CEFET/RJ
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO

AJUSTE FINO DE CRITÉRIOS DE DETECÇÃO PARA APRIMORAR A IDENTIFICAÇÃO DE ANOMALIAS EM SÉRIES TEMPORAIS

Edson Paulo da Silva Pinto Sobrinho

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
Kele Teixeira Belloze

Rio de Janeiro,

03-03-2026

Ajuste Fino de Critérios de Detecção para Aprimorar a Identificação de Anomalias em Séries Temporais

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

Edson Paulo da Silva Pinto Sobrinho

Aprovada por:

Presidente, Prof. Eduardo Soares Ogasawara, D.Sc. (Orientador)

Kele Teixeira Belloze, D.Sc. (Coorientadora)

Rafaelli de Carvalho Coutinho, D.Sc.

Fabio André Machado Porto, D.Sc.

Rio de Janeiro,

03-03-2026

FICHA CATALOGRÁFICA A SER SOLICITADA NA BIBLIOTECA DO CEFET/RJ APÓS A REVISÃO FINAL DO TEXTO.

MAIS INFORMAÇÕES EM:
https://eic.cefet-rj.br/ppcic/index.php/ficha-catalografica-nada-consta/

DEDICATÓRIA

Dedico esta Dissertação aos meus pais, Luís Paulo e Vera Regina, pelo apoio incondicional, pela confiança e por sempre valorizarem a educação.

AGRADECIMENTOS

Aos meus irmãos, sobrinhos e à minha família, pelo incentivo e suporte contínuos.

As pessoas especiais em minha vida, pelo apoio, compreensão e parceria ao longo desta jornada.

Aos meus orientadores, Prof. Eduardo Ogasawara e Profa. Kele Belloze, e aos meus colegas do DAL Lab, bem como aos demais membros e professores do PPCIC, pelas contribuições técnicas, acadêmicas e institucionais ao desenvolvimento desta pesquisa.

À Petrobras, pelo apoio institucional e financeiro, bem como aos meus colegas de trabalho e gerentes, pela compreensão e apoio ao longo desta jornada acadêmica.

À CAPES, ao CNPq e à FAPERJ, pelo apoio financeiro essencial para a realização desta pesquisa.

RESUMO

Ajuste Fino de Critérios de Detecção para Aprimorar a Identificação de Anomalias em Séries Temporais

Edson Paulo da Silva Pinto Sobrinho

Orientadores:
Eduardo Soares Ogasawara
Kele Teixeira Belloze

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.

A Detecção de Anomalias (DA) é o problema de identificar observações que não correspondem às típicas em uma série temporal. Os métodos de DA definem implicitamente critérios de detecção, tais como medidas de desvio, limiares de filtragem e estratégias de seleção de candidatos a anomalia. A escolha inadequada desses critérios resulta em saídas imprecisas, gerando alertas espúrios ou deixando de identificar eventos. O ajuste desses critérios é essencial para sistemas de monitoramento. Para enfrentar esse desafio, este estudo explora o ajuste fino de medidas de desvio, limiares de filtragem e estratégias de seleção de candidatos. Os resultados experimentais mostram que a escolha adequada dos critérios melhora significativamente o desempenho da DA, muitas vezes com impacto maior que a mudança dos próprios métodos de detecção.

Palavras-chave: Detecção de Anomalias em Séries Temporais; Aprendizado de Máquina; Medidas de Desvio; Limiares de Filtragem; Estratégias de Seleção de Candidatos; *Benchmarking*; Métricas de Avaliação.

Rio de Janeiro,

03-03-2026

ABSTRACT

Fine-Tuning Detection Criteria for Enhancing Anomaly Detection in Time Series

Edson Paulo da Silva Pinto Sobrinho

Advisors:
Eduardo Soares Ogasawara
Kele Teixeira Belloze

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.

Anomaly Detection (AD) is the problem of identifying observations that do not conform to typical ones in a time series. Detection methods implicitly define detection criteria, such as deviation measures, filter thresholds, and candidate anomaly selection strategies. Choosing inappropriate criteria results in inaccurate outputs, generating spurious alerts or missing events. Adjusting these criteria is essential for monitoring systems. To address this challenge, this study explores the fine-tuning of deviation measures, filter thresholds, and candidate selection strategies. Experimental results show that the proper choice of criteria significantly improves AD performance, often with greater impact than changing the detection methods.

Key-words: Anomaly Detection in Time Series; Machine Learning; Deviation Measures; Filter Thresholds; Candidate Selection Strategies; Benchmarking; Evaluation Metrics.

Rio de Janeiro,

03-03-2026

Sumário

I Introdução **1**

II Revisão da Literatura **5**

II.1 Contexto 5

II.1.1 Medidas de Desvio 6

II.1.2 Critérios de Filtro 7

II.1.3 Seleção de Candidato 7

II.2 Trabalhos Relacionados 7

II.2.1 Domínios e Contexto 8

II.2.2 Critérios de Detecção Aplicados nos Métodos de Identificação de Anomalias 9

II.2.3 Comparações 12

II.2.4 Resultados e Lacunas 13

II.3 Considerações 14

III Método de *Benchmark* **15**

III.1 Visão Geral do *Framework* de *Benchmark* 15

III.2 Medidas de Desvio 17

III.3 Critérios de Filtro 17

III.4 Seleção de Candidato 18

III.5 Implementação no *Harbinger* 19

III.5.1 Módulo Utilitário `harutils` 19

III.5.2 Configurações *default* e evolução do modelo de critérios 20

III.5.3 Especializações de métodos e decisões metodológicas 21

III.5.4 Customização dos critérios no uso 21

IV Avaliação Experimental **22**

IV.1 Configuração Experimental 22

IV.2 Análise Geral da Detecção 24

IV.3 Análise Detalhada do Conjunto de Dados: YAHOO 26

IV.4 Análise Detalhada do Conjunto de Dados: GECCO 30

IV.5 Análise Detalhada do Conjunto de Dados: NAB 32
IV.6 Análise Aprofundada da Série A3-TS43 34

V Conclusão 38

Referências 40

Lista de Figuras

I.1 Anomalias identificadas a partir da análise dos resíduos de um modelo *Autoregressive Integrated Moving Average* (ARIMA) em uma série temporal (adaptado de Ogasawara et al., 2025). 2

III.1 Medidas de Desvio, Limiares de Filtragem e Estratégias de Seleção de Candidatos 15

IV.1 Detecção de anomalias na série *A3Benchmark-TS43* (Yahoo). 35

Lista de Tabelas

IV.1 Resultado geral pelos critérios de detecção. 24
IV.2 Escala qualitativa de magnitude por critério e métrica de desempenho. 25
IV.3 Desempenho no Conjunto de Dados Yahoo 27
IV.4 Caracterização do núcleo competitivo por métrica no conjunto Yahoo. 28
IV.5 Desempenho no Conjunto de Dados Gecco. 31
IV.6 Desempenho no Conjunto de Dados: NAB 33

Lista de Acrônimos

Recall	*Recall*.
A1Benchmark	*A1 Benchmark (real series)*.
A2Benchmark	*A2 Benchmark (synthetic)*.
A3Benchmark	*A3 Benchmark (synthetic)*.
A4Benchmark	*A4 Benchmark (synthetic)*.
ABS	*Absolute Deviation*.
AD	Anomaly Detection.
ARIMA	*Autoregressive Integrated Moving Average*.
BP	*Boxplot Filter*.
Conv1D	*One-Dimensional Convolutional Neural Network*.
CPU	*Central Processing Unit*.
DA	Detecção de Anomalias.
ELM	*Extreme Learning Machine*.
EMD	*Empirical Mode Decomposition*.
F	*First Detected Anomaly*.
F1	*F1-score*.
FBIAD	*Forward and Backward Inertial Anomaly Detector*.
FN	*False Negative*.
FP	*False Positive*.
GD	*Gaussian Distribution*.
Gecco	*GECCO 2018 Challenge*.
H	*Highest Deviation Observed*.
IQR	*Interquartile Range*.

L1	*Manhattan Distance*.
L2	*Euclidean Norm*.
LSTM	*Long Short-Term Memory*.
MAE	*Mean Absolute Error*.
MSE	*Mean Squared Error*.
NAB	*Numenta Anomaly Benchmark*.
Precisão	*Precision*.
RBF	*Radial Basis Function*.
REMD	*Residual Empirical Mode Decomposition*.
RMSE	*Root Mean Squared Error*.
RT	*Ratio Threshold*.
SQD	*Squared Deviation*.
SVM	*Support Vector Machine*.
VN	*True Negative*.
VP	*True Positive*.
Yahoo	*Yahoo Labs Dataset*.

Capítulo I Introdução

A Detecção de Anomalias (DA) é o problema de identificar observações que não estão em conformidade com os padrões típicos de uma série temporal [Chandola et al., 2009]. Essas anomalias estão associadas a problemas em sistemas monitorados, como falhas em equipamentos, mudanças no comportamento dos usuários ou alterações nas condições ambientais [Ariyaluran Habeeb et al., 2019].

Em séries temporais, entendidas como sequências de observações ordenadas no tempo, diversos problemas analíticos podem ser investigados, entre eles a identificação de anomalias. Nesse contexto, a DA manifesta-se na identificação de padrões atípicos, nos quais desvios em relação ao comportamento esperado podem indicar eventos relevantes ou indesejados. Esses padrões anômalos emergem da análise da evolução temporal dos dados e constituem a base para sistemas de monitoramento e apoio à decisão em diferentes contextos [Ogasawara et al., 2025].

Nesse contexto, a relevância da DA estende-se a diversos domínios nos quais a identificação antecipada de eventos anômalos pode evitar prejuízos financeiros, melhorar a tomada de decisão e aumentar a confiabilidade de sistemas automatizados, incluindo monitoramento industrial, análise da qualidade da água, cibersegurança, análise financeira e exploração e perfuração de poços de petróleo. No contexto da qualidade da água, por exemplo, a DA em séries temporais de parâmetros físico-químicos é empregada para identificar episódios de contaminação que representam riscos à saúde pública [Sela et al., 2012].

Apesar de sua ampla aplicabilidade, um desafio central da DA consiste no projeto de métodos capazes de se adaptar a diferentes características dos dados, com redução da ocorrência de alertas espúrios [Lima et al., 2024; Ariyaluran Habeeb et al., 2019; Ogasawara et al., 2021]. Nesse cenário, *False Positive* (FP) estão associados a intervenções desnecessárias, enquanto *False Negative* (FN), correspondentes a anomalias não detectadas, estão associados a perdas ou falhas significativas [Ibidunmoye et al., 2015]. O balanceamento entre FP e FN depende, portanto, de decisões metodológicas adequadas e de uma compreensão aprofundada das propriedades dos dados analisados.

Para a avaliação desse equilíbrio, métricas de detecção, como o *F1-score* (F1), são frequentemente empregadas, pois permitem a ponderação simultânea de FP e FN. Ainda assim, um método pode apresentar valores aparentemente balanceados sem alcançar níveis satisfatórios de Precisão e *Recall* em domínios específicos [Dixit et al., 2022; Cook et al., 2020; Ahmed et al., 2016].

De forma geral, os métodos de DA em séries temporais podem ser compreendidos como processos estruturados em etapas recorrentes, independentemente da abordagem específica adotada. Em linhas gerais, essas abordagens envolvem a modelagem ou a caracterização do comportamento esperado da série, o cálculo de desvios em relação a esse comportamento, a aplicação de critérios para filtrar desvios relevantes e, quando necessário, a seleção de observações representativas entre múltiplas detecções próximas [Chandola et al., 2009; Aggarwal, 2016]. Embora métodos estatísticos, baseados em distância, em aprendizado de máquina ou híbridos implementem essas etapas de maneiras distintas, a definição desses critérios, de forma explícita ou implícita, constitui uma necessidade comum à maioria das estratégias de detecção de anomalias.

Para ilustrar esse processo e o papel dos critérios de decisão na DA, apresenta-se na Figura I.1 um exemplo baseado na análise de resíduos de uma série temporal.

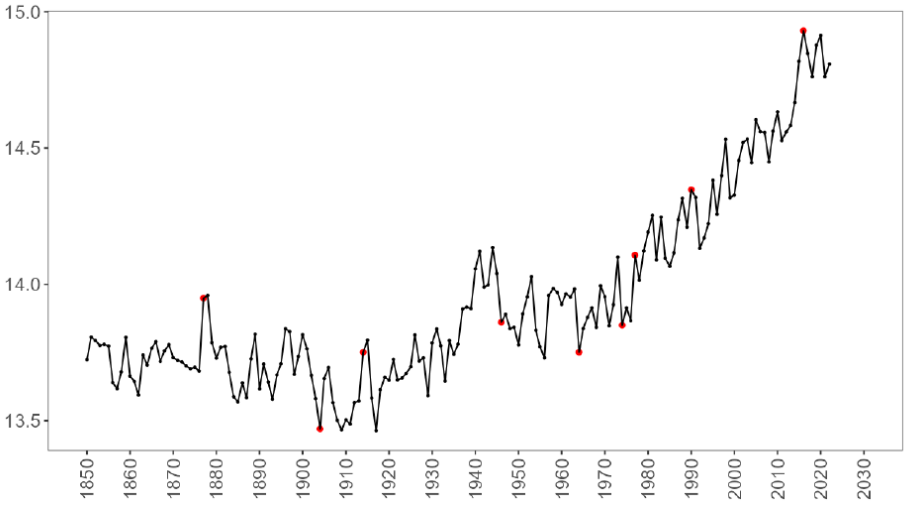


Figura I.1: Anomalias identificadas a partir da análise dos resíduos de um modelo *Autoregressive Integrated Moving Average* (ARIMA) em uma série temporal (adaptado de Ogasawara et al., 2025).

A Figura I.1 exemplifica o processo geral de detecção de anomalias em séries temporais, envolvendo a modelagem do comportamento esperado e a aplicação de um critério de decisão sobre os desvios observados.

Grande parte das variações observadas no desempenho da DA decorre de decisões implícitas nos métodos de detecção. Entre elas, destacam-se a escolha das Medidas de Desvio, a definição dos Critérios de Filtragem e a estratégia adotada para a seleção de Anomalias Candidatas quando múltiplas observações próximas são sinalizadas. Embora nem sempre discutidos explicitamente, esses elementos desempenham papel determinante sobre o desempenho da DA [Ogasawara et al., 2025].

As medidas de desvio definem como a diferença entre os valores observados e os esperados é quantificada, o que influencia diretamente o comportamento da detecção. Na literatura, os desvios quadráticos apresentam maior adequação a dados com distribuição aproximadamente normal,

enquanto os desvios absolutos, como o erro absoluto médio, apresentam maior robustez quando a distribuição não segue a normalidade [Gorard, 2005; Hodson, 2022].

De forma complementar, os critérios de filtragem estabelecem o ponto a partir do qual um desvio é considerado suficiente para a caracterização de uma anomalia. Esses critérios costumam basear-se em pressupostos sobre a distribuição dos dados, como limites derivados de *boxplots* ou de distribuições gaussianas [Han et al., 2022; Aggarwal, 2016]. Apesar de aparecerem de forma mais explícita nos métodos de detecção, sua sensibilidade e seu impacto no desempenho ainda recebem atenção limitada.

Após o cálculo das medidas de desvio e a aplicação dos critérios de filtragem, podem surgir grupos de anomalias candidatas. Nesses casos, torna-se necessário definir quais observações devem ser confirmadas como anomalias. As estratégias mais recorrentes incluem a seleção da primeira anomalia detectada em um grupo ou daquela associada ao maior desvio [Lima et al., 2022]. Essa escolha afeta diretamente o desempenho da DA, embora seja, em geral, pré-definida e pouco adaptativa.

