CEFET/RJ
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO

ANÁLISE COMPARATIVA DE MÉTODOS PARA DETECÇÃO DE EVENTOS EM SÉRIES TEMPORAIS

Luciana Escobar Gonçalves Vignoli

Dissertação apresentada 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 título de mestre.

Orientador(a): Laura Silva de Assis
Coorientador(a): Eduardo Soares Ogasawara

Rio de Janeiro,
Fevereiro de 2021

Análise Comparativa de Métodos para Detecção de Eventos em Séries Temporais

Dissertação apresentada 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 título de mestre.

Luciana Escobar Gonçalves Vignoli

Banca Examinadora:

Laura Silva de Assis

Presidente, Professora. D.Sc. Laura Silva de Assis (CEFET/RJ)

Professor D.Sc. Eduardo Soares Ogasawara (CEFET/RJ)

Rafaelli de Carvalho Coutinho

Professora D.Sc. Rafaelli de Carvalho Coutinho (CEFET/RJ)

Professor D.Sc. Fábio André Machado Porto (LNCC)

Rio de Janeiro,

Fevereiro de 2021

Ficha catalográfica elaborada pela Biblioteca Central do CEFET/RJ

V686 Vignoli, Luciana Escobar Gonçalves
Analise comparativa de métodos para detecção de eventos em series temporais / Luciana Escobar Gonçalves Vignoli — 2021.
77f : il. color. , enc.

Dissertação (Mestrado) Centro Federal de Educação Tecnológica Celso Suckow da Fonseca , 2021.
Bibliografia : f. 70-77
Orientadora: Laura Silva de Assis
Coorientador: Eduardo Soares Ogasawara

1. Análise de séries temporais. 2. Análise espectral. 3. Algoritmos. 4. Processamento eletrônico de dados. I. Assis, Laura Silva de (Orient.). II. Ogasawara, Eduardo Soares (Coorient.). III. Título.

CDD 519.55

Elaborada pela bibliotecária Tania Mello – CRB/7 nº 5507/04

DEDICATÓRIA

"*Não ande apenas pelo caminho traçado, pois ele conduz somente até onde os outros já foram.*"

– Alexander Graham Bell –

AGRADECIMENTOS

..

A Deus, por estar sempre a meu lado em minhas orações nos momentos mais difíceis, e nos momentos de agradecimento também.

A meu marido Henrique, pela paciência e compreensão nos vários momentos em que, mesmo em casa, eu estava ausente. Pelo apoio e dedicação nas horas em que eu pensei que não dava para continuar e queria desistir... porém ele estendeu a mão e ajudou-me a seguir em frente, incentivando e acreditando em mim.

A minha filha Isabela, mesmo tão pequenina, precisando entender que em certos momentos sua mãe necessitava de silêncio para estudar e participar de reuniões remotas, enquanto ela só queria brincar e um pouquinho de atenção e amor da mamãe.

A meu pai Eduardo, de modo especial, por sempre me incentivar e apoiar na busca e realização dos meus sonhos. Sei que esse momento seria de grande orgulho para ele, e faria o impossível para ter apenas mais um abraço e sorriso dele nesse importante momento de conquista.

A minha mãe Ana e minha irmã Simone, por estarem sempre presente, me abrigando e ajudando a cuidar da minha filha em horas de aula e estudo (que não foram poucas).

A minha tia e madrinha Marilda, por me ajudar enquanto minha bolsa não era aprovada. Sem ela seria impossível continuar as idas e vindas tão custosas no deslocamento Rio x Região dos Lagos.

A minha orientadora professora Laura Assis, por todo apoio dado nesse período de estudos. Serei sempre grata pelos conselhos, sugestões, paciência e ensino. Um exemplo de pessoa, que com sua calma, me ajudou a fluir com textos e mais textos mesmo em dias difíceis, quando eu achava que não dava mais.

A meu coorientador professor Dr. Eduardo Ogasawara, pelos ensinamentos e direcionamentos nos estudos, sempre com clareza e objetividade, me mostrando a importância de ter foco e determinação. Serei eternamente grata por todo aprendizado.

Aos demais professores do PPCIC, com quem fiz disciplinas e muito aprendi. Minha gratidão por tudo!

Aos servidores da secretaria da Pós, por serem sempre prestativos e gentis no atendimento.

RESUMO

Análise Comparativa de Métodos para Detecção de Eventos em Séries Temporais

Grandes volumes de dados são coletados e armazenados diariamente, necessitando de um tratamento adequado para retornar informações valiosas durante uma análise. Esses dados, quando obedecem a uma ordem cronológica de tempo, consistem em séries temporais. Detectar eventos nessas séries é uma tarefa importante em diversas áreas de conhecimento, não se restringindo apenas à Tecnologia da Informação. Eventos podem representar uma anormalidade, uma mudança de comportamento ou um padrão que se repete na série. Diversos métodos presentes na literatura buscam identificar um único tipo de evento, entretanto, uma quantidade menor aborda essa detecção de uma maneira mais generalizada. Esta dissertação propõe uma análise comparativa de diferentes métodos para detecção de eventos em séries temporais, envolvendo identificação de anomalias e pontos de mudança. Tal comparação é realizada através de métodos estatísticos baseados na média móvel, processo de decomposição e técnicas baseadas em vizinhança. Foram realizados experimentos com dados sintéticos e reais envolvendo *datasets* de diferentes áreas de conhecimento como monitoramento da qualidade da água, tráfego de dados do *Yahoo* e processos de exploração de petróleo. Os resultados obtidos foram promissores e mostraram que cada conjunto de dado tem sua particularidade, e é muito importante analisar qual método se adequa melhor a um conjunto específico, onde uma boa escolha pode resultar em até $0,99$ de precisão na detecção de dados reais.

Palavras-chave: Detecção de Eventos; Séries Temporais; Detecção de Anomalias; Detecção de Pontos de Mudança.

ABSTRACT

Comparative Analysis of Methods for Events Detection in Time Series

Large volumes of data are collected and stored daily, requiring adequate treatment to return valuable information during analysis. These data, when obeying a chronological order of time, consist of time series. Detecting events in these series is an important task in several areas of knowledge, not being restricted to Information Technology. Events can represent an abnormality, a change in behavior, or a pattern that is repeated in the series. Several methods present in the literature seek to identify a single type of event, however, a smaller amount addresses this detection in a more generalized way. This dissertation proposes a comparative analysis of different methods for detecting events in time series, involving the identification of anomalies and change points. This comparison is performed through statistical methods based on the moving average, decomposition process, and neighborhood-based techniques. Computational experiments were performed with synthetic and real data involving datasets from different areas of knowledge such as water quality monitoring, data traffic from Yahoo, and oil exploration processes. The results obtained were promising and showed that each data set has its particularity, and it is very important to analyze which method is best suited to a specific set, where a good choice can result in up to 0.99 precision in detecting real data.

Key-words: Event Detection; Time Series; Anomaly Detection; Change Point Detection.

Lista de Ilustrações

2.1 Série temporal com dados variando ao longo do tempo. 5
2.2 Exemplo de uma série temporal com variação anual, adaptada [Shumway and Stoffer, 2017]. 6
2.3 Exemplo de uma série temporal multivariada. Adaptada de [Liu et al., 2015]. 7
2.4 Exemplos de estacionariedade e não estacionariedade em séries temporais. Adaptado de [Salles et al., 2019]. 9
2.5 Série temporal e sua decomposição em tendência, sazonalidade e restante, adaptada de [Silva et al., 2016]. 12
2.6 Exemplo de séries temporais, subsequências, e janelas deslizantes [Dutra, 2016]. . . . 14

3.1 *Boxplot*, adaptado de Han et al. [2011]. 16
3.2 Exemplo de série temporal com anomalias pontuais. 17
3.3 Exemplo de série temporal com anomalias contextuais. 18
3.4 Exemplo de série temporal com anomalia coletiva, adaptado de Chandola et al. [2009]. 19
3.5 Ilustração da estratégia adotada pelo método SCP [Guralnik and Srivastava, 1999]. . 24
3.6 Ilustração da estratégia adotada pelo método *ChangeFinder*, adaptado de [Takeuchi and Yamanishi, 2006]. 25
3.7 O exemplo ilustra a detecção de *motifs* em uma série temporal, adaptado de [Lin et al., 2002]. 28

5.1 Sub-*datasets test* e $reference$. 46
5.2 Exemplo de *dataframe* contendo o resultado de detecção de eventos. 46

6.1 Série temporal não estacionaria utilizada para avaliação dos métodos. 49
6.2 Detecções de eventos na série didática por diferentes métodos. 54
6.3 Detecções de eventos na série **pH** através dos 11 métodos apresentados nesta pesquisa. 60

Lista de Tabelas

4.1 Critérios de Inclusão. 29
4.2 Critérios de Exclusão. 30
4.3 Classificação das referências bibliográficas por tipos de eventos, cenário e dimensão. . 35

5.1 Variáveis e suas respectivas siglas. 40
5.2 Parâmetros necessários a definir em cada método 40
5.3 Matriz de confusão 41

6.1 Variáveis do conjunto de dados relacionado à qualidade da água. 50
6.2 Variáveis do conjunto de dados do *Yahoo* nos *benchmarks* $A1$, $A2$, $A3$ e $A4$. 51
6.3 Tipos de eventos indesejáveis do conjunto de dados $3W$. 51
6.4 Descrição dos *datasets* selecionados. 52
6.5 Testes para definição do tamanho da janela w. 53
6.6 Resultados obtidos com $F1$ para o *dataset* da Água. 55
6.7 Resultados obtidos com acurácia balanceada para o *dataset* da Água. 56
6.8 Resultados contendo os acertos para o *dataset* da Água. 57
6.9 Resultados obtidos com a distância *a posteriori* para o *dataset* da Água. 57
6.10 Resultados obtidos com a distância *a priori* para o *dataset* da Água 58
6.11 Resultados referentes ao *dataset* do *Yahoo* 62
6.12 Resultados obtidos com a distância *a priori* para o 3W *dataset*. 64
6.13 Resultados obtidos com a distância *a posteriori* para o 3W *dataset*. 65
6.14 Comparação da qualidade das detecções de eventos nos conjuntos de dados selecionados produzidas por diferentes métodos com base na métrica F1 e no tempo de execução em segundos. 65
6.15 Comparação dos métodos SD-EWMA e TSSD-EWMA. 66
6.16 Análise geral por *dataset* e método. 67

Sumário

1 Introdução **1**

2 Séries Temporais **4**

2.1 Série Temporal 4

2.1.1 Séries Temporais Univariadas 5

2.1.2 Séries Temporais Multivariadas 6

2.1.3 Séries Temporais Estacionárias 7

2.1.4 Séries Temporais Não Estacionárias 8

2.2 Componentes de uma Série Temporal 8

2.2.1 Tendência 9

2.2.2 Ciclos 11

2.2.3 Sazonalidade 11

2.2.4 Ruídos 11

2.3 Segmentação de uma Série Temporal 13

2.3.1 Subsequências de Séries Temporais 13

2.3.2 Janelas Deslizantes 13

3 Detecção de Eventos **15**

3.1 Anomalias 15

3.1.1 Tipos de Anomalias 17

3.1.2 Métodos de Detecção de Anomalias 19

3.2 Pontos de Mudança 22

3.2.1 Métodos de Detecção de Pontos de Mudança 23

3.3 Descoberta de Padrões ou *Motifs* 27

4 Trabalhos Relacionados **29**

5 Metodologia **37**

5.1 Aquisição dos Dados 38

5.2 Escolha dos Métodos 38

5.3 Definição dos Parâmetros 39
5.4 Execução dos Métodos 41
5.5 Métricas para Avaliar a Detecção de Eventos 41
5.6 Comparação dos Métodos 44

6 Resultados 48
6.1 Ambiente de Desenvolvimento e Testes 48
6.2 *Datasets* 48
6.2.1 Conjunto de Dados Sintético: Série Não Estacionária 49
6.2.2 Conjunto de Dados Sintético: Qualidade da Água 49
6.2.3 Conjunto de Dados Sintético: Dados do *Yahoo* 50
6.2.4 Conjunto de Dados Reais: 3W 51
6.3 Definição dos Parâmetros 52
6.4 Análise e Testes dos Resultados 53
6.4.1 Análise da Série Temporal Didática 53
6.4.2 Análise do *dataset* Água 55
6.4.3 Análise do *dataset Yahoo* 61
6.4.4 Análise do *dataset* 3W 63
6.4.5 Análise de todos os *datasets* 64

7 Conclusão 68
Referências Bibliográficas 70

Capítulo 1

Introdução

Diante de um cenário onde grandes volumes de dados são coletados e armazenados diariamente, ferramentas de análise são necessárias para extrair o máximo de informação relevante, transformando esses dados em conhecimento [Han et al., 2011]. Assim, origina-se a mineração de dados. Embora dados sejam encontrados por toda parte na internet, frequentemente, pouco conhecimento relevante é extraído dos mesmos. Portanto, há uma necessidade de abordar métodos e ferramentas adequadas, com o desafio de encontrar, extrair e vincular informações úteis, advindas de várias fontes de dados, através da mineração [Dao et al., 2015]. Entre os desafios de minerar está a dificuldade em se analisar dados complexos, como aqueles que variam ao longo do tempo. Esses dados podem ser representados como séries temporais.

Alguns aspectos da descoberta de conhecimento, através de séries temporais, com técnicas de mineração envolvem a análise da mudança significativa do comportamento da série. Essas mudanças são denominadas eventos. Um evento pode ocorrer instantaneamente em um determinado momento ou pode constituir um intervalo de tempo, sendo a identificação de eventos um aspecto importante na mineração de dados temporais [Guralnik and Srivastava, 1999; Chakravarthy et al., 1994]. Um evento detectado em dados de séries temporais pode representar a ocorrência de um fenômeno com significado específico e definido em um determinado domínio de conhecimento.

Um evento, do mundo real, é definido como uma ocorrência que está associada ao local onde o evento ocorreu e, também ao tempo, ou seja quando ele ocorreu. Em séries temporais, o objetivo da detecção de eventos envolve identificar ocorrências específicas de interesse em um ponto ou em intervalos de tempo. Dado que eventos estão associados a pontos específicos no tempo, observações temporais próximas ao momento em que ocorreu o evento geralmente são as mais importantes para prever dados futuros. O objetivo da previsão é aprender a detectar com precisão a ocorrência de eventos em instâncias futuras [Batal et al., 2015; Alkhamees and Fasli, 2017].

Comumente, eventos detectados em séries temporais se apresentam como anomalias ou pontos de mudança, e se caracterizam como observações que não estão em conformidade com o padrão de comportamento esperado dentro do conjunto de dados em análise [Ben-Gal, 2005]. Por sua vez, pontos de mudança (*change points*) separam diferentes estados no processo que gera a série

temporal, enquanto que padrões (*motifs*) são eventos que ocorrem em um determinado intervalo e se repetem ao longo da série [Keogh and Kasetty, 2003]. O problema de detecção de pontos de mudança está relacionado ao problema de detecção de mudanças de conceito (*drifts*) em séries temporais. Neste caso, a detecção de pontos de mudança objetiva encontrar o instante (ou intervalo) específico no tempo que marca a ocorrência da mudança de conceito detectada.

Uma característica importante a ser considerada é a geração dos valores que constituem a série temporal. Dados coletados em um tempo anterior e que estão disponíveis para análise são mais simples do que dados que chegam na série em tempo real. Nesse último caso, as séries são chamadas *streams* (transmissão contínua), ou seja, um fluxo contínuo de dados.

Algoritmos que analisam séries temporais são classificados como *batch* ou *online*. Um algoritmo caracterizado como *batch* considera que todos os dados da série foram coletados antes do início da análise [Guralnik and Srivastava, 1999]. Tais algoritmos examinam conjuntos de dados já coletados, sem que ocorra a chegada de novos dados. Esse conceito é diferente dos algoritmos classificados como *online*, onde os dados são analisados ponto a ponto, em tempo real, conforme o decorrer de uma janela deslizante de tamanho predeterminado. Essa abordagem é muito útil por exemplo em sistemas de segurança, onde é possível rastrear anormalidades através de sensores de monitoramento [Ma and Perkins, 2003].

Muitos métodos limitam-se a identificar eventos de forma univariada, sem considerar dados de outros atributos em decorrência da complexidade do tema. Uma detecção de forma multivariada normalmente é um desafio, devido ao tamanho e variedade dos eventos em dados do mundo real [Cappers and van Wijk, 2018], cada vez mais presente em diversos tipos de aplicações e em diferentes áreas de conhecimento. Estes dados são provenientes, por exemplo, de áreas como medicina, obtidos através de informações sobre pacientes [Moskovitch et al., 2014], química, onde são analisadas medições sobre a qualidade da água [Mao et al., 2017; Perelman et al., 2012], telecomunicações, em registros de chamadas telefônicas [Cappers and van Wijk, 2018], sísmica, através de monitoramento de duração de terremotos [Wu et al., 2019] e exploração de petróleo [Bach et al., 2014].

O objetivo deste trabalho é apresentar uma comparação de métodos para detecção de eventos em séries temporais, através da detecção de pontos de mudança e da detecção de anomalias. A literatura relata uma diversidade de métodos, e a escolha adequada para uma determinada série temporal não é uma tarefa trivial. Métodos especializados na detecção de uma categoria de eventos podem negligenciar a ocorrência de eventos de outra espécie, ou ainda, identificá-los de forma incorreta. No entanto, falhas em um processo de identificação de eventos podem afetar uma tomada de decisão ou levar à produção de falsos positivos.

Foram selecionados onze métodos de detecção de eventos em séries temporais para comparação,

análise e discussão dos resultados. Estes compreendem busca por anomalias e pontos de mudança através de técnicas distintas como métodos estatísticos, decomposição, volatilidade e proximidade. Tais métodos foram testados em *datasets* sintéticos, onde foram analisadas um total de 376 séries sintéticas, além de 214 séries obtidas de conjuntos de dados reais. Das 590, foram excluídas 12 séries nulas, totalizando 578 séries que foram submetidas a análise nos 11 métodos de detecção, e para cada método foram computadas 7 métricas para estruturar a análise comparativa. Foi gerado um total 44.506 valores resultantes coletados e processados para alimentar o capítulo referente aos resultados obtidos. Os *datasets* utilizados nesta pesquisa foram submetidos a análise do tipo *batch*, onde os experimentos envolveram a análise da série completa, não sendo os métodos aqui abordados aplicados a um cenário de fluxo contínuo, que são casos onde dados chegam na série em tempo real.

Assim, essa dissertação busca apoiar uma tomada de decisão sobre qual o método mais adequado para um determinado conjunto de dados. Para isso, é apresentado um estudo abrangente sobre três *datasets* de áreas e aplicações diferentes. Buscando uma comparação mais sólida, foram utilizadas sete métricas com distintos objetivos para avaliar a detecção. Resultados promissores foram alcançados, sendo possível identificar em um cenário quais variáveis respondem melhor à detecção, e qual método propõe melhores resultados. Em dados sintéticos e reais foi alcançado até precisão de $1,00$ na detecção.

Além da introdução, o texto desta dissertação é composto de mais seis capítulos. No Capítulo 2 é apresentado um *background* sobre séries temporais e outras definições importantes para o desenvolvimento desta pesquisa. O Capítulo 3 aborda as diferentes formas de eventos que podem ser encontradas em séries temporais e seus respectivos métodos de detecção. O Capítulo 4 discorre sobre os trabalhos relacionados, apresentando diferentes técnicas e abordagens para comparação de métodos de detecção de eventos presente na literatura. A apresentação da metodologia é realizada no Capítulo 5. O Capítulo 6 apresenta os resultados computacionais dos experimentos realizados. O Capítulo 7 resume as principais contribuições deste trabalho, apresentando as conclusões da pesquisa e possíveis desdobramentos.

Capítulo 2

Séries Temporais

As últimas décadas registraram um aumento considerável na quantidade de dados coletados e armazenados. A fim de extrair informações relevantes destes dados é importante definir técnicas eficientes que auxiliem ou possibilitem a tomada de decisão da melhor forma possível. Dados que são registrados ao longo do tempo levam a uma representação denominada série temporal [Dutra, 2016].

A análise de séries temporais vem despertando interesse de pesquisadores, pois pode ser utilizada em diversos campos e tipos de aplicações, tais como *(i)* finanças, na análise orçamentária anual ou no registro de valores diários de ações; *(ii)* medicina, na análise de um eletroencefalograma ou eletrocardiograma; *(iii)* sísmica, onde dados são coletados por sensores a cada minuto; *(iv)* epidemiologia, em que casos mensais de sarampo são registrados no país; dentre tantos outros [Mueen et al., 2009; Shumway and Stoffer, 2017]. Ao examinar uma série, espera-se que nela exista uma causa relacionada com o tempo, a qual influenciou os dados no passado e que pode continuar influenciando no futuro. Como consequência, estes dados devem ser bem avaliados para descoberta de informações relevantes. Neste capítulo são apresentadas informações básicas necessárias para a compreensão dos conceitos discorridos no restante desta dissertação.

2.1 Série Temporal

Uma série temporal consiste em um conjunto de observações de dados que foi gerado e armazenado obedecendo uma ordem cronológica de tempo. Para analisar uma série são aplicadas técnicas estatísticas com o objetivo de resumir as informações presentes nos dados, considerando duas abordagens: *(i)* no domínio do tempo e *(ii)* no domínio da frequência. A abordagem no domínio do tempo explora a influência que pontos passados no tempo podem ter em pontos futuros, como por exemplo, o que aconteceu hoje pode influenciar o que irá acontecer amanhã. Já a abordagem no domínio da frequência investiga a ocorrência de ciclos nos pontos de dados, sendo o ciclo um determinado comportamento que se repete ao longo do tempo [Shumway and Stoffer, 2017].

De modo geral, a análise de séries temporais ainda se divide em duas categorias, como algoritmos de previsão e algoritmos para detecção de eventos em séries temporais. Algoritmos de previsão tem como objetivo prever o valor de uma determinada observação y no tempo $t+1$ dado que valores anteriores a $t+1$ são conhecidos. Algoritmos para detecção de eventos compreendem uma busca por padrões ou anomalias nos dados. A representação gráfica de uma série temporal pode ser visualizada através da Figura 2.1, onde são plotados os valores das variáveis aleatórias no eixo das ordenadas e a escala de tempo no eixo das abscissas [Aggarwal, 2017].

