Resultados para as séries simuladas

Série	Precisão	Revocação	F1	Acurácia balanceada
Pontual	1	1	1	1
Sequência	0.5	0.05	0.1	0.95
Falsa anomalia	1	0.5	0.66	0.75
Tendência	0.8	0.57	0.66	0.78
Média	1	0.5	0.66	0.78
Variância	1	0.3	0.5	0.66

desnecessários ou errados, ou seja, a maioria das marcações de anomalias são corretas, garantindo um alto valor de precisão. Essa característica foi consistente com as observações encontradas nos dataset do Yahoo, NAB e Gecco. No entanto, a principal fraqueza do método é a baixa revocação, pois, apesar de não gerar falsos positivos, há um número considerável de falsos negativos – falhando em identificar anomalias quando deveria.

Os valores de F1 e acurácia balanceada são intermediários. O F1 leva em consideração tanto a precisão quanto a revocação, enquanto a acurácia balanceada ajusta o desequilíbrio entre as classes.

Essa análise fortalece a hipótese de que sistemas que exigem alta precisão, ou seja, confiabilidade elevada nos alertas gerados, são os que mais se beneficiam da aplicação do REMD, devido à minimização do número de falsos positivos. Por exemplo, no monitoramento de dados de saúde para detectar epidemias, a prioridade dada à precisão evita a sobrecarga do sistema de saúde e o desperdício de recursos com exames e tratamentos desnecessários [Buckeridge et al., 2005]. Além dos custos econômicos, há também consequências psicossociais significativas: um diagnóstico falso de doença pode causar ansiedade e impactos negativos na vida pessoal e profissional dos indivíduos, reforçando a importância de maximizar a especificidade [Apple et al., 2021]. Outros exemplos existem no acompanhamento de infraestruturas críticas para prevenir interrupções desnecessárias de serviços essenciais [Amin and Wollenberg, 2005; Sun et al., 2022]; na detecção de fraudes financeiras, visando evitar bloqueios indevidos [Bolton and Hand, 2002; Ahmed et al., 2019]; e na segurança cibernética, onde há uma alta variância e ruído nos dados [Cresci et al., 2020].

Outra conclusão importante extraída desse diagnóstico é a diferença de desempenho do método em relação aos distintos conjuntos de dados utilizados. O *Yahoo Labs* caracteriza-se por anomalias pontuais, com variações estatísticas entre as séries, demonstrando um melhor desempenho geral. O Gecco Challenge *Gecco Challenge*, por sua vez, é dominado por anomalias sequenciais, resultando em um desempenho inferior. Já o *Numenta Anomaly Benchmark* apresenta uma combinação de ambos os tipos de anomalias, exibindo um comportamento intermediário, conforme mostrado na tabela V.6.

V.6 Análise geral de processamento

Uma etapa importante para novos métodos é o tempo de execução. A Tabela V.8 exibe a duração, em segundos, da execução de cada método para diferentes conjuntos de dados. Neste ponto, os métodos REMD e o EMD apresentam uma desvantagem. Embora não sejam os piores métodos em termos de tempo de execução, seu desempenho é inferior ao dos concorrentes que não se baseiam em aprendizado profundo, devido ao alto custo computacional no processo de decomposição das IMF's. Por outro lado, FBIAD e ARIMA apresentam consistentemente o melhor desempenho em termos de tempo de execução.

Tabela V.8: Tempo de execução (s) por Método e Dataset

Method	**Yahoo Labs**	**Numenta**	**Gecco Challenge**
REMD	3.230	766.4	5.513
EMD	2.987	698.3	5.146
FBIAD	**0.012**	**0.044**	0.016
ARIMA	0.812	0.205	0.033
LSTM	2.432	5.309	3.962
ELM	0.015	0.032	**0.013**
Conv1D	275.3	1083	818.4
SVM	0.266	1.293	0.644

Em resumo, o método REMD apresenta um desempenho excelente em termos de precisão e F1, especialmente nos conjuntos de dados do Yahoo e da Numenta. No entanto, ele tem um tempo de execução mais longo em comparação com outros métodos, destacando que há espaço para melhorias em termos de eficiência de tempo.

Capítulo VI Conclusão

Este trabalho apresenta um novo método de detecção de anomalias denominado Decomposição de Modo Empírico Refinada (REMD). O método combina o EMD com o modelo ARIMA. Ele é composto por quatro fases: decomposição por EMD, separação de frequência através da curva de rugosidade, ajuste do modelo ARIMA e detecção de *outliers*.