Em conjunto, Medidas de Desvio, Critérios de Filtragem e estratégias de Seleção de Anomalias Candidatas são frequentemente aplicados de forma fixa, sem consideração explícita das características específicas dos dados ou dos objetivos da aplicação [Lima et al., 2022; Souza et al., 2024; Talagala et al., 2020]. Essa rigidez contribui para a degradação do desempenho quando as propriedades dos dados, em particular as características estatísticas e dinâmicas observadas ao longo das séries temporais, diferem daquelas consideradas no desenvolvimento original dos métodos.

Diante desse cenário, três perguntas de pesquisa orientam este estudo: (Q1) a alteração da Medida de Desvio impacta a DA? (Q2) as detecções são sensíveis às mudanças nos Critérios de Filtragem? (Q3) as detecções são sensíveis à escolha da Anomalia Candidata quando múltiplas observações próximas são detectadas?

Com base nessas questões, este trabalho propõe um *framework* de *benchmark* para o ajuste e a análise sistemática dos critérios de DA no processo de identificação de anomalias em séries temporais, e as contribuições científicas e metodológicas desta dissertação são apresentadas a seguir.

A contribuição central consiste na organização explícita dos critérios envolvidos no processo de identificação de anomalias, incluindo Medidas de Desvio, Critérios de Filtragem e estratégias de Seleção de Anomalias Candidatas, como componentes do processo decisório. Esse *framework* de *benchmark* fornece uma base conceitual para a análise desses critérios no desempenho dos métodos de DA, complementando abordagens experimentais que se concentram predominantemente na comparação entre métodos de detecção.

Além disso, é apresentada uma análise sistemática de diferentes configurações desses critérios no desempenho da DA. Para isso, é conduzida uma avaliação experimental na qual os critérios são ajustados de forma controlada e combinada, permitindo comparar seu impacto relativo em distintos

métodos e conjuntos de dados.

Por fim, o trabalho indica que o ajuste desses critérios de DA, de acordo com os objetivos da aplicação, pode resultar em ganhos de desempenho superiores aos obtidos pela simples substituição do método de detecção, abordando uma lacuna recorrente na literatura.

Além desta introdução, o trabalho está organizado em quatro capítulos adicionais. O Capítulo II apresenta os fundamentos da detecção de anomalias e a revisão da literatura, incluindo a análise dos trabalhos relacionados. O Capítulo III descreve o método de *benchmark* adotado, detalhando a modelagem e a adaptação dos critérios de detecção. O Capítulo IV apresenta a avaliação experimental, incluindo a descrição dos conjuntos de dados, a configuração dos experimentos e a análise dos resultados. Por fim, o Capítulo V apresenta as conclusões do estudo.

Capítulo II Revisão da Literatura

Este capítulo revisa a literatura sobre DA em séries temporais, dando continuidade à discussão introduzida no capítulo anterior, na qual a identificação de padrões atípicos foi caracterizada como um elemento central de sistemas de monitoramento e apoio à decisão. O foco da revisão recai sobre os critérios de DA que aprimoram o desempenho dos métodos, em particular medidas de desvio, critérios de filtro e estratégias de seleção de candidatos, frequentemente tratados de forma implícita na literatura. Para orientar a leitura, a Seção II.1 apresenta o contexto e os fundamentos da DA em séries temporais. Em seguida, a Seção II.2 organiza e discute os trabalhos relacionados, destacando como esses critérios são aplicados aos principais métodos e aos *benchmarks*. Por fim, a Seção II.3 sintetiza os principais achados e as lacunas identificadas.

II.1 Contexto

A DA em séries temporais constitui um problema central na análise de dados temporais, cujo objetivo é identificar observações ou intervalos que desviam da distribuição esperada. As anomalias assumem um significado específico de acordo com o domínio da aplicação [Talagala et al., 2020; Ren et al., 2019] e estão associadas a desvios em tendências, volatilidade ou outros aspectos temporais [Ariyaluran Habeeb et al., 2019; Ogasawara et al., 2021].

Os métodos de DA abrangem diversas abordagens, como análise de distribuição, regressão e classificação [Chandola et al., 2009; Truong et al., 2020]. Um problema recorrente nesses métodos é o uso de filtros que estabelecem limites a partir dos quais uma observação é considerada anômala. Em geral, a detecção baseia-se na diferença entre os valores observados e os previstos ou ajustados por um modelo [Lima et al., 2024].

Nos métodos baseados em predição, o próximo valor da série é estimado e comparado com a observação real, de modo que desvios mais elevados indicam anomalias. Por outro lado, nas abordagens ajustadas por modelo de DA, o modelo é calibrado diretamente à série, e as anomalias são identificadas com base no desvio entre os valores observados e os ajustados [Talagala et al., 2020; Zhang et al., 2020; Ren et al., 2019].

Os métodos de identificação de anomalias em séries temporais podem ser compreendidos como um *pipeline* composto por etapas recorrentes, independentemente da abordagem adotada. Essas

etapas incluem a modelagem ou caracterização do comportamento esperado da série, o cálculo de desvios em relação a esse comportamento, a aplicação de critérios para identificar observações atípicas e, quando necessário, a seleção de observações representativas entre múltiplas detecções próximas [Chandola et al., 2009; Aggarwal, 2016; Talagala et al., 2020].

Os termos sensibilidade e seletividade são utilizados para caracterizar o comportamento dos critérios de detecção em relação às métricas de avaliação. A sensibilidade refere-se à capacidade do método de identificar corretamente as anomalias presentes na série temporal, estando associada a maiores valores de *Recall*. Por sua vez, seletividade corresponde à capacidade de restringir detecções indevidas, estando associada a maiores valores de Precisão. O compromisso entre sensibilidade e seletividade é comumente analisado por meio do F1, que combina precisão e revocação em uma única medida de desempenho [Chandola et al., 2009; Ibidunmoye et al., 2015; Cook et al., 2020].

Nesse contexto, o desempenho da detecção é sensível à Medida de Desvio, ao Critério de Filtragem e à estratégia de Seleção de Anomalias Candidatas adotados. Essas escolhas influenciam diretamente o equilíbrio entre sensibilidade e seletividade do método. Como o método de cálculo do desvio impacta diretamente a eficácia da detecção, a Seção II.1.1 aborda as medidas de desvio. Dada uma medida de desvio, o critério de filtragem que separa as anomalias das observações típicas é discutido na Seção II.1.2. Quando múltiplas anomalias candidatas próximas são detectadas, a estratégia de seleção dos candidatos é apresentada na Seção II.1.3.

II.1.1 Medidas de Desvio

As medidas de desvio quantificam a diferença entre as observações e os valores previstos ou ajustados por métodos de DA [Han et al., 2022; Gorard, 2005]. De forma geral, essas medidas são classificadas em dois tipos: (i) *Absolute Deviation* (ABS), e (ii) *Squared Deviation* (SQD).

O ABS é calculado como a diferença absoluta entre uma observação x_i e o valor esperado $\hat{x_i}$: $w_i = |x_i - \hat{x_i}|$. Por sua vez, o SQD amplifica as discrepâncias maiores ao elevar a diferença ao quadrado: $w_i = (x_i - \hat{x_i})^2$.

A escolha da medida de desvio influencia os resultados da DA. Estatisticamente, o SQD está alinhado à suposição de erros normalmente distribuídos, o que o torna sensível a valores extremos. Por outro lado, ABS apresenta maior robustez diante de distribuições assimétricas ou contendo *outliers*, cenário comum em séries temporais com ruído ou anomalias [Gorard, 2005; Hodson, 2022].

As medidas de desvio são utilizadas como componentes centrais dos métodos de DA, seja de forma explícita em técnicas baseadas em desvio [Aggarwal, 2016; Han et al., 2022], seja de forma implícita em abordagens baseadas em classificação [Cauteruccio et al., 2021], agrupamento [Li et al., 2021] e estatística [Salles et al., 2020; Lima et al., 2022].

II.1.2 Critérios de Filtro

O critério de filtro define os limiares que distinguem as observações típicas das anomalias. Ele atua como uma fronteira de decisão para sinalizar anomalias com base nos resíduos entre os valores observados e os esperados [Ahmad et al., 2017; Hasani, 2017].

Uma técnica amplamente utilizada é o critério baseado em *Boxplot Filter* (BP), que emprega o *Interquartile Range* (IQR) para estabelecer o limiar: $T_{\mathrm{BP}}(\omega) = Q_3(\omega) + 1.5 \times (Q_3(\omega) - Q_1(\omega))$, em que $Q_1(\omega)$ e $Q_3(\omega)$ são, respectivamente, o primeiro e o terceiro quartis dos resíduos ω [Han et al., 2022; Lima et al., 2022].

Quando a distribuição dos dados é aproximadamente normal, os limiares podem ser definidos com base em pressupostos gaussianos por meio da *Gaussian Distribution* (GD). Observações que ultrapassam três desvios-padrão da média são consideradas anômalas: $T_{GD}(Z_\omega) = 3, \quad \text{com} \quad Z_{\omega_i} = \frac{\omega_i - \mu(Z_\omega)}{\sigma(Z_\omega)}$, em que μ e σ representam, respectivamente, a média e o desvio-padrão dos resíduos.

O *Ratio Threshold* (RT) normaliza os resíduos pelo seu valor máximo e calcula um desvio relativo: $\omega' = 1 - \frac{\omega}{\max(\omega)}$. Uma anomalia é sinalizada quando $T_{\mathrm{RT}}(\omega) = \overline{\omega'} + \sigma_{\omega'}$, em que $\overline{\omega'}$ e $\sigma_{\omega'}$ correspondem, respectivamente, à média e ao desvio-padrão de ω' [Souza et al., 2024; Scharf and Demeure, 1991].

II.1.3 Seleção de Candidato

Quando múltiplas anomalias candidatas são detectadas em sequência temporal, a escolha da anomalia apropriada contribui para evitar alarmes redundantes [Lima et al., 2024; Salles et al., 2024]. Duas estratégias comuns de seleção de candidatos são: (i) *First Detected Anomaly* (F) e (ii) *Highest Deviation Observed* (H).

A estratégia F seleciona a anomalia detectada mais cedo, enquanto a estratégia H seleciona a observação associada ao maior desvio em relação aos valores esperados [Lima et al., 2022, 2024]. A escolha entre essas estratégias impacta métricas de desempenho como Precisão, *Recall* e F1. Em geral, F tende a oferecer melhor equilíbrio entre Precisão e *Recall*, enquanto H está associada a maior Precisão, mas com maior proporção de FP [Lima et al., 2024].

II.2 Trabalhos Relacionados

A metodologia PICO [Sackett et al., 1996] é utilizada para estruturar a análise dos critérios de DA em séries temporais e para orientar a organização dos trabalhos relacionados a esta dissertação. No elemento P (*Population*), consideram-se as séries temporais provenientes dos domínios descritos na seção II.2.1, incluindo indústria, meio ambiente, cibersegurança, finanças e exploração de petróleo, cada um exigindo escolhas adequadas de medidas de desvio, filtros e estratégias de seleção. O

elemento I (*Intervention*) corresponde aos critérios apresentados na seção II.2.2, abrangendo medidas de desvio ABS e SQD, filtros BP, GD e RT, e estratégias de seleção F e H. Já o elemento C (*Comparison*), discutido na seção II.2.3, envolve o exame de diferentes combinações desses critérios aplicadas a distintas famílias de métodos e avaliadas em *benchmarks* como Yahoo, NAB e Gecco. Por fim, o elemento O (*Outcome*), conforme a seção II.2.4, explicita que essas escolhas afetam diretamente métricas de desempenho como Precisão, *Recall* e F1, o que revela lacunas quanto ao tratamento explícito dos critérios de detecção na literatura e motiva a análise modular proposta nesta dissertação.

A seleção dos estudos é conduzida com o apoio da ferramenta *Elicit*[1], utilizada para organizar e sintetizar as publicações com base em critérios alinhados à estratégia PICO adotada neste trabalho. Os estudos são analisados quanto à sua relevância para a questão de pesquisa, ao domínio de aplicação (por exemplo, séries temporais industriais, ambientais, financeiras e de cibersegurança), às intervenções investigadas, incluindo medidas de desvio, critérios de filtragem e estratégias de seleção de anomalias candidatas, às abordagens de comparação empregadas (como critérios fixos ou métodos de detecção sem adaptação aos dados) e aos resultados reportados, fundamentados em métricas de desempenho como Precisão, *Recall* e F1, bem como na ocorrência de FP e FN.

As buscas automáticas são realizadas por meio de consultas estruturadas a partir de palavras-chave relacionadas à Detecção de Anomalias em séries temporais, ao ajuste fino de critérios de detecção e aos principais domínios de aplicação considerados. Além dessas buscas, são incorporados trabalhos de referência indicados pelo orientador, em especial aqueles já integrados ao *Harbinger* [2] [Salles et al., 2020], como os detectores híbridos REMD (*Residual Empirical Mode Decomposition*) e FBIAD (*Forward and Backward Inertial Anomaly Detector*).

O método REMD combina a decomposição por EMD (*Empirical Mode Decomposition*) com modelos de previsão baseados em ARIMA (*Autoregressive Integrated Moving Average*) para a modelagem de resíduos [Souza et al., 2024]. Já o FBIAD explora medidas de inércia associadas à dinâmica dos dados para caracterizar desvios anômalos [Lima et al., 2022]. A inclusão desses métodos reforça a contextualização dos critérios analisados e fornece base conceitual e metodológica para a comparação sistemática desenvolvida nas seções subsequentes.

II.2.1 Domínios e Contexto

A DA em séries temporais é estudada em diversos domínios, como o monitoramento industrial, a qualidade ambiental, a cibersegurança, as finanças e a exploração ou a perfuração de poços de petróleo. Em cada um desses contextos, os critérios de detecção não são apenas detalhes técnicos,

[1] `https://elicit.com`
[2] https://github.com/cefet-rj-dal/harbinger/tree/master

mas também fatores que condicionam a utilidade prática dos métodos [Chandola et al., 2009; Ariyaluran Habeeb et al., 2019].

No setor industrial, séries temporais de sensores são utilizadas para identificar falhas, degradação do desempenho e gargalos operacionais. Nesses cenários, a DA depende diretamente de como o desvio entre o comportamento observado e o esperado é mensurado, de quais limiares são adotados para definir anomalias e de como os candidatos são selecionados quando múltiplos pontos são sinalizados. Essas escolhas impactam decisões de manutenção, de continuidade operacional e de redução do tempo de inatividade [Ibidunmoye et al., 2015; Ogasawara et al., 2021; Ariyaluran Habeeb et al., 2019].

Em contextos ambientais, séries temporais são empregadas para monitorar variáveis como a qualidade da água, a poluição atmosférica e indicadores climáticos [Perelman et al., 2012]. A DA é empregada para identificar episódios críticos, como contaminações, picos de poluição ou variações abruptas de temperatura. Nesses cenários, a forma como o desvio é mensurado, os limiares de filtragem são definidos e os candidatos são selecionados impacta diretamente a sensibilidade do sistema à detecção de eventos críticos. No monitoramento da qualidade da água, por exemplo, a identificação antecipada de episódios de contaminação é essencial para reduzir riscos à saúde pública [Sela et al., 2012].

Na área de cibersegurança, séries temporais multivariadas provenientes de *logs* de sistemas, tráfego de rede ou métricas de desempenho requerem critérios de detecção capazes de lidar com a não estacionariedade, picos transitórios e a alta variabilidade. Abordagens em tempo real combinam medidas de desvio com filtros para priorizar ocorrências e mitigar FP [Ahmad et al., 2017; Ariyaluran Habeeb et al., 2019; Lima et al., 2022].

Nos sistemas financeiros, séries de transações e indicadores financeiros empregam critérios de desvio e de filtragem para identificar comportamentos atípicos e fraude. A escolha desses critérios afeta diretamente as taxas de FP e FN em cenários operacionais [Ahmed et al., 2016; Cook et al., 2020; Dixit et al., 2022].