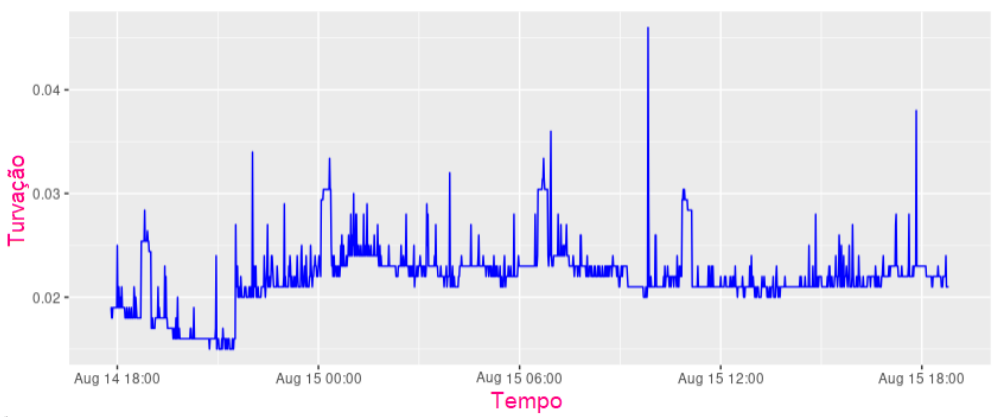


Figura 2.1: Série temporal com dados variando ao longo do tempo.

Uma série temporal é referenciada como uma série univariada quando essas observações estão relacionadas a uma única variável, conforme é apresentado na Seção 2.1.1. Comumente, o comportamento de uma série temporal univariada é estudado como uma função de seus dados passados. Entretanto, se a série apresenta mais de uma variável, estando ou não relacionadas entre si, ela é denotada por série temporal multivariada. A definição de séries temporais multivariadas é apresentada na Seção 2.1.2. Quando o comportamento das séries sugere um tipo de regularidade ao longo do tempo, tem-se o conceito chamado estacionariedade tratado na Seção 2.1.3. No entanto, a maioria das séries temporais do mundo real não apresentam regularidade em seu comportamento, sendo definidas assim como não estacionárias, as quais são abordadas na Seção 2.1.4.

2.1.1 Séries Temporais Univariadas

Uma série temporal univariada refere-se a uma série temporal que consiste de observações únicas registradas sequencialmente ao longo do tempo em uma variável. A Figura 2.2 ilustra um exemplo de uma série temporal univariada com dados de desvios globais de temperatura, sendo que estes valores variam anualmente.

Definição 1. *Uma série temporal* ***univariada*** *y é uma sequência de n observações aleatórias, $<y_1, y_2, y_3, \ldots, y_n>$, onde y_1 representa o valor assumido pela série no primeiro instante de tempo (mais antigo) e y_n representa o valor da série no instante mais recente. O comprimento n de uma*

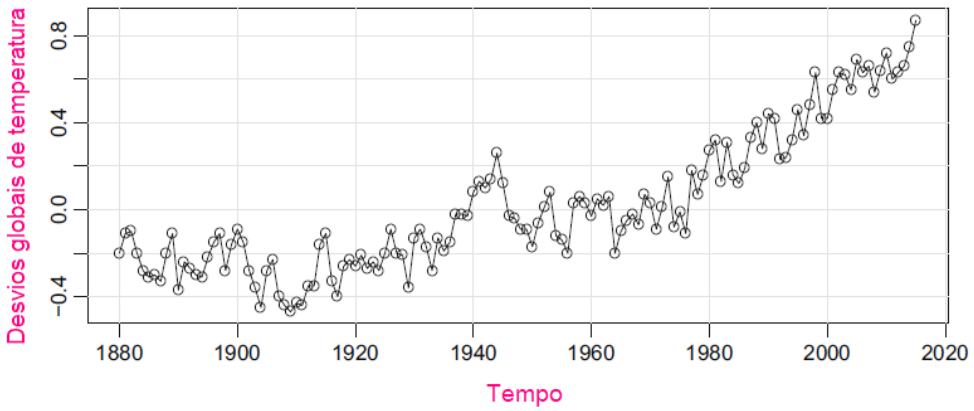


Figura 2.2: Exemplo de uma série temporal com variação anual, adaptada [Shumway and Stoffer, 2017].

série temporal y é representado por $|y|$ e uma observação específica de uma série temporal é referenciada como y_t, indexada no tempo t, sendo que $t = \{1, \ldots, n\}$ [Esling and Agon, 2012; Shumway and Stoffer, 2017]. Uma série temporal y pode ser definida como apresentado na Equação 2.1.

$$y = <y_1, y_2, y_3, \ldots, y_n>, \quad y \in \mathbb{R} \tag{2.1}$$

2.1.2 Séries Temporais Multivariadas

Uma série temporal multivariada consiste de uma sequência de dados composta por múltiplas observações ao longo do tempo, que podem estar ou não relacionadas. Existem situações nas quais esse relacionamento entre as observações medidas em conjunto são interessantes em uma análise pois tendem a caracterizar dependência entre as observações de diferentes variáveis. Uma determinada condição cardíaca, por exemplo, pode ser diagnosticada não apenas pelo eletrocardiograma, mas também pelos valores da pressão arterial e outras informações no mesmo instante de tempo. Em diversas séries temporais multivariadas, existe uma relação de dependência entre as variáveis [Xie et al., 2012].

Definição 2. *Uma série temporal* ***multivariada*** y_m *é um conjunto de séries temporais* $<y_{d_1}, y_{d_2}, y_{d_3}, \ldots, y_{d_{dim}}>$, *sendo que* $d_i \in \{1, 2, \ldots, dim\}$. *dim representa a quantidade de dimensões associadas e* d_i *refere-se a uma dimensão específica da série temporal. Então* y_{d_i} *representa a série temporal correspondente. Uma série temporal* y_m *pode ser definida conforme a Equação 2.2.*

$$y_m = <y_{d_1}, y_{d_2}, y_{d_3}, \ldots, y_{d_{dim}}>, \quad y_m \in \mathbb{N} \tag{2.2}$$

Um exemplo de série temporal multivariada é dado através da Figura 2.3. A simulação da contaminação de um sistema de distribuição de água por nitrato de cádmio é representada nesta

figura. As observações desta série representam sinais de um sensor colhidos nesta simulação. Os valores foram divididos em duas partes, a e b. Sendo que as observações em a representam os dados limpos, i.e., sem contaminação da água, e as observações em b são referentes aos dados que sofreram contaminação por nitrato de cádmio no instante 60.

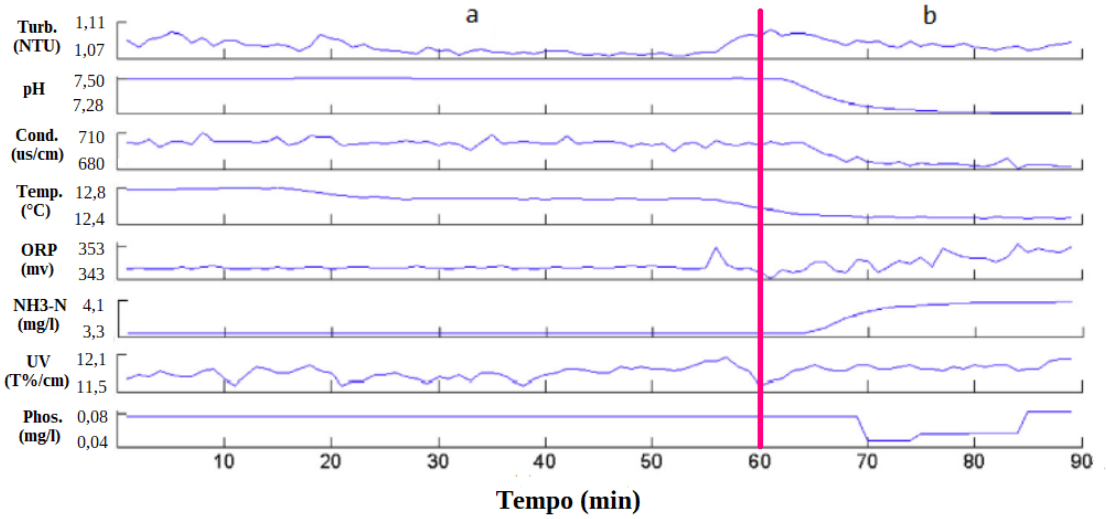


Figura 2.3: Exemplo de uma série temporal multivariada. Adaptada de [Liu et al., 2015].

2.1.3 Séries Temporais Estacionárias

Séries temporais estacionárias sugerem uma regularidade no seu comportamento ao longo do tempo. Podem ser definidas como estritamente estacionárias e fracamente estacionárias, sendo esta última comumente referenciada apenas por estacionária. Uma série temporal estacionária apresenta média e variância constantes.

Definição 3. *Uma série temporal é definida* ***estritamente estacionária*** *quando o comportamento probabilístico de todas as sequências possíveis de valores* $<y_1, y_2, y_3, \ldots, y_n>$ *se apresenta de maneira semelhante ao da sequência deslocada no tempo* $<y_{1+h}, y_{2+h}, y_{3+h}, \ldots, y_{n+h}>$*, sendo* h *a mudança de tempo. As sequências* $<y_1, y_2, y_3, \ldots, y_n >e <y_{1+h}, y_{2+h}, y_{3+h}, \ldots, y_{n+h}>$ *são idênticas para todo inteiro positivo* n*, sendo* n *um período de tempo com mesma duração. Cabe definir que* y_n *e* y_{n+h} *são igualmente distribuídas para qualquer valor* h *[Shumway and Stoffer, 2017; Salles et al., 2019].*

No entanto, o conceito de série estritamente estacionária dificilmente é encontrado em séries temporais obtidas a partir de dados reais. Uma versão mais suave, a de estacionariedade fraca impõe condições apenas nos dois primeiros momentos da série, diferente da série estritamente estacionária onde são impostas condições a todas as distribuições possíveis de uma série temporal.

Definição 4. *Uma série temporal pode ser definida como* ***fracamente estacionária*** *quando compreender um processo de variação finita, tal que: (i) sua função média* $E(y_t) = \mu_t = \mu$ *é constante*

e não depende do tempo t; e (ii) sua função de autocovariância $\nu(s,t)$ entre y_t e o valor da série temporal deslocada no tempo y_s depende apenas da diferença $|s-t|$ [Shumway and Stoffer, 2017; Salles et al., 2019].

2.1.4 Séries Temporais Não Estacionárias

Uma série temporal é considerada não estacionária quando viola uma ou mais restrições impostas por um processo estacionário e pode se apresentar de diferentes maneiras: *(i)* não estacionária de tendência, *(ii)* não estacionária de nível, *(iii)* séries heterocedásticas e, *(iv)* não estacionária por diferenciação.

Uma série temporal que apresenta comportamento de **tendência** ocorre quando a média sofre uma alteração, aumentando ou diminuindo ao longo do tempo. A não estacionariedade em uma série temporal também pode ser causada por quebras estruturais, que podem eventualmente ocasionar **mudanças de nível** na série, resultando em uma função média diferente para diferentes partes da série. Uma mudança da variação ao longo do tempo também causada por quebras estruturais é a **heterocedasticidade**, que surgem quando as mudanças aumentam ou diminuem a volatilidade das observações de séries temporais ao longo do tempo. Essas séries são conhecidas por heterocedásticas. A presença de uma raiz unitária garante que as observações retornem a um nível histórico e não fiquem vagando em qualquer direção e é chamada **não estacionária por diferenciação**. Sem uma raiz unitária, as observações de séries temporais tendem a flutuar em torno de componentes determinísticos, como média ou tendência. A Figura 2.4 retrata exemplos de séries temporais que apresentam as propriedades de (a) estacionariedade. É possível observar a não estacionariedade na forma de (b) não estacionariedade de tendência, (c) não estacionariedade de nível, (d) série heterocedástica e (e) não estacionariedade por diferenciação. As linhas em rosa escuro e rosa claro representam as funções de média e variância das séries temporais, respectivamente.

2.2 Componentes de uma Série Temporal

Séries temporais consistem em um conjunto de observações ordenadas no tempo. Normalmente, essas medições são igualmente espaçadas, como por exemplo a cada ano, mês, dia, hora, minuto ou segundo. A propriedade mais importante de uma série temporal é que as observações ordenadas são dependentes ao longo do tempo, e a natureza dessa dependência é de interesse na análise da série. Um objetivo importante na análise de séries temporais é a decomposição de uma série em um conjunto de componentes não observáveis, que podem ser associados a diferentes tipos de variações temporais [McLaren et al., 2018].

A ideia da decomposição de séries temporais é muito antiga e foi usada para o cálculo de órbitas planetárias por astrônomos do século XVII. A decomposição de uma série permite a identificação

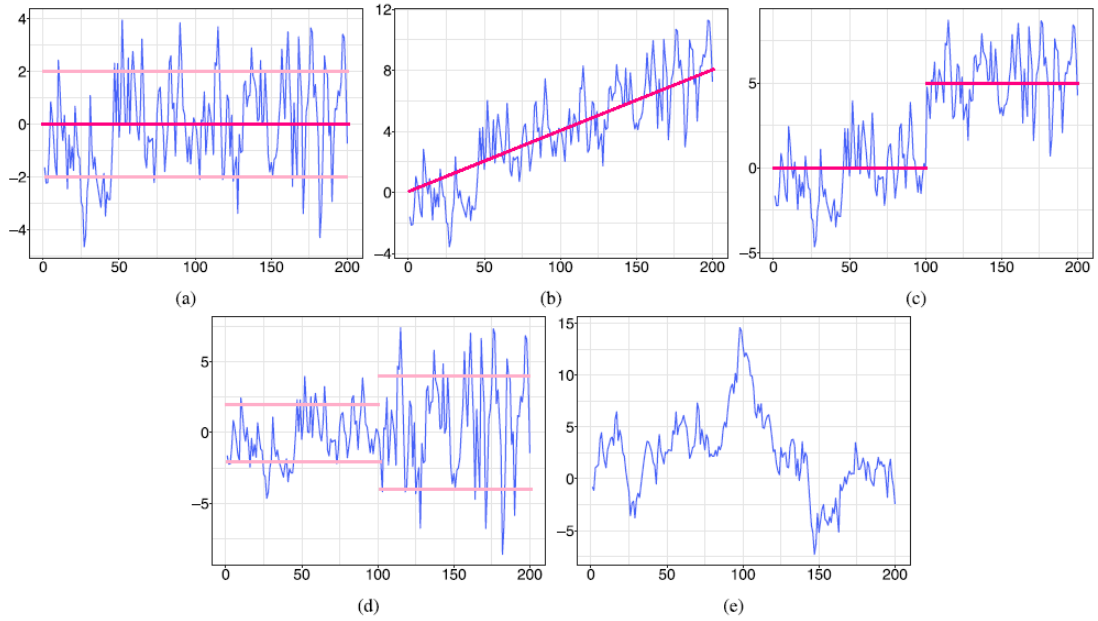


Figura 2.4: Exemplos de estacionariedade e não estacionariedade em séries temporais. Adaptado de [Salles et al., 2019].

.

de quais componentes estão presentes, sendo muito importante na análise. Persons [1919] foi o primeiro a declarar de maneira explícita os componentes de uma série temporal não observados categorizando-os em quatro tipos de componentes:

1. Uma tendência de longo prazo.
2. Movimentos cíclicos sobrepostos à tendência de longo prazo.
3. Um movimento sazonal dentro de cada ano.
4. Variações residuais devido às mudanças que afetam variáveis individuais.

2.2.1 Tendência

Tendência é o comportamento a longo prazo de uma série temporal que pode ser causado por diversos fatores, como por exemplo, o crescimento demográfico. Para analisar as mudanças que ocorrem nesse prazo, torna-se viável calcular a tendência da série. Esse cálculo tem como resultado uma série temporal que explica as tendências subjacentes através de uma versão suavizada da série original. Três objetivos básicos estão envolvidos na estimativa de tendência: *(i)* avaliação do comportamento para calcular previsões; *(ii)* identificação do nível que a série pode assumir (crescente ou decrescente); e *(iii)* remoção da tendência na série buscando facilitar a visualização de outras componentes.

Existe um grande número de modelos determinísticos e estocásticos que foram propostos para estimativa de tendências. Os modelos determinísticos são baseados na suposição de que a tendência

pode ser bem aproximada por funções matemáticas de tempo, como polinômios de baixo grau. Os modelos de tendências estocásticas assumem que a tendência pode ser melhor modelada por diferenças de ordem inferior juntamente com erros autorregressivos e de média móvel [McLaren et al., 2018]. Nesta dissertação são abordadas a média móvel e a média móvel ponderada exponencialmente.

Média Móvel

Através deste método, um componente de tendência (ou média móvel) η_t é definido como a média dos n pontos anteriores ao tempo t, conforme mostrado na Equação 2.3. O resultado é posicionado no período central dos valores calculados. Posteriormente, um novo ponto é inserido, enquanto que o primeiro ponto da média imediatamente anterior é desprezado. Assim são calculadas novas médias que se movem até o fim da série. A média móvel é uma das maneiras mais intuitivas para o cálculo da estimativa de tendência.

$$\eta_t = \frac{y_t + y_{t-1} + \cdots + y_{t-n}}{n} = \frac{1}{n}\sum_{i=t}^{t-n} y_i \tag{2.3}$$

Mediana Móvel

Semelhante ao cálculo da média móvel, a estimativa de tendência por mediana móvel é intuitiva, compreendendo substituir a média por mediana. Formalmente, estimar a média móvel de uma série temporal y no tempo t é dada pela mediana dos n pontos anteriores, conforme a Equação 2.4.

$$\eta_t = md(y_t, y_{t-1}, \cdots, y_{t-n}) \tag{2.4}$$

Ajuste Exponencial

O ajuste exponencial apresenta algumas vantagens em relação às estimativas de tendência citadas anteriormente, que envolvem média móvel e mediana móvel, para calcular a tendência de uma série temporal. Este método está relacionado à média móvel, porém utiliza um coeficiente de suavização θ no lugar de uma série de n pontos anteriores. Assim, este método dá origem à média móvel exponencialmente ponderada. Em síntese, cada valor ajustado depende de todos os valores passados e os pesos atribuídos a cada observação obedecem a uma estrutura decrescente ao longo do tempo.

A média móvel exponencialmente ponderada é a média de todos os pontos passados, onde a influência desses pontos passados decai exponencialmente com o tempo. Seja y_t o valor da série temporal no tempo t e $\theta \in [0, 1]$ a constante de suavização, a média móvel exponencialmente ponderada é definida através da Equação 2.5.

$$\eta_t = \theta y_{t-1} + (1 - \theta)\eta_{t-1} \tag{2.5}$$

Dado que uma tendência foi identificada na série temporal, ela pode ser removida buscando facilitar a visualização das outras componentes presentes. Assim, uma série temporal poderá consistir ainda dos seguintes itens: *(i)* componente de sazonalidade e *(ii)* componente ruído. O ruído representa o que sobrou na série temporal ao se extrair a tendência e a sazonalidade [Prema and Rao, 2015; Hyndman and Athanasopoulos, 2013]. O componente ciclo, que será apresentado a seguir, normalmente está associado à tendência, ocorrendo de forma conjunta e gerando o ciclo de tendência.

2.2.2 Ciclos

Variações cíclicas são padrões de longo prazo (superiores a um ano), como por exemplo períodos de crescimento e recessão da economia. Normalmente são analisados conjuntamente com a tendência, onde são conhecidos como ciclos de tendência, como citado na Subseção 2.2.1. Alguns autores não mencionam as variações cíclicas porque em certos casos a série temporal precisa abranger décadas para que seja possível identificar um comportamento cíclico.

2.2.3 Sazonalidade

A sazonalidade tem origem no clima e nas estações do ano, pois são períodos que se repetem anualmente. Épocas do ano como o Natal, a Páscoa e fim do ano letivo têm um grande impacto no comércio e no consumo de determinados bens e serviços, nomeadamente viagens de avião, ocupação hoteleira e consumo de gasolina. Em resumo, variações sazonais são oscilações nos valores das observações de curto prazo e que ocorrem sempre dentro de um ano, e se repetem a cada ano.

Um padrão sazonal existe quando uma série é influenciada por fatores sazonais, por exemplo, o trimestre do ano, o mês ou o dia da semana. A sazonalidade é sempre de um período fixo e conhecido. Portanto, as séries temporais sazonais às vezes são chamadas de séries temporais periódicas. As componentes de sazonalidade são obtidas através da razão de médias móveis. Esta técnica consiste em fazer a razão entre o valor observado na série e o valor da média móvel.

2.2.4 Ruídos

O componente ruidoso (ou restante) em qualquer modelo de decomposição representa variações relacionadas a eventos imprevisíveis de todos os tipos. A maioria dos valores irregulares apresentam um padrão estável, mas algumas anomalias podem estar presentes nos dados. Esses valores anômalos podem frequentemente serem atribuídos a causas identificáveis como greves, secas, inun-

dações ou erros de processamento de dados [McLaren et al., 2018].

As variações restantes R_t são obtidas através da remoção das componentes de tendência η_t e sazonalidade S_t da série temporal original y, conforme apresentado na Equação 2.6. A Figura 2.5 apresenta um exemplo de decomposição de uma série temporal.

$$R_t = y - \eta_t - S_t \tag{2.6}$$

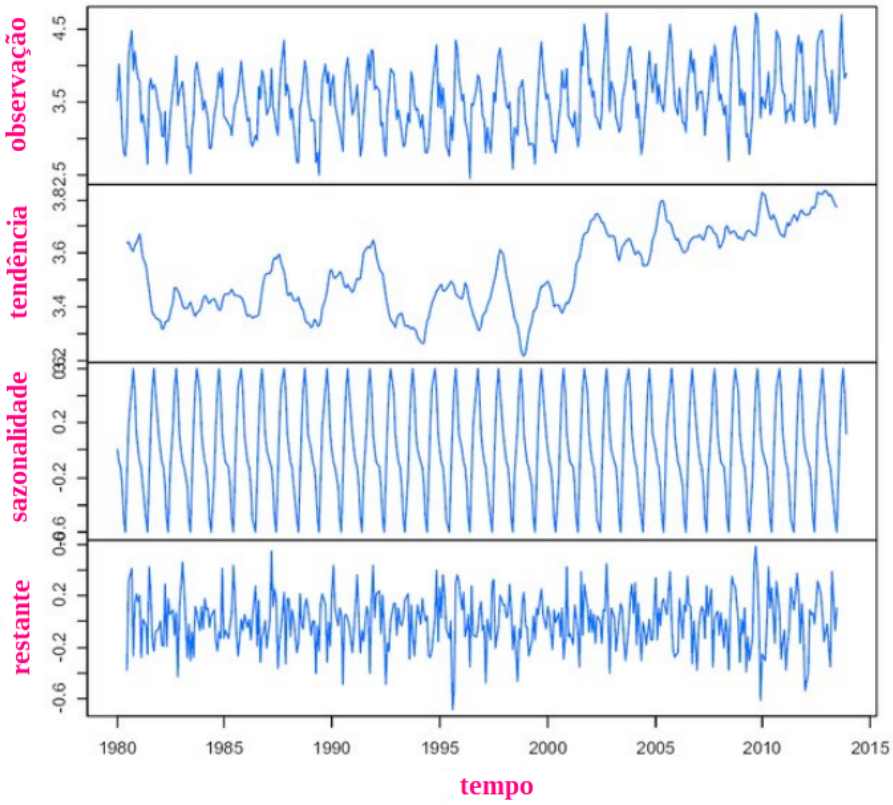


Figura 2.5: Série temporal e sua decomposição em tendência, sazonalidade e restante, adaptada de [Silva et al., 2016].