Realizamos uma avaliação experimental extensiva do REMD em três conjuntos de dados diversos: Yahoo Labs, Numenta Anomaly Benchmark e Desafio GECCO 2018. Comparamos o REMD com outros métodos representativos de detecção de anomalias, incluindo abordagens estatísticas, como EMD, FBIAD e ARIMA, e abordagens de aprendizado de máquina, como LSTM, ELM, Conv1D e SVM.

Os resultados experimentais demonstraram que o REMD consistentemente superou outros métodos, especialmente em relação à precisão e à pontuação F1. Ele mostrou uma capacidade superior de detectar anomalias em séries temporais de diferentes características, incluindo aquelas de padrões complexos e vários tipos de anomalias.

A separação por frequência é uma das grandes contribuições do trabalho. Ela levou à robustez do método, mesmo quando aplicado a conjuntos de dados muito diferentes entre si. Apesar de seu desempenho superior na detecção de anomalias, o REMD exibiu uma desvantagem em termos de tempo de execução quando comparado com outros métodos estatísticos, particularmente FBIAD e ARIMA. A eficiência computacional do REMD, seja em tempo, em custo ou técnicas de otimização, poderá ser um tema para trabalhos futuros.

As características de alta precisão do método REMD podem trazer benefícios significativos em aplicações onde a precisão é crítica. Áreas como monitoramento de infraestruturas críticas, detecção de fraudes financeiras, cibersegurança e controle de qualidade industrial podem se beneficiar ao minimizar falsos alarmes, evitando pertubações desnecessárias e alocando de forma eficiente os recursos. Embora o método apresente uma revocação moderada, seu uso em contextos que priorizam precisão pode resultar em operações confiáveis e eficazes, destacando sua utilidade prática em ambientes que exigem alta confiança na detecção de anomalias.

Por fim, esta pesquisa resultou na publicação de um artigo apresentado no congresso *International Joint Conference on Neural Networks* (IJCNN) 2024, intitulado *REMD: A Novel Hybrid Anomaly Detection Method Based on EMD and ARIMA* [Souza et al., 2024b], que trata do tema

deste estudo. E na publicação de um artigo à revista *BMC*, intitulado *Evaluation of Statistical Process Control Charts for Infant Mortality Monitoring in Brazilian Cities with Different Population Sizes*[Souza et al., 2024a]. Nesse artigo, é analisado o desempenho do método de controle estatístico de processos para monitoramento da mortalidade infantil em cidades brasileiras de diferentes tamanhos populacionais, sendo esse método o principal utilizado para análise e acompanhamento de dados temporais em epidemiologia.

Como trabalhos futuros, existem inúmeras oportunidades para expandir e evoluir o método REMD e suas aplicações. Em primeiro lugar, é essencial conduzir um diagnóstico detalhado para analisar o impacto marginal de cada fase do algoritmo REMD, o que pode oferecer informações valiosas sobre as contribuições específicas de cada etapa para o desempenho geral do método. Isso poderá ajudar a identificar áreas que poderiam ser otimizadas ou aprimoradas. Em segundo lugar, o desenvolvimento de uma versão do REMD que seja compatível com grandes conjuntos de dados é uma direção promissora. Implementações incrementais ou baseadas em janelas deslizantes podem ser exploradas para melhorar a eficiência computacional e viabilizar a aplicação do método em ambientes que lidam com fluxos contínuos de dados em larga escala. Por último, dada a precisão robusta do REMD, investigar sua aplicação em dados de saúde é particularmente interessante. Setores como o monitoramento de epidemias e doenças raras podem se beneficiar significativamente de um método que minimiza falsos positivos e maximiza a precisão, promovendo decisões mais assertivas e confiáveis em um contexto onde a precisão é extremamente crítica.

Referências

Ahmed, S., Mahmood, A., and Hu, M. Machine learning for credit card fraud detection: review, challenges and open research opportunities. *Expert Systems with Applications*, 167:114121, 2019.

Akoglu, L. and Tong, H. Outlier detection in complex event sequences. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 10–18. ACM, 2019.

Ambrosino, F., Thinová, L., Briestenský, M., Šebela, S., and Sabbarese, C. Detecting time series anomalies using hybrid methods applied to Radon signals recorded in caves for possible correlation with earthquakes. *Acta Geodaetica et Geophysica*, 55(3):405 – 420, 2020.