Na área de exploração e produção de petróleo, os dados de sensores de perfuração de poços apresentam variações operacionais e mudanças de regime. Critérios explícitos de desvio e de filtragem, bem como decisões de seleção de candidatos, são necessários para detectar eventos em fluxo contínuo e sob não estacionariedade [Lima et al., 2024; Talagala et al., 2020; Ogasawara et al., 2021].

II.2.2 Critérios de Detecção Aplicados nos Métodos de Identificação de Anomalias

Na literatura, a decisão sobre o que constitui uma anomalia é orientada por três componentes fundamentais: a forma como o desvio é mensurado, o critério adotado para estabelecer o limiar de filtragem e a estratégia adotada para selecionar candidatos quando múltiplos pontos são identifica-

dos.

As medidas de desvio definem como a discrepância entre o valor observado e o esperado é quantificada. Modelos estatísticos, como o ARIMA, utilizam predominantemente medidas baseadas em SQD, como a *Root Mean Squared Error* (RMSE) entre valores previstos e observados, tratada como resíduo [Scharf and Demeure, 1991]. Essa mesma lógica é adotada em muitas abordagens de aprendizado profundo, em que *autoencoders* utilizam o erro de reconstrução frequentemente calculado com funções de *Mean Squared Error* (MSE) [Pang et al., 2021; Geiger et al., 2020]; redes recorrentes como *Long Short-Term Memory* (LSTM) utilizam resíduos de previsão também baseados em SQD [Zamanzadeh Darban et al., 2024]; e arquiteturas convolucionais como o *One-Dimensional Convolutional Neural Network* (Conv1D) seguem o mesmo princípio ao comparar valores previstos e observados [Zhang et al., 2023].

Por outro lado, estudos como [Gorard, 2005] e [Hodson, 2022] destacam que medidas baseadas em ABS, como a *Mean Absolute Error* (MAE) e a *Manhattan Distance* (L1), tendem a ser mais robustas em cenários com ruído, presença de *outliers* ou distribuições assimétricas.

Em métodos híbridos que integram a decomposição, como o REMD, a identificação de anomalias baseia-se no desvio entre os componentes de alta frequência extraídos por EMD e os valores previstos pelo ARIMA. Esse desvio é quantificado a partir dos resíduos definidos como $\xi = H - \hat{H}$, avaliados em termos de sua magnitude absoluta e normalizados pelo maior valor absoluto observado ao longo da série, caracterizando uma medida de desvio baseada na distância $L1$, isto é, no desvio absoluto (ABS) [Souza et al., 2024].

Critérios baseados em quartis, como o filtro BP, são amplamente utilizados para identificar valores extremos em séries temporais sem assumir pressupostos paramétricos sobre a distribuição dos dados. Essa abordagem, centrada no IQR, delimita limites superiores e inferiores que permitem isolar outliers com robustez, sendo eficaz em contextos com distribuições assimétricas ou com ruído estrutural [Aggarwal, 2016; Boniol et al., 2024]. Variantes, como o *Boxplot-Mean*, também são exploradas para aumentar a sensibilidade à média. No campo da detecção de anomalias em séries temporais, métodos como o FBIAD aplicam o BP diretamente às medidas de inércia resultantes das variações entre as direções *forward* e *backward*, realçando pontos que destoam significativamente [Lima et al., 2022].

A simplicidade computacional do BP decorre de sua implementação baseada apenas em medidas de posição, como medianas e quartis, sem necessidade de ajustes paramétricos, o que o torna apropriado para aplicações em tempo real ou em cenários com restrições computacionais [Aggarwal, 2016]. Sua resiliência a suposições estatísticas rígidas contribui para sua ampla adoção em *pipelines* de DA em séries temporais [Boniol et al., 2024].

O critério de filtro GD define limiares com base na suposição de que os desvios ou resíduos

seguem uma distribuição aproximadamente normal. A partir de estatísticas como média e desvio-padrão, pontos situados a múltiplos desvios da média são considerados anomalias, o que fundamenta técnicas clássicas de detecção baseadas em resíduos [Chandola et al., 2009]. O GD é amplamente utilizado em modelos de previsão, incluindo regressão estatística, ARIMA e redes neurais profundas, especialmente em configurações nas quais os resíduos tendem a apresentar distribuição aproximadamente simétrica após normalização [Geiger et al., 2020; Pang et al., 2021]. Benchmarks modernos, como o TSB-UAD, aplicam padronização via *z-score* para definir limiares gaussianos, reforçando sua relevância [Boniol et al., 2024].

Essa abordagem é particularmente eficaz em cenários em que os modelos são treinados com funções de perda gaussianas ou com regularizações que promovem simetria residual, o que favorece a aplicação direta de limiares paramétricos [Pang et al., 2021]. A padronização facilita a definição de limites e permite comparações entre séries em escalas distintas [Boniol et al., 2024]. Como consequência, o GD é amplamente empregado em experimentos que demandam reprodutibilidade e controle estatístico rigoroso.

O critério RT, originalmente introduzido no método REMD, define candidatos a anomalia como a razão entre cada ponto da série e seu valor máximo [Souza et al., 2024]. Essa abordagem parte da premissa de que valores anômalos em séries normalizadas, ou com eventos de alta magnitude, tendem a apresentar proporções inferiores ao pico, realçando desvios abruptos. Inspirado em princípios da engenharia eletrônica, o RT captura variações súbitas de sinais, como picos ou quedas de amplitude, sendo eficaz em domínios como o monitoramento industrial, a energia e os sinais biomédicos [Zhang et al., 2021; Souza, 2024]. Ao evitar pressupostos paramétricos, o RT se mostra leve, interpretável e robusto, com desempenho destacado em experimentos envolvendo o REMD.

As estratégias de seleção de candidatos a anomalias resolvem o problema de decisão entre múltiplos pontos que ultrapassam os limiares estabelecidos por BP, GD ou RT. Dois critérios heurísticos prevalecem na literatura: a estratégia F, que seleciona o primeiro ponto excedente e favorece detecções precoces, e a estratégia H, que escolhe o ponto de maior intensidade de desvio e prioriza robustez. Ambos são frequentemente aplicados implicitamente em diversos métodos.

Modelos estatísticos, como ARIMA, costumam adotar implicitamente o critério F ao sinalizar anomalias assim que os resíduos excedem o limiar definido [Zare Moayedi and Masnadi-Shirazi, 2008; Pena et al., 2013]. Em contrapartida, abordagens de aprendizado profundo, como LSTM, *autoencoders* e Conv1D, tendem a adotar o critério H, ao selecionar o ponto com o maior erro de reconstrução ou de previsão [Pang et al., 2021; Geiger et al., 2020; Zhang et al., 2023; Zamanzadeh Darban et al., 2024]. Em métodos baseados em decomposição, como EMD, observa-se maior variação; no caso do REMD, a seleção do tipo H é usada de forma explícita [Souza et al., 2024]. Em *Support Vector Machine* (SVM), a decisão baseia-se na distância ao hiperplano, o que também

privilegia pontos mais atípicos [Ma and Perkins, 2003].

Esses padrões indicam que, mesmo quando não explicitados, os critérios de seleção influenciam o comportamento dos métodos. Estratégias do tipo F são comuns em modelos sequenciais e reativos, enquanto estratégias do tipo H predominam em métodos orientados por escores contínuos. Assim, a combinação entre ABS ou SQD, filtros BP, GD ou RT e estratégias F ou H molda o comportamento da detecção e explica diferenças expressivas entre métodos avaliados sob condições semelhantes.

II.2.3 Comparações

A literatura de DA em séries temporais costuma organizar as comparações em três eixos: métodos clássicos *versus* modelos de *deep learning*, desempenho em *benchmarks* padronizados e análise das métricas de avaliação. Nesta dissertação, contudo, o foco das comparações está nos critérios de detecção previamente apresentados, com ênfase nos critérios de desvio ABS e SQD, nos critérios de filtro BP, GD e RT, e nas estratégias de seleção de candidatos F e H. As análises a seguir discutem como esses critérios são aplicados ou assumidos implicitamente nos principais métodos empregados nos experimentos.

Os métodos de DA em séries temporais abrangem diferentes paradigmas, incluindo modelos estatísticos, funções de decisão, redes neurais profundas e abordagens híbridas. Entre os modelos estatísticos, destacam-se aqueles baseados em ARIMA, amplamente utilizados para previsão univariada por meio de regressão temporal e de análise de resíduos [Zare Moayedi and Masnadi-Shirazi, 2008]. Já os modelos baseados em função de decisão, como SVM e ELM, operam com fronteiras aprendidas em espaços de alta dimensionalidade e são aplicados em configurações *one-class* ou semissupervisionadas [Ma and Perkins, 2003; Aggarwal, 2016].

Modelos de *deep learning*, como LSTM e Conv1D, são amplamente utilizados em arquiteturas de reconstrução ou previsão de sinal, baseando-se na diferença entre valores observados e estimados para inferir desvios significativos [Pang et al., 2021; Zamanzadeh Darban et al., 2024]. Essas abordagens frequentemente adotam medidas como ABS e filtros robustos, como BP ou GD, adequados a contextos ruidosos e a variações locais [Blázquez-García et al., 2021; Salles et al., 2020].

Métodos híbridos como REMD, EMD, FBIAD e ELM combinam técnicas de decomposição de sinais, de previsão e de critérios de decisão, proporcionando maior flexibilidade na detecção de padrões sazonais, rupturas abruptas ou oscilações complexas [Souza et al., 2024; Lima et al., 2022; Salles et al., 2020]. Em particular, o REMD se destaca por integrar explicitamente ABS, RT e H, permitindo o refinamento das decisões com base na variação do desvio, na filtragem proporcional e na escolha do ponto mais crítico [Souza, 2024].

Em termos de *datasets*, os métodos de DA são avaliados em três *benchmarks* amplamente utilizados: Yahoo, NAB e Gecco. Para manter a consistência entre as comparações, as métricas adotadas

incluem Precisão, *Recall* e F1, sendo esta última a principal métrica empregada na avaliação final, conforme práticas estabelecidas em estudos anteriores [Blázquez-García et al., 2021; Pang et al., 2021]. A análise revela que diferentes combinações entre medidas de desvio ABS e SQD, filtros BP, GD e RT e estratégias de seleção F e H podem impactar significativamente o desempenho dessas métricas, mesmo quando aplicadas ao mesmo modelo de DA. Esse comportamento reforça que não apenas o método importa, mas também os mecanismos internos que o operacionalizam.

Por fim, embora a literatura costume tratar os métodos como caixas fechadas, esta dissertação adota uma perspectiva mais granular ao explorar e modular os critérios subjacentes à detecção. Ao explicitar esses mecanismos e suas variações, abre-se espaço para estratégias de ajuste fino adaptadas ao domínio de aplicação e ao perfil das séries temporais analisadas, em consonância com abordagens recentes de configuração personalizada para tarefas de DA [Souza et al., 2024; Aggarwal, 2016; Blázquez-García et al., 2021].

II.2.4 Resultados e Lacunas

A literatura revisada evidencia que os critérios de DA exercem influência decisiva sobre o desempenho dos métodos. No caso das medidas de desvio, diferentes funções apresentam desempenhos contrastantes; enquanto os desvios SQD tendem a ser mais adequados a distribuições próximas da normalidade, os desvios ABS demonstram maior robustez em cenários caracterizados por ruído elevado ou distribuições assimétricas [Gorard, 2005; Hodson, 2022].

De forma análoga, os critérios de filtragem constituem um fator sensível no processo de detecção. Estudos indicam que pequenas variações nos filtros BP, GD ou RT podem alterar significativamente as taxas de FP e FN, inclusive em *benchmarks* padronizados como Yahoo, NAB e Gecco [Wenig et al., 2022; Paparrizos et al., 2022]. Em particular, abordagens baseadas em limiares adaptativos, como o filtro GD, mostram-se mais flexíveis ao incorporar estatísticas locais da série, enquanto limiares fixos tendem a apresentar limitações de generalização entre diferentes domínios [Chandola et al., 2009; Aggarwal, 2016].

No que se refere à seleção de candidatos, a literatura aponta que a priorização de anomalias pontuais (F) ou de segmentos anômalos (H) impacta diretamente as métricas finais de desempenho. Métodos como REMD e FBIAD explicitam essa escolha em sua formulação, evidenciando que a estratégia adotada influencia o conjunto de anomalias reportadas [Souza et al., 2024; Lima et al., 2022].

Embora esses estudos confirmem a relevância dos critérios de DA, observa-se que, na maior parte da literatura, tais decisões são tratadas de forma isolada, implícita ou fortemente acoplada a configurações específicas de cada método. Em contraste, esta dissertação propõe uma avaliação integrada e sistemática dos critérios de DA, analisando de forma conjunta os efeitos das medidas

de desvio, dos critérios de filtragem e das estratégias de seleção de candidatos sobre o desempenho dos métodos, sob múltiplas métricas e diferentes *benchmarks*, com apoio em testes de significância e medidas de efeito.

Com isso, o trabalho avança sobre as lacunas identificadas na literatura ao oferecer uma abordagem modular, comparativa e reprodutível para o ajuste fino dos critérios de DA, contribuindo para a compreensão dos efeitos isolados e combinados desses critérios na identificação de anomalias em séries temporais.

II.3 Considerações

De modo geral, a revisão da literatura indica que a pesquisa em DA tem se concentrado predominantemente no desenvolvimento de novos métodos e na comparação entre diferentes abordagens. Nesse contexto, os critérios de DA costumam ser incorporados como escolhas auxiliares, muitas vezes tratadas de forma implícita ou secundária nos modelos propostos.

Essa característica da literatura evidencia a necessidade de estudos que investiguem de maneira estruturada o papel desses critérios, considerando seu impacto sobre o desempenho e a interpretação dos resultados. Diferentemente dos *benchmarks* experimentais tradicionais, que priorizam a comparação entre métodos, o *framework* proposto neste trabalho organiza esses critérios como dimensões do processo de decisão, permitindo avaliar como diferentes configurações influenciam o comportamento dos métodos em termos de sensibilidade e seletividade. Além disso, a ausência de um método universalmente superior, bem como os *trade-offs* recorrentes entre métricas de avaliação, reforçam a importância de análises comparativas conduzidas sob diferentes configurações experimentais.

Capítulo III Método de *Benchmark*

Este capítulo apresenta o método de *benchmark* adotado nesta dissertação, cujo objetivo é avaliar métodos de DA a partir de uma perspectiva mais granular do que aquela usualmente encontrada na literatura. Em vez de tratar os métodos como caixas fechadas, a abordagem proposta explicita os principais critérios de DA, permitindo sua combinação, especialização e ajuste fino no processo de identificação de anomalias.

Para esse fim, o método de *benchmark* organiza o processo de DA em três componentes centrais: o Critério de Medida de Desvio, o Critério de Filtro e o Critério de Seleção de Candidatos, conforme ilustrado na Figura III.1. Essa decomposição possibilita a análise sistemática do papel desempenhado por cada critério e de suas interações no desempenho dos métodos avaliados.

III.1 Visão Geral do *Framework* de *Benchmark*

Durante a execução do *benchmark*, diferentes configurações dos critérios de DA são consideradas e combinadas ao processo de detecção de anomalias. Embora a estrutura geral do método permaneça a mesma, a adoção de estratégias distintas para os critérios de Medida de Desvio, Filtro e Seleção de Candidatos pode resultar em comportamentos significativamente diferentes.

Essa separação conceitual entre o método de DA e os critérios internos de detecção permite avaliar, de forma controlada, o impacto individual e combinado desses componentes no desempenho final, sem pressupor alterações na formulação do método subjacente.