A primeira série mostrada na Figura 2.5 representa a série temporal original, enquanto que no segundo bloco é possível observar sua tendência. A terceira parte da figura retrata a sazonalidade, enquanto que na última série são apresentados os componentes restantes. É possível observar que a sazonalidade muda lentamente ao longo do tempo, considerando que a série temporal do exemplo compreende mais de 10 anos.

Existe uma tendência quando há um aumento ou diminuição a longo prazo no comportamento dos dados, não necessariamente linear, podendo representar uma mudança crescente ou decrescente na série. A sazonalidade é sempre afetada por fatores sazonais, como épocas do ano ou dias da semana, e geralmente possuem frequência fixa e conhecida.

Séries temporais costumam ser muito longas, contendo milhares de observações. Usualmente,

o interesse não está nas propriedades globais da série, mas se restringe a pequenas partes dela, conhecidas por subsequências [Chiu et al., 2003].

2.3 Segmentação de uma Série Temporal

A segmentação de séries temporais é uma ferramenta útil para a detecção de eventos. O objetivo de um algoritmo de segmentação é particionar uma série temporal y em subséries menores denominadas subsequências. Esta seção fornece uma visão geral de alguns dos métodos de segmentação existentes, como o particionamento em subsequências de tamanho semelhante.

2.3.1 Subsequências de Séries Temporais

Encontrar subsequências em séries temporais têm se tornado uma tarefa objetiva em mineração de dados. Subsequência é uma amostra contínua de uma série temporal que possui comprimento definido, e é necessariamente menor que a série. Permite analisar um subconjunto dos dados com o objetivo de avaliar algumas propriedades locais [Chiu et al., 2003].

Definição 5. *Uma subsequência é definida como a q-ésima sequência de tamanho m em uma série temporal y, representada por* $seq_{m,q}(y)$*, que é uma sequência ordenada de valores* $<y_q, y_{q+1}, \ldots, y_{q+m-1}>$*, onde* $|seq_{m,q}(y)| = m$ *e* $1 \leq q \leq |y| - m$*. Tal subsequência pode ser definida como mostrado na Equação 2.7:*

$$y_{m,q} = subseq_{m,q}(y) \tag{2.7}$$

Para que seja possível extrair todas as subsequências de uma série temporal, utilizamos o conceito de janela deslizante, onde cada janela é deslocada sobre a série temporal y, e então extraída uma subsequência com o mesmo tamanho da janela.

2.3.2 Janelas Deslizantes

Uma janela deslizante é a técnica que consiste em estipular o valor do tamanho m, permitindo obter todas as subsequências possíveis deste tamanho a partir da série temporal de origem [Keogh and Lin, 2005; Lampert et al., 2008]. Desta forma é obtido um conjunto de subsequências de comprimento m. Janelas deslizantes são muito utilizadas em estudos sobre séries temporais, especialmente nos casos onde dados são coletados em tempo real.

Definição 6. *Janelas deslizantes de tamanho m para uma série temporal y de tamanho n são definidas por Keogh and Lin [2005] como sendo uma função* sw_y *que produz uma matriz* ω *de dimensão* $(n-m+1) \times m$*, onde cada linha* ω_i *com* $i \in \{1, 2, \ldots, (n-m+1)\}$ *representa a i-ésima*

subsequência de y de tamanho m. A matriz ω é definida pela Equação 2.8. Tal matriz contém todas as subsequências possíveis ao deslizar uma janela de tamanho pré-definido m na série y.

$$\omega = sw_y, \quad \forall\ w_i \in \omega, \quad w_i = seq_{m,i}(y) \tag{2.8}$$

Este conceito é amplamente utilizado em análises de séries temporais para fazer comparações entre subsequências com o objetivo de encontrar suas semelhanças [Hoan and Exbrayat, 2013]. A Figura 2.6 mostra a aplicação das definições acima para uma série temporal. A série temporal y está representada pela curva em azul, a parte em vermelho representa uma subsequência desta série temporal e as linhas em verdes tracejadas são exemplos de algumas das subsequências extraídas da série temporal baseada em janelas deslizantes.

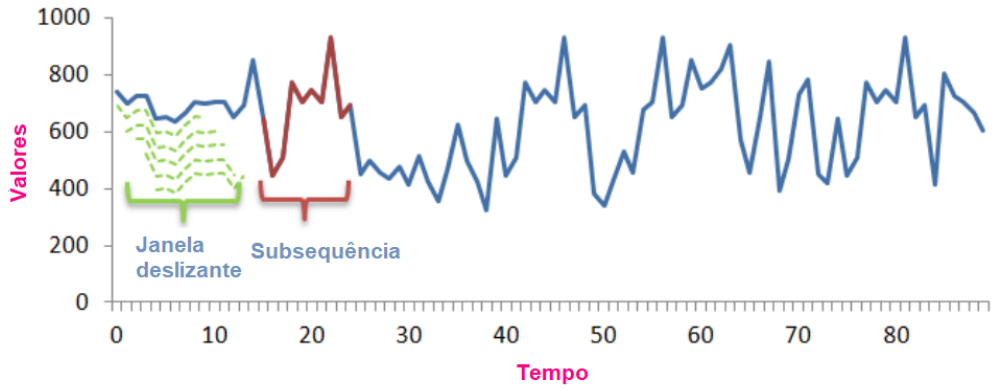


Figura 2.6: Exemplo de séries temporais, subsequências, e janelas deslizantes [Dutra, 2016].

Capítulo 3

Detecção de Eventos

Dados de séries temporais são sequências de medição ao longo do tempo que obedecem a um determinado comportamento. Esse comportamento pode mudar, devido a ocorrência de um evento. Detectar eventos em séries temporais consiste em identificar ocorrências específicas em uma ou mais linhas do tempo, como por exemplo, uma variação repentina na temperatura de uma caldeira industrial ou uma anormalidade em um eletrocardiograma de um paciente.

A detecção de eventos é um aspecto importante na mineração de dados temporais. Geralmente eventos são pontos de uma série temporal que podem ser picos, alterações de nível, mudanças repentinas, entre outros. Um evento pode ser tão simples quanto uma alteração em uma variável, ou tão complexo quanto grandes alterações em duas ou mais variáveis [Xie et al., 2012].

Uma questão a ser considerada na identificação de eventos é o seu período de tempo. Em dados coletados por sensores a cada minuto, um evento pode aparecer como uma mudança repentina em um ou dois pontos de valores obtidos, caracterizando a ocorrência de uma anomalia, ou durar até mesmo semanas, apresentando uma característica de ponto de mudança. A detecção de anomalias é diferente da detecção de pontos de mudança, pois enquanto a detecção de pontos de mudança busca por uma alteração de um comportamento para outro na série, a detecção de anomalias procura uma saída temporária do comportamento normal [Boriah, 2010]. As seções a seguir abordam os tipos de eventos encontrados em séries temporais: anomalias, pontos de mudança e padrões. Além disso, discorrem sobre seus significados e importância no contexto da análise de dados em séries temporais.

3.1 Anomalias

A mineração de dados é um processo que consiste em analisar dados de diferentes fontes de informação e transformá-los em dados úteis, e nesse contexto, a presença de anomalias representa um grande desafio. Anomalias podem ser qualquer informação nova, anormal, inconsistente, irrelevante, falsa ou ruidosa. A detecção de anomalias refere-se ao problema de identificar padrões nos dados que não estão em conformidade com o comportamento esperado. Hawkins [1980] definiu

anomalia como uma observação que se desvia tanto das demais que levanta suspeitas de que foi gerada por um mecanismo diferente. É um ponto de dado diferente dos demais em um conjunto de dados.

Alguns processos de mineração de dados descartam anomalias do conjunto de dados, considerando-as como ruídos ou exceções. No entanto, esses eventos raros podem ser mais interessantes na análise do que os valores que aparecem com maior frequência [Han et al., 2011]. Assim, o problema de detecção de anomalias se aplica a diversos domínios, onde se deseja encontrar eventos interessantes e incomuns nos dados. A ideia por trás dos métodos de detecção é a criação de um modelo de probabilidade, estatístico ou um algoritmo que caracterize quais dados são normais. Os desvios desse modelo são usados para identificar os valores discrepantes, ou seja, as anomalias.

Sendo observações que se destacam por parecerem não terem sido geradas pelo mesmo processo que as demais em uma série temporal, anomalias podem ser consideradas ruídos a serem removidos do conjunto de dados por não representá-los. Neste caso, podem ser modeladas como observações isoladas dos dados restantes com base em funções de similaridade ou distância [Gupta et al., 2014].

A identificação de anomalias pode ser resumida a aplicação de um *boxplot*. Esta técnica consiste em resumir os dados de uma série temporal univariada em cinco atributos. O menor valor não anômalo (Min), primeiro quartil ou quartil inferior (Q_1), segundo quartil ou mediana, terceiro quartil ou quartil superior (Q_3) e maior valor não anômalo (Max). O intervalo interquartil (IQR) é o resultado da subtração (Q_3 - Q_1). Assim é possível definir por meio da Equação 3.1 anomalias em uma série temporal y, onde os valores que estiverem fora do intervalo Min-Max são marcados como valores anômalos na referida série.

$$anomalia(y) = \{i, \forall\ i \mid y_i \notin [Q_1(y) - 1.5 \cdot IQR(y), Q_3(y) + 1.5 \cdot IQR(y)]\} \tag{3.1}$$

A Figura 3.1 ilustra este conceito, onde toda a distribuição dos dados está no formato de um *boxplot*, sendo as anomalias valores que ficaram fora dos limites inferior e superior.

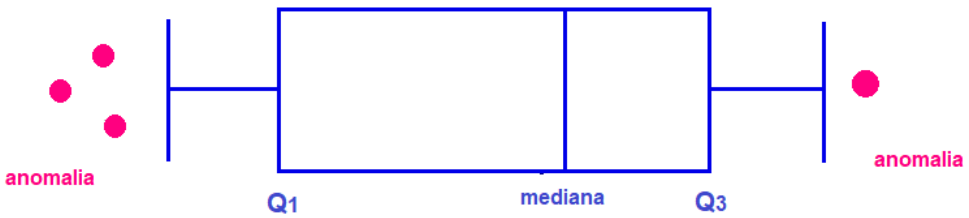


Figura 3.1: *Boxplot*, adaptado de Han et al. [2011].

As anomalias também podem fornecer informações relevantes para um determinado fim e não devem ser ignoradas ou descartadas [Han et al., 2011]. No contexto de séries temporais, existe um interesse particular na detecção de anomalias que podem representar a ocorrência de um evento

que foge ao comportamento de uma determinada série y. De modo geral, anomalias são "poucas" e "diferentes" em um conjunto de dados. Poucas porque normalmente a quantidade de valores anormais é menor que a quantidade de dados normais, e diferentes porque se comportam de maneira distinta em comparação com os dados normais. Diferentes tipos de dados podem conter diferentes tipos de anomalias e um conjunto de dados pode possuir mais de um tipo de anormalidade [Poonsirivong and Jittawiriyanukoon, 2017].

3.1.1 Tipos de Anomalias

Uma consideração importante em uma técnica de detecção é a natureza da anomalia. Chandola et al. [2009] classificam as anomalias em três categorias: anomalias pontuais, anomalias contextuais e anomalias coletivas, que são detalhadas na sequência.

Anomalias Pontuais

Uma anomalia pontual é o tipo mais simples de anomalia encontrado, sendo assim objeto de estudo da maioria das pesquisas. Em um conjunto de dados, uma anomalia é pontual se desviar significativamente dos demais valores do conjunto. Um exemplo real a ser considerado envolve fraudes em cartões de crédito, onde um determinado usuário possui um certo histórico de compras, e em um dado momento, toda sua rotina de gastos é alterada de forma significativa. A Figura 3.2 ilustra um exemplo de série temporal contendo anomalias pontuais nos últimos pontos de dados entre 1980 e 1990.

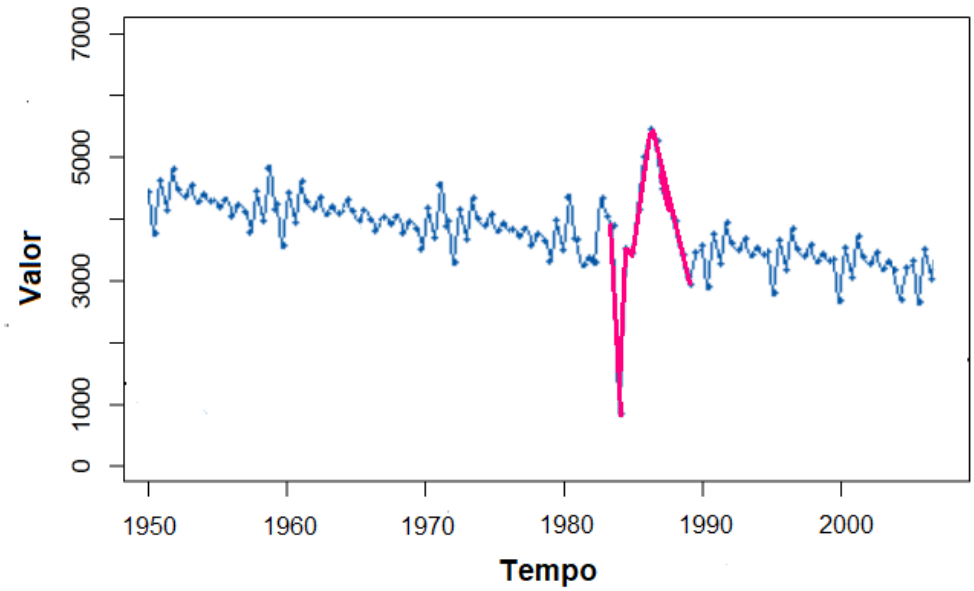


Figura 3.2: Exemplo de série temporal com anomalias pontuais.

Anomalias Contextuais

Um ou mais pontos de dados são caracterizados anomalias contextuais quando analisados dentro de um contexto especificado como parte da definição do problema. Um exemplo dado é a medição da temperatura de 35° *Celsius* no Rio de Janeiro. Considerando um período de verão, essa temperatura é definida normal. Porém, se o dado foi medido na estação de inverno, é considerado uma anomalia. Esse tipo de análise depende do contexto no qual o dado está inserido. Normalmente, os atributos dos pontos de dados em questão são divididos em dois grupos na detecção de anomalias contextuais:

- **Atributos Contextuais:** os atributos contextuais definem o contexto dos pontos de dados. No caso do exemplo da temperatura no Rio de Janeiro, os atributos contextuais podem ser o horário e a data da medição.
- **Atributos Comportamentais:** os atributos comportamentais definem as características não contextuais dos pontos de dados, sendo usado para avaliar se o valor é uma anomalia ou não. No exemplo da temperatura, os atributos comportamentais podem ser a pressão e umidade do ar.

A Figura 3.3 mostra um exemplo de uma série temporal de valores de temperatura por mês, em uma determinada região nos últimos anos, com uma anomalia contextual denominada t_1. É possível observar que a temperatura no tempo t_2 é a mesma que no tempo t_1, mas ocorre em um contexto diferente e, portanto, não é considerada uma anomalia.

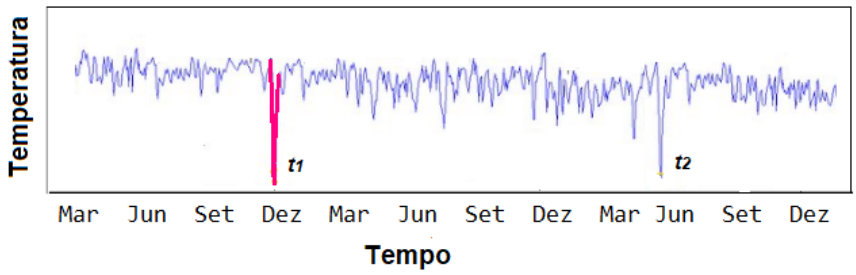


Figura 3.3: Exemplo de série temporal com anomalias contextuais.

Anomalias Coletivas

Anomalias coletivas ocorrem quando uma coleção de pontos de dados relacionados diferem em relação ao conjunto de dados. A Figura 3.4 mostra um eletrocardiograma humano, onde a região destacada denota uma anomalia porque o mesmo valor baixo existe por um tempo anormalmente longo, o que corresponde a uma contração prematura atrial.

É importante observar que essa coleção de eventos é uma anomalia, mas os eventos individuais não são anomalias quando ocorrem em outros locais na sequência de dados. Ao contrário da

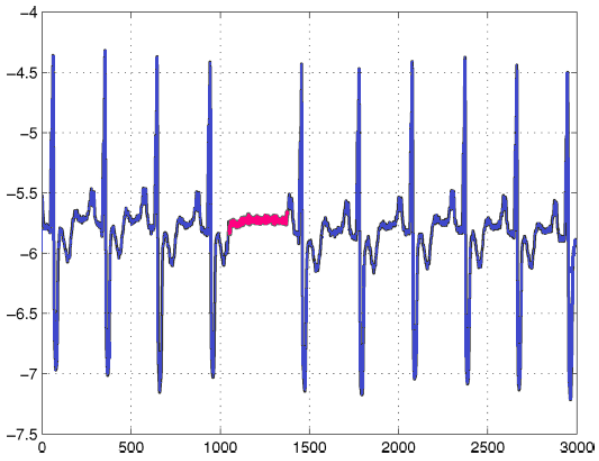


Figura 3.4: Exemplo de série temporal com anomalia coletiva, adaptado de Chandola et al. [2009].

anomalia pontual e da anomalia contextual, na anomalia coletiva é necessário observar os grupos de dados, e não apenas cada ocorrência individualmente.

3.1.2 Métodos de Detecção de Anomalias

No contexto de séries temporais, existe um interesse particular na detecção de anomalias que podem representar a ocorrência de um evento. Uma das possibilidades para detectar anomalias ocorre através do conceito de proximidade, onde o objetivo é identificar as anomalias baseadas em sua relação com outros pontos nos dados. Um outro conceito muito utilizado é o baseado em distância, sendo este mais comum em dados de baixa dimensionalidade.

A literatura apresenta diversos métodos para a detecção de anomalias, desde o método mais simples de detecção que envolve a Análise do Valor Extremo (AVE), até métodos que se baseiam em outras técnicas para fundamentar o processo, como os baseados em decomposição, Normalização Adaptativa (NA), o GARCH (*Generalized Autoregressive Conditional Heteroscedasticity*), o KNN-CAD (*K-Nearest Neighbors Conformal Anomaly Detector*) e o KNN-LDCD (*K-Nearest Neighbors Lazy Drifting Conformal Detector*). O método de decomposição adota uma abordagem que consiste em decompor a série temporal nas componentes de tendência, sazonalidade e restante [Dancho and Vaughan, 2019]. No método NA é feito uso de inércia para abordar as séries não estacionarias heterocedásticas através do cálculo da média móvel [Ogasawara et al., 2010]. O GARCH trata da detecção de anomalias de volatilidade, enquanto que o KNN-CAD e o KNN-LDCD são métodos que utilizam abordagens envolvendo a distância em relação aos k vizinhos mais próximos [Gammerman and Vovk, 2007].

Análise do Valor Extremo (AVE)

Um dos métodos mais básicos para identificar anomalias em séries temporais ocorre através da análise de valores extremos. Um valor extremo normalmente é caracterizado por valores exageradamente pequenos ou grandes, de ocorrências irregulares e raras. A teoria dos valores extremos pertence a um ramo da teoria das probabilidades que se relaciona ao comportamento das estatísticas de ordem extrema em uma amostra [Ding et al., 2019].

A modelagem de valor extremo desempenha um papel importante na maioria dos algoritmos de detecção de anomalias, quantificando com uma pontuação numérica os desvios dos pontos de dados dos padrões considerados normais. Frequentemente, muitos modelos de valor extremo usam modelos probabilísticos para pontuar a probabilidade de um ponto de dado ser ou não um valor extremo [Aggarwal, 2017].

Método da Decomposição (DE)

O método utilizado para a detecção de anomalias através da decomposição adota uma abordagem univariada. Tal abordagem consiste em um procedimento de decompor a série temporal em três componentes: tendência, sazonalidade e restante. A remoção de tendência envolve a determinação e remoção de uma tendência inerente observada no comportamento da série temporal. A remoção da sazonalidade consiste em decompor uma série em componentes, ou seja, sinais com diferentes escalas [Salles et al., 2019]. Ao extrair da série original a tendência e a sazonalidade, obtêm-se o componente restante, e sobre este componente é realizada a busca por anomalias. A remoção de tendência é descrita na Equação 3.2, onde y_t é a série temporal original, η_t é o componente de tendência e $\hat{y}_t$ é a série residual após extrair a tendência. Na sequência é extraída a componente de sazonalidade de $\hat{y}_t$, conforme a Equação 3.3, resultando em uma série final apenas de resíduos.

$$\hat{y}_t = y_t - \eta_t \tag{3.2}$$

$$\hat{y}_t = \hat{y}_t - S_t \tag{3.3}$$

Normalização Adaptativa (NA)

O método de Normalização Adaptativa é uma variação da técnica conhecida por janelas deslizantes, para lidar com séries temporais heterocedásticas não estacionárias. A abordagem NA apresenta a vantagem de retratar diferentes volatilidades, sendo um diferencial em relação às janelas deslizantes que quando normalizadas, apresentam a mesma volatilidade [Ogasawara et al., 2010].

O método consiste em efetuar três estágios: *(i)* transformação de séries temporais não estacionárias em uma sequência estacionária, criando uma sequência de janelas deslizantes que não se sobrepõe, ou seja, disjuntas; *(ii)* remoção dos *outliers*; *(iii)* aplicação da normalização de dados. O resultado desse processo consiste em dados que servirão de entrada para um método de aprendizagem, como uma rede neural artificial.