Amin, M. A. and Wollenberg, B. F. Toward a smart grid: Power delivery for the 21st century. *IEEE Power and Energy Magazine*, 3(5):34–41, 2005.

Aminikhanghahi, S. and Cook, D. J. A survey of methods for time series change point detection. *Knowledge and Information Systems*, 51(2):339 – 367, 2017.

Apple, F. S., Sandoval, Y., Jaffe, A. S., and Ordonez-Llanos, J. Clinical implications of false-positive cardiac troponin results. *Journal of the American College of Cardiology*, 78(4):358–360, 2021.

Atluri, G., Karpatne, A., and Kumar, V. Spatio-temporal data mining: A survey of problems and methods. *ACM Computing Surveys*, 51(4), 2018.

Bach, K., Gundersen, O., Knappskog, C., and Öztürk, P. Automatic case capturing for problematic drilling situations. In *Lecture Notes in Computer Science*, volume 8765, pages 48–62, 2014.

Baroni, L., Ogasawara, E., Castro, A., Mello, A., Paixão, E., Fraga, F., Borges, H., Lima, J., Souza, J., Tavares, L., Salles, R., Carvalho, D., Bezerra, E., Coutinho, R., Pacitti, E., and Porto, F. *harbinger: A Unified Time Series Event Detection Framework*, 2024.

Bates, D. M. and Watts, D. G. Relative Curvature Measures of Nonlinearity. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(1):1–16, 1980.

Bei An, Yanan Jiang, C. W. P. S. T. S. C. H. and Liu, K. Ground infrastructure monitoring in coastal areas using time-series insar technology: the case study of pudong international airport, shanghai. *International Journal of Digital Earth*, 16(1):355–374, 2023.

Bell, W. R., Holan, S. H., and McElroy, T. S. *Economic Time Series: Modeling and Seasonality.* CRC Press, 2018.

Bisgaard, S. and Kulahci, M. *Time Series Analysis and Forecasting by Example.* John Wiley & Sons, 2011.

Blázquez-García, A., Conde, A., Mori, U., and Lozano, J. A. A Review on Outlier/Anomaly Detection in Time Series Data. *ACM Computing Surveys*, 54(3), 2021.

Bolton, R. J. and Hand, D. J. Statistical fraud detection: A review. *Statistical Science*, 17(3):235–255, 2002.

Braei, M. and Wagner, S. Anomaly detection in univariate time-series: A survey on the state-of-the-art. *CoRR*, abs/2004.00433, 2020.

Breuniq, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. LOF: Identifying density-based local outliers. *SIGMOD Record (ACM Special Interest Group on Management of Data)*, 29(2):93 – 104, 2000.

Brillinger, D. R. *Time Series: Data Analysis and Theory*, 1981.

Buckeridge, D. L. et al. Surveillance for influenza and respiratory syncytial virus, ontario, canada, 2003-2005. *Emerging Infectious Diseases*, 11(7):914–922, 2005.

Campos, G. O. and Zimek, A. A framework for the evaluation of anomaly detection in streaming heterogeneous data. In *Proceedings of the 19th International Conference on Extending Database Technology*, pages 47–58, 2016.

Cao, L. Support vector machines experts for time series forecasting. *Neurocomputing*, 51:321 – 339, 2003.

Çelik, M., Dadaşer-Çelik, F., and Dokuz, A. Ş. Anomaly detection in temperature data using DBSCAN algorithm. In *INISTA 2011 - 2011 International Symposium on INnovations in Intelligent SysTems and Applications*, pages 91 – 95, 2011.

Chandola, V., Banerjee, A., and Kumar, V. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3), 2009.

Chandola, V., Banerjee, A., and Kumar, V. Anomaly detection for discrete sequences: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 24(5):823 – 839, 2012.

Chen, Z. D. and Malli, M. Early warning of complex equipment with multiple operating variables– a case study of ac aeration equipment in water treatment. *IEEE Transactions on Reliability*, 64(2):917–927, 2015.

Christy Jackson, J., Prassanna, J., Abdul Quadir, M., and Sivakumar, V. Withdrawn: Stock market analysis and prediction using time series analysis. *Materials Today: Proceedings*, 2021.

Cleveland, R. B., Cleveland, W. S., McRae, J. E., and Terpenning, I. Stl: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, 6(3):3–33, 1990.

Clifford, G. D. and McSharry, P. E. A robust classifier of ekg events using social network theory algorithms. *IEEE Transactions on biomedical engineering*, 58(12):3179–3186, 2011.