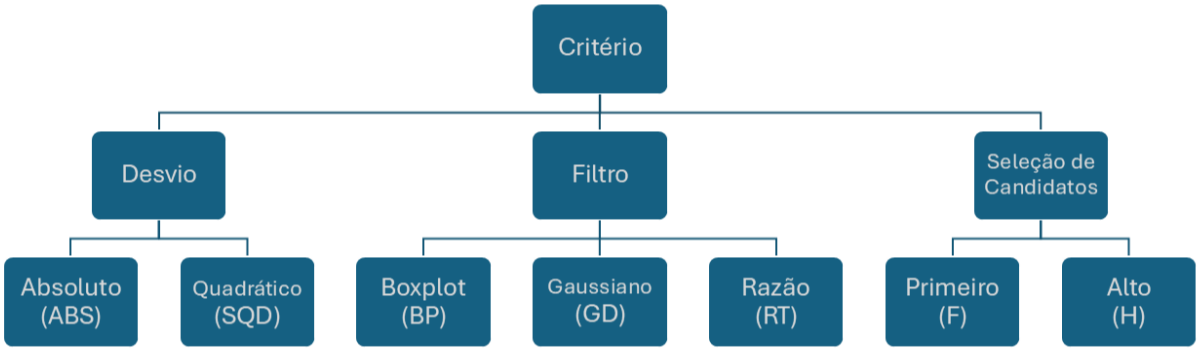


Figura III.1: Medidas de Desvio, Limiares de Filtragem e Estratégias de Seleção de Candidatos

A Figura III.1 sintetiza do processo de identificação de anomalias em três componentes centrais.

Conforme apresentação na seção II.1, o método de DA é incialmente ajustado à série temporal de entrada, produzindo resíduos que representam a discrepância entre os valores observados e estimados. Em seguida, uma Medida de Desvio é aplicada a esses resíduos para quantificar a magnitude das discrepâncias. Os valores resultantes são então submetidos a um Critério de Filtro, responsável por identificar observações candidatas a anomalias com base em limiares definidos. Por fim, um Critério de Seleção de Candidatos é empregado para confirmar quais observações devem ser efetivamente classificadas como anômalas.

No Algoritmo 1, a função de *benchmark* recebe como entrada uma série temporal Y, um método de DA adm e os componentes correspondentes aos critérios de Medida de Desvio ($devm$), Filtro ($filter$) e Seleção de Candidatos ($candidate$). Esses elementos definem o método de detecção e os critérios de DA responsáveis pelo ajuste fino do processo de identificação de anomalias, ao controlar a quantificação dos desvios, os mecanismos de filtragem e as estratégias de seleção de candidatos.

Algorithm 1 Algoritmo de Benchmark

```
1:  function benchmark(Y, adm, devm, filter, candidate)
2:      model ← setup(adm, devm, filter, candidate)
3:      model ← fit(model, Y)                          ▷ Ajusta o modelo com os dados
4:      anomalies ← detect(model, Y)                   ▷ Detecta anomalias
5:      return anomalies
6:  function detect(model, Y)
7:      ω ← residuals(model, Y)
8:      ω̇ ← devm(model, ω)
9:      canom ← filter(model, ω̇)
10:     anom ← candidate(model, canom)
11:     return anom
```

As etapas iniciais do *benchmark* correspondem à configuração e ao ajuste do modelo. Na linha 2, o modelo é configurado a partir do método adm e dos critérios selecionados; na linha 3, o modelo é ajustado à série temporal Y; e, na linha 4, a função de detecção é invocada com base no modelo ajustado.

As operações específicas de detecção são implementadas no corpo da função `detect`, apresentada nas linhas 6 a 11. Nessa etapa, os resíduos ω são calculados a partir da aplicação do modelo ajustado à série temporal Y (linha 7). Em seguida, na linha 8, esses resíduos são transformados em uma medida de desvio $\dot{\omega}$ de acordo com o Critério de Medida de Desvio. A partir dessa medida, candidatos a anomalias $canom$ são identificados por meio do Critério de Filtro (linha 9) e, por fim, as anomalias finais $anom$ são determinadas pelo Critério de Seleção de Candidatos (linha 10), sendo então retornadas pela função (linha 11).

III.2 Medidas de Desvio

Conforme descrito na Seção III.1 e ilustrado na Figura III.1, o Critério de Medida de Desvio (*devm*) é utilizado para quantificar a magnitude dos desvios na série temporal. No processo de DA, essa etapa estabelece a ligação entre o comportamento estimado pelo método e as discrepâncias observadas nos dados, operando diretamente sobre os resíduos gerados após o ajuste do modelo à série temporal.

O *framework* viabiliza a escolha entre diferentes estratégias de cálculo, conforme discutido na Seção II.1.1, incluindo o ABS e o SQD. Essas medidas fornecem uma visão sobre a variância dos resíduos e estão associadas ao resultado da detecção. Ao explicitar esse critério no *benchmark*, torna-se possível analisar, de forma controlada, como diferentes formas de quantificação dos resíduos influenciam o comportamento global do processo de detecção.

Do ponto de vista metodológico, a Medida de Desvio atua como um elemento de padronização conceitual entre métodos distintos de DA. Conforme formalizado no Algoritmo 1, o cálculo dos resíduos produz, independentemente do paradigma adotado pelo método subjacente, uma sequência de valores que expressa o afastamento entre observações e estimativas. A aplicação de uma Medida de Desvio sobre esses resíduos transforma essa informação em uma representação numérica comparável, que serve de base para as etapas de filtragem e de seleção de candidatos.

A escolha da Medida de Desvio afeta diretamente a sensibilidade do processo de detecção a ruído, *outliers* e variações locais na série temporal. Diferentes estratégias de cálculo podem gerar distribuições distintas de valores de desvio, impactando o conjunto de observações submetidas ao Critério de Filtro e, consequentemente, a quantidade e a distribuição dos candidatos a anomalias identificados nas etapas seguintes do *pipeline*.

Ao isolar o Critério de Medida de Desvio como um componente modular do *framework* de *benchmark*, conforme descrito na Seção III.1, este trabalho possibilita a avaliação sistemática do impacto dessa escolha no desempenho final da DA. Mantendo constantes o método de detecção e os demais critérios do *pipeline*, a variação controlada da Medida de Desvio permite investigar como diferentes estratégias de quantificação dos resíduos se refletem nos resultados obtidos, fornecendo subsídios para análises comparativas mais precisas nas seções subsequentes.

III.3 Critérios de Filtro

Conforme descrito na Seção III.1 e formalizado no Algoritmo 1, o Critério de Filtro ($filter$) é utilizado para definir os limiares que determinam se os valores de desvio são suficientemente elevados para que observações sejam consideradas candidatas a anomalias. No *pipeline* da DA, esse critério atua diretamente sobre os valores de desvio produzidos na etapa anterior, controlando quais

observações serão consideradas candidatas à análise subsequente.

O *framework* de *benchmark* viabiliza a escolha entre diferentes estratégias de filtragem, conforme descrito na Seção II.1.2, incluindo BP, GD e RT. Essas estratégias permitem estabelecer limiares distintos para a identificação de candidatos a anomalias, influenciando o conjunto de observações retidas após a aplicação do filtro e, consequentemente, o comportamento global do processo de DA.

Do ponto de vista metodológico, o Critério de Filtro exerce papel fundamental no equilíbrio entre FP e FN. Filtros mais permissivos tendem a reter um maior número de candidatos, aumentando a sensibilidade do processo de detecção, enquanto filtros mais conservadores reduzem o conjunto de candidatos, privilegiando decisões mais restritivas. Esse *trade-off* impacta diretamente as etapas subsequentes do *pipeline*, em particular a aplicação do Critério de Seleção de Candidatos.

Ao definir explicitamente o Critério de Filtro como um componente modular do *framework* de *benchmark*, torna-se possível avaliar, de maneira sistemática, o impacto de diferentes estratégias de limiarização no processo de DA. Mantendo constantes o método de detecção e a Medida de Desvio, a variação controlada do Critério de Filtro permite analisar como a filtragem afeta a quantidade, a distribuição temporal e a relevância dos candidatos a anomalias.

Além disso, a interação entre o Critério de Filtro e a Medida de Desvio reforça a importância da modularidade adotada neste trabalho. Diferentes combinações entre estratégias de quantificação dos resíduos e critérios de filtragem podem produzir comportamentos substancialmente distintos, mesmo quando aplicadas ao mesmo método de DA, evidenciando a necessidade de tratar o Critério de Filtro como um componente explícito e ajustável no processo de *benchmark*.

III.4 Seleção de Candidato

Conforme descrito na Seção III.1 e formalizado no Algoritmo 1, o Critério de Seleção de Candidatos (*candidate*) é aplicado quando múltiplos candidatos a anomalias temporalmente próximos são identificados após a etapa de filtragem. No *pipeline* da DA, essa etapa atua após a filtragem dos desvios, consolidando as decisões de detecção e evitando a marcação redundante de observações associadas a um mesmo evento anômalo.

O *framework* de *benchmark* viabiliza a escolha entre diferentes estratégias de seleção de candidatos, conforme apresentado na Seção II.1.3, incluindo F e H. Essas estratégias definem como o conjunto de candidatos produzido pelo Critério de Filtro é reduzido a um conjunto final de anomalias, influenciando diretamente o resultado do processo de detecção.

Do ponto de vista metodológico, o Critério de Seleção de Candidatos exerce impacto direto sobre o equilíbrio entre Precisão e *Recall*. Estratégias que priorizam a confirmação antecipada de anomalias tendem a favorecer a identificação rápida de eventos anômalos individuais, enquanto estratégias que privilegiam a seleção do desvio mais representativo em uma região temporal podem

reduzir a ocorrência de detecções redundantes. Essas diferenças afetam a forma como o processo de DA responde a regiões da série temporal caracterizadas por múltiplos desvios consecutivos.

Ao definir explicitamente o Critério de Seleção de Candidatos como um componente modular do *framework* de *benchmark*, torna-se possível analisar, de maneira controlada, como diferentes estratégias de consolidação de candidatos influenciam o desempenho final da identificação de anomalias. Mantendo constantes o método de DA, a Medida de Desvio e o Critério de Filtro, a variação isolada do Critério de Seleção de Candidatos permite investigar seu impacto sobre a quantidade, a distribuição temporal e a relevância das anomalias confirmadas.

III.5 Implementação no *Harbinger*

Esta seção descreve como o método de *benchmark* apresentado ao longo deste capítulo foi operacionalizado no *framework Harbinger*. Em particular, são detalhadas as decisões de implementação adotadas para materializar os critérios de Medida de Desvio, Critério de Filtro e Critério de Seleção de Candidatos, definidos nas Seções III.2, III.3 e III.4. Dessa forma, evidencia-se como as escolhas metodológicas propostas foram incorporadas ao ambiente experimental, assegurando que não permaneçam apenas em nível conceitual, mas sejam efetivamente aplicadas em experimentos reprodutíveis na linguagem *R*.

O *Harbinger* é um *framework* voltado à análise de séries temporais que oferece suporte integrado à detecção de anomalias, à detecção de pontos de mudança e à descoberta de motivos [Salles et al., 2020]. Sua arquitetura é baseada em objetos do tipo `S3`, um mecanismo de orientação a objetos leve do *R* no qual o comportamento dos detectores é definido por funções genéricas e é especializado dinamicamente conforme a classe do objeto, e no conceito de *Experiment Lines*, sendo construída sobre o *DAL Toolbox*. Essa organização permite a composição flexível de métodos de detecção com funções auxiliares responsáveis pela quantificação dos desvios, pela aplicação de critérios de filtragem e pela consolidação dos eventos detectados, o que se mostrou aderente ao método de *benchmark* adotado nesta dissertação.

III.5.1 Módulo Utilitário `harutils`

A implementação dos critérios de DA no *Harbinger* é centralizada no módulo utilitário `harutils`[1], responsável por agrupar as funções associadas às Medidas de Desvio, aos Critérios de Filtro e às Estratégias de Seleção de Candidatos definidos neste capítulo. Esse módulo constitui o principal ponto de materialização, no nível da implementação, da abstração metodológica adotada no método de *benchmark*, permitindo configurar, substituir e avaliar isoladamente cada componente interno do *pipeline* de detecção.

[1] `https://github.com/cefet-rj-dal/harbinger/blob/master/R/harbinger_utils.R`

Ao concentrar essas funcionalidades em um único módulo utilitário, o *Harbinger* assegura a aplicação consistente dos critérios de detecção ao longo dos experimentos, independentemente do método de DA subjacente. Essa organização reflete diretamente a separação conceitual estabelecida na Seção III.1, na qual os critérios de Medida de Desvio, Filtro e Seleção de Candidatos são tratados como mecanismos explícitos de ajuste fino do processo de identificação de anomalias.

As Medidas de Desvio discutidas na Seção III.2 são implementadas por funções do tipo `har_distance_*`, que contemplam agregações baseadas nas normas *Manhattan Distance* (L1) e *Euclidean Norm* (L2), associadas, respectivamente, aos critérios ABS e SQD. Os Critérios de Filtro apresentados na Seção III.3 são operacionalizados por funções do tipo `har_outliers_*`, abrangendo estratégias baseadas em BP, GD e RT. De forma análoga, as Estratégias de Seleção de Candidatos descritas na Seção III.4 são implementadas por funções do tipo `har_outliers_checks_*`, incluindo as estratégias F e H.

Essa organização evita a duplicação de código entre métodos distintos e reforça o caráter modular do *framework*, viabilizando a análise controlada do impacto individual e combinado de cada critério de DA, conforme proposto nesta dissertação. A implementação do módulo `harutils` encontra-se disponível no repositório do projeto.

III.5.2 Configurações *default* e evolução do modelo de critérios

A implementação dos critérios de DA no *Harbinger* é centralizada no módulo utilitário `harutils`, responsável por agrupar as funções associadas às Medidas de Desvio, aos Critérios de Filtro e às Estratégias de Seleção de Candidatos definidos neste capítulo.

Historicamente, em métodos propostos antes da consolidação da abordagem de ajuste fino de critérios, como o FBIAD [Lima et al., 2022] e o REMD [Souza et al., 2024], essas escolhas estavam presentes, porém eram, em grande parte, implícitas e raramente explicitadas de forma isolada. Com o avanço das investigações sobre o impacto dos critérios de DA no desempenho da detecção, particularmente após a publicação dos estudos de ajuste fino de critérios de DA [Sobrinho et al., 2025], o *Harbinger* passou a adotar explicitamente um modelo de explicitação dos critérios, no qual essas decisões se tornam visíveis e configuráveis no nível da implementação.

Do ponto de vista metodológico, essa evolução é fundamental, pois evidencia que o comportamento de um método não depende apenas de seu modelo estatístico ou de aprendizado, mas também das combinações específicas de critérios que operacionalizam a decisão de anomalia. Tornar essas escolhas explícitas permite interpretar de forma mais precisa os resultados experimentais e fundamenta o método de *benchmark* adotado nesta dissertação, no qual tais critérios são avaliados de forma controlada e comparativa.

III.5.3 Especializações de métodos e decisões metodológicas

Além das configurações *default*, alguns métodos implementados no *Harbinger* redefinem internamente os critérios de detecção como parte de sua própria concepção metodológica. Nesses casos, as escolhas de Medida de Desvio, Critério de Filtro e Estratégia de Seleção de Candidatos não decorrem de intervenções do usuário final, mas refletem decisões inerentes ao método.

O REMD constitui um exemplo representativo desse comportamento, ao integrar explicitamente combinações específicas de critérios discutidas na literatura, como ABS, RT e H, com o objetivo de capturar variações proporcionais e selecionar os desvios mais representativos em agrupamentos contíguos [Souza et al., 2024]. Essas escolhas estão diretamente alinhadas às análises comparativas apresentadas na Subseção II.2.3 e ilustram como critérios frequentemente tratados de forma implícita podem ser formalizados e analisados no nível da implementação.