A partir de uma sequência y_t de comprimento n, com $t = \{1, 2, \cdots, n\}$, sua média móvel η_t de comprimento $n - m + 1$ e um tamanho de janela deslizante m, uma nova sequência η é definida e dividida em $n - m + 1$ janelas deslizantes disjuntas, onde ocorrerá a busca por anomalias.

Heteroscedasticidade Condicional Autoregressiva Generalizada (GARCH)

A maioria das séries temporais financeiras exibe propriedades não lineares que não podem ser capturadas pelos modelos lineares existentes. Esse fato ocorre devido à volatilidade dessas séries que varia bastante ao longo do tempo. Surge assim uma demanda para o estudo da volatilidade das séries temporais. Modelos econométricos aparecem para tratar a não linearidade dos dados, incluindo volatilidade estocástica, como o ARCH (*Autoregressive Conditional Heteroscedasticity*) e GARCH (*Generalized Autoregressive Conditional Heteroscedasticity*), sendo o último o mais conhecido e aplicado [Berlinger et al., 2015]. Na área financeira, a volatilidade está associada a risco, que no contexto de séries temporais pode indicar um evento.

Os modelos do tipo GARCH envolvem a estimativa da volatilidade com base em observações anteriores. O GARCH é um modelo de série temporal não linear, onde uma série y_t é explicada conforme a Equação 3.4, sendo μ_t o componente médio e σ_t a variância condicional. A sequência de ruído R_t é i.i.d. N$(0, 1)$, de modo que a distribuição condicional de $\eta_t = y_t - \mu_t$, dado $\eta_{t-1}, \eta y_{t-2}, \ldots$ é $N(0, \sigma_t^2)$ [Carmona, 2014].

$$y_t = \mu_t + \sigma_t R_t \tag{3.4}$$

A Equação 3.5 define a variância condicional em p períodos anteriores. Esse termo autorregressivo p modela a variância condicional dos erros quadráticos, enquanto q modela a variação do processo. A variável Φ indica quanto a última observação têm de influência na variância condicional atual, ou seja, de hoje, enquanto que Θ aponta quanto a volatilidade do período anterior deve influenciar a volatilidade hoje. Um maior impacto nos dados é sentido quanto maior for Φ, e a duração desse impacto é maior quanto mais alto for o valor de Θ.

$$\sigma_t^2 = \sigma_t + \sum_{j=1}^{p} \Phi_j \sigma_{t-j}^2 + \sum_{j=1}^{q} \Theta_j \eta_{t-j}^2 \tag{3.5}$$

Detector de Anomalia Conforme Baseado no KNN (KNN-CAD)

O KNN-CAD (*K-Nearest Neighbors Conformal Anomaly Detector*) é um método de detecção de anomalias sem modelo, que se adapta a não estacionariedade no fluxo de dados. Baseado no algoritmo KNN, tem como objetivo definir uma pontuação que é medida através da distância entre um ponto p ao seu k-ésimo vizinho mais próximo. Quanto mais distante o ponto está de seu vizinho, maior a possibilidade de ser uma anomalia. Combinado ao KNN o método também engloba o algoritmo ICAD (*Inductive Conformal Anomaly Detection*), que ignora a referência fornecida e utiliza uma janela de treinamento adequada. Assim, calcula a dissimilaridade como a soma das distâncias dos k-vizinhos mais próximos [Gammerman and Vovk, 2007].

Considerando como valores de entrada y_t, com $\{t = 1, 2, \cdots, n\}$, uma série temporal de tamanho n. A detecção de anomalias baseada no KNN usa uma distância d para quantificar o grau de dissimilaridade entre as observações. A partir de cada observação, é definida uma pontuação em relação a distância aos seus vizinhos mais próximos. Então, um valor $\epsilon \in (0, 1)$ é estimado, e se estiver abaixo de um limite de anomalia predefinido, a observação é classificada como uma anomalia.

Detector de Anomalia Conforme Baseado no KNN (KNN-LDCD)

Partindo de uma modificação no algoritmo ICAD, surge o *Lazy Drifting Conformal Detector* (LDCD), que também se adapta a não estacionariedade dos dados. É um método variante do KNN-CAD, que assim como este tem como base o algoritmo KNN, onde calcula-se uma pontuação medida através da distância entre um ponto p ao seu k-ésimo vizinho mais próximo. Porém, diferente do KNN-CAD, que calcula a dissimilaridade como a soma das distâncias dos k-vizinhos mais próximos, no KNN-LDCD a dissimilaridade é computada como a média das distâncias dos k-vizinhos mais próximos para pontuar cada observação. Nesse método as pontuações vão sendo atualizadas de acordo com uma fila de calibração, que possui o mesmo tamanho da janela de treinamento [Ishimtsev et al., 2017].

Esses métodos baseados no KNN postulam que observações consideradas normais geralmente estão mais próximas de seus vizinhos, em oposição a observações remotas, que estão mais distantes. Em ambos, a distância de *Mahalanobis* é utilizada, pois apresenta melhores resultados em séries temporais univariadas, onde considera a correlação nos dados.

3.2 Pontos de Mudança

Dados de séries temporais normalmente exibem alterações em sua estrutura, como anomalias em um eletrocardiograma ou mudanças repentinas no mercado financeiro. A análise de detecção de

mudanças busca identificar não apenas se a série temporal é uma concatenação de vários segmentos, mas também quantos pontos de mudança existem [Ding et al., 2016].

A detecção de pontos de mudança é útil na modelagem e previsão de séries temporais, além de ser encontrada em aplicações como monitoramento de condições médicas e detecção de mudanças climáticas. Um ponto de mudança representa uma transição entre diferentes estados em um processo de geração dos dados da série temporal.

Pontos de mudança são observações no tempo em que as propriedades de uma série são alteradas. Essas propriedades podem ser a média, a distribuição de probabilidade ou a variância, por exemplo. Um critério importante para a detecção de pontos de mudança é a capacidade em identificar o ponto em tempo real ou quase real [Aminikhanghahi and Cook, 2017].

Dada uma sequência de variáveis de séries temporais, a detecção de ponto de mudança pode ser definida como um problema de teste de hipóteses entre duas alternativas, onde H_0: "Nenhuma mudança ocorre" é a hipótese nula e H_A: "Uma mudança ocorre" é a hipótese alternativa. Seja $\{y_t, y_{t+1}, \ldots, y_n\}$ uma subsequência de observações de uma série temporal [Chen and Zhang, 2015; Harchaoui et al., 2009]:

- $H_0 : \mathbb{P}_{y_t} = \cdots = \mathbb{P}_{y_k} = \cdots = \mathbb{P}_{y_n}$.
- H_A : Existe $t < k^* < n$ tal que $\mathbb{P}_{y_t} = \cdots = \mathbb{P}_{y_{k^*}} \neq \mathbb{P}_{y_{k^*+1}} = \cdots = \mathbb{P}_{y_n}$.

onde $\mathbb{P}_{y_i}$ é a função densidade de probabilidade da janela deslizante iniciada no ponto y_i, e k^* é um ponto de mudança.

3.2.1 Métodos de Detecção de Pontos de Mudança

Existem diversas causas para que anomalias apareçam em um conjunto de dados: mau funcionamento de uma máquina, fraude em transações financeiras, mudança brusca no clima ou até mesmo um erro humano na inserção de dados em um sistema. Porém, existem casos em que um evento causa uma alteração no comportamento da série, divergindo do conceito de anormalidade, e essa alteração é conhecida como ponto de mudança.

O problema de detecção de ponto de mudança tem recebido muita atenção da comunidade acadêmica, pois abrange uma diversidade de problemas do mundo real. Pontos de mudança são aqueles pontos no tempo que dividem um conjunto de dados em segmentos homogêneos distintos. O problema de identificação de pontos de mudança é um dos problemas estatísticos mais desafiadores, uma vez que tanto o número de pontos de mudança quanto suas localizações são desconhecidos [Bulunga, 2012].

Ponto de Mudança Seminal (SCP)

O método SCP (*Seminal Change Point*) proposto por Guralnik and Srivastava [1999] tornou-se referência na literatura relacionada à detecção de eventos que envolvem pontos de mudança. Este adota uma abordagem univariada e segue a estratégia subsequente: para qualquer ponto no tempo, são ajustados modelos aos segmentos de dados antes e depois do ponto. Então, determina-se a existência de um ponto de mudança se o total de erros de ajuste for significativamente reduzido em comparação com o caso em que não há ponto de mudança.

Em sua versão mais simples, o método adota uma função polinomial, como uma regressão linear. Essa função é usada como modelo para ajustar os dados. O erro de ajuste é medido em termos de erros quadrados [Takeuchi and Yamanishi, 2006]. A Figura 3.5 ilustra a estratégia adotada pelo método. Neste caso, foram observados resultados promissores e computacionalmente menos custosos ao se fixar $v = |sw|/2$, sendo v o ponto central da janela e $|sw|$ o seu tamanho.

A Figura 3.5 representa uma janela sw obtida de uma série temporal y. O ponto central v da janela é selecionado, e então a série é modelada através de uma regressão linear onde são medidos os erros de ajuste antes e depois de v (erro em azul). Uma modelagem da janela inteira também é calculada (erro em vermelho). Se o valor do erro da janela inteira (vermelho) for maior que o erro da metade da janela (azul), então v provavelmente é um ponto de mudança.

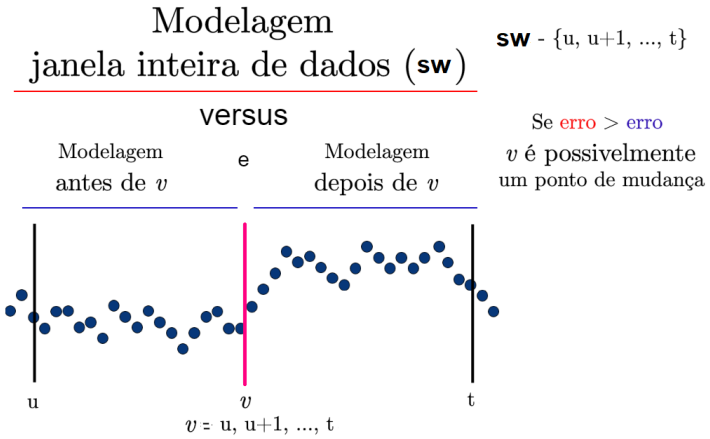


Figura 3.5: Ilustração da estratégia adotada pelo método SCP [Guralnik and Srivastava, 1999].

ChangeFinder (CF)

Takeuchi and Yamanishi [2006] propuseram um método capaz de detectar anomalias e pontos de mudança em dados de séries temporais univariadas, o qual é baseado no método de [Guralnik and Srivastava, 1999]. O trabalho apresenta a possibilidade de utilização do método em cenários *online*, especializado para dados em *streaming*, *i.e.*, dados em constante atualização. Essa especialização é particularmente interessante para aplicações baseadas em sensores que coletam dados

continuamente ao longo do tempo.

A estratégia geral do método, ilustrada na Figura 3.6, é composta de duas fases, onde na primeira são detectadas as anomalias. Um modelo de aprendizado incremental é ajustado à série temporal e é atribuída uma pontuação para cada observação. A pontuação é calculada em termos de seu desvio do modelo aprendido, que é baseado em erros quadráticos. Pontuações mais altas indicam anomalias. Na segunda fase são detectados pontos de mudança. É produzida uma nova série temporal que consiste das médias móveis das pontuações calculadas na primeira fase. Novamente um modelo de aprendizado incremental é ajustado a essa nova série, e uma pontuação dada em termos do desvio do modelo aprendido. Dessa forma, a detecção de pontos de mudança é reduzida ao problema de detectar anomalias nessa série.

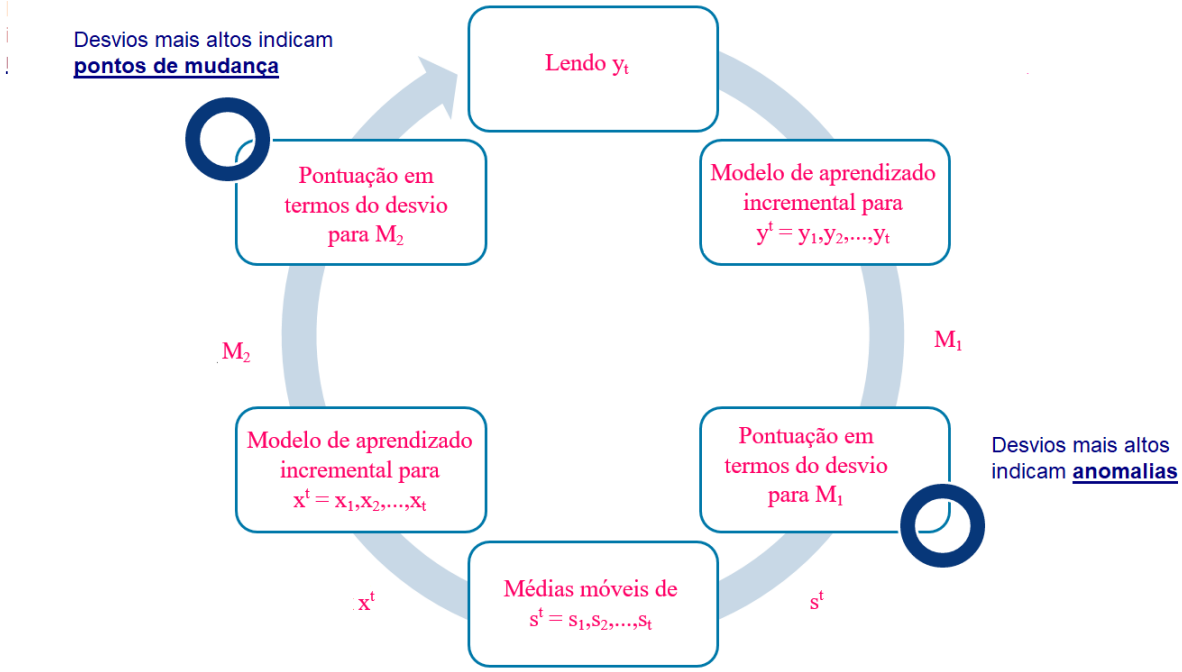


Figura 3.6: Ilustração da estratégia adotada pelo método *ChangeFinder*, adaptado de [Takeuchi and Yamanishi, 2006].

Média Móvel Exponencialmente Ponderada (EWMA)

O método EWMA (*Exponentially Weighted Moving Average*) consiste em obter uma média ponderada de todas as observações, sendo atribuído maior peso aos dados recentes, enquanto que os dados mais antigos recebem um peso menor. Este método é usado para detectar pequenas mudanças nos dados de séries temporais, e a importância das observações atuais e históricas é definida pela constante de ponderação λ. O EWMA é definido conforme a Equação 3.6, onde λ é a constante de ponderação, η_t é a média móvel exponencialmente ponderada (EWMA) e y_t é a observação [Raza et al., 2015]:

$$\eta_t = \lambda y_t + (1 - \lambda)\eta_{t-1}. \tag{3.6}$$

Média Móvel Exponencialmente Ponderada Probabilística - PEWMA

O método EWMA é comumente utilizado para detecção de pontos de mudança, o qual suaviza pequenas variações do fluxo de dados. Partindo deste conceito, Carter and Streilein [2012] desenvolveram um método probabilístico baseado no EWMA, o qual foi denominado PEWMA (*Probabilistic Exponentially Weighted Moving Average*). Tal método ajusta de forma dinâmica os limites de anomalia conforme a probabilidade da observação para detectar pontos de mudança. O PEWMA considera tanto uma média na forma de uma média móvel, quanto um desvio padrão das observações.

Como uma extensão do EWMA, o PEWMA tem uma variável de peso exponencialmente decrescente λ. O trabalho de Carter and Streilein [2012] introduziu um peso probabilístico adicional β à média móvel das amostras anteriores. Dessa forma $(1 - \beta P)$ foi escolhido para ser colocado como peso em λ, onde P representa a probabilidade. Assim, as amostras com menor probabilidade de observação pouco influenciam na estimativa atualizada. O cálculo de μ_t é apresentado na Equação 3.7, onde $0 < \lambda < 1$, P_t é a probabilidade de y_t e β é o peso atribuído a P_t.

$$\mu_t = \lambda(1 - \beta P_t)\mu_{t-1} + (1 - \lambda(1 - \beta P_t))y_t \tag{3.7}$$

Detecção de Mudança Baseada no EWMA (SD-EWMA)

O método SD-EWMA (*Shift-Detection Based on EWMA*) usa um gráfico de controle baseado no modelo EWMA para detectar ponto de mudança em séries temporais estacionárias. Um gráfico de controle mostra a representação gráfica estatística da amostra para um CEP (Controle Estatístico do Processo). O gráfico de controle EWMA reúne dados presentes e passados de maneira que uma pequena mudança na série pode ser detectada com mais facilidade e rapidez [Raza et al., 2015].

O algoritmo possui duas fases distintas: *(i)* treinamento, onde são calculados os parâmetros; e *(ii)* testes, onde ocorre a detecção de pontos de mudança. A fase de teste consiste em calcular, para cada observação da série, as estatísticas EWMA, os erros de previsão, a variância estimada, além dos limites dentro dos quais as observações devem estar. Todos valores que estiverem fora destes limites são sinalizados como pontos de mudança.

Detecção de Mudança em Dois Estágios Baseada no EWMA (TSSD-EWMA)

O método TSSD-EWMA (*Two-Stage Shift-Detection Based on EWMA*) é uma versão do SD-EWMA aplicável à séries temporais não estacionárias [Raza et al., 2015]. O TSSD-EWMA possui

duas etapas: a primeira utiliza o método SD-EWMA de modo online, onde processa de forma contínua os dados futuros do fluxo de dados. Já na segunda etapa, é utilizado um teste de hipótese estatística para validar a mudança detectada na primeira etapa, visando reduzir a quantidade de falso positivos. O teste valida a estacionariedade das subsequências, e as estatísticas são descritas na Equação 3.8.

$$D_n = sup_y \, |F_y \, - \, S_y|, \tag{3.8}$$

onde D_n é o menor limite superior de todas as diferenças pontuais, F_y é a função de distribuição acumulada e S_y é a função de distribuição esperada. Em resumo, o teste estatístico usa a diferença absoluta máxima para comparar as funções acumuladas *versus* esperadas para testar a hipótese nula.

3.3 Descoberta de Padrões ou *Motifs*

O termo *motifs* em séries temporais é usado para determinar subsequências muito semelhantes entre si. Muitos trabalhos na literatura têm se concentrado em propor algoritmos rápidos que encontram soluções aproximadas para a descoberta de *motifs*. O objetivo na descoberta de *motifs* é encontrar padrões frequentes, até então desconhecidos em uma série temporal, sem qualquer informação prévia sobre sua posição [Lin et al., 2002; Chiu et al., 2003].

Definição 7. *Segundo Campisano et al. [2018], dada uma sequência q e uma série temporal y, q é um motif em y com suporte ψ, se e somente se, q for incluído em y pelo menos ψ vezes. O comprimento de um motif q, denotado por $|q|$, é também conhecido como tamanho da palavra. A Equação 3.9 define que:*

$$\omega = sw_{|q|}(y), \quad motif(q, y, \psi) \quad \leftrightarrow \quad \exists \; R \subseteq \omega, \quad (|R| \geq \psi), \quad \forall \; w_i \in R, \quad w_i = q \tag{3.9}$$

A Equação 3.9 contém diversas informações sobre a definição de um *motif* em uma série temporal. Primeiro, é definido que a matriz ω é o conjunto de subsequências de tamanho q produzidas pela janela deslizante sobre a série temporal y. Depois, $motif(q, y, \psi)$ estabelece q como um *motif* presente em uma série temporal y, com suporte ψ se e somente se q ocorrer em y pelo menos ψ vezes.

A descoberta de *motifs* pode ser produzida sobre dados unidimensionais ou multidimensionais, além da possibilidade de ser aplicada em diferentes tipos de sequências, como temporais ou espaço-temporais. Uma propriedade importante em relação aos *motifs* é que a subsequência repetida não

é conhecida anteriormente e é descoberta ao explorar os dados por completo. Um *motifs* pode ser descoberto ao se fazer uma comparação entre subsequências obtidas a partir de janelas deslizantes. A Figura 3.7 é um exemplo de *motif* que ocorre três vezes em um determinado conjunto de dados. Em A, B e C é destacada a semelhança entre as subsequências.

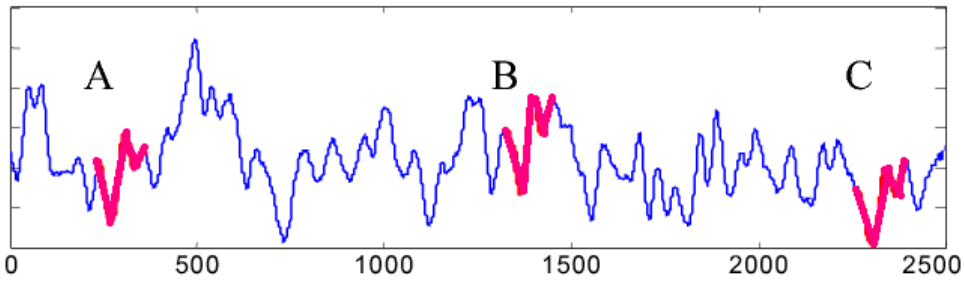


Figura 3.7: O exemplo ilustra a detecção de *motifs* em uma série temporal, adaptado de [Lin et al., 2002].

Capítulo 4

Trabalhos Relacionados

O material para o presente estudo foi obtido através de um levantamento de artigos publicados em periódicos por um mapeamento sistemático sobre detecção de eventos em séries temporais. Para conduzir a procura por trabalhos relacionados foram utilizadas *strings* de busca contendo palavras-chave sobre o assunto, visando mapear artigos de referência e contribuição ao tema abordado. Um refinamento foi aplicado visando priorizar documentos que enfatizassem a comparação de métodos de detecção de eventos. O objetivo desta revisão foi obter uma boa base para fundamentar conceitos importantes em relação ao tema em estudo, perceber os tipos de trabalhos que estão sendo conduzidos e auxiliar na escrita desta dissertação.

A base de dados *Scopus* foi utilizada para a pesquisa, a partir da seguinte *string* de busca: TITLE-ABS-KEY(“event detection” AND “time series”) AND (LIMIT-TO(LANGUAGE, “English”)) AND (LIMIT-TO(DOCTYPE, “cp”) OR LIMIT-TO(DOCTYPE, “ar”)). A busca retornou 346 artigos, para os quais alguns critérios de inclusão e exclusão foram aplicados. Os critérios de inclusão são apresentados na Tabela 4.1.