Cresci, S., Lillo, F., Sharma, K., Testa, A., and Tessone, C. Adversarial machine learning in network intrusion detection: Current status and challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(6):e1387, 2020.

Cui, W. and Wang, H. Anomaly detection and visualization of school electricity consumption data. In *2017 IEEE 2nd International Conference on Big Data Analysis, ICBDA 2017*, pages 606 – 611, 2017.

Du, M., Chen, Z., Liu, C., Oak, R., and Song, D. Lifelong anomaly detection through unlearning. In *Proceedings of the ACM Conference on Computer and Communications Security*, pages 1283 – 1297, 2019.

Fu, T.-C. A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24(1):164 – 181, 2011.

Gabarda, S. and Cristóbal, G. Detection of events in seismic time series by time-frequency methods. *IET Signal Processing*, 4(4):413 – 420, 2010.

Goldstein, M. and Uchida, S. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PLoS ONE*, 11(4), 2016.

Gupta, M., Gao, J., Aggarwal, C. C., and Han, J. Outlier Detection for Temporal Data: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, 26(9):2250 – 2267, 2014.

Hamilton, J. D. *Time Series Analysis*. Princeton University Press, 2020.

Han, J., Pei, J., and Tong, H. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, Cambridge, MA, 4th edition edition, 2022.

Huang, G.-B., Zhou, H., Ding, X., and Zhang, R. Extreme learning machine for regression and multiclass classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 42(2):513 – 529, 2012.

Huang, N., Shen, Z., Long, S., Wu, M., Shih, H., Zheng, Q., Yen, N.-C., Tung, C.-C., and Liu, H. The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 454:903–995, 1998a.

Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Snin, H. H., Zheng, Q., Yen, N.-C., Tung, C. C., and Liu, H. H. The empirical mode decomposition and the Hubert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 454(1971):903 – 995, 1998b.

James, S. and Woodall, W. H. Anomaly detection in smart grids: Handling temporal dependence. *Quality and Reliability Engineering International*, 36(5):1732–1746, 2020.

Jolliffe, I. T. *Principal Component Analysis*. Springer, New York, second edition edition, 2002.

Kriegel, H.-P., Kröger, P., Schubert, E., and Zimek, A. LoOP: Local outlier probabilities. In *International Conference on Information and Knowledge Management, Proceedings*, pages 1649 – 1652, 2009.

Lei, Y., Lin, J., He, Z., and Zuo, M. J. A review on empirical mode decomposition in fault diagnosis of rotating machinery. *Mechanical Systems and Signal Processing*, 35(1-2):108 – 126, 2013.

Lima, J., Salles, R., Porto, F., Coutinho, R., Alpis, P., Escobar, L., Pacitti, E., and Ogasawara, E. Forward and Backward Inertial Anomaly Detector: A Novel Time Series Event Detection Method. In *Proceedings of the International Joint Conference on Neural Networks*, volume 2022-July, pages 1–8, 2022.

Lin, S., Clark, R., Birke, R., Schonborn, S., Trigoni, N., and Roberts, S. Anomaly Detection for Time Series Using VAE-LSTM Hybrid Model. In *ICASSP*, volume 2020-May, pages 4322 – 4326, 2020.

Luo, X., Zhang, S., and Webb, D. Manufacturing surface topography prediction for precision machining. *International Journal of Advanced Manufacturing Technology*, 37:402–409, 2008.

Ma, J. and Perkins, S. Time-series Novelty Detection Using One-class Support Vector Machines. In *Proceedings of the International Joint Conference on Neural Networks*, volume 3, pages 1741 – 1745, 2003.

Mainsah, E., Greenwood, J. A., and Chetwynd, D. G. *Metrology and Properties of Engineering Surfaces*. Springer Science & Business Media, 2001.

Moayedi, H. Z. and Masnadi-Shirazi, M. Arima model for network traffic prediction and anomaly detection. In *Proceedings - International Symposium on Information Technology 2008, ITSim*, volume 3, 2008.

Moody, G. B. and Mark, R. G. Detection and classification of ekg events in real time using multilayer perceptrons. *Computers in Cardiology*, 1984:141–144, 1984.

Moritz, S., Rehbach, F., Chandrasekaran, S., Rebolledo, M., and Bartz-Beielstein, T. GECCO Industrial Challenge 2018 Dataset. Technical report, https://zenodo.org/record/3884398, 2018.

Nanduri, A. and Sherry, L. Anomaly detection in aircraft data using Recurrent Neural Networks (RNN). In *ICNS 2016: Securing an Integrated CNS System to Meet Future Challenges*, 2016.