III.5.4 Customização dos critérios no uso

Embora os métodos de DA possuam configurações *default* e, em alguns casos, especializações internas, o *Harbinger* permite que os critérios de detecção sejam customizados no momento do uso por meio da substituição das funções associadas aos componentes internos do método. Essa customização é realizada diretamente a partir das implementações disponibilizadas pelo módulo utilitário `harutils`, sem a necessidade de modificações no código-fonte dos detectores.

Do ponto de vista metodológico, essa possibilidade de customização operacionaliza, no nível do uso, a separação conceitual entre o método de detecção e os critérios de DA estabelecida ao longo deste capítulo. Ao permitir a substituição controlada das Medidas de Desvio, dos Critérios de Filtro e das Estratégias de Seleção de Candidatos, o *framework* viabiliza a aplicação prática do ajuste fino de critérios, mantendo inalterado o núcleo do método avaliado.

Exemplos práticos de customização e avaliação do *pipeline* do *Harbinger* encontram-se disponíveis na documentação do projeto, como material complementar de apoio ao uso e à reprodutibilidade experimental[2].

Essa flexibilidade constitui um elemento central do método de *benchmark* adotado nesta dissertação, pois possibilita a avaliação sistemática do impacto individual e combinado dos critérios de DA sob diferentes configurações experimentais, como explorado no Capítulo IV.

[2] `https://github.com/cefet-rj-dal/harbinger/tree/master/examples/general`

Capítulo IV Avaliação Experimental

Este Capítulo analisa os resultados experimentais para avaliar o impacto da variação dos três principais critérios de detecção: (i) Critério de Medida de Desvio, (ii) Critério de Filtro e (iii) Critério de Seleção de Candidato. Esses critérios são avaliados utilizando diferentes métodos de DA em diversos tipos de séries temporais, respondendo às questões de pesquisa propostas no Capítulo I.

IV.1 Configuração Experimental

Três conjuntos de dados com características distintas foram selecionados para avaliar a abordagem proposta: *Yahoo Labs Dataset* (Yahoo), *Numenta Anomaly Benchmark* (NAB) e *GECCO 2018 Challenge* (Gecco).

O conjunto de dados Yahoo contém observações horárias com anomalias rotuladas manualmente, englobando séries temporais reais e sintéticas. Ele inclui cinco séries temporais reais (subconjunto A1Benchmark) relacionadas ao tráfego de serviços do Yahoo e quinze séries temporais sintéticas (subconjuntos A2Benchmark, A3Benchmark e A4Benchmark) [Salles et al., 2020].

O conjunto NAB inclui séries temporais rotuladas para tarefas de DA, com dados sintéticos e reais. Ele é utilizado para monitorar serviços em nuvem, como uso da Unidade Central de Processamento (*Central Processing Unit*, CPU), tráfego de rede e leituras de disco [Lima et al., 2022; Souza et al., 2024].

O conjunto de dados Gecco é composto por 1.500 observações horárias relacionadas ao monitoramento da qualidade da água, com 72 eventos rotulados em nove variáveis, incluindo temperatura, pH, potencial redox, condutividade elétrica e turbidez [Lima et al., 2022; Souza et al., 2024].

Embora os conjuntos de dados utilizados possuam anomalias previamente rotuladas, os métodos avaliados baseiam-se na modelagem do comportamento esperado da série temporal para a obtenção de resíduos, sem empregar procedimentos de treinamento supervisionado orientados pelos rótulos. As séries completas são utilizadas no processo de detecção, e as anotações de anomalia são consideradas apenas na etapa de avaliação do desempenho dos métodos.

Cada método de DA foi avaliado sob todas as combinações de:

- Medida de Desvio: ABS e SQD,
- Critério de Filtro: BP, GD e RT,

- Critério de Seleção de Candidatos: F e H.

Os métodos empregados incluíram abordagens estatísticas [Souza et al., 2024], EMD, FBIAD [Lima et al., 2022] e ARIMA, além de modelos de aprendizado de máquina (LSTM, ELM, Conv1D e SVM) [Ogasawara et al., 2025]. Esses métodos foram escolhidos com base em sua relevância para tarefas de DA e na disponibilidade no *Harbinger framework* [Salles et al., 2020].

Os experimentos foram conduzidos em um ambiente computacional com processador Xeon (12 núcleos), 256 GB de RAM e Ubuntu 22.04 LTS. Foram executados métodos estatísticos e baseados em decomposição e modelagem temporal, incluindo EMD, REMD, FBIAD e ARIMA, bem como modelos de aprendizado de máquina, nomeadamente LSTM, ELM, Conv1D e SVM.

Os hiperparâmetros desses métodos foram definidos da seguinte forma: EMD e REMD utilizaram cinco execuções com parâmetro de ruído de 0,1. O FBIAD empregou uma janela deslizante de 30 observações. Os modelos LSTM e Conv1D foram treinados por 10.000 épocas. O ELM utilizou a função de ativação Purelin, enquanto o SVM adotou o kernel de função de base radial (*Radial Basis Function* (RBF)).

O desempenho foi avaliado utilizando métricas de classificação derivadas da matriz de confusão, VP, VN, FP e FN [Salles et al., 2024]. As medidas derivadas foram: (i) Precisão, a proporção de anomalias corretamente identificadas entre todas as detecções, (ii) *Recall*, a proporção de anomalias reais que são detectadas, e (iii) F1, a média harmônica entre Precisão e *Recall* [Han et al., 2022].

As diferentes combinações de Medida de Desvio, Critério de Filtro e Seleção de Candidatos resultam em comportamentos distintos dos métodos quanto à sensibilidade e à seletividade da detecção, refletidos diretamente nas métricas de Precisão, *Recall* e F1.

Seguindo essa configuração, a Seção IV.2 apresenta a análise de desempenho geral dos critérios definidos na Seção III. Na sequência, os resultados são discutidos de forma detalhada para cada conjunto de dados: Yahoo (Seção IV.3), Gecco (Seção IV.4) e NAB (Seção IV.5). Adicionalmente, a Seção IV.6 apresenta um estudo aprofundado sobre uma série específica do conjunto Yahoo.

Para orientar a leitura das tabelas, valores destacados em negrito indicam o melhor desempenho médio observado, enquanto valores sublinhados representam desempenhos estatisticamente equivalentes ao melhor, conforme os testes estatísticos aplicados (Wilcoxon, Kruskal–Wallis e pós-hoc de Dunn). A adoção desses testes não paramétricos é motivada pelo fato de que, em avaliações comparativas envolvendo múltiplos métodos e conjuntos de dados, os pressupostos de normalidade e homocedasticidade raramente são satisfeitos, conforme indicado por testes clássicos de verificação de normalidade e comparação de distribuições [Shapiro and Wilk, 1965; Friedman, 1937]. Esse cenário é particularmente recorrente em tarefas de DA, nas quais os resultados de desempenho são obtidos a partir de séries temporais heterogêneas, frequentemente caracterizadas por distribuições assimétricas e elevada variabilidade entre instâncias.

Nesse contexto, os testes de Wilcoxon [Wilcoxon, 1945] e de Kruskal–Wallis [Kruskal and Wallis, 1952] constituem alternativas robustas para a comparação de métodos sob distribuições não normais, enquanto o procedimento pós-hoc de Dunn [Dunn, 1964] permite identificar pares de configurações com diferenças estatisticamente significativas ou estatisticamente equivalentes. A padronização adotada está alinhada às boas práticas consolidadas na literatura de comparação experimental, segundo as quais diferenças numéricas isoladas não devem ser interpretadas como evidência suficiente de superioridade entre métodos ou configurações. Dessa forma, o uso combinado de testes de significância estatística e critérios de equivalência estatística reduz o risco de conclusões baseadas em variações marginais de desempenho, assegurando maior rigor e consistência na comparação entre os critérios de detecção avaliados.

IV.2 Análise Geral da Detecção

Esta seção analisa o desempenho médio de Precisão, *Recall* e F1 nos três conjuntos de dados avaliados, com foco nos critérios internos que estruturam o processo de detecção de anomalias apresentados na Seção III: Medida de Desvio, Critério de Filtro e Critério de Seleção de Candidatos. A Tabela IV.1 apresenta as médias globais calculadas a partir de todas as combinações testadas, permitindo avaliar o impacto relativo de cada critério no desempenho geral da detecção.

Tabela IV.1: Resultado geral pelos critérios de detecção.

	Medidas de Desvio		**Filtros**			**Candidatos**	
Métrica	ABS	SQD	BP	GD	RT	F	H
Precisão	**0,219**	0,217	0,070	0,239	**0,345**	**0,225**	0,210
Recall	**0,315**	0,290	**0,376**	0,331	0,200	**0,323**	0,281
F1	**0,201**	0,178	0,089	**0,246**	0,237	**0,199**	0,179

Valores em negrito indicam o melhor desempenho médio por critério e métrica. Valores sublinhados indicam desempenho estatisticamente equivalente ao melhor.

A partir da Tabela IV.1, observa-se que os critérios internos exercem impactos distintos sobre as métricas de desempenho, sendo o Critério de Filtro o principal fator de diferenciação entre as configurações avaliadas. Em contraste, as variações associadas às Medidas de Desvio e à Seleção de Candidatos apresentam influência limitada no comportamento global da detecção.

Para complementar a análise baseada em médias globais, a Tabela IV.2 apresenta uma classificação qualitativa da magnitude das diferenças observadas entre critérios de detecção. Essa classificação foi fundamentada na combinação de três evidências: (i) os resultados dos testes estatísticos de significância, (ii) as estimativas de tamanho de efeito reportadas pelos testes não paramétricos (por exemplo, ϵ^2 no caso do teste de Kruskal–Wallis) e (iii) as diferenças percentuais médias em relação à melhor configuração (*diff_pct*), calculadas a partir das médias globais.

Diferenças classificadas como *desprezíveis* correspondem a cenários nos quais os testes estatís-

ticos não indicaram significância e os tamanhos de efeito foram nulos ou muito reduzidos, além de diferenças percentuais médias inferiores a aproximadamente 10% em relação ao melhor desempenho. A classificação *pequena* foi atribuída a comparações nas quais, embora as diferenças não alterem substancialmente o regime de detecção, observaram-se efeitos estatísticos ou diferenças percentuais moderadas, sugerindo impacto prático limitado, porém não negligenciável. Comparações não incluídas no núcleo competitivo de cada métrica foram explicitamente marcadas como "–", evitando interpretações infladas baseadas em contrastes triviais.

Tabela IV.2: Escala qualitativa de magnitude por critério e métrica de desempenho.

Comparação	**Precisão**	***Recall***	***F1-score***
Medida de Desvio (ABS vs. SQD)	Desprezível	Desprezível	Pequena
Filtro (GD vs. RT)	Desprezível	–	Desprezível
Filtro (BP vs. GD)	–	Pequena	–
Filtro (BP vs. RT)	–	–	–
Seleção de Candidatos (F vs. H)	Desprezível	Pequena	Desprezível

A classificação indica a magnitude das diferenças entre critérios no núcleo competitivo de cada métrica; "–" indica comparações não consideradas.

No que se refere às Medidas de Desvio, as diferenças observadas entre ABS e SQD são numericamente pequenas em todas as métricas avaliadas. A comparação entre essas duas medidas, conduzida por meio do teste não paramétrico de Wilcoxon, não aponta diferenças estatisticamente significativas. Embora ABS apresente valores ligeiramente superiores em Precisão, *Recall* e F1, os resultados dos testes indicam equivalência estatística.

Conforme sintetizado na Tabela IV.2, as magnitudes dessas diferenças são classificadas como desprezíveis ou pequenas, evidenciando que ambas as medidas produzem comportamentos semelhantes do ponto de vista prático. Esse resultado mostra que a escolha entre desvios absolutos ou quadráticos, por si só, não altera de forma consistente o equilíbrio entre detecções corretas e alarmes espúrios, exercendo um papel secundário no desempenho geral da detecção.

Em relação ao Critério de Filtro, as médias apresentadas na Tabela IV.1 revelam padrões distintos de comportamento conforme a métrica analisada. As diferenças entre os filtros foram avaliadas por meio de testes para múltiplos grupos (Kruskal–Wallis ou ANOVA, conforme a métrica), seguidos de análises pós-hoc, permitindo identificar regimes distintos de sensibilidade e seletividade. Esses padrões refletem compromissos clássicos entre sensibilidade e seletividade, os quais se manifestam de maneira diferente em Precisão, *Recall* e F1.

Em Precisão, os filtros GD e RT apresentam valores médios superiores, operando em um regime mais conservador, associado à redução de FP. A Tabela IV.2 caracteriza a diferença entre essas duas configurações como de magnitude desprezível, situando ambas em um patamar semelhante de seletividade. O filtro BP, por sua vez, apresenta valores inferiores nessa métrica, comportamento

compatível com seu caráter mais permissivo.

Em *Recall*, observa-se o comportamento complementar. O filtro BP alcança os maiores valores médios, refletindo sua maior capacidade de retenção de anomalias reais, ainda que ao custo de menor seletividade. O filtro GD exibe desempenho estatisticamente equivalente, com diferença de pequena magnitude (Tabela IV.2), indicando um compromisso intermediário entre sensibilidade e controle de FP. O filtro RT opera em um regime mais restritivo, priorizando a seletividade e resultando em valores inferiores de *Recall*.

No caso do F1, que expressa o equilíbrio entre Precisão e *Recall*, os filtros GD e RT formam o núcleo competitivo, apresentando desempenhos estatisticamente equivalentes e diferença de magnitude desprezível (Tabela IV.2). Esse resultado evidencia que ambas as estratégias alcançam um equilíbrio semelhante entre sensibilidade e seletividade, refletindo uma compensação mais estável entre FP e FN. O filtro BP apresenta valores inferiores nessa métrica, indicando que o aumento de sensibilidade observado em *Recall* implica maior incidência de FP, o que compromete o desempenho global quando ambos os tipos de erro são considerados simultaneamente.

No que se refere à Seleção de Candidatos, a comparação entre as estratégias F e H, conduzida por meio do teste de Wilcoxon, indica diferenças de pequena ou desprezível magnitude nas métricas avaliadas. Esses resultados mostram que esse critério atua predominantemente como um mecanismo de refinamento, influenciando de forma mais sutil o conjunto final de detecções, sem alterar de maneira significativa os regimes definidos pelo critério de filtro.

De forma geral, a leitura conjunta das Tabelas IV.1 e IV.2 mostra que as Medidas de Desvio ABS e SQD apresentam comportamentos largamente equivalentes, enquanto o Critério de Filtro está associado a variações mais pronunciadas entre regimes de detecção, dependendo da métrica considerada. A Seleção de Candidatos exerce influência adicional, atuando como um mecanismo de refinamento sobre conjuntos de candidatos já definidos pelo filtro, com impacto mais sutil no desempenho médio.

As seções seguintes estendem essa análise aos conjuntos Yahoo (Seção IV.3), Gecco (Seção IV.4) e NAB (Seção IV.5), bem como a uma análise aprofundada da série A3-TS43 do conjunto Yahoo (Seção IV.6). Essa organização permite examinar como os critérios internos de DA se manifestam sob diferentes características de conjuntos de dados de séries temporais, articulando os padrões globais identificados nesta seção com análises contextualizadas por domínio e por série.

IV.3 Análise Detalhada do Conjunto de Dados: YAHOO

Esta seção apresenta a análise detalhada dos experimentos conduzidos no conjunto de dados Yahoo, um *benchmark* amplamente utilizado na avaliação de métodos de DA. O conjunto é composto por séries temporais univariadas, com amostragem horária e anomalias rotuladas em nível de

ponto, abrangendo tanto séries reais associadas ao monitoramento de tráfego e infraestrutura de serviços (A1Benchmark) quanto séries sintéticas geradas para simular diferentes padrões de comportamento anômalo (A2Benchmark, A3Benchmark e A4Benchmark). Essa composição heterogênea permite avaliar o desempenho dos métodos sob cenários variados de regularidade, ruído, sazonalidade e densidade de anomalias, constituindo um ambiente desafiador e representativo para a análise comparativa.