Tabela 4.1: Critérios de Inclusão.

Código	Descrição
CI-1	Trabalhos da área de Ciência da Computação e demais áreas afins que tratam detecção de eventos em séries temporais.
CI-2	Trabalhos de outras áreas que abordam de forma coerente e concisa os conceitos e aplicações de eventos.
CI-3	Trabalhos da área de Ciência da Computação e demais áreas afins que envolvam comparação de métodos de detecção de anomalias (ou *outliers*) em séries temporais.
CI-4	Trabalhos da área de Ciência da Computação e demais áreas afins que envolvam comparação de métodos de detecção de pontos de mudança em séries temporais.

Além dos critérios de inclusão, foram definidos alguns critérios de exclusão buscando filtrar melhor os resultados da busca, mirando apenas em artigos que apresentaram uma abordagem completa e com metodologias que contribuíssem para esta pesquisa. A Tabela 4.2 apresenta os critérios de exclusão considerados.

Na seleção dos artigos foram lidos os resumos dos 346 artigos resultantes da busca e selecionados

Tabela 4.2: Critérios de Exclusão.

Código	Descrição
CE-1	O artigo é uma versão mais antiga de um outro já considerado.
CE-2	Não foi possível ter acesso ao trabalho completo.
CE-3	Publicações em que o conteúdo dispõe apenas de conceitos.
CE-4	Artigos que não satisfaçam aos critérios de inclusão.

os mais coerentes com a pesquisa desta dissertação. Após um primeiro refinamento, cerca de 50% dos trabalhos foram eliminados. A exclusão desses trabalhos foi por incompatibilidade dos métodos utilizados, ou por serem trabalhos muitos semelhantes a outros já selecionados. O próximo passo foi ler a introdução e conclusão dos 121 artigos restantes e novamente separar aqueles mais relacionados a esta dissertação. Nesta segunda etapa restaram 32 artigos para leitura completa. Após esse processo, mais 11 trabalhos foram incluídos através da técnica *Snowballing*, pela necessidade de um material complementar de apoio. Nesse estágio foram consultados artigos e livros citados pelo artigo base anteriormente selecionado. Consequentemente tivemos um total de 43 trabalhos para leitura completa e suporte a pesquisa em questão. Na sequência são apresentados os artigos que compõem a revisão da literatura, os quais tiveram grande importância na elaboração desta dissertação.

Hawkins [1980] definiu anomalia como uma observação que se desvia tanto das demais chegando a sugerir que foi produzida por um mecanismo diferente. O processo de identificar anomalias consiste em analisar dados que apresentam um comportamento que se desvia do padrão esperado. Então, detectar anomalias consiste em definir uma região que apresente um comportamento considerado normal nos dados, e rotular como sendo qualquer observação que fique fora dessa região.

A comparação de diferentes métodos de detecção de anomalias em séries temporais tem sido bastante explorada na comunidade científica. Chandola et al. [2009] apresentaram um agrupamento de técnicas de detecção de anomalias em diferentes categorias, fornecendo uma compreensão fácil e sucinta das mesmas, que em sua maioria envolvem classificação baseada em vizinhos mais próximos, agrupamento e estatísticas. De forma semelhante é apresentado em [Malik et al., 2014] uma comparação de técnicas de detecção de anomalias baseada em diferentes métodos que aparecem na forma de uma pesquisa ampla e abrangente envolvendo métodos estatísticos, proximidade, distância, agrupamento, densidade e redes neurais. Essa pesquisa conclui que muitos algoritmos associados a essas técnicas apresentam bons resultados na identificação de valores anômalos.

Braei and Wagner [2020] relatam uma comparação de vinte métodos de detecção de anomalias em séries temporais univariadas, que foram divididos em três categorias: métodos estatísticos, métodos clássicos de aprendizado de máquina e métodos usando redes neurais. Os resultados obtidos apresentaram melhor desempenho em métodos estatísticos, e mostraram que a propriedade dos dados afeta o desempenho dos algoritmos. Para fornecer um melhor entendimento das diversas

técnicas de mineração de dados para detecção de anomalias, algumas abordagens híbridas são analisadas. Estas fornecem melhores resultados e superam a desvantagem de uma abordagem isolada sobre a outra, dado que uma abordagem híbrida consiste na combinação de dois métodos existentes [Agrawal and Agrawal, 2015].

Freitas [2019], em seu trabalho, apresenta um estudo comparativo de diferentes métodos para detecção de anomalias baseados na classificação dos dados. A pesquisa abrange diferentes tipos de domínio dos dados, apresentam medidas de desempenho dos métodos comparados e é discutida a relação entre diferentes domínios além dos algoritmos utilizados. Foi medida a acurácia da detecção antes e depois da remoção de anomalias para fundamentar a discussão, e os métodos aplicados para detecção foram baseados em estatística, proximidade e agrupamento.

A análise de ponto de mudança é uma ferramenta estatística que busca identificar se dados de uma série temporal possuem homogeneidade. Um método para detecção de eventos em séries temporais baseados em pontos de mudança é apresentado por Guralnik and Srivastava [1999], cujo trabalho se tornou referência na literatura envolvendo tal detecção. A abordagem ajusta modelos aos segmentos de dados antes e depois do ponto. Em seguida, a existência de um ponto de mudança é determinada se o total de erros de ajuste é significativamente reduzido em comparação com o caso em que não há ponto de mudança. Foram realizados experimentos em *datasets* sintéticos e reais, e o algoritmo apresentou bom desempenho, identificando pontos de mudança com grande precisão na versão *batch*.

A partir da metodologia de Guralnik and Srivastava [1999], um método unificador denominado *ChangeFinder* foi proposto em [Takeuchi and Yamanishi, 2006] visando detectar anomalias e pontos de mudança em séries temporais univariadas. Este método é apto a cenários de monitoramento *online*. O *ChangeFinder* é dividido em duas etapas, onde a primeira consiste em identificar *outliers* através de uma pontuação calculada sobre os desvios do modelo aprendido. Depois, são identificados os pontos de mudança, a partir de uma nova série temporal formada pelas médias móveis das pontuações obtidas na etapa anterior.

O *survey* descrito por Aminikhanghahi and Cook [2017] enumera, categoriza e compara muitos dos métodos propostos para detectar pontos de mudança em séries temporais. Os métodos examinados incluem algoritmos supervisionados e não supervisionados que foram introduzidos e avaliados. Fearnhead and Rigaill [2016] apresentam um algoritmo para detecção de pontos de mudança que é robusto à presença de *outliers*. O método minimiza o custo penalizado medindo o ajuste aos dados por uma função de perda que é quadrática por partes. Com isso, se torna robusto à presença de *outliers* arbitrariamente grandes. Um outro algoritmo, chamado *SwiftEvent* é proposto para detecção de eventos em séries temporais, o qual aprende critérios a partir de amostras rotuladas. A marcação de eventos nas séries temporais permite um aprendizado supervisionado para busca de

eventos em outras séries temporais [Gensler and Sick, 2017].

Alguns trabalhos apresentam um *framework* para aplicação dos métodos de detecção de anomalias e pontos de mudança. De Paepe et al. [2020] descreveram o *framework* SDM (*Series Distance Matrix*), que utiliza distância entre subsequências através do *Matrix Profile*, combinando diferentes medidas de distância e métodos de processamento para descobrir anomalias. Eriksson et al. [2010] propuseram o *BasisDetect*, um *framework* para detecção de anomalias em diferentes tipos de dados de redes, onde é medido o tráfego e estabelecido um comportamento considerado normal nos dados. As etapas do *BasisDetect* compreendem o aprendizado, a decomposição e a detecção das anomalias através de técnicas estatísticas. O trabalho de Hahn et al. [2020] apresenta um *framework* baseado na abordagem *bayesiana* para detectar pontos de mudança multivariados, através da projeção de uma série multivariada em uma série univariada. Calikus et al. [2020] implementaram vinte detectores obtidos através da combinação de doze métodos em um *framework* chamado SAFARI (*Streaming Anomaly Detection Framework using Reference Instances*). Talagala et al. [2020] propuseram um *framework* onde são calculados limites baseados na teoria dos valores extremos para a detecção de anomalias.

Zhang et al. [2017] propõem um *framework* cp3o (procedimento de ponto de mudança via objetivos podados) e dois algoritmos para detecção de pontos de mudança: um utilizando Estatística-E, e outro utilizando estatística de *Kolmogorov-Smirnov*. Resultados experimentais destacam o desempenho desses algoritmos em comparação com métodos de ponto de mudança paramétricos e não paramétricos, pois fornecem um bom equilíbrio entre velocidade e precisão. Lu et al. [2016] desenvolveram um *framework* para detectar pontos de mudança composto de duas fases: *(i)* estimativa de periodicidade através de um modelo AR (*Auto Regressive*) para medir anomalias, e *(ii)* a detecção. Na fase de detecção um teste estatístico é aplicado às medições. Xiong et al. [2015] apresentaram um *framework* de três etapas para detectar pontos de mudança em séries hidrológicas multivariadas: detecção de ponto de mudança para séries univariadas individuais, estimativa de distribuições marginais e detecção de ponto de mudança para estrutura de dependência.

Em um contexto online, todos os dados não estão imediatamente disponíveis, mas chegam continuamente. Algumas propostas para detecção de eventos em dados de *streaming* foram encontradas na literatura. Sreevidya [2014] apresenta uma revisão dos métodos de detecção de anomalias. Essa revisão se concentra em esclarecer o problema sobre o fluxo de dados e técnicas específicas usadas para detectar anormalidades em dados de *streaming* durante um processo de mineração. Comprovou-se que métodos individuais não são eficientes na transmissão de dados. Além disso, o método baseado em suposição pode funcionar muito bem se a suposição feita anteriormente sobre os dados estiver correta. Se as informações anteriores sobre os dados não forem conhecidas, é mais eficiente usar uma abordagem combinada para detecção de anomalias.

Métodos capazes de detectar com precisão pontos de mudança em dados de fluxo contínuo utilizam a divergência de *Kullback-Leibler* para sinalizar mudanças [Plasse and Adams, 2019; Alippi et al., 2015]. O artigo de Papataxiarhis and Hadjief. [2018] tem como objetivo o processamento *on-line* de eventos originados em redes de sensores. A abordagem analisa as variáveis monitoradas pelo sistema buscando relações entre os eventos ocorridos durante a operação.

Devido a complexidade das séries com mais de uma dimensão, muitos métodos de detecção de eventos encontrados são projetados para lidar com séries temporais univariadas. Entretanto, alguns métodos voltados para a detecção de forma multivariada são abordados na literatura e apresentados nesta revisão. Um método baseado em programação genética para detecção de eventos multivariados em fluxos de séries temporais é proposto por Xie et al. [2012]. A abordagem utiliza programas baseados em árvores com operadores especializados em fluxos de séries temporais. Cappers and van Wijk [2018] fornecem uma abordagem para análise de eventos multivariados combinando análise sequencial e de atributos em um sistema unificado. O Eventpad permite ao analista inspecionar seleções de interesse entre eventos e atributos para buscar padrões de interesse, onde são aplicados conceitos de regras, agregação de padrões e seleções.

A detecção de eventos em dados de séries temporais está presente em diversas aplicações, tanto na detecção de anomalias quanto em pontos de mudança. Na área de dados médicos, Hunter and McIntosh [1999] descrevem uma abordagem para detectar eventos multivariados em dados complexos. Os dados são provenientes de um monitor que captura o reposicionamento de uma sonda transcutânea de O_2/CO_2 em cada bebê na unidade de terapia intensiva neonatal (UTI). Resultados demonstraram 89% de precisão na identificação dos eventos. Batal et al. [2012] apresentam um *framework* para encontrar padrões preditivos de séries temporais multivariadas. Os dados clínicos foram captados de pacientes diabéticos, através da abstração temporal. Primeiro séries temporais são convertidas em sequências de abstrações temporais. Depois são construídos padrões temporais mais complexos, voltando no tempo através de operadores temporais. Resultados demonstraram que a estrutura é útil para encontrar padrões importantes na prevenção de vários tipos de distúrbios associados ao diabetes. Batal et al. [2015] desenvolveram um método para mineração de padrão em dados temporais complexos, como registros eletrônicos de saúde.

Na parte que abrange o setor de energia elétrica, métodos de detecção de eventos em monitoramento do consumo de energia elétrica são avaliados por Anderson et al. [2012]. O artigo de Yadav et al. [2018] propõe um método para detecção precisa, localização temporal e classificação de múltiplos eventos em tempo real em sistemas de energia. Em dados de séries climáticas, Naoki and Kurths [2010a,b] aplicam uma Transformação do Espectro Singular baseada na análise de componentes principais para detectar pontos de mudança em séries temporais. Bai et al. [2013] apresenta um novo *framework*, *Sevent*, para detecção automática de eventos de co-anomalias climáticos em

várias séries de temperatura. A busca por eventos em dados ambientais multivariados é proposta por García et al. [2018], onde é aplicada uma média móvel autoregressiva para medir anomalias.

Ainda dentro das aplicações, diversos trabalhos apresentaram métodos de detecção de eventos visando medir níveis de qualidade da água. Uma das referências, Liu et al. [2015] utiliza a distância euclidiana para avaliar a presença de anomalias. Em Ba and McKenna [2015] os autores comparam 4 métodos de detecção de pontos de mudança aplicando uma curva ROC (*Receiver Operating Characteristic Curve*), que distingue entre dados normais e dados contaminados, para avaliar o melhor desempenho. Em Perelman et al. [2012] é proposto um *framework* para detectar falhas na qualidade de distribuição da água a partir de séries temporais multivariadas e utilizam redes neurais artificiais para detectar possíveis *outliers*. Alkhamees and Fasli [2017] propõem um *cluster* de *outliers* para detectar eventos anômalos em dados de qualidade da água, baseados em uma distribuição binomial para determinar a probabilidade de um evento existir.

Na sísmica, Gabarda and Cristóbal [2010] apresentam um método para detectar com precisão eventos significativos em sinais sísmicos usando uma medida não supervisionada baseada na entropia local do sinal. Wu et al. [2019] desenvolveram um método de detecção de eventos baseado em aprendizado profundo, buscando identificar eventos em sinais sísmicos através de uma nova rede neural convolucional baseada em região em cascata. O problema de detecção de eventos em dados financeiros através de um limite dinâmico foi apresentado em [Alkhamees and Fasli, 2017], onde uma abordagem de mudança direcional foi utilizada para ajustar os preços baseados no dia anterior.

Diferentes métodos são apresentados na literatura para detectar eventos. Um novo método de detecção de *outliers* baseado em *cluster* semi-supervisionado (SCOD) é proposto por Liu et al. [2019]. Este método combina técnicas baseadas em densidade e agrupamento. A abordagem começa com um agrupamento que divide os dados em diferentes grupos de acordo com suas características, e depois usa a densidade para calcular o fator de anomalia, obtendo precisão na detecção. Um algoritmo para detectar pontos de mudança baseado na divergência relativa de *Pearson* é descrito em [Liu et al., 2013], o qual utiliza uma estimativa direta de razão de densidade. Truong et al. [2020] apresentam uma seleção de algoritmos para a detecção *offline* de múltiplos pontos de mudança em séries temporais multivariadas, baseados na estimativa de máxima verossimilhança, regressão linear por partes, métrica do tipo *Mahalanobis*, detecção baseada em classificação e detecção baseada em *kernel*.

A Tabela 4.3 apresenta as referências de artigos e *surveys* referentes à revisão da literatura apresentada neste trabalho, especificando o tipo de evento identificado no trabalho, o cenário a dimensão dos dados utilizados em seus resultados experimentais.

O objetivo desta dissertação é a comparação de métodos para a detecção de eventos em séries temporais, que envolvem detecção de anomalias, de pontos de mudança e ambos. A partir dos

Tabela 4.3: Classificação das referências bibliográficas por tipos de eventos, cenário e dimensão.

	Tipo				Cenário		Dimensão	
	Referência	**Ev.**	**Anom.**	**P.M.**	***Batch***	***Streams***	**Univar.**	**Multivar.**
	Frameworks							
1	Eriksson et al. [2010]		X		X		X	
2	Calikus et al. [2020]		X			X	X	
3	Talagala et al. [2020]		X			X		X
4	Batal et al. [2012]	X			X			X
5	Bai et al. [2013]	X			X			X
6	Xiong et al. [2015]			X	X			X
7	Perelman et al. [2012]		X		X			X
8	Lu et al. [2016]			X	X		X	
9	Zhang et al. [2017]			X	X		X	
10	Hahn et al. [2020]			X	X			X
11	De Paepe et al. [2020]		X		X		X	
	Comparação de Métodos							
12	Malik et al. [2014]		X		X		X	
13	Chandola et al. [2009]		X		X		X	
14	Ba and McKenna [2015]			X		X	X	
15	Agrawal and Agrawal [2015]		X		X		X	
16	Braei and Wagner [2020]		X		X		X	
17	Freitas [2019]		X		X		X	
18	Truong et al. [2020]			X	X			X
	Detecção							
19	Hunter and McIntosh [1999]	X			X			X
20	Aminikhanghahi and Cook [2017]			X	X		X	
21	Guralnik and Srivastava [1999]			X	X		X	
22	Takeuchi and Yamanishi [2006]		X	X	X		X	
23	Gabarda and Cristóbal [2010]	X			X		X	
24	Naoki and Kurths [2010b]			X	X		X	
25	Naoki and Kurths [2010a]			X	X		X	
26	Anderson et al. [2012]		X			X	X	
27	Xie et al. [2012]	X			X			X
28	Liu et al. [2013]			X	X		X	
29	Sreevidya [2014]		X			X	X	
30	Alippi et al. [2015]			X		X		X
31	Batal et al. [2015]	X			X			X
32	Liu et al. [2015]	X			X			X
33	Fearnhead and Rigaill [2016]			X		X	X	
34	Alkhamees and Fasli [2017]	X				X	X	
35	Gensler and Sick [2017]	X			X		X	
36	Cappers and van Wijk [2018]	X			X			X
37	García et al. [2018]		X		X			X
38	Yadav et al. [2018]		X			X	X	
39	Liu et al. [2019]		X		X		X	
40	Plasse and Adams [2019]			X		X	X	
41	Wu et al. [2019]	X			X		X	
42	Carreño et al. [2019]		X		X		X	
43	Papataxiarhis and Hadjief. [2018]		X			X		X

resultados obtidos e tendo em vista toda a revisão bibliográfica realizada, busca-se encontrar o método mais adequado para determinado tipo de dado. Diferentes *datasets*, sintéticos e reais, são testados, visando uma análise fundamentada da aplicação dos métodos.

Com base em toda a revisão bibliográfica realizada, foi possível analisar muitos trabalhos presentes na literatura que abordam a detecção de apenas um tipo de evento. Poucos abrangem o

assunto da detecção de eventos de forma a captar mais de um tipo, como é o objetivo deste trabalho. O diferencial desta dissertação está também em avaliar o resultado das detecções através de sete métricas, sendo duas delas implementadas para esta pesquisa. Diversas referências na literatura apresentam resultados com base apenas na métrica $F1$, ou na *acurácia* da detecção.

Capítulo 5

Metodologia

Muitos fenômenos podem ser observados e organizados como uma sequência de dados em uma linha do tempo. Esta sequência pode ser modelada como uma série temporal possibilitando descobertas de informações interessantes através da mesma. Uma área relevante na análise de séries temporais envolve a identificação de eventos, onde é observada a mudança do comportamento da série.

A aplicabilidade em diversos domínios torna o problema da detecção de eventos uma tarefa de grande importância. Entretanto, aplicar um método a um determinado conjunto de dados e avaliar o resultado dessa detecção não é uma tarefa trivial. Configurar parâmetros, validar o método e medir a qualidade dos resultados são tarefas que vão muito além de executar uma função com variáveis de valor e tempo. Para isso, torna-se necessária uma metodologia que estruture uma análise comparativa para definir qual método se aplica melhor a um determinado tipo de dado.

A metodologia proposta intenta comparar diferentes métodos de detecção de eventos em séries temporais, através de seis métricas que buscam avaliar a qualidade da detecção envolvendo anomalias, pontos de mudança e ambas. Uma metodologia válida visa tornar possível a parametrização dos métodos, de modo que a avaliação do desempenho individual seja de maneira justa e uniforme. Para a metodologia proposta nesta dissertação, cinco etapas foram adotadas:

1. Aquisição dos dados.
2. Escolha dos métodos.
3. Definição dos parâmetros.
4. Execução dos métodos
5. Métricas para avaliação dos resultados.

O desenvolvimento de cada uma dessas etapas é abordado nas seções a seguir.

5.1 Aquisição dos Dados

A aquisição dos dados envolve a escolha e definição dos *datasets* que serão utilizados na composição dos experimentos e avaliação. Neste processo, foi decidido utilizar e obter conjuntos de dados sintéticos e reais contendo uma referência rotulada. Assim, foram selecionados *datasets* que continham uma variável composta da marcação do evento, para fins de comparação do desempenho dos métodos. Os *datasets* escolhidos foram os apresentados abaixo, e serão referenciados resumidamente daqui por diante como Água, *Yahoo* e 3W.

1. **Água:** referente a dados sintéticos de qualidade da água.
2. ***Yahoo*:** referente a dados sintéticos e reais do *Yahoo*.
3. **3W:** referente a dados reais de exploração de petróleo.

Uma etapa de pré-processamento foi necessária, pois algumas séries temporais presentes nos *datasets* supracitados, contendo dados reais, possuem comportamento linear. Por este motivo não apresentam grande variação nos seus respectivos dados e com tal característica precisaram ser removidas. Ainda visando garantir o efeito comparativo desta metodologia, também foram removidas as séries que não apresentaram resultados para pelo menos três dos métodos considerados. Valores ausentes foram tratados através de uma função de omissão, que mantém os dados nas suas posições originais para preservar a continuidade da série temporal.

5.2 Escolha dos Métodos

Esta etapa envolve a escolha de métodos para detecção de eventos em séries temporais. Onze métodos de detecção foram selecionados e utilizados na avaliação experimental. Tais métodos são listados a seguir e foram explicados no Capítulo 3.