Pena, E. H. M., De Assis, M. V. O., and Proença, M. L. Anomaly Detection Using Forecasting Methods ARIMA and HWDS. In *Proceedings - International Conference of the Chilean Computer Science Society, SCCC*, volume 0, pages 63 – 66, 2013.

Perelman, L., Arad, J., Housh, M., and Ostfeld, A. Event detection in water distribution systems from multivariate water quality time series. *Environmental Science and Technology*, 46(15):8212 – 8219, 2012.

Provotar, O. I., Linder, Y. M., and Veres, M. M. Unsupervised Anomaly Detection in Time Series Using LSTM-Based Autoencoders. In *2019 IEEE International Conference on Advanced Trends in Information Theory, ATIT 2019 - Proceedings*, pages 513 – 517, 2019.

Ray, S., McEvoy, D. S., Aaron, S., Hickman, T.-T., and Wright, A. Using statistical anomaly detection models to find clinical decision support malfunctions. *Journal of the American Medical Informatics Association*, 25(7):862 – 871, 2018.

Salles, R., Lima, J., Baroni, L., Castro, A., Souza, J., Carvalho, D., Coutinho, R., Bezerra, E., Pacitti, E., Porto, F., and Ogasawara, E. harbinger: An Unified Time Series Event Detection Framework, 2023a.

Salles, R., Pacitti, E., Bezerra, E., Marques, C., Pacheco, C., Oliveira, C., Porto, F., and Ogasawara, E. TSPredIT: Integrated Tuning of Data Preprocessing and Time Series Prediction Models. *Transactions on Large-Scale Data- and Knowledge-Centered Systems*, 14160 LNCS:41–55, 2023b.

Schmidl, S., Wenig, P., and Papenbrock, T. Anomaly Detection in Time Series: A Comprehensive Evaluation. *Proceedings of the VLDB Endowment*, 15(9):1779 – 1797, 2022.

Shumway, R. H. and Stoffer, D. S. *Time Series Analysis and Its Applications: With R Examples*. Springer, 2017.

Souza, J., Boccolini, C., Baroni, L., et al. Evaluation of statistical process control charts for infant mortality monitoring in brazilian cities with different population sizes. *BMC Research Notes*, 17:299, 2024a.

Souza, J., Paixão, E., Fraga, F., Baroni, L., Alves, R. F. S., Belloze, K., dos Santos, J., Bezerra, E., Porto, F., and Ogasawara, E. Remd: A novel hybrid anomaly detection method based on emd and arima. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2024b.

Stewart, J. *Calculus: Early Transcendentals*. Cengage Learning, 9th edition, 2020.

Stout, k. and Blunt, l. Three dimensional surface topography, 2000.

Sun, C., Guo, W., and Liu, G. Using machine learning for the resilience of smart grid systems. *Energy Reports*, 8:1256–1266, 2022.

Takeishi, N. and Yairi, T. Anomaly detection from multivariate time-series with sparse representation. In *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics*, volume 2014-January, pages 2651 – 2656, 2014.

Thiyagarajan, K., Kodagoda, S., and Van Nguyen, L. Predictive analytics for detecting sensor failure using autoregressive integrated moving average model. In *Proceedings of the 2017 12th IEEE Conference on Industrial Electronics and Applications, ICIEA 2017*, volume 2018-February, pages 1926 – 1931, 2017.

Thomas, T. Characterization of surface roughness. *Precision Engineering*, 3(2):97 – 104, 1981.

Tonatzin, M. and Rabatel, G. Real-time multivariate statistical control of contaminated surface waters. *Journal of Environmental Management*, 68(1):49–61, 2003.

Tsay, R. S. *Analysis of Financial Time Series*. Wiley, 2010.

Whitehouse, D. J. *Handbook of Surface and Nanometrology*. CRC Press, 2010.

Wu, Y., Lin, Y., Zhou, Z., Bolton, D. C., Liu, J., and Johnson, P. DeepDetect: A Cascaded Region-Based Densely Connected Network for Seismic Event Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 57(1):62 – 75, 2019.

Yingst, J. and Henden, A. Detection of seismic anomalies for hydrocarbon exploration. *Geophysics*, 71(2):L19–L26, 2006.

Zhang, M., Li, X., and Wang, L. An Adaptive Outlier Detection and Processing Approach towards Time Series Sensor Data. *IEEE Access*, 7:175192 – 175212, 2019.