A análise explorou a interação entre os métodos de DA e os critérios internos de detecção, Medida de Desvio, Critério de Filtro e Seleção de Candidatos, a partir de 96 combinações distintas, resultantes da aplicação de todas as configurações possíveis de critérios a cada método avaliado no conjunto Yahoo. Para cada combinação, as métricas Precisão, *Recall* e F1 foram calculadas por série e, em seguida, agregadas como médias sobre o conjunto de séries temporais avaliadas. A Tabela IV.3 resume os resultados ao apresentar, para cada método, as combinações que alcançaram o melhor desempenho médio em cada métrica de avaliação experimental, incluindo configurações estatisticamente equivalentes ao melhor resultado observado.

Observa-se que alguns métodos são representados por menos de três linhas na Tabela IV.3. Isso ocorre quando uma mesma combinação de critérios concentra os melhores desempenhos do método em mais de uma métrica. Esse comportamento é verificado em FBIAD, LSTM, ELM e Conv1D, cujas configurações mais competitivas em Precisão e F1 coincidem, restando apenas uma entrada adicional associada ao melhor desempenho em *Recall*.

Tabela IV.3: Desempenho no Conjunto de Dados Yahoo

Método	**Critérios**	**Precisão**	***Recall***	**F1**	**Método**	**Critérios**	**Precisão**	***Recall***	**F1**
REMD	ABS-RT-F	0,714	0,482	0,643	**LSTM**	ABS-GD-F	0,456	0,578	0,457
	ABS-RT-H	0,708	0,482	0,691		SQD-BP-F	0,039	0,657	0,070
	ABS-BP-H	0,133	0,629	0,245					
EMD	SQD-RT-H	0,205	0,092	0,103	**ELM**	ABS-GD-F	0,414	0,581	0,447
	SQD-RT-F	0,155	0,130	0,131		SQD-BP-F	0,039	0,653	0,069
	ABS-RT-H	0,138	0,170	0,112					
FBIAD	ABS-GD-H	0,582	0,523	0,597	**Conv1D**	SQD-RT-H	**0,785**	0,474	0,524
	ABS-RT-H	0,378	0,626	0,345		SQD-BP-F	0,038	0,599	0,066
ARIMA	ABS-RT-F	0,632	0,422	0,685	**SVM**	ABS-RT-F	0,630	0,432	0,470
	ABS-RT-H	0,526	0,361	**0,694**		SQD-GD-F	0,615	0,636	0,576
	ABS-BP-F	0,141	**0,758**	0,235		ABS-BP-F	0,111	0,701	0,172

Valores em negrito indicam o melhor desempenho absoluto por métrica entre todos os métodos. Valores sublinhados indicam desempenho estatisticamente equivalente ao melhor, com magnitude desprezível ou pequena.

A Tabela IV.4 evidencia que a composição do núcleo competitivo variou conforme a métrica considerada. Em Precisão e *Recall*, observa-se que apenas uma combinação atinge o melhor desempenho por métrica, sem que outras configurações apresentem equivalência estatística ao melhor resultado observado. Em contraste, o F1 apresenta um núcleo competitivo mais amplo, no qual diferentes combinações alcançam desempenhos estatisticamente equivalentes ao melhor, com dife-

renças de magnitude desprezível, tanto em comparações intra-método quanto inter-método. Esse comportamento indica que, em métricas de avaliação de desempenho, métodos distintos de DA atingem níveis comparáveis de desempenho quando operam sob combinações semelhantes de critérios de detecção.

Tabela IV.4: Caracterização do núcleo competitivo por métrica no conjunto Yahoo.

Métrica	**Combinações no Núcleo Competitivo**	**Magnitude**
Precisão	Conv1D: SQD-RT-H	–
Recall	ARIMA: ABS-BP-F	–
F1	ARIMA: ABS-RT-H vs. (ARIMA: ABS-RT-F)	Desprezível
	ARIMA: ABS-RT-H vs. (REMD: ABS-RT-H)	Desprezível

A tabela apresenta o núcleo competitivo por métrica, destacando apenas comparações estatisticamente equivalentes ao melhor resultado observado. A coluna "Magnitude" indica a escala qualitativa da diferença; "–" denota ausência de equivalência estatística.

No que se refere às Medidas de Desvio, a comparação entre ABS e SQD foi avaliada por meio do teste não paramétrico de Wilcoxon. A leitura dos resultados por método indica que ambas as medidas sustentam combinações competitivas no conjunto Yahoo, ainda que com diferentes padrões de preferência conforme o método de DA. Na Tabela IV.3, observa-se a presença recorrente de ABS em métodos como REMD, FBIAD e ARIMA, enquanto SQD aparece de forma consistente em combinações associadas a EMD, LSTM, ELM, Conv1D e SVM.

Esse comportamento está alinhado com os resultados da análise geral da Seção IV.2, conforme evidenciado na Tabela IV.2, na qual as diferenças entre ABS e SQD não se mostraram estatisticamente significativas, com magnitudes classificadas como desprezíveis ou pequenas. No contexto do conjunto Yahoo, a Medida de Desvio contribui como um mecanismo de ajuste fino, modulando o comportamento dos métodos sem, de forma isolada, redefinir os regimes de DA.

No que se refere ao Critério de Filtro, a análise considera a atuação dos filtros BP, GD e RT no contexto de cada método de DA, avaliados por meio do teste de Kruskal–Wallis. A leitura dos resultados por método evidencia que diferentes abordagens tendem a se beneficiar de configurações distintas de filtragem, refletindo preferências condicionadas à forma como cada método modela e explora o comportamento das séries temporais.

Observa-se, na Tabela IV.3, que filtros mais conservadores, como GD e RT, aparecem de forma recorrente em combinações competitivas associadas a métodos tradicionais e baseados em modelos estatísticos, como REMD, EMD, FBIAD, ARIMA e SVM. Nessas abordagens, a filtragem mais restritiva atua como mecanismo de ajuste fino ao controlar a incidência de falsos positivos, em consonância com escores de anomalia sensíveis a variações locais da série temporal.

Em contraste, o filtro BP surge com maior frequência em combinações competitivas associadas a métodos baseados em aprendizado, como LSTM, ELM e Conv1D, bem como em configurações

específicas do ARIMA, especialmente quando a métrica priorizada é o *Recall*. Nesse caso, a filtragem mais permissiva favorece a retenção de candidatos, explorando a maior capacidade desses métodos em absorver variabilidade e capturar padrões complexos.

Do ponto de vista estatístico, o teste de Kruskal–Wallis apontou diferenças significativas associadas ao Critério de Filtro como critério interno de DA, especialmente para Precisão ($p < 0{,}01$), com tamanho de efeito moderado (epsilon^2). Esse resultado indica que a filtragem participa do refinamento do processo de DA, promovendo uma calibração do compromisso entre sensibilidade e seletividade em função da métrica priorizada. No conjunto Yahoo, essa atuação ocorre de forma integrada aos métodos e aos demais critérios internos, caracterizando o filtro como um elemento de aprimoramento incremental do comportamento de detecção, sem redefinir isoladamente os regimes de DA.

No que se refere à Seleção de Candidatos, a análise considera a atuação das estratégias F e H no contexto de cada método de DA, avaliadas por meio do teste não paramétrico de Wilcoxon. A leitura dos resultados por método indica que ambas as estratégias sustentam combinações competitivas no conjunto Yahoo, com padrões de utilização condicionados à forma como cada abordagem estrutura e interpreta seus escores de anomalia.

Observa-se, na Tabela IV.3, que a estratégia F aparece associada a configurações específicas de métodos como ARIMA, REMD, LSTM e ELM, sobretudo quando combinada a filtros mais permissivos e quando a métrica priorizada é o *Recall*. Nesses casos, a seleção do primeiro ponto excedente favorece a detecção antecipada de eventos anômalos dentro do conjunto de candidatos previamente filtrado.

Em contrapartida, a estratégia H se manifesta de forma recorrente em combinações competitivas associadas a métodos orientados por escores contínuos de anomalia, como REMD, EMD, FBIAD, SVM e Conv1D, bem como em configurações do ARIMA. Nessas abordagens, a priorização do ponto de maior intensidade de desvio atua como um mecanismo complementar de refinamento, favorecendo decisões mais robustas dentro do conjunto de candidatos já definido pelas etapas anteriores do processo de DA.

Do ponto de vista estatístico, os testes não indicaram diferenças significativas entre as estratégias F e H nas métricas avaliadas, com magnitudes classificadas como pequenas ou desprezíveis. Em consonância com os resultados agregados na Tabela IV.1, a Seleção de Candidatos atua como um critério de refinamento final do processo de DA no conjunto Yahoo, promovendo ajustes marginais no balanço entre FP e FN dentro das configurações já estabelecidas pelas escolhas de Medida de Desvio e Critério de Filtro.

De forma integrada, a análise do conjunto Yahoo indicou que o desempenho observado resulta da articulação entre os métodos de DA e os critérios internos de detecção, nos quais a Medida de

Desvio, o Critério de Filtro e a Seleção de Candidatos contribuíram de maneira complementar para o ajuste fino do processo de identificação de anomalias. Cada decisão atua aprimorando aspectos específicos do comportamento de detecção, sem redefinir isoladamente os regimes impostos pelo método de DA.

IV.4 Análise Detalhada do Conjunto de Dados: GECCO

Esta seção apresenta a análise detalhada dos experimentos conduzidos no conjunto de dados Gecco, oriundo do Desafio GECCO 2018 [Moritz et al., 2018]. O conjunto é composto por séries temporais multivariadas que descrevem a composição físico-química da água potável, a partir de nove variáveis monitoradas com resolução temporal de um minuto. Esse contexto impõe desafios adicionais aos métodos de DA, uma vez que as anomalias tendem a se manifestar de forma mais sutil, distribuída ao longo do tempo e menos pontual, exigindo maior sensibilidade a desvios locais e menor tolerância a flutuações inerentes ao processo.

Para os experimentos desta dissertação, foi selecionado um subconjunto contendo aproximadamente 1 500 observações, correspondente a um dia completo de monitoramento, incluindo 72 anomalias rotuladas. Esse recorte forneceu uma base de referência consistente para a avaliação quantitativa do desempenho dos métodos, permitindo examinar como os critérios internos de detecção se comportam em um domínio de monitoramento ambiental de alta granularidade temporal.

De forma análoga ao procedimento adotado na análise do conjunto Yahoo (Seção IV.3), a avaliação no Gecco explora a interação entre os métodos de DA e os critérios internos de detecção, Medida de Desvio, Critério de Filtro e Seleção de Candidatos. Para cada método, são consideradas todas as combinações possíveis desses critérios, e as métricas de Precisão, *Recall* e F1 foram calculadas por série e, posteriormente, agregadas como médias sobre o conjunto avaliado.

A Tabela IV.5 sintetiza os resultados ao apresentar, para cada método, as combinações de critérios que alcançaram o melhor desempenho médio em cada métrica de avaliação, restringindo-se às configurações mais representativas observadas no conjunto Gecco. Diferentemente do observado no Yahoo, não se verificou a formação de um núcleo competitivo amplo ou estável entre métodos distintos, sendo cada métrica dominada por combinações pontuais e específicas.

Observa-se ainda que alguns métodos são representados por menos de três linhas na Tabela IV.5. Esse comportamento ocorre quando uma mesma combinação de critérios concentra os melhores desempenhos do método em mais de uma métrica. Tal padrão é verificado, por exemplo, em EMD, LSTM e ELM, cujas configurações mais competitivas se sobrepõem parcialmente entre Precisão, *Recall* e F1. Essa característica reflete a maior sensibilidade do conjunto Gecco às escolhas internas de critérios e antecipa a análise detalhada por critério apresentada a seguir.

No que se refere às Medidas de Desvio, a comparação entre ABS e SQD no conjunto Gecco,

Tabela IV.5: Desempenho no Conjunto de Dados Gecco.

Método	Critérios	Precisão	*Recall*	F1	Método	Critérios	Precisão	*Recall*	F1
REMD	SQD-BP-F	0,072	0,085	0,076	**LSTM**	SQD-BP-F	0,073	0,073	0,071
	ABS-GD-H	0,160	0,032	0,064		SQD-RT-H	0,207	0,017	0,029
	SQD-RT-H	0,127	0,012	0,101					
EMD	ABS-RT-F	0,125	**0,102**	0,084	**ELM**	SQD-BP-F	0,070	0,090	0,076
	SQD-RT-H	**0,306**	0,014	0,026		ABS-GD-H	0,178	0,043	0,066
FBIAD	SQD-BP-F	0,079	0,076	0,070	**Conv1D**	SQD-BP-H	0,060	0,090	0,067
	ABS-GD-F	0,269	0,054	**0,160**		ABS-BP-F	0,138	0,076	0,090
	ABS-GD-H	0,276	0,054	0,134		SQD-RT-H	0,206	0,022	0,032
ARIMA	SQD-BP-F	0,078	0,093	0,081	**SVM**	SQD-BP-F	0,051	0,079	0,060
	ABS-BP-F	0,110	0,074	0,089		ABS-BP-F	0,080	0,063	0,062
	SQD-RT-H	0,170	0,017	0,080		ABS-GD-H	0,152	0,035	0,049

Valores em negrito indicam o melhor desempenho dentro de cada grupo de métodos.

avaliada por meio do teste não paramétrico de Wilcoxon, não indicou diferenças estatisticamente significativas nas métricas analisadas. Ainda assim, a leitura dos resultados por método revela padrões distintos de preferência conforme a métrica priorizada, evidenciando o papel desse critério como mecanismo de ajuste fino do processo de DA.

Observa-se, na Tabela IV.5, que o EMD apresenta seu melhor resultado em Precisão sob a configuração SQD-RT-H, enquanto o maior *Recall* ocorre com ABS-RT-F. Essa alternância é coerente com o efeito esperado das medidas de desvio: SQD, ao amplificar desvios mais intensos, tende a favorecer decisões mais seletivas, ao passo que ABS promove maior sensibilidade a variações locais, ampliando a retenção de anomalias candidatas. Em termos de F1, a configuração baseada em ABS-F apresenta desempenho superior às alternativas com H, indicando que o ganho em Precisão associado à seleção mais restritiva não se traduz, nesse caso, em melhor equilíbrio global entre FP e FN.

Comportamento análogo é observado no FBIAD, cujas melhores configurações em F1 e Precisão também se apoiam em ABS, sugerindo que medidas menos sensíveis a *outliers* extremos favorecem um balanço mais estável no domínio do Gecco. Em contraste, métodos como REMD, LSTM e ELM apresentam combinações competitivas distribuídas entre ABS e SQD, sem um padrão dominante consistente, reforçando o caráter modulador da Medida de Desvio, sem redefinir isoladamente os regimes de detecção.

No que se refere ao Critério de Filtro, a análise estatística por meio do teste de Kruskal–Wallis apontou diferenças significativas, especialmente para Precisão ($p < 0{,}001$) e *Recall* ($p < 0{,}01$), com tamanhos de efeito classificados como moderados. Esses resultados indicam que a filtragem exerce influência substancial no desempenho dos métodos no conjunto Gecco.

A leitura por método evidencia que filtros mais conservadores, como GD e RT, aparecem de forma recorrente em configurações associadas a maiores valores de Precisão e F1, como observado no

EMD, FBIAD e em configurações específicas do ARIMA e do SVM. Nessas abordagens, a filtragem mais restritiva contribui para a contenção de falsos positivos em um cenário marcado por elevada variabilidade e ruído frequente.

Em contraste, o filtro BP surge com maior frequência em combinações associadas a maiores valores de *Recall*, particularmente em métodos como REMD, LSTM, ELM e Conv1D. Nesse caso, a filtragem mais permissiva favorece a retenção de anomalias candidatas, ampliando a sensibilidade do processo de detecção, ainda que com impacto negativo sobre a seletividade.