1. Decomposição (DE);
2. Normalização Adaptativa (NA);
3. *KNN based Conformal Anomaly Detector - sum* (KNN-CAD);
4. *KNN based Conformal Anomaly Detector - average* (KNN-LDCD);
5. Heteroscedasticidade condicional autoregressiva generalizada (GARCH);
6. Média Móvel Exponencialmente Ponderada (EWMA);
7. Média Móvel Exponencialmente Ponderada Probabilística (PEWMA);

8. Detecção de mudança baseada no EWMA (SD-EWMA);
9. Detecção de mudança em dois estágios baseada no EWMA (TSSD-EWMA);
10. Ponto de Mudança Seminal (SCP);
11. *ChangeFinder* (CF).

Todas as séries que passaram pela etapa de pré-processamento foram submetidas a testes em cada um dos métodos listados, buscando encontrar o melhor método para cada tipo de dado contido em cada na série temporal em estudo. Os métodos NA, SCP e CF foram implementados, e os demais obtidos de pacotes da linguagem estatística *R*.

5.3 Definição dos Parâmetros

Os métodos apresentados nesta dissertação possuem diferentes tipos de parâmetros a serem ajustados. NA e SCP necessitam apenas da definição do tamanho da janela, enquanto que no método DE é preciso ajustar tanto o valor da janela quanto o limite máximo de anomalias por conjunto de dado. Os métodos KNN-CAD e KNN-LDCD também possuem o tamanho da janela em sua configuração de parâmetros, sendo ainda essencial definir o valor da quantidade de vizinhos candidatos, o tamanho do conjunto de treinamento e um limite de anomalias fixado com um valor numérico entre 0 e 1.

Métodos dentro do contexto de média móvel, como EWMA, PEWMA, SD-EWMA e TSSD-EWMA, assim como os métodos citados no parágrafo anterior, também requerem a definição do tamanho da janela. EWMA e PEWMA ainda requisitam outros parâmetros: ponderação máxima, peso atribuído a probabilidade de uma determinada observação e limites de controle. Estes últimos também necessários nos métodos SD-EWMA e TSSD-EWMA. O peso atribuído à observação deve ser definido como zero ao executar o EWMA, pois este é uma característica específica do PEWMA. O erro de suavização constante é outro parâmetro a ser definido nos métodos SD-EWMA e TSSD-EWMA. O tamanho das subsequências a serem submetidas ao teste estatístico é exclusivo do TSSD-EWMA.

Uma correta definição e configuração dos parâmetros reflete no sucesso da busca por eventos. Para isso, é necessário ajustar corretamente o tamanho da janela, assim como definir um modelo de aprendizado incremental adequado quando necessário. O tamanho da média móvel é necessário no CF, e além deste parâmetro é possível escolher um modelo como regressão linear, ARIMA, AR ou ETS. O ETS consiste em um modelo de suavização exponencial. No GARCH é necessário a realização da definição de diversos parâmetros comparado aos demais métodos. É necessária a definição de um modelo médio, que pode ser constante, AR ou ARCH por exemplo; um modelo de

variância, tal como $sGARCH$, $fGARCH$ ou $NGARCH$; modelo de densidade condicional, como uma distribuição normal ou uma distribuição de erro generalizada [Ghalanos, 2014].

A Tabela 5.1 enumera as variáveis e suas respectivas siglas, enquanto que a Tabela 5.2 sumariza a relação método *versus* parâmetros, ou seja, quais parâmetros devem ser definidos nos métodos em estudo.

Tabela 5.1: Variáveis e suas respectivas siglas.

Sigla	Descrição
w	tamanho da janela
max_anoms	limite máximo de anomalias
mdl	modelo
m	tamanho da média móvel
$alpha0$	ponderação máxima
$beta$	peso atribuído à probabilidade da observação
l_c	limites de controle
$threshold$	erro de suavização constante
m_s	tamanho das subsequências para teste estatístico
$n.train$	tamanho do conjunto de treinamento
k	quantidade de vizinhos candidatos
l	tamanho da janela para KNNs
$threshold_a$	limite de anomalias
$mean.model$	modelo médio
$distribution.model$	modelo de densidade condicional
$variance.model$	modelo de variância

Tabela 5.2: Parâmetros necessários a definir em cada método

Métodos	**Variáveis**			
DE	max_anoms	-	-	-
NA	w	-	-	-
GARCH	$distribution.model$	$variance.model$	$mean.model$	
SCP	w	-	-	-
CF	mdl	m	-	-
EWMA	$n.train$	$alpha0$	$beta$	l_c
PEWMA	$n.train$	$alpha0$	$beta$	l_c
SD-EWMA	$n.train$	$threshold$	l_c	
TSSD-EWMA	$n.train$	$threshold$	l_c	m_s
KNN-CAD	$n.train$	$threshold_a$	l	k
KNN-LDCD	$n.train$	$threshold_a$	l	k

5.4 Execução dos Métodos

Após a definição dos parâmetros, os métodos foram executados através do *framework Harbinger* Salles et al. [2020]. Este *framework* possui a facilidade de inclusão de métodos visando uma detecção unificada de eventos de diferentes tipos em séries temporais, assim como a análise comparativa dos desempenhos de diferentes métodos de detecção aplicados. O *Harbinger* implementa e combina os resultados de alguns dos principais métodos de detecção de eventos disponíveis na literatura. Este *framework* além de permitir a inclusão de novos métodos, também oportuniza a realização da otimização de seus respectivos parâmetros. As detecções feitas através do *Harbinger* podem ser avaliadas tanto por meio de visualização gráfica dos resultados, quanto pela computação de diversas métricas de qualidade. Tais características permitem a condução de análises comparativas entre os diferentes métodos de detecção selecionados.

5.5 Métricas para Avaliar a Detecção de Eventos

Para avaliação e comparação do desempenho dos métodos em estudo (ver Seções 3.1.2 e 3.2.1) foram usadas métricas que são frequentemente utilizadas para analisar a detecção de eventos em séries temporais. Uma matriz de confusão indica os erros e acertos da predição e os compara aos resultados esperados, i.e. aos valores de referência. A Tabela 5.3 exemplifica a composição da matriz de confusão.

Tabela 5.3: Matriz de confusão

	Referência Falsa	Referência Verdadeira
Predição Falsa	VN	FN
Predição Verdadeira	FP	VP

É importante ressaltar os quatro termos apresentados que formam a matriz de confusão (Tabela 5.3), pois são fundamentais para a compreensão das métricas e entendimento de seus significados.

- **VN:** Verdadeiro Negativo - ocorrências classificadas corretamente como negativo.
- **FN:** Falso Negativo - ocorrências positivas classificadas incorretamente como negativas.
- **FP:** Falso Positivo - ocorrências negativas classificadas incorretamente como positivas.
- **VP:** Verdadeiro Positivo - ocorrências classificadas corretamente como positivo.

Neste trabalho são consideradas 7 métricas para avaliação dos métodos de detecção de eventos em estudo, pois uma maior variedade de métricas permite uma visão mais abrangente do desempenho. A saber:

- tempo de execução;
- F1;
- *balanced_accuracy*;
- acertos;
- falso positivo;
- distância *a posteriori*;
- distância *a priori*.

Cada métrica utilizada será discutida a seguir, e além destas, outras métricas fundamentais para o entendimento e compreensão das que estão sendo abordadas neste trabalho são apresentadas.

1. ***Accuracy:*** *Accuracy*, ou acurácia é uma métrica que avalia a proporção entre o número de previsões corretas e o número total de observações. Apresenta bons resultados apenas quando o número de observações positivas e negativas está equilibrado. A *Accuracy* pode ser definida como apresentado na Equação 5.1:

$$accuracy = \frac{VP + VN}{(VP + FP + VN + FN)} \tag{5.1}$$

   A métrica mais simples é a *accuracy*, que é simplesmente a taxa de valores classificados corretamente. Uma precisão de 100% significa que as previsões são exatamente as mesmas que as referências verdadeiras. Esta parece ser uma escolha natural, mas existem várias desvantagens em usar apenas a acurácia como métrica de avaliação, dado que é muito sensível a dados desequilibrados.

2. ***Precision:*** *Precision*, ou precisão é definida como o número de verdadeiros positivos (VP) sobre o número de verdadeiros positivos mais o número de falsos positivos (FP), conforme mostrado na Equação 5.2. Pode ser descrito como a veracidade dos eventos detectados sendo realmente eventos, e não falsos eventos. Em outras palavras, é uma métrica que avalia dentre todas as observações identificadas como positivas, quantas estavam corretas.

$$precision = \frac{VP}{(VP + FP)} \tag{5.2}$$

3. ***Recall:*** *Recall* é definida como o número de verdadeiros positivos (VP) sobre o número de verdadeiros positivos mais o número de falsos negativos (FN), conforme exibido na Equação 5.3. Outra palavra para descrever *Recall* pode ser completude, que representa quantos dos

eventos reais podem ser identificadas por um modelo específico. Em resumo, é a métrica que avalia dentre todas as ocorrências positivas marcadas como VP, quantas estavam corretas.

$$recall = \frac{VP}{(VP + FN)} \tag{5.3}$$

4. ***F1:*** *F1* consiste em uma média harmônica entre *precision* e *recall*, conforme visto na Equação 5.4. O valor retornado varia entre 0 e 1, sendo que quanto maior o valor de F1 melhor é o resultado.

$$\frac{2 * precision * recall}{precision + recall} \tag{5.4}$$

   O $F1$ retorna um valor compreendido no intervalo $[0, 1]$, e demonstra o quão preciso e robusto é o método, ou seja, quantas observações ele classifica corretamente e quantas ele deixa de classificar porque eram difíceis de rotular. No entanto, é preciso ter cuidado ao avaliar um método com base apenas no $F1$, pois torna-se inapropriado em casos onde há um desequilíbrio entre valores de referência positivos e negativos.

5. ***Sensitivity:*** *Sensitivity*, ou sensitividade, é a quantidade de verdadeiros positivos (VP) sobre o número de verdadeiros positivos mais o número de falsos negativos (FN), conforme mostrado na Equação 5.5. Seu cálculo é semelhante ao *Recall*, onde ambos apresentam os mesmos resultados. De modo geral, descreve-se como a proporção de ocorrências positivas identificadas corretamente.

$$sensitivity = \frac{VP}{(VP + FN)} \tag{5.5}$$

6. ***Specificity:*** *Specificity*, ou especificidade, é a quantidade de verdadeiros negativos (VN) sobre o número de falsos positivos (FP) mais o número de verdadeiros negativos (VN), conforme mostrado na Equação 5.6. Em síntese, representa a proporção de ocorrências negativas identificadas corretamente.

$$specificity = \frac{VN}{(FP + VN)} \tag{5.6}$$

7. ***Balanced_accuracy:*** *balanced_accuracy* representa uma média simples entre *sensitivity* e *specificity*.

$$balanced\_accuracy = \frac{(sensitivity + specificity)}{2} \tag{5.7}$$

Além das métricas listadas, propomos a avaliação de outras duas métricas baseadas na distância do evento detectado ao evento real presente na série. Seja $eventos = \{e_1, e_2, \cdots, e_v\}$ o conjunto de eventos detectados pelo método selecionado, $D = \{D_1, D_2, \cdots, D_j\}$ o conjunto contendo todos os *datasets* selecionados e $D_{ref} = sub\text{-}dataset\{D, tempo, referencia\}$ o conjunto de eventos reais presentes na série temporal, buscamos encontrar os eventos detectados mais próximos do primeiro

evento que existe na série. Escolher o primeiro evento para medir essas distâncias significa entender melhor onde o método consegue detectar quando o problema está começando. Por exemplo, o momento em que uma máquina começa a apresentar defeitos, ou mesmo o instante em que uma pessoa com problemas de saúde apresenta um agravamento (ou mudança) em seu estado clínico.

Para isso, o conjunto de eventos detectados denominado *eventos* foi dividido em duas partes: uma contendo os eventos detectados antes da ocorrência do primeiro evento de referência na série, e a outra contendo os eventos que estão localizados após. Como consequência, ficou definido como $D_{priori} = D_{priori(1)}, D_{priori(2)}, \ldots, D_{priori(u)}$ o conjunto contendo os eventos detectados antes da referência e $D_{poster} = D_{poster(u+1)}, D_{poster(u+2)}, \ldots, D_{poster(v)}$ o conjunto dos eventos detectados após a referência. A seguir descrevemos essas distâncias, denominadas *a posteriori* e *a priori*.

1. **Distância a *posteriori*:** A métrica denominada distância a *posteriori* foi incluída com o objetivo de medir o valor absoluto (va) entre a distância do primeiro evento de referência marcado na série (e_1) até o evento posterior detectado pelo método ($poster(u+1)$), conforme Equação 5.8.

$$posteriori = va[D_{poster(u+1)} - e_1] \tag{5.8}$$

2. **Distância a *priori:*** De modo semelhante a medida anterior, a métrica da distância a *priori* tem como objetivo calcular a distância do primeiro evento de referência (e_1) até o primeiro evento anterior a ele detectado pelo método ($priori(u)$), conforme Equação 5.9.

$$priori = va[e_1 - D_{priori(u)}] \tag{5.9}$$

5.6 Comparação dos Métodos

Para avaliação e validação dos métodos de detecção de eventos, foram selecionados onze métodos que envolvem detecção de anomalias, detecção de pontos de mudança e ambas. Também foram selecionadas sete métricas para medir o desempenho dos métodos, além do tempo computacional. Para a avaliação foram selecionados três conjuntos de dados, de contextos e aplicações diferentes: Água, com dados medidos por sensores para avaliar a qualidade da água, *Yahoo*, com dados reais e sintéticos do tráfego de dados do *Yahoo* e 3W, um *dataset* real contendo eventos indesejáveis em processos de exploração de petróleo, onde os eventos foram marcados por especialistas.

Uma metodologia para comparar diferentes métodos de detecção de eventos pode ser estruturada da seguinte maneira: seja $D = \{D_1, D_2, \cdots, D_{max}\}$ o conjunto contendo todos os *datasets* selecionados para o estudo de diferentes domínios, sendo max a quantidade de conjuntos de dados

selecionados. Seja $met = \{met_1, met_2, \cdots, met_{maxk}\}$ o conjunto de diferentes métodos de detecção de eventos, sendo $maxk$ a quantidade de métodos que serão utilizados na análise. O resultado da execução dos métodos é dado através do conjunto $eventos = \{e_1, e_2, \cdots, e_v\}$, dado que v corresponde a quantidade de eventos detectados pelo método selecionado. O conjunto de métricas disponíveis é dado por $mtr = \{mtr_1, mtr_2, \cdots, mtr_{maxz}\}$, sendo $maxz$ a quantidade de métricas disponíveis. Por fim, o conjunto $resultado = \{r_1, r_2, \cdots, r_z\}$, onde z representa a quantidade de métricas a serem calculadas. O Algoritmo 1 apresenta o processo abordado na metodologia.

Algorithm 1 EDMC(D, met, mtr)

```
1:  for i ← 1 to max do
2:      D_i ← preprocessamento(D)
3:      D_t(i) ← subdataset(D_i, tempo, valor)
4:      D_ref(i) ← subdataset(D_i, tempo, referencia)
5:      for j ← 1 to maxk do
6:          eventos_i,j ← calculaMet(D_t(i), D_ref(i), met_j)
7:          for z ← 1 to maxz do
8:              resultado_z ← calculaMtr(eventos_i,j, mtr_z)
9:              df ← combina(resultado_z)
10:         end for
11:     end for
12: end for
13: return df
```

O Algoritmo 1 apresenta o pseudocódigo com o processo de comparação de métodos para detecção de eventos proposta nesta dissertação. Como dados de Entrada, são recebidos o conjunto de *datasets* D, os métodos met que serão comparados e as métricas mtr que serão utilizadas na avaliação de desempenho dos resultados. A função retorna um *dataframe* df com os eventos detectados e as métricas calculadas. Na linha 1 uma estrutura de repetição é criada para percorrer todos os *datasets* em estudo. Uma fase de pré-processamento é aplicada na linha 2. Nessa fase foram excluídas do estudo comparativo séries que apresentam comportamento linear ou aquelas, para as quais, são impossível de se obter resultados de pelo menos três métodos.

Na linha 3, um *sub-dataset* $D_{t(i)}$ é selecionado a partir de D_i, onde as variáveis presentes consistem em uma de tempo e outra do valor a ser analisado. Seguindo a mesma ideia, a linha 4 apresenta outra seleção do *dataset* D_i, porém contendo valores de tempo e os eventos rotulados (referência) conforme ilustrado na Figura 5.1. Após a preparação dos parâmetros necessários, um método é selecionado na linha 5. Então, é passado para a execução na linha 6 onde a função $calculaMet$ executa o método selecionado juntamente com os *sub-datasets* $D_{t(i)}$ e $D_{ref(i)}$. Na linha 8 o resultado obtido após a execução do método é passado para a função seguinte, $CalculaMtr$, onde serão computadas as métricas. A Figura 5.2 mostra um resultado de detecção obtido pelo algoritmo EDMC (*Event Detection Methods Comparison*). Estes resultados mostram a classificação

dos eventos em *anomaly* (anomalias) e *change point* (ponto de mudança).

test 24908 obs. of 2 variables

Filter

	timestamp	P.JUS.CKGL
1	2014-03-19 04:04:53	16570620
2	2014-03-19 04:04:54	16570620
3	2014-03-19 04:04:55	16570620
4	2014-03-19 04:04:56	16570620
5	2014-03-19 04:04:57	16570620
6	2014-03-19 04:04:58	16570620
7	2014-03-19 04:04:59	16570620
8	2014-03-19 04:05:00	16570620
9	2014-03-19 04:05:01	16570620
10	2014-03-19 04:05:02	16570620

reference 24908 obs. of 2 variables

Filter

	timestamp	event
1	2014-03-19 04:04:53	FALSE
2	2014-03-19 04:04:54	FALSE
3	2014-03-19 04:04:55	FALSE
4	2014-03-19 04:04:56	FALSE
5	2014-03-19 04:04:57	FALSE
6	2014-03-19 04:04:58	FALSE
7	2014-03-19 04:04:59	FALSE
8	2014-03-19 04:05:00	FALSE
9	2014-03-19 04:05:01	FALSE
10	2014-03-19 04:05:02	FALSE

Figura 5.1: Sub-*datasets test* e $reference$.

Filter

	time	serie	type
1	2018-04-26 14:51:09	P.PDG	anomaly
2	2018-04-26 14:51:10	P.PDG	anomaly
3	2018-04-26 14:51:11	P.PDG	anomaly
4	2018-04-26 14:51:12	P.PDG	anomaly
5	2018-04-26 14:51:13	P.PDG	anomaly
6	2018-04-26 14:51:14	P.PDG	anomaly
7	2018-04-26 14:51:15	P.PDG	anomaly
8	2018-04-26 14:51:16	P.PDG	change point
9	2018-04-26 14:51:17	P.PDG	anomaly

Figura 5.2: Exemplo de *dataframe* contendo o resultado de detecção de eventos.

Após a obtenção dos resultados, o *loop* mais interno (linha 7) percorre as métricas, do conjunto de métricas definido, para a avaliação do desempenho dos métodos. Então, cada métrica f é aplicada aos valores armazenados em $eventos_{i,j}$. O resultado é armazenado conforme a linha 8. Por fim, os resultados são armazenados em um *dataframe* para a comparação posterior (linha 9).

A partir dos resultados que este processo fornece, é possível comparar qual método obteve melhor desempenho através de uma ou mais métricas. Selecionar mais de uma métrica permite uma comparação mais coerente e de melhor entendimento dos resultados, outorgando maior confiabilidade na análise do comportamento de determinado método em um *dataset* definido. Balancear e analisar diferentes métricas permite avaliar de maneira mais abrangente a qualidade das detecções de eventos. Para complementar, uma análise visual através dos gráficos ajuda a trazer perspectivas diferentes dos resultados obtidos com a análise quantitativa obtida a partir das métricas, enquanto que a análise qualitativa busca uma melhor compreensão do desempenho dos métodos.

Neste capítulo foi apresentada a metodologia proposta para comparação dos métodos de detecção de eventos em séries temporais abordados. O objetivo é avaliar o desempenho dos métodos selecionados considerando diferentes tipos de dados e diferentes comportamentos de série. Nesta dissertação abordamos onze métodos (Seção 5.2) e sete métricas (Seção 5.5) que foram utilizadas para auxiliar na compreensão e análise dos resultados.

Capítulo 6

Resultados

Neste capítulo serão apresentados os resultados dos testes visando a comparação dos diferentes métodos para detecção de eventos. Os métodos foram submetidos a uma análise em diferentes *datasets* sintéticos e reais, onde foi avaliado o desempenho desses métodos. As próximas subseções apresentam os conjuntos de dados utilizados, os parâmetros selecionados e ainda descrevem os resultados sob as seguintes perspectivas: *(i)* análise de uma única série temporal didática, *(ii)* análise do *dataset* da qualidade da água, *(iii)* análise do *dataset* do *Yahoo*, *(iv)* análise do *dataset* $3W$ e *(v)* uma análise geral de todos os *datasets*. Após são discutidos alguns dos principais resultados obtidos durante a análise.

6.1 Ambiente de Desenvolvimento e Testes

Os testes foram executados em uma máquina compartilhada, cuja configuração consiste em um processador Intel Core $i7$ com 16 cores, 128GB de RAM e sistema operacional *Ubuntu* 20.04.

6.2 *Datasets*

Para avaliar os métodos de detecção de eventos, três *datasets* de referência da literatura foram selecionados:

1. *dataset* de qualidade da água, contendo dados de séries temporais coletados por sensores a cada minuto;
2. dados reais do *Yahoo* sobre tráfego de serviços com medições observadas a cada hora, e sintéticos com sazonalidade variados por hora;
3. dados reais do 3W contendo eventos reais indesejáveis em processos de exploração poços de petróleo.

Além dos *datasets* selecionados, uma série sintética não estacionária foi utilizada para apresentar resultados no formato de uma única série temporal.

6.2.1 Conjunto de Dados Sintético: Série Não Estacionária

A série não estacionária foi desenvolvida para facilitar a visualização e entendimento das mudanças de comportamento de uma série temporal. É composta por 1000 observações, como pode ser observado através da Figura 6.1. A sequência de 0 a 200 representa uma série estacionária, que compreende um processo estocástico onde a média e a variância são constantes. A partir da observação 201, as demais sequências violam as restrições de estacionariedade e, por isso, são consideradas não estacionárias. De 201 a 400 é observado um exemplo onde a média cresce linearmente, dando origem a sequência de série temporal com tendência linear determinística [Salles et al., 2019].

Quando a média é diferente em alguns momentos da série ocorrem quebras estruturais conhecidas por mudanças de nível, como é possível observar no intervalo de 401 a 600. De maneira análoga às mudanças de nível, as séries heterocedásticas apresentam diferentes propriedades de variação ao longo da série temporal, como de 601 a 800. A heterocedasticidade surge de mudanças no ambiente que aumentam ou diminuem a volatilidade das observações de séries temporais ao longo do tempo. Por fim, de 801 a 1000 é apresentada uma sequência contendo uma série temporal não estacionária por diferenciação, sendo séries não estacionárias que apresentam raízes unitárias. Uma raiz unitária implica que as séries sofrem a influência de componentes de longo prazo ou tendências estocásticas, e essas tendências são removidas através do processo de diferenciação [Salles et al., 2019]. A Figura 6.1 ilustra a série completa, onde é possível observar os diferentes tipos de não estacionariedade.

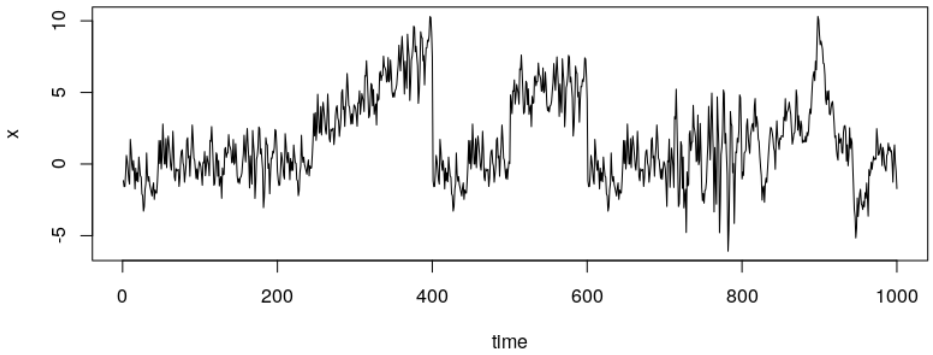


Figura 6.1: Série temporal não estacionaria utilizada para avaliação dos métodos.

6.2.2 Conjunto de Dados Sintético: Qualidade da Água

Os exemplos apresentados a seguir utilizam um conjunto de dados criado para o *GECCO Challenge 2018* [Rehbach et al., 2018], uma competição anual para pesquisadores e profissionais de Inteligência Computacional (CI) presentes na conferência GECCO (*The Genetic and Evolutionary Computation Conference*), organizada pela *Association for Computing Machinery* (ACM). O desafio GECCO de 2018 teve como tema a Internet das Coisas (IoT), particularmente o problema de Detecção Online de Anomalias para Qualidade da Água Potável. O conjunto de dados utilizado é

originado de sensores que medem variáveis relacionadas à qualidade da água obtidos da empresa *Thüringer Fernwasserver-sorgung*, onde um intervalo de 1500 pontos de dados foi selecionado e modificado para compor o *dataset* utilizado nos experimentos. Tal variação nos valores tornou-se necessária dado que as medições originais dos sensores apresentaram muita estabilidade. O conjunto de dados contém 72 eventos identificados e está disponível no pacote-R **EventDetectR** [Moritz and Rehbach, 2020]. A Tabela 6.1 mostra as informações coletadas.

Tabela 6.1: Variáveis do conjunto de dados relacionado à qualidade da água.

Nome	**Descrição**
Time	Momento da medição, no formato: aaaa-mm-dd HH:MM:SS
Tp	Temperatura da água, em graus C
Cl	Quantidade de dióxido de cloro na água, em mg/L
pH	Valor de PH da água
Redox	Potencial Redox, em mV
Leit	Condutividade elétrica da água, em μS/cm
Trueb	Turvação da água, em NTU
Cl_2	Segunda medição de quantidade de dióxido de cloro, em mg/L
Fm	Fluxo de água na linha 1
Fm_2	Fluxo de água na linha 2
EVENT	Marcação do evento

Resultados experimentais foram obtidos com todas as variáveis (séries), e serão discutidos nas próximas seções seu comportamento diante dos diferentes métodos.

6.2.3 Conjunto de Dados Sintético: Dados do *Yahoo*

Criado pela equipe de Ciências da Mídia do *Yahoo Labs* para testar e validar algoritmos, o *dataset* do *Yahoo* é disponibilizado através do programa de compartilhamento de dados *Webscope*. Consiste de séries temporais com anomalias marcadas, onde uma parte do *dataset* é sintético, e a outra parte é baseada no tráfego real dos serviços do *Yahoo* contendo anomalias rotuladas manualmente por editores [Webscope, 2015]. Os *datasets* são divididos em quatro *benchmarks*:

1. *A*1**:** composto pelos dados reais;
2. *A*2**:** composto por dados sintético contendo anomalias marcadas;
3. *A*3**:** composto por dados sintéticos com sazonalidade também possui marcações de anomalias;
4. *A*4**:** composto por dados sintéticos com anomalias e pontos de mudança marcados.

A Tabela 6.2 mostra as informações contidas nos arquivos que contém as séries temporais.

Tabela 6.2: Variáveis do conjunto de dados do *Yahoo* nos *benchmarks* $A1$, $A2$, $A3$ e $A4$.

Benchmark	Nome	Descrição
$A1$, $A2$	timestamp	Variável de contagem numérica, começando por 1.
	value	Valor da medição que compõe a série temporal.
	is_anomaly	Valor de referência, indicando se existe (1) ou não (0) anomalia.
$A3$, $A4$	timestamp	Variável de tempo no formato *timestamp*.
	value	Valor da medição que compõe a série temporal.
	anomaly	Valor de referência, indicando se existe (1) ou não (0) anomalia.
	changepoint	Valor de referência, indicando se existe (1) ou não (0) ponto de mudança.
	trend	Variável referente a tendência.
	noise	Variável referente ao ruído
	seasonality1	Valor de sazonalidade.
	seasonality2	Valor de sazonalidade.
	seasonality3	Valor de sazonalidade.

6.2.4 Conjunto de Dados Reais: 3W

O conjunto de dados 3W [Vargas et al., 2019] é o primeiro conjunto de dados realista e público, contendo raros eventos reais indesejáveis em processos de exploração de poços de petróleo. É composto por três tipos de séries temporais que são determinadas por suas fontes: real, simulada e desenhada à mão. Séries reais representam o realmente ocorrido nos poços da Petrobras durante a exploração de petróleo. O uso de séries simuladas e desenhadas à mão visa fundamentalmente diminuir o desequilíbrio do conjunto de dados formado inicialmente apenas por dados reais. As observações destas séries foram classificadas por profissionais da área como representativas de um evento indesejável ou não. Cada série apresenta-se rotulada com um único código associado à operação normal ou à categoria de evento indesejável observado na série. Nenhuma série contém mais de um evento indesejável. A Tabela 6.3 apresenta os diferentes tipos de eventos indesejáveis.

Tabela 6.3: Tipos de eventos indesejáveis do conjunto de dados $3W$.

Tipo	**Descrição do evento**
1	Aumento abrupto da BSW
2	Fechamento espurioso da DHSV
5	Perda rápida de produtividade
6	Restrição rápida em PCK
7	Escala em PCK
8	Hidrato na linha de produção

Os conjuntos de dados selecionados para esta pesquisa são resumidos na Tabela 6.4, onde é descrita para cada *dataset* a quantidade de observações, o número de variáveis, assim como o número referente às observações que são classificadas como eventos em cada *benchmark* e a porcentagem de

eventos em relação a quantidade de observações. Para os *datasets Yahoo* e 3W foi calculada uma média desses valores, dado que tais conjuntos de dados possuem uma grande quantidade de séries para cada tipo.

Tabela 6.4: Descrição dos *datasets* selecionados.

	Variáveis	**Observações**	**Eventos**	**% de eventos**	**MinMax Obs.**	**MinMax Ev.**
			Água			
	11	1500	72	5%	1500-1500	72-72
			Yahoo			
A1	3	1416	25	1%	741-1461	0-227
A2	3	1421	5	0,35%	1421-1421	1-9
A3	9	1680	9	0,53%	1680-1680	1-16
A4	9	1680	8	0,48%	1680-1680	1-16
			3W			
Tipo 1	10	23659	1	0,004%	10751-49394	1-1
Tipo 2	10	7120	1	0,014%	1703-17923	1-1
Tipo 5	10	20554	1	0,005%	1440-48905	1-1
Tipo 6	10	9035	1	0,011%	1079-24370	1-1
Tipo 7	10	79847	1	0,001%	44150-126263	1-1
Tipo 8	10	30364	1	0,003%	15688-51199	1-1

6.3 Definição dos Parâmetros

Os parâmetros foram definidos para que a comparação ocorresse da forma mais justa, tanto quanto possível, entre todos os métodos. Para isso, três diferentes tamanhos de janela foram testados nas duas primeiras séries válidas de cada conjunto de dados: $w = \{10, 50, 100\}$. Conforme os resultados apresentados na Tabela 6.5, o valor de w que apresentou melhor resultado foi o escolhido para tamanho da janela nos experimentos realizados. A escolha do tamanho da janela seguiu a seguinte estrutura: três tamanhos distintos foram avaliados em três diferentes métodos sendo um de cada tipo. Os métodos selecionados envolvem: média móvel através da NA; distância, sob a execução do KNN-CAD e média móvel exponencialmente ponderada, com o EWMA. Foram elegidas as duas primeiras séries temporais de cada conjunto de dados para este estudo, e as métricas que fundamentaram a comparação foi a quantidade de acertos e a distância *a priori*. Para cada série em cada método foi destacado o melhor valor obtido, e então a soma destes valores destacados indicam o valor de janela vencedor. Campos marcados com asterisco (*) significam que nesta combinação (série, método, tamanho de janela) o método não executou.

Ao analisar a Tabela 6.5, foram destacados os melhores valores para cada série, e agrupados por tamanho da janela w com o objetivo de definir qual o tamanho mais adequado para estes dados. Janelas de tamanho 10 tiveram 5 marcações, enquanto que as de tamanho 50 resultaram em 8 melhores valores obtidos na execução. Janelas de tamanho 100 mostraram não ser as mais adequadas, pois apenas 4 melhores resultados foram computados.

Tabela 6.5: Testes para definição do tamanho da janela w.

			NA			**KNN-CAD**			**EWMA**		
			Tamanho da janela								
		Série	**10**	**50**	**100**	**10**	**50**	**100**	**10**	**50**	**100**
Acertos	**Água**	1	22	30	16	*	0	0	2	1	1
		2	0	72	72	*	3	3	4	4	3
	Yahoo	1	2	2	2	*	1	0	2	2	2
		2	16	15	14	*	1	1	6	6	6
	3W	1	0	0	0	*	1	0	0	0	0
		2	0	0	0	*	0	0	0	0	0
			NA			**KNN-CAD**			**EWMA**		
			Tamanho da janela								
		Série	**10**	**50**	**100**	**10**	**50**	**100**	**10**	**50**	**100**
Distância a priori	**Água**	1	55	55	55	*	6	-	-	-	-
		2	-	-	-	*	6	-	-	-	-
	Yahoo	1	3	9	8	*	24	9	35	35	35
		2	216	433	445	*	19	280	89	89	89
	3W	1	-	-	-	*	11	1	19	19	19
		2	-	-	-	*	226	213	32	32	32

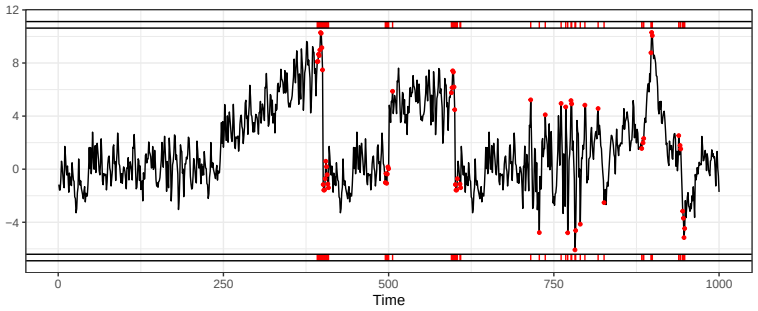
6.4 Análise e Testes dos Resultados

Os conjuntos de dados sintéticos e reais adotados foram sujeitos a um processo de detecção de anomalias e pontos de mudança através da execução dos 11 métodos abordados nesta dissertação. Através da matriz de confusão e das métricas $F1$, *balanced_accuracy*, distância *a posteriori*, distância *a priori* e *tempo* de execução é possível avaliar a qualidade da detecção. Os testes foram implementados no *framework Harbinger*, sendo que nos gráficos apresentados as marcações em vermelho indicam anomalias detectadas pelo método adotado, linhas tracejadas cinza indicam pontos de mudança detectados, marcações azuis indicam os eventos reais identificados no conjunto de dados e marcações verdes apontam para a coincidência entre a anomalia detectada e o evento real.

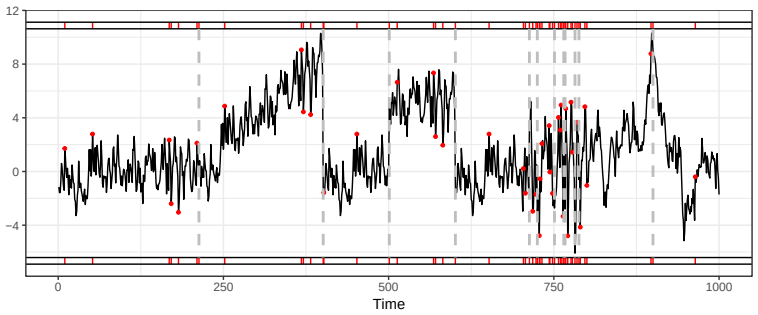
6.4.1 Análise da Série Temporal Didática

Para a análise de uma série temporal didática foi escolhida uma série não estacionária, pois apresenta inicialmente um comportamento estacionário, depois mudanças de níveis e não estacionariedade. Essa série contém eventos marcados que ajudam a avaliar o desempenho dos métodos.

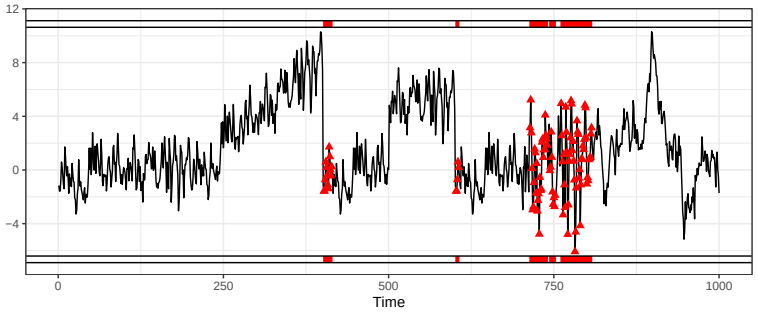
Através da Figura 6.2(a) é possível observar anomalias detectadas pelo método NA. Este método é especializado em séries temporais não estacionárias e heterocedásticas, conseguindo detectar eventos onde ocorreram mudanças significativas na média ou variância de cada subsequência. A análise do gráfico representado pela Figura 6.2(b) permite observar as marcações em linhas tracejadas cinzas, que assinalam as mudanças de níveis identificadas como pontos de mudança no método CF. Anomalias também são detectadas, sendo o diferencial deste método a detecção de dois tipos de



(a) Eventos detectados por NA.



(b) Eventos detectados por CF.



(c) Eventos detectados por GARCH.

Figura 6.2: Detecções de eventos na série didática por diferentes métodos.

eventos.

Ao se comparar as detecções nas Figuras 6.2(a) e 6.2(b) é possível observar que, apesar de adotarem abordagens diferentes, NA e CF apresentam detecções coincidentes. Neste caso, NA conseguiu detectar pontos de mudança que foram identificados de forma incorreta como anomalias. A análise dos resultados de CF contribui, portanto, para uma melhor compreensão da natureza dos eventos detectados por NA. O mesmo pode ser dito ao comparar as detecções de CF com as produzidas pelo método GARCH, apresentadas na Figura 6.3(c). Analogamente, a análise dos resultados obtidos pelo GARCH possibilita a compreensão de que o conjunto de pontos de mudança detectados por CF em torno da observação y_{750} são na verdade anomalias de volatilidade.

6.4.2 Análise do *dataset* Água

A análise do *dataset* da qualidade da água baseou-se em testes com todas as 9 séries presentes. Para apresentar os resultados obtidos na comparação dos métodos, os valores foram agrupados em diferentes tabelas. Cada tabela contém os valores resultantes para cada uma das 7 métricas selecionada. A escolha da melhor métrica para um determinado conjunto de dados depende muito do interesse na análise que está sendo conduzida. As 7 métricas aqui apresentadas servem para apoiar uma tomada de decisão para escolher o melhor método de detecção, pois cada uma propõe um objetivo diferente: $F1$ e *balanced_accuracy* medem a precisão da detecção, enquanto que *acertos* determina a quantidade de VP detectados. As medidas referentes as distâncias (*a posteriori* e *a priori*) ajudam a perceber o atraso na detecção, e o quanto o método foi capaz de prever a ocorrência do evento. Desse modo, esta dissertação apresenta diversas métricas que expõem diferentes percepções visando uma abordagem mais ampla dos resultados e possibilitando a condução de diferentes análises. A Tabela 6.6 apresenta os resultados através da métrica $F1$.

Tabela 6.6: Resultados obtidos com $F1$ para o *dataset* da Água.

Método	**Variáveis**								
	Tp	Cl	pH	Redox	Leit	Trueb	Cl_2	Fm	Fm_2
DE	0,8821	*	1,0000	1,0000	0,8804	0,9901	*	*	0,9573
NA	0,8771	0,9402	1,0000	0,9865	0,9043	0,9862	0,9740	0,9740	0,9591
GARCH	0,9510	*	0,9381	0,9716	0,8855	0,9588	*	0,9238	0,9440
SCP	0,9393	0,9580	0,9280	0,9215	0,9075	0,9284	0,9402	0,9267	0,9023
CF	0,9431	0,9597	1,0000	1,0000	0,8791	0,9460	0,9726	0,9680	0,9744
EWMA	0,9691	0,9630	0,9764	0,9732	0,9680	0,9654	0,9662	0,9666	0,9712
PEWMA	0,9608	0,9572	0,9571	0,9600	0,9583	0,9487	0,9529	0,9626	0,9645
SD-EWMA	0,9658	0,9521	0,9753	0,9756	0,9612	0,9693	0,9691	0,9670	0,9700
TSSD-EWMA	0,9683	0,9684	0,9753	0,9763	0,9659	0,9728	0,9723	0,9698	0,9704
KNN-CAD	0,9623	0,9640	0,9619	0,9636	0,9641	0,9643	0,9630	0,9626	0,9626
KNN-LDCD	0,9623	0,9647	0,9619	0,9636	0,9641	0,9643	0,9630	0,9626	0,9626

Ao analisar os resultados obtidos com a Tabela 6.6 através da métrica $F1$ é possível observar que 3 métodos apresentaram melhores resultados que os demais. Cada um teve melhor desempenho em três variáveis: *(i)* **DE** apresentou maior valor de $F1$ em pH, *Redox* e $Trueb$; *(ii)* **NA** mostrou melhores resultados para as variáveis pH, Cl_2 e Fm; e *(iii)* **CF** apontou bons valores de $F1$ para pH, *Redox* e Fm_2. Com essa análise é possível visualizar que duas variáveis (pH e *Redox*) devem ser observadas cuidadosamente nos próximos passos, pois pH apresentou melhores resultados de $F1$ nos 3 métodos, enquanto que *Redox* mostrou bons índices em 2 dos 3 métodos.

Os resultados obtidos através da métrica acurácia balanceada ($balanced_accuracy$) são exibidos na Tabela 6.7. Para tal métrica, os resultados alcançados foram bem satisfatórios, sendo possível visualizar com mais facilidade os métodos que apresentaram melhores ou piores desempenhos pois a escala dos resultados possui maior variação. Na Tabela 6.7 é possível ver que o método **DE** obteve destaque, sendo aquele que expôs 5 melhores resultados dentre as 9 variáveis analisadas. Por outro ângulo, é possível ver que as variáveis *pH* e *Redox* apresentaram melhores resultados tanto com o método **DE** quanto com o **NA** (*pH*) e o método **CF** em ambas variáveis.

Tabela 6.7: Resultados obtidos com acurácia balanceada para o *dataset* da Água.

Método	**Variáveis**								
	Tp	Cl	pH	Redox	Leit	Trueb	Cl_2	Fm	Fm_2
DE	0,6211	*	1,0000	1,0000	0,5264	0,9902	*	*	0,4821
NA	0,6104	0,4858	1,0000	0,9867	0,4401	0,9863	0,4986	0,4986	0,4839
GARCH	0,4828	*	0,9152	0,9526	0,4639	0,5498	*	0,4774	0,5358
SCP	0,5049	0,4828	0,9328	0,9272	0,5161	0,9332	0,5455	0,4535	0,5515
CF	0,5747	0,5044	1,0000	1,0000	0,5253	0,8427	0,5038	0,4926	0,4989
EWMA	0,5003	0,5009	0,5274	0,5243	0,4992	0,5166	0,4975	0,4978	0,5024
PEWMA	0,5054	0,4887	0,5085	0,5113	0,5030	0,5071	0,4845	0,5006	0,4891
SD-EWMA	0,5037	0,4970	0,5264	0,5531	0,4992	0,5402	0,4937	0,4916	0,5277
TSSD-EWMA	0,5062	0,4930	0,5264	0,5538	0,4905	0,5437	0,4968	0,4944	0,5215
KNN-CAD	0,4870	0,5086	0,5065	0,5214	0,4954	0,5221	0,4943	0,4940	0,5072
KNN-LDCD	0,4870	0,5159	5065	0,5214	0,4954	0,5221	0,4943	0,4940	0,5072

A acurácia balanceada é uma métrica importante para ser analisada e dar suporte a discussão dos resultados. A Tabela 6.8 apresenta a quantidade de acertos de cada método nas variáveis em análise. É importante que o método apresente uma boa quantidade de acertos em relação a quantidade total de eventos que existem na série, dado que o objetivo é identificá-los com uma quantidade razoável de falsos positivos.

A quantidade de acertos em uma detecção de eventos têm um papel importante na análise quantitativa de um método. A Tabela 6.8 apresenta os resultados obtidos no *dataset* da água, que contém 72 eventos rotulados. Através da quantidade de acertos obtida, os métodos **DE** e **SCP** obtiveram melhores resultados para 5 variáveis. Métodos baseados em vizinhança (**KNNs**) e baseados em **EWMA**, inclusive, não obtiveram resultados satisfatórios na detecção de eventos das séries do *dataset* da Água.