No que se refere à Seleção de Candidatos, a comparação entre as estratégias F e H, conduzida por meio do teste de Wilcoxon, não indicou diferenças estatisticamente significativas nas métricas avaliadas. Ainda assim, a leitura dos resultados por método revela padrões consistentes com os efeitos esperados desse critério como etapa final de refinamento.

Observa-se que a estratégia F aparece associada a maiores valores de *Recall* em métodos como EMD, REMD e FBIAD, especialmente quando combinada a filtros mais permissivos. Essa estratégia favorece a detecção antecipada de eventos anômalos dentro do conjunto de candidatos previamente filtrado. Por outro lado, a estratégia H manifesta-se com maior frequência em configurações associadas a maiores valores de Precisão, como observado no EMD e em configurações do SVM e do ARIMA, refletindo a priorização de pontos com maior intensidade de desvio.

Em conjunto, a Seleção de Candidatos atua no Gecco como um mecanismo de ajuste final, promovendo deslocamentos marginais no balanço entre FP e FN, sem redefinir isoladamente o comportamento dos métodos. Assim como observado para as demais etapas, sua atuação é condicionada pelas escolhas anteriores de Medida de Desvio e Critério de Filtro, reforçando a caracterização dos critérios internos como mecanismos de ajuste fino do processo de DA.

IV.5 Análise Detalhada do Conjunto de Dados: NAB

Esta seção apresenta a análise detalhada dos experimentos conduzidos no conjunto de dados NAB, amplamente adotado como referência para a avaliação de métodos de DA em séries temporais rotuladas. O conjunto reúne séries reais e sintéticas, representativas de cenários de monitoramento em ambientes de computação em nuvem, nos quais anomalias podem se manifestar tanto de forma abrupta quanto progressiva. Esse contexto impõe desafios específicos à detecção, particularmente no equilíbrio entre sensibilidade à detecção precoce e controle de alarmes espúrios.

A Tabela IV.6 apresenta, para cada método avaliado, as combinações de critérios que alcançaram o melhor desempenho médio em Precisão, *Recall* e F1. Assim como observado no conjunto Gecco (Seção IV.4), os resultados no NAB não indicaram a formação de um núcleo competitivo amplo ou estável entre métodos distintos, sendo os melhores resultados concentrados em combinações pontuais e fortemente dependentes das escolhas internas de critérios.

Tabela IV.6: Desempenho no Conjunto de Dados: NAB

Método	Critérios	Precisão	*Recall*	F1	Método	Critérios	Precisão	*Recall*	F1
REMD	ABS-GD-F	0,015	0,500	0,059	**LSTM**	SQD-RT-F	0,118	0,066	0,083
	SQD-BP-H	0,003	**1,000**	0,006		ABS-BP-F	0,015	0,489	0,028
EMD	SQD-RT-F	0,000	0,000	0,000	**ELM**	ABS-RT-H	0,438	0,198	0,263
						SQD-RT-H	**0,500**	0,182	0,263
						ABS-BP-H	0,023	0,604	0,042
FBIAD	ABS-RT-H	0,006	0,500	0,023	**Conv1D**	SQD-RT-H	0,389	0,134	0,196
	ABS-RT-F	0,006	0,500	0,023		ABS-BP-H	0,010	0,491	0,020
ARIMA	ABS-RT-H	0,300	0,108	**0,522**	**SVM**	ABS-RT-H	0,412	0,172	0,233
	SQD-BP-H	0,021	0,675	0,057		SQD-RT-H	0,471	0,157	0,231
						SQD-BP-H	0,011	0,607	0,022

Observa-se que alguns métodos são representados por menos de três linhas na Tabela IV.6. Isso ocorre quando uma mesma combinação de critérios concentra os melhores desempenhos do método em mais de uma métrica, como verificado em REMD, FBIAD, ARIMA, LSTM e Conv1D, cujas configurações mais competitivas em Precisão, *Recall* ou F1 coincidem parcialmente.

No caso do EMD, não foram apresentadas combinações associadas às métricas de Precisão, *Recall* e F1, uma vez que todas as configurações avaliadas resultaram em valores nulos nessas métricas. Ainda assim, o método é reportado por completude experimental, registrando-se a combinação SQD-RT-F, a qual não alterou as métricas analisadas nesta seção.

No que se refere às Medidas de Desvio, a comparação entre ABS e SQD, avaliada por meio do teste de Wilcoxon, não indicou diferenças estatisticamente significativas nas métricas avaliadas. Ainda assim, a leitura dos resultados por método evidencia padrões consistentes de comportamento. Em métodos como REMD e FBIAD, observa-se que configurações baseadas em SQD tendem a favorecer valores elevados de *Recall*, refletindo maior sensibilidade à detecção de eventos anômalos, ainda que com prejuízo à Precisão. Em contraste, métodos como ARIMA e SVM apresentam melhores valores de F1 quando combinados a ABS, indicando um equilíbrio mais estável entre detecção correta e contenção de falsos positivos.

Esse comportamento está alinhado à análise geral apresentada na Seção IV.2, na qual as diferenças entre ABS e SQD se mostram de magnitude desprezível ou pequena (Tabela IV.2). No NAB, a Medida de Desvio atua, portanto, como um mecanismo de ajuste fino, modulando a sensibilidade dos métodos a diferentes padrões de desvio, sem redefinir de forma isolada os regimes de detecção.

No que se refere ao Critério de Filtro, a comparação entre BP, GD e RT, avaliada por meio do teste de Kruskal–Wallis, não indicou diferenças estatisticamente significativas para Precisão e *Recall*, embora tenha indicado diferença estatisticamente significativa para F1. A leitura dos resultados por método sugere, contudo, padrões recorrentes na forma como BP, GD e RT interagem com diferentes abordagens de DA no conjunto NAB.

Em métodos como REMD, observa-se que o maior *Recall* do método ocorre sob BP, enquanto uma alternativa com GD apresenta menor sensibilidade, mas maior controle do conjunto de candidatos. Em FBIAD, as melhores combinações do método são obtidas com RT, mantendo *Recall* constante nas duas estratégias de seleção, o que sugere que, nessa abordagem, a filtragem atua sobretudo estabilizando o conjunto de candidatos em um regime mais conservador.

Em métodos como ARIMA, SVM e nas arquiteturas baseadas em aprendizado, RT aparece com maior frequência nas combinações associadas a maior seletividade, ao passo que BP reaparece em configurações voltadas a maior *Recall*. No caso do ELM, por exemplo, as configurações mais competitivas concentram-se em RT, enquanto uma alternativa com BP eleva *Recall* com perda de seletividade, ilustrando o deslocamento do compromisso entre sensibilidade e Precisão induzido pela filtragem.

No que se refere à Seleção de Candidatos, a comparação entre as estratégias F e H, avaliada por meio do teste de Wilcoxon, não indicou diferenças estatisticamente significativas nas métricas de Precisão, *Recall* e F1. Ainda assim, a leitura dos resultados por método revela padrões consistentes com o papel desse critério como etapa final de refinamento do processo de DA no conjunto NAB.

Em métodos como REMD e FBIAD, a estratégia F aparece associada a configurações com maior *Recall*, refletindo a priorização da detecção antecipada de eventos anômalos dentro do conjunto de candidatos previamente filtrado. Nessas abordagens, a seleção do primeiro ponto excedente favorece maior sensibilidade, ainda que com impacto limitado sobre a seletividade global.

Em contraste, a estratégia H se manifesta de forma recorrente em combinações associadas a maior equilíbrio entre Precisão e *Recall*, como observado em ARIMA, SVM e em arquiteturas baseadas em aprendizado. Nesses casos, a priorização do ponto com maior intensidade de desvio atua como um mecanismo complementar de refinamento, favorecendo decisões mais seletivas sem alterar substancialmente o conjunto de candidatos herdado das etapas anteriores.

De forma integrada, a análise do conjunto NAB indica que o desempenho observado resulta da articulação entre os métodos de DA e do ajuste fino dos critérios de detecção, os quais contribuem de forma complementar para aprimorar o processo de identificação de anomalias em séries temporais.

IV.6 Análise Aprofundada da Série A3-TS43

Esta seção apresenta uma análise aprofundada dos resultados de detecção de anomalias no conjunto de dados Yahoo, com foco na série temporal *A3Benchmark-TS43*. Essa série pertence ao subconjunto sintético A3 e é caracterizada por tendência crescente, forte componente sazonal e eventos anômalos injetados de forma pontual, configurando um cenário intrinsecamente não estacionário. Nesse contexto, variações abruptas ocorrem sobre uma dinâmica global não constante, o que impõe desafios adicionais à detecção de anomalias e torna particularmente relevante a escolha

dos métodos de DA e dos critérios internos que estruturam o processo de detecção, já analisados de forma agregada na Seção IV.2.

A Figura IV.1 apresenta a detecção de anomalias ao longo da série, evidenciando padrões concretos de acerto e erro para diferentes combinações de método e critérios de DA. As anomalias de referência (*ground truth*) são indicadas em azul-escuro; os Verdadeiros Positivos (VP) em verde; os Falsos Positivos (FP) em vermelho; e os Falsos Negativos (FN) em azul-claro. Essa visualização permite observar, em nível de série individual, como os efeitos médios discutidos na Seção IV.2 se manifestam localmente ao longo do tempo.

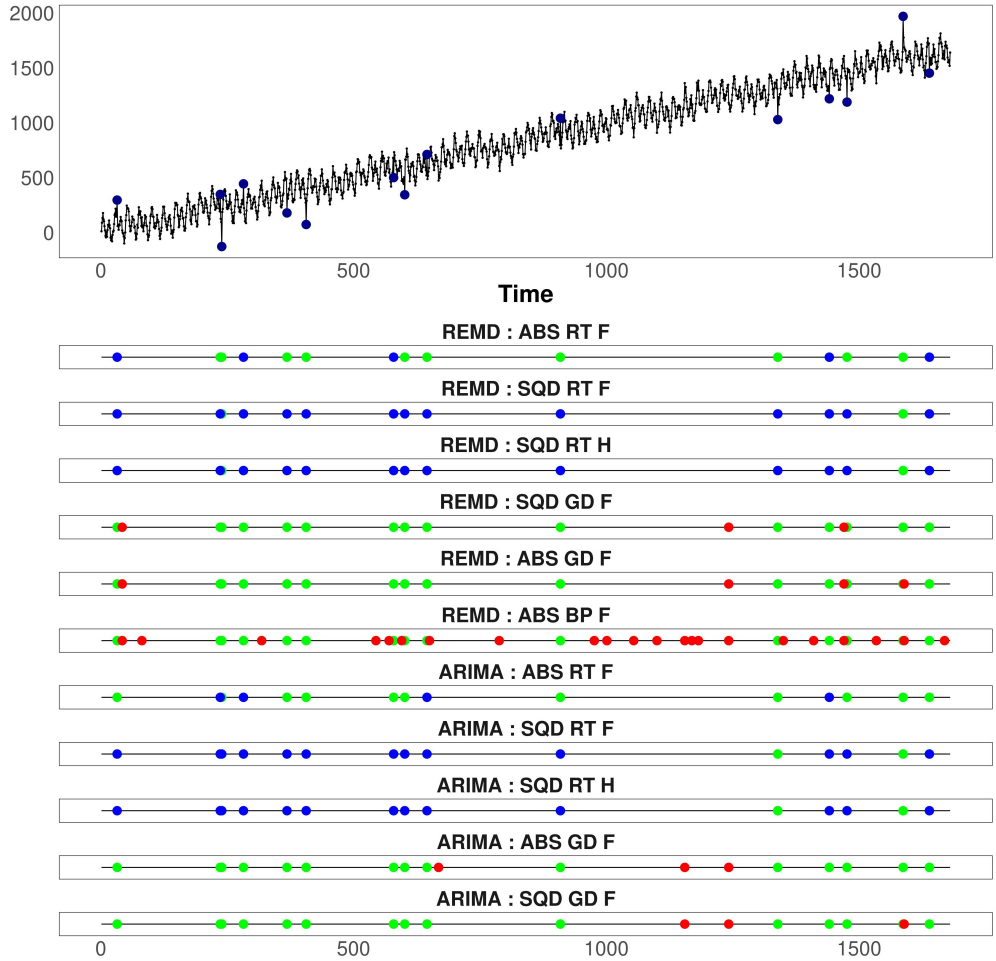


Figura IV.1: Detecção de anomalias na série *A3Benchmark-TS43* (Yahoo).

A Figura IV.1 evidencia que o impacto da Medida de Desvio sobre o comportamento da detecção se manifesta de forma condicionada ao método de DA adotado, explicitando diferentes *trade-offs* entre sensibilidade e seletividade. No caso do ARIMA, combinações baseadas em ABS, quando associadas a critérios de filtro mais seletivos, tendem a concentrar as detecções em intervalos correspondentes a desvios abruptos, reduzindo a dispersão de FP ao longo da série. Esse comportamento é coerente com a análise global (Seção IV.2), na qual ABS apresentou desempenho médio superior a SQD. Em contraste, configurações com SQD acentuam desvios de maior magnitude, aumentando

a sensibilidade à identificação de eventos extremos e favorecendo padrões de detecção compatíveis com filtros de natureza gaussiana, ao custo de maior incidência de alarmes espúrios.

Para o método REMD, cuja modelagem se baseia em decomposição empírica e adaptação local, a interação entre Medida de Desvio e estrutura do modelo torna-se ainda mais pronunciada. Em regimes mais seletivos, o uso de ABS promove maior alinhamento entre as detecções e as anomalias de referência, indicando maior robustez frente a flutuações locais sobre a tendência não estacionária da série. Por outro lado, a adoção de SQD enfatiza desvios mais acentuados, favorecendo a identificação de picos isolados, mas também induzindo maior dispersão de detecções em determinados regimes, refletindo, em escala local, os *trade-offs* observados na avaliação global dos critérios.

No que diz respeito ao Critério de Filtro, a comparação entre BP, GD e RT, para ambos os métodos, revela regimes de detecção claramente distintos. Filtros do tipo BP produzem padrões caracterizados por elevada sensibilidade, com detecções amplamente distribuídas ao longo da série e acúmulo de FP, refletindo um regime no qual o *Recall* é priorizado em detrimento da Precisão. Esse comportamento é consistente com os resultados da Seção IV.2, que apontaram o filtro BP como associado a maiores níveis médios de *Recall*. Em oposição, o filtro RT induz um regime mais conservador, no qual as detecções se concentram em regiões de maior desvio em relação ao comportamento esperado, privilegiando a seletividade e reduzindo a ocorrência de alarmes espúrios, em consonância com seu melhor desempenho médio em Precisão na análise geral. O filtro GD estabelece um regime intermediário, no qual se observa um compromisso mais estável entre sensibilidade e controle de erros, configurando um equilíbrio mais consistente entre *Recall* e Precisão, também identificado no núcleo competitivo da análise global.

A análise por método reforça essas distinções. No ARIMA, cuja modelagem assume uma estrutura temporal paramétrica, critérios de filtro mais permissivos amplificam a propagação de resíduos locais, enquanto estratégias mais seletivas restringem as detecções a eventos de maior discrepância em relação ao modelo ajustado. No REMD, a interação com o critério de filtro evidencia maior sensibilidade à forma como os componentes são agregados e avaliados, resultando em padrões de detecção fortemente dependentes da combinação entre filtro e Medida de Desvio, reforçando o papel desses critérios como mecanismos de ajuste fino adaptativos, conforme discutido de forma agregada na Seção IV.2.

Quanto à Seleção de Candidatos, a análise desta subseção concentra-se nas configurações baseadas na estratégia F, conforme ilustrado na Figura IV.1. Nesse recorte, o critério de seleção atua como mecanismo de consolidação das detecções produzidas pelos métodos e pelos critérios anteriores, influenciando a organização temporal de VP, FP e FN, sem redefinir os regimes estabelecidos pela interação entre método de DA e Critério de Filtro. Tal papel complementar é consistente com a vantagem média observada para F em relação a H na análise global (Seção IV.2), ainda que com

efeitos de magnitude limitada.

De forma integrada, a análise da série *A3Benchmark-TS43* evidencia que os padrões de detecção observados na Figura IV.1 decorrem da articulação entre os critérios de DA e os métodos de detecção, na qual Medida de Desvio, Critério de Filtro e Seleção de Candidatos desempenham papéis de ajuste fino para o aprimoramento do processo de identificação de anomalias em séries temporais.

Capítulo V Conclusão

Este trabalho examina o impacto do ajuste fino dos critérios de detecção no desempenho de métodos de DA em séries temporais, com foco nos critérios de Medida de Desvio, Filtro e Seleção de Candidatos. Partindo do reconhecimento de que decisões metodológicas implícitas nesses critérios influenciam diretamente o equilíbrio entre FP e FN, frequentemente resultando em alertas espúrios ou na omissão de eventos relevantes, o estudo adota uma abordagem analítica que prioriza a análise sistemática dos critérios responsáveis pela identificação de anomalias, em detrimento de comparações exclusivamente baseadas nos métodos de DA.

Para responder às questões de pesquisa propostas, foi empregado um *framework* modular de *benchmark*, implementado no *Harbinger*, que permite combinar e avaliar, de forma controlada, diferentes critérios de detecção em múltiplos métodos e conjuntos de dados. Essa abordagem possibilita isolar o efeito individual e combinado dos critérios, evitando interpretações baseadas apenas no comportamento agregado dos métodos, tratados como caixas fechadas.

A análise experimental conduzida neste trabalho responde às questões de pesquisa apresentadas no Capítulo I, fornecendo a base empírica para a avaliação do impacto dos critérios de detecção no desempenho da DA.

No que se refere às Medidas de Desvio, a análise experimental responde à Q1 ao indicar que ABS e SQD apresentam comportamentos amplamente equivalentes no desempenho médio da detecção, com diferenças de magnitude desprezível ou pequena nas métricas de Precisão, *Recall* e F1. Embora variações pontuais sejam observadas em conjuntos de dados específicos, os resultados mostram que a escolha do Critério de Medida de Desvio influencia a distribuição dos resíduos, mas exerce impacto limitado, quando considerada de forma isolada, sobre os regimes globais de detecção, atuando de modo complementar aos demais critérios.

Em relação aos Critérios de Filtro, a análise experimental responde à Q2 ao indicar que esse critério constitui o principal fator de diferenciação entre regimes de detecção. Filtros como BP, GD e RT conduzem a compromissos distintos entre sensibilidade e seletividade, refletindo-se de maneira consistente nas métricas avaliadas. Enquanto filtros mais permissivos favorecem o *Recall*, filtros mais conservadores priorizam a Precisão, e configurações intermediárias tendem a produzir melhor equilíbrio global, conforme evidenciado pelo F1. Esses resultados reforçam que a definição do limiar de filtragem exerce papel central no comportamento da DA.

Quanto à Seleção de Candidatos, a análise experimental responde à Q3 ao indicar que F e H atuam predominantemente como mecanismos de refinamento das decisões já induzidas pelo Critério de Filtro. As diferenças observadas entre essas estratégias apresentam magnitudes pequenas ou desprezíveis no desempenho médio, indicando que sua influência é mais sutil e dependente do contexto, sem redefinir, de forma isolada, os regimes de detecção estabelecidos nas etapas anteriores do *pipeline*.

De forma integrada, este trabalho demonstra que o desempenho da DA não é determinado exclusivamente pelo método adotado, mas pela articulação entre os critérios de DA e os métodos de identificação de anomalias. A explicitação e o ajuste fino desses critérios ampliam a interpretabilidade dos resultados, permitem maior adaptação às características específicas das séries temporais e contribuem para avaliações experimentais mais reprodutíveis e transparentes. Nesse sentido, o ajuste fino dos critérios de DA emerge como um mecanismo eficaz para aprimorar a identificação de anomalias, frequentemente com impacto comparável ou superior à simples substituição do método subjacente, sem pretensão de universalidade e sempre condicionado às propriedades dos dados analisados.

Como principais contribuições, esta dissertação apresenta uma análise sistemática do papel dos critérios de DA, um *framework* modular para avaliação comparativa e evidências empíricas que qualificam os critérios de DA como mecanismos adaptativos no processo de identificação de anomalias.

Por fim, são apontados como trabalhos futuros a extensão dessa abordagem para ambientes em tempo real e o desenvolvimento de estratégias de ajuste automático dos critérios com base em propriedades observáveis das séries temporais, bem como a integração de mecanismos adaptativos de ajuste fino em *frameworks* de DA. Como desdobramentos dessa linha, destaca-se a investigação do papel do sinal do desvio, da adoção de limiares assimétricos, da persistência temporal dos desvios e da correlação entre variáveis, visando aprimorar a robustez da detecção e reduzir a ocorrência de alarmes espúrios em diferentes domínios de aplicação.

Referências

Aggarwal, C. C. *Outlier Analysis*. Springer International Publishing, 2016.

Ahmad, S., Lavin, A., Purdy, S., and Agha, Z. Unsupervised real-time anomaly detection for streaming data. *Neurocomputing*, 262:134 – 147, 2017. https://doi.org/10.1016/j.neucom.2017.04.070

Ahmed, M., Mahmood, A. N., and Islam, M. R. A survey of anomaly detection techniques in financial domain. *Future Generation Computer Systems*, 55:278 – 288, 2016. https://doi.org/10.1016/j.future.2015.01.001

Ariyaluran Habeeb, R. A., Nasaruddin, F., Gani, A., Targio Hashem, I. A., Ahmed, E., and Imran, M. Real-time big data processing for anomaly detection: A Survey. *International Journal of Information Management*, 45:289 – 307, 2019. https://doi.org/10.1016/j.ijinfomgt.2018.08.006

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. https://doi.org/10.1145/3444690

Boniol, P., Liu, Q., Huang, M., Palpanas, T., and Paparrizos, J. Dive into time-series anomaly detection: A decade review, 2024.

Cauteruccio, F., Cinelli, L., Corradini, E., Terracina, G., Ursino, D., Virgili, L., Savaglio, C., Liotta, A., and Fortino, G. A framework for anomaly detection and classification in Multiple IoT scenarios. *Future Generation Computer Systems*, 114:322 – 335, 2021. https://doi.org/10.1016/j.future.2020.08.010

Chandola, V., Banerjee, A., and Kumar, V. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3), 2009. https://doi.org/10.1145/1541880.1541882

Cook, A. A., Misirli, G., and Fan, Z. Anomaly Detection for IoT Time-Series Data: A Survey. *IEEE Internet of Things Journal*, 7(7):6481 – 6494, 2020. https://doi.org/10.1109/JIOT.2019.2958185

Dixit, P., Bhattacharya, P., Tanwar, S., and Gupta, R. Anomaly detection in autonomous electric vehicles using AI techniques: A comprehensive survey. *Expert Systems*, 39(5), 2022. https://doi.org/10.1111/exsy.12754

Dunn, O. J. Multiple comparisons using rank sums. *Technometrics*, 6(3):241–252, 1964.

Friedman, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200):675–701, 1937.

Geiger, A., Liu, Z., Alnegheimish, S., Cuesta-Infante, A., and Veeramachaneni, K. Tadgan: Time series anomaly detection using generative adversarial networks. *Proceedings of the 2020 IEEE International Conference on Big Data (Big Data)*, pages 1–12, 2020. https://doi.org/10.1109/BigData50022.2020.9378497

Gorard, S. Revisiting A 90-year-old debate: The advantages of the mean deviation. *British Journal of Educational Studies*, 53(4):417 – 430, 2005. https://doi.org/10.1111/j.1467-8527.2005.00304.x

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

Hasani, Z. Robust anomaly detection algorithms for real-time big data: Comparison of algorithms. In *2017 6th Mediterranean Conference on Embedded Computing, MECO 2017 - Including ECYPS 2017, Proceedings*, 2017. https://doi.org/10.1109/MECO.2017.7977130

Hodson, T. O. Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. *Geoscientific Model Development*, 15(14):5481 – 5487, 2022. https://doi.org/10.5194/gmd-15-5481-2022

Ibidunmoye, O., Hernández-Rodriguez, F., and Elmroth, E. Performance anomaly detection and bottleneck identification. *ACM Computing Surveys*, 48(1), 2015. https://doi.org/10.1145/2791120

Kruskal, W. H. and Wallis, W. A. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621, 1952.

Li, J., Izakian, H., Pedrycz, W., and Jamal, I. Clustering-based anomaly detection in multivariate time series data. *Applied Soft Computing*, 100, 2021. https://doi.org/10.1016/j.asoc.2020.106919

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 *2022 International Joint Conference on Neural Networks (IJCNN)*, volume 2022-July, pages 1–8, 2022. https://doi.org/10.1109/IJCNN55064.2022.9892088

Lima, J., Tavares, L. G., Pacitti, E., Ferreira, J. E., Santos, I., Siqueira, I. G., Carvalho, D., Porto, F., Coutinho, R., and Ogasawara, E. Online Event Detection in Streaming Time Series: Novel Metrics and Practical Insights. In *Proceedings of the International Joint Conference on Neural Networks*, pages 1–8, 2024. https://doi.org/10.1109/IJCNN60899.2024.10650809

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, 2003.*, volume 3, pages 1741–1745 vol.3, 2003. https://doi.org/10.1109/IJCNN.2003.1223670

Moritz, S., Rehbach, F., Chandrasekaran, S., Rebolledo, M., and Bartz-Beielstein, T. Gecco industrial challenge 2018 dataset: A water quality dataset for the 'internet of things: Online anomaly detection for drinking water quality' competition at the genetic and evolutionary computation conference 2018, kyoto, japan., 2018.

Ogasawara, E., Salles, R., Escobar, L., Baroni, L., Lima, J., and Porto, F. Online event detection for sensor data. In *XLII Ibero-Latin American Congress on Computational Methods in Engineering*, volume 3, pages 1–7, Rio de Janeiro, RJ, 2021.

Ogasawara, E., Salles, R., Porto, F., and Pacitti, E. *Event Detection in Time Series*. Springer, 2025 edition, 2025.

Pang, G., Shen, C., Cao, L., and Van Den Hengel, A. Deep Learning for Anomaly Detection: A Review. *ACM Computing Surveys*, 54(2), 2021. https://doi.org/10.1145/3439950

Paparrizos, J., Kang, Y., Boniol, P., Tsay, R. S., Palpanas, T., and Franklin, M. J. TSB-UAD: An End-to-End Benchmark Suite for Univariate Time-Series Anomaly Detection. *Proceedings of the VLDB Endowment*, 15:1697 – 1711, 2022. https://doi.org/10.14778/3529337.3529354

Pena, E. H. M., de Assis, M. V. O., and Proença, M. L. Anomaly detection using forecasting methods arima and hwds. In *2013 32nd International Conference of the Chilean Computer Science Society (SCCC)*, pages 63–66, 2013. https://doi.org/10.1109/SCCC.2013.18

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

Ren, H., Xu, B., Wang, Y., Yi, C., Huang, C., Kou, X., Xing, T., Yang, M., Tong, J., and Zhang, Q. Time-series anomaly detection service at Microsoft. In *Proceedings of the ACM SIGKDD*

International Conference on Knowledge Discovery and Data Mining, pages 3009 – 3017, 2019. https://doi.org/10.1145/3292500.3330680

Sackett, D. L., Rosenberg, W. M., Gray, J. M., Haynes, R. B., and Richardson, W. S. Evidence-based medicine: What it is and what it isn't. *BMJ*, 312(7023):71–72, 1996. https://doi.org/10.1136/bmj.312.7023.71

Salles, R., Escobar, L., Baroni, L., Zorrilla, R., Ziviani, A., Kreischer, V., Delicato, F., Pires, P., Maia, L., Coutinho, R., Assis, L., and Ogasawara, E. Um framework para integração e análise de métodos de detecção de eventos em séries temporais. In *Anais do Simpósio Brasileiro de Banco de Dados (SBBD)*. SBC, 2020.

Salles, R., Lima, J., Reis, M., Coutinho, R., Pacitti, E., Masseglia, F., Akbarinia, R., Chen, C., Garibaldi, J., Porto, F., and Ogasawara, E. SoftED: Metrics for soft evaluation of time series event detection. *Computers and Industrial Engineering*, 198, 2024. https://doi.org/10.1016/j.cie.2024.110728

Scharf, L. L. and Demeure, C. *Statistical Signal Processing: Detection, Estimation, and Time Series Analysis*. Addison-Wesley Publishing Company, 1991.

Sela, L., Arad, J., Housh, M., and Ostfeld, A. Event detection in water distribution systems from multivariate water quality time series. *Environmental Science & Technology*, 46:8212–8219, 2012. https://doi.org/10.1021/es3014024

Shapiro, S. S. and Wilk, M. B. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.

Sobrinho, E. P., Souza, J., Lima, J., Tavares, L., and Ogasawara, E. Fine-tuning detection criteria for enhancing anomaly detection in time series. In *Anais do XL Simpósio Brasileiro de Bancos de Dados (SBBD)*. Sociedade Brasileira de Computação, 2025.

Souza, J., Pãixao, 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 *Proceedings of the International Joint Conference on Neural Networks*, pages 1–8, 2024. https://doi.org/10.1109/IJCNN60899.2024.10651192

Souza, J. G. d. REMD: Um método híbrido para detecção de anomalias baseado em decomposição por modos empíricos e modelagem arima. Dissertação de mestrado, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (CEFET/RJ), Rio de Janeiro, RJ, 2024.

Talagala, P. D., Hyndman, R. J., Smith-Miles, K., Kandanaarachchi, S., and Muñoz, M. A. Anomaly Detection in Streaming Nonstationary Temporal Data. *Journal of Computational and Graphical Statistics*, 29(1):13 – 27, 2020. https://doi.org/10.1080/10618600.2019.1617160

Truong, C., Oudre, L., and Vayatis, N. Selective review of offline change point detection methods. *Signal Processing*, 167, 2020. https://doi.org/10.1016/j.sigpro.2019.107299

Wenig, P., Schmidl, S., and Papenbrock, T. TimeEval: A benchmarking toolkit for time series anomaly detection algorithms. *Proceedings of the VLDB Endowment*, 15(12):3678–3681, 2022. https://doi.org/10.14778/3554821.3554873

Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

Zamanzadeh Darban, Z., Webb, G. I., Pan, S., Aggarwal, C., and Salehi, M. Deep Learning for Time Series Anomaly Detection: A Survey. *ACM Computing Surveys*, 57(1):15:1–15:42, 2024. https://doi.org/10.1145/3691338

Zare Moayedi, H. and Masnadi-Shirazi, M. Arima model for network traffic prediction and anomaly detection. In *2008 International Symposium on Information Technology*, volume 4, pages 1–6, 2008. https://doi.org/10.1109/ITSIM.2008.4631947

Zhang, J. E., Wu, D., and Boulet, B. Time series anomaly detection for smart grids: A survey. In *2021 IEEE Electrical Power and Energy Conference (EPEC)*. IEEE, 2021. https://doi.org/10.1109/EPEC52095.2021.9621752

Zhang, M., Guo, J., Li, X., and Jin, R. Data-driven anomaly detection approach for time-series streaming data. *Sensors (Switzerland)*, 20(19):1 – 17, 2020. https://doi.org/10.3390/s20195646

Zhang, Y., Wang, L., and Huang, X. Comparative evaluation of deep learning architectures for time series anomaly detection. *Expert Systems with Applications*, 213:118849, 2023. https://doi.org/10.1016/j.eswa.2022.118849