As métricas referentes à distância *a priori* e distância *a posteriori* são importantes para definir o quão próximo ou o quão distante a detecção está do evento de referência presente na série. A distância *a posteriori*, apresentada na Tabela 6.9, mostra os resultados ocorridos após o evento.

Tabela 6.8: Resultados contendo os acertos para o *dataset* da Água.

Método	**Variáveis**								
	Tp	Cl	pH	Redox	Leit	Trueb	Cl_2	Fm	Fm_2
DE	31	*	72	72	17	72	*	*	0
NA	30	3	72	72	1	72	0	0	0
GARCH	1	*	68	69	7	10	*	4	10
SCP	6	0	72	72	12	72	12	0	18
CF	16	3	72	72	17	56	1	0	0
EWMA	1	2	4	4	1	4	1	1	1
PEWMA	3	1	4	4	3	5	1	2	0
SD-EWMA	2	3	4	8	2	7	0	0	5
TSSD-EWMA	2	0	4	8	0	7	0	0	4
KNN-CAD	0	3	3	5	1	5	1	1	3
KNN-LDCD	0	4	3	5	1	5	1	1	3

Essa marcação sinaliza um atraso na detecção.

Tabela 6.9: Resultados obtidos com a distância *a posteriori* para o *dataset* da Água.

Método	**Variáveis**								
	Tp	Cl	pH	Redox	Leit	Trueb	Cl_2	Fm	Fm_2
DE	0	*	0	0	106	0	*	*	-
NA	9	42	0	0	106	0	592	515	409
GARCH	39	*	1	1	0	908	*	61	0
SCP	310	42	0	0	0	0	0	42	0
CF	9	42	0	0	106	6	106	400	529
EWMA	34	20	0	0	106	0	131	60	0
PEWMA	9	20	0	0	86	0	25	60	52
SD-EWMA	216	20	0	0	106	0	267	60	0
TSSD-EWMA	303	88	0	0	106	0	285	60	0
KNN-CAD	41	31	67	0	48	0	3	8	5
KNN-LDCD	41	31	67	0	48	0	3	8	5

Analisando a Tabela 6.9 é possível a percepção de 3 métodos com melhores resultados: **SCP**, que obteve 6 melhores resultados. **EWMA** e **SD-EWMA** também apresentaram resultados satisfatórios, onde cada um mostrou 5 melhores números. Em uma avaliação que envolve tanto a distância *a posteriori* quanto a distância *a priori*, quanto mais próximos de zero for o resultado, melhor. Isso significa que a detecção está mais próxima da referência. Algumas variáveis se destacam, como pH, *Redox* e $Trueb$. Analisando a tabela verticalmente, temos diversas marcações exatas no evento

(i.e. zeros). Dos 11 métodos utilizados nesta pesquisa, a quantidade de métodos que teve uma detecção exata foi: 8 para a variável pH, 10 para *Redox* e 9 para $Trueb$. Nesses casos, o atraso foi igual a zero. A seguir são apresentados os resultados com a métrica da distância *a priori* através da Tabela 6.10.

Tabela 6.10: Resultados obtidos com a distância *a priori* para o *dataset* da Água

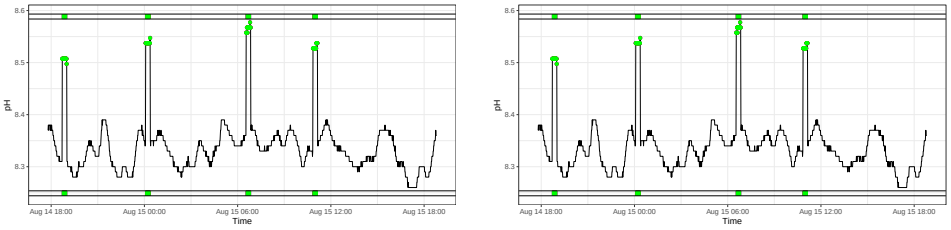
Método	**Variáveis**								
	Tp	Cl	pH	Redox	Leit	Trueb	Cl_2	Fm	Fm_2
DE	54	*	-	-	32	43	*	*	55
NA	55	50	-	42	55	43	10	-	44
GARCH	32	*	55	55	31	-	*	-	55
SCP	27	30	23	24	30	21	24	-	24
CF	46	50	-	-	32	-	18	54	-
EWMA	-	-	-	-	-	-	-	-	1
PEWMA	-	-	-	-	-	-	-	-	-
SD-EWMA	-	-	-	-	3	-	-	-	1
TSSD-EWMA	-	-	-	-	3	-	-	-	1
KNN-CAD	6	6	6	-	3	-	-	5	6
KNN-LDCD	6	6	6	-	3	-	-	5	6

A Tabela 6.10 permite avaliar as detecções ocorridas antes do evento de referência. É possível observar que os métodos **EWMA** e **PEWMA** não apresentaram medidas de distância *a priori* (exceto para a variável Fm_2, que o **EWMA** identificou 1 ponto antes). De maneira análoga, **SD-EWMA** e **TSSD-EWMA** apresentaram medições *a priori* apenas nas variáveis *Leit* e Fm_2. Como o objetivo desta métrica é avaliar as detecções ocorridas antes do evento, essas variáveis não são adequadas. As identificações mais próximas do evento foram identificadas pelos métodos **KNN-CAD** e **KNN-LDCD**. Um resumo das Tabelas 6.6, 6.7, 6.8, 6.9 e 6.10 é apresentado a seguir, demonstrando os melhores métodos segundo seu desempenho medido pelas métricas:

- **F1**: DE, NA e CF
- **bal_acc**: DE
- **acertos**: DE e SCP
- ***a prosteriori***: SCP, EWMA e SD-EWMA
- ***a priori***: KNN-CAD e KNN-LDCD

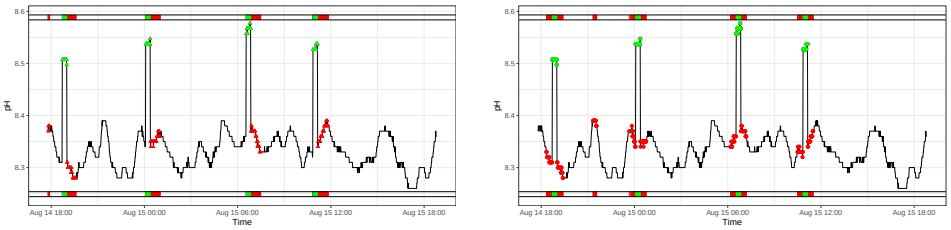
Uma avaliação quantitativa, como apresentada nesta seção, ajuda a compreender todo o processo de detecção de eventos por diferentes perspectivas através das métricas. No entanto, pode não ser

suficiente para visualizar o potencial de desempenho de um método. Com isso surge a necessidade de uma avaliação qualitativa através da visualização gráfica dos resultados. Essa análise busca contribuir com a quantitativa, trazendo uma visão complementar. A Figura 6.3 ilustra exemplos de detecção dos 11 métodos com uma variável do *dataset* da Água. Esse *dataset* apresenta 4 intervalos de eventos facilmente identificados. A variável escolhida foi a **pH**, por apresentar um padrão no comportamento desses intervalos de eventos presentes. Com isso tornou-se possível uma percepção rápida do método que obteve o melhor desempenho apenas pela análise visual dos gráficos.



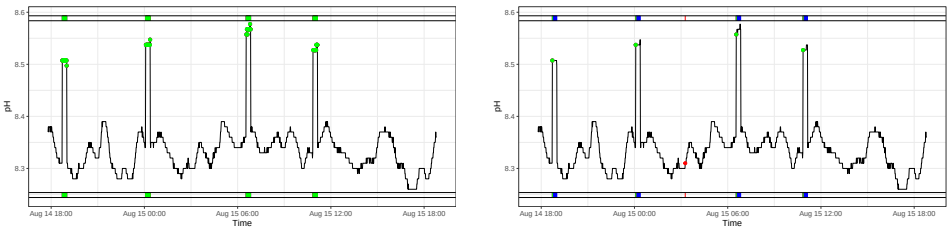
(a) Eventos detectados por DE.

(b) Eventos detectados por NA.



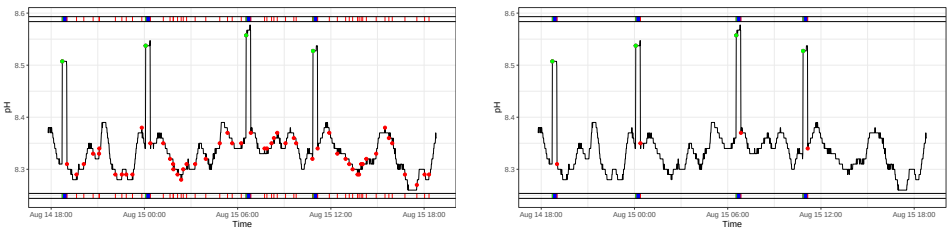
(c) Eventos detectados por GARCH.

(d) Eventos detectados por SCP.



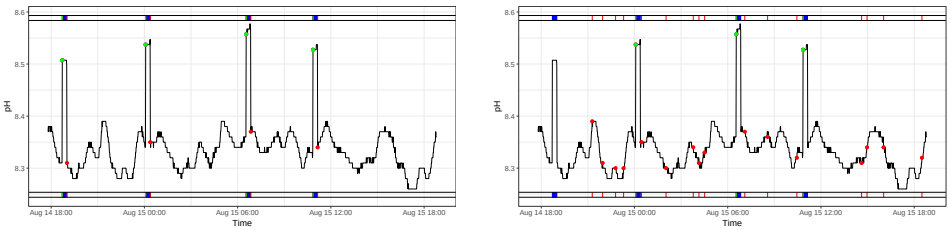
(e) Eventos detectados por CF.

(f) Eventos detectados por EWMA.



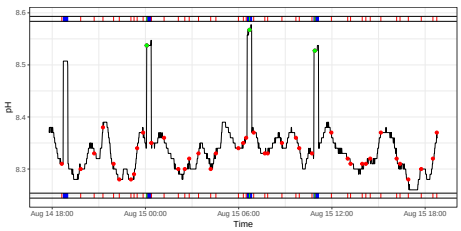
(g) Eventos detectados por PEWMA.

(h) Eventos detectados por SD-EWMA.



(i) Eventos detectados por TSSD-EWMA.

(j) Eventos detectados por KNN-CAD.



(k) Eventos detectados por KNN-LDCD.

Figura 6.3: Detecções de eventos na série **pH** através dos 11 métodos apresentados nesta pesquisa.

A Figura 6.3 apresenta o resultado da detecção, sendo visível o bom desempenho dos métodos através das Figuras 6.3(a), 6.3(b), 6.3(e). Nesses casos, DE, NA e CF atingiram 100% de precisão, pois detectaram todos os eventos não marcando nenhum instante como falso positivo. O método GARCH apontou anomalias de volatilidade, onde o evento detectado está associado ao risco. Marcações foram feitas quando o comportamento da série subiu (grandes picos), e depois quando houve uma queda brusca.

6.4.3 Análise do *dataset Yahoo*

O conjunto de dados do *Yahoo* é composto por 4 *benchmarks*, onde o $A1$ compreende dados reais coletados do tráfego de dados do *Yahoo*, o $A2$ composto de séries sintéticas, o $A3$ e $A4$ também composto de séries sintéticas, no entanto com sazonalidade. Todos os *benchmarks* contém anomalias rotuladas de forma manual pelos editores. Os resultados referentes aos dados do *Yahoo* estão agrupados conforme os *benchmarks*, dos quais foram obtidas as médias e tais valores estão apresentados na Tabela 6.11.

Tabela 6.11: Resultados referentes ao *dataset* do *Yahoo*

Bench.	**Método**	**Tempo**	**F1**	**bal_acc**	**FP**	**Acertos**	***Posteriori***	***Priori***
A1	DE	*	*	*	*	*	*	*
	NA	0,0771	0,9740	0,8399	56	15	66	155
	GARCH	0,3499	0,9452	0,7867	134	17	24	166
	SCP	2,1432	0,9560	0,6714	101	12	109	33
	CF	0,1144	0,9808	0,8557	40	14	82	158
	EWMA	0,3813	0,9863	0,6459	12	3	93	126
	PEWMA	0,3421	0,9816	0,6335	27	2	34	64
	SD-EWMA	0,4891	0,9862	0,6852	15	4	57	96
	TSSD-EWMA	0,4838	0,9889	0,5701	4	2	255	194
	KNN-CAD	0,4447	0,9775	0,5796	37	2	14	21
	KNN-LDCD	0,4601	0,9775	0,5796	37	2	14	21
A2	DE	*	*	*	*	*	*	*
	NA	0,0823	0,9996	0,9760	1	5	1	-
	GARCH	0,3155	0,9522	0,6713	126	3	1	317
	SCP	2,2348	0,9745	0,8214	68	4	15	16
	CF	0,0404	0,9986	0,8873	3	4	10	48
	EWMA	0,3573	0,9985	0,7801	1	2	5	372
	PEWMA	0,3689	0,9942	0,7951	14	2	1	81
	SD-EWMA	0,5212	0,9982	0,8301	3	2	5	334
	TSSD-EWMA	0,5232	0,9981	0,7487	2	2	31	344
	KNN-CAD	0,4788	0,9861	0,7322	36	2	6	31
	KNN-LDCD	0,4995	0,9861	0,7322	36	2	6	31
A3	DE	*	*	*	*	*	*	*
	NA	0,0851	0,9980	0,6892	0	3	264	-
	GARCH	0,4161	0,9535	0,4980	136	1	11	69
	SCP	2,6713	0,9320	0,4873	205	1	11	12
	CF	0,0340	0,9891	0,7818	32	5	15	25
	EWMA	0,4653	0,9980	0,6927	0	3	301	-
	PEWMA	0,4633	0,9890	0,7612	2	5	170	93
	SD-EWMA	0,6821	0,9973	0,9495	8	8	54	218
	TSSD-EWMA	0,6774	0,9967	0,5740	1	1	621	41
	KNN-CAD	0,5927	0,9856	0,7253	42	4	19	24
	KNN-LDCD	0,6017	0,9856	0,7253	42	4	19	24
A4	DE	*	*	*	*	*	*	*
	NA	0,0867	0,9815	0,6966	53	4	239	97
	GARCH	0,4359	0,9629	0,5026	111	1	91	70
	SCP	2,6818	0,9448	0,4966	167	84	135	22
	CF	0,0375	0,9783	0,6890	66	3	99	53
	EWMA	0,4580	0,9980	0,6908	1	3	310	333
	PEWMA	0,4514	0,9978	0,7499	3	4	170	229
	SD-EWMA	0,6822	0,9971	0,9325	8	7	45	197
	TSSD-EWMA	0,6863	0,9971	0,5736	2	1	551	313
	KNN-CAD	0,5862	0,9856	0,7228	43	4	14	24
	KNN-LDCD	0,5996	0,9856	0,7228	43	4	14	24

O *benchmark* $A1$ contém dados reais do Yahoo, e os melhores resultados apresentados foram através dos métodos **TSSD-EWMA**, **KNN-CAD** e **KNN-LDCD**, com 2 melhores resultados em cada um dentro das 7 métricas analisadas. A dificuldade na detecção está relacionada à complexidade dos dados reais. Em contra partida o bom resultado obtido pelos métodos baseados em **KNN** indicam a sensibilidade de métodos de aprendizado de máquina e a sua adequabilidade para dados coletados em fluxo contínuo. No *benchmark* $A2$, que contém dados sintéticos, 4 melhores resultados foram obtidos com o NA, indicando possibilidade de inércia nos dados. No conjunto de dados sintéticos com sazonalidade $A3$, os métodos **NA**, **SCP**, **EWMA**, **SD-EWMA** forneceram 2 melhores resultados cada. Esses resultados confirmam a eficiência dos métodos baseados em média móvel no caso de dados com sazonalidade. Em $A4$, os métodos baseados em **KNN** apresentaram 2 menores distâncias (*a posteriori* e *a priori*) em relação aos outros métodos.

6.4.4 Análise do *dataset* 3W

O *3W dataset* consiste de dados reais indesejáveis em processos de exploração de poços de petróleo e gás. Pela natureza dos dados, é de suma importância que os eventos sejam detectados antes da ocorrência. A métrica que identifica uma detecção antes da ocorrência é a distância *a priori*, e através dela os resultados serão apresentados por tipo de evento indesejado. Também é necessário um menor valor possível de atraso na detecção, para que seja possível uma intervenção humana, sendo este valor computado pela distância *a posteriori*.

As Tabelas 6.12 e 6.13 foram construídas com base nos experimentos de detecção de eventos contidos nas séries temporais do conjunto de dados 3W. Os valores apresentados são computados com base na média dos resultados de dados oriundos de diferentes poços de exploração de petróleo. Os resultados estão organizados por tipo de evento não desejável contido nas séries. Métricas com marcações iguais a "*" ou "-" indicam impossibilidade de execução e de detecção, respectivamente.

Através da Tabela 6.12 percebe-se que o método **KNN-LDCD** obteve 3 das distâncias anteriores mais próximas ao evento da série. Isso confirma o bom desempenho dos métodos KNN em aplicações que utilizam dados reais. Demais métodos tiveram uma distância muito grande em relação ao evento da série, com algumas exceções como DE no Tipo 2 e **GARCH** no Tipo 8, que apresentaram bom desempenho na detecção antecedente ao evento. **TSSD-EWMA**, apesar de não identificar antes em todos os tipos, não obtiveram distâncias tão ruins. Eventos do Tipo 1 pelo método **CF** tiveram um atraso médio de 1 hora e 30 minutos, dado que as medições são coletadas a cada segundo, e as medidas de distância seguem as unidades de medida da série. Eventos detectados antes são importantes quando estão nas proximidades do evento, pois nesse caso uma intervenção pode reverter determinado problema que está por vir. Por exemplo, no caso desse *dataset* de exploração de petróleo, um evento do Tipo 1 representa um aumento abrupto da BSW

Tabela 6.12: Resultados obtidos com a distância *a priori* para o 3W *dataset*.

Método	**Tipos de Eventos Indesejáveis**					
	Tipo1	Tipo2	Tipo5	Tipo6	Tipo7	Tipo8
DE	288	1	588	*	2042	334
NA	1343	123	815	753	1177	172
GARCH	2963	2230	555	979	1002	9
SCP	113	161	235	445	662	13
CF	5178	278	1157	1982	790	773
EWMA	1945	481	997	1179	1054	279
PEWMA	2142	266	675	1097	496	234
SD-EWMA	954	258	696	1017	574	268
TSSD-EWMA	303	88	696	1168	106	268
KNN-CAD	101	55	85	103	219	37
KNN-LDCD	64	54	85	103	219	37

(*Basic Sediment and Water*), que representa um aumento de água e sedimentos básicos no poço. Durante o ciclo de um poço é normal que ela aumente, mas um aumento repentino pode levar a diversos problemas relacionados à produtividade do poço. Então, como no resultado obtido pelo **CF**, um atraso de 1 hora e 30 minutos pode ser tarde demais para tentar se reverter o problema. O ideal seria detectar alguns minutos antes para ser possível tomar uma decisão de parar a produção antes que o problema propague e afete todo o processo. A Tabela 6.13 a seguir apresenta os resultados médios obtidos com a métrica distância *a posteriori*.

A Tabela 6.13 confirma que os métodos baseados em **KNN** são mais adequados para dados de séries temporais reais, pois apresentaram 5 melhores resultados entre os 6 tipos de eventos indesejáveis. Essa métrica define o atraso da detecção em relação a referência. Um valor *a posteriori* muito grande significa que o evento já aconteceu, e não há mais nada a fazer. Um pequeno atraso, em alguns casos, pode simbolizar que alguma medida ainda pode ser tomada em relação ao evento visando evitar um problema maior.

6.4.5 Análise de todos os *datasets*

Na análise envolvendo todos os *datasets* é apresentada uma comparação das métricas de qualidade $F1$ sobre as detecções de eventos produzidas por diferentes métodos com diferentes técnicas de detecção. É possível observar na Tabela 6.14, de maneira unificada, os desempenhos da detecção através dos diversos métodos discutidos com base nas séries temporais contidas nos conjuntos de dados da Água, *Yahoo* e 3W.

Tabela 6.13: Resultados obtidos com a distância *a posteriori* para o 3W *dataset*.

Método	**Tipos de Eventos Indesejáveis**					
	Tipo1	Tipo2	Tipo5	Tipo6	Tipo7	Tipo8
DE	9365	20	13993	*	16651	0
NA	5777	171	2939	363	11548	871
GARCH	7362	437	416	1017	973	19
SCP	91	70	632	181	1605	157
CF	8069	548	4824	1026	9606	5255
EWMA	6467	715	4440	1241	7262	3552
PEWMA	7468	552	2475	1685	6789	952
SD-EWMA	4852	371	766	734	2036	772
TSSD-EWMA	4852	371	766	754	2036	772
KNN-CAD	56	26	34	70	594	37
KNN-LDCD	54	26	34	70	594	37

Tabela 6.14: Comparação da qualidade das detecções de eventos nos conjuntos de dados selecionados produzidas por diferentes métodos com base na métrica F1 e no tempo de execução em segundos.

	F1			**Tempo**		
	Água	*Yahoo*	3W	Água	*Yahoo*	3W
DE	0,9517	*	0,9448	0,23	*	0,81
NA	0,9557	0,9889	0,9447	0,06	0,08	0,82
GARCH	0,9390	0,9542	0,9872		0,38	5,51
SCP	0,9280	0,9510	0,9231	2,31	2,47	44,63
CF	0,9603	0,9873	0,9867	0,03	0,05	2,65
EWMA	0,9688	0,9961	0,9956	0,36	0,42	30,43
PEWMA	0,9580	0,9915	0,9925	0,36	0,41	27,82
SD-EWMA	0,9673	0,9955	0,9968	0,54	0,60	239,89
TSSD-EWMA	0,9711	0,9956	0,9970	0,54	0,58	268,62
KNN-CAD	0,9641	0,9843	0,9872	1,04	0,53	9,19
KNN-LDCD	0,9632	0,9843	0,9872	0,48	0,55	9,35

A Tabela 6.14 indica através da análise do $F1$ que o método **TSSD-EWMA** obteve um melhor desempenho dentro de um contexto geral. No entanto, apenas essa análise não é suficiente para apoiar uma escolha definitiva, pois na comparação do $F1$ entre as variáveis do *dataset* da Água, na Tabela 6.6, foi possível observar que os métodos **DE**, **NA** e **CF** obtiveram melhores valores para a

referida métrica. Uma análise visual dentro de um contexto mais específico permite compreender que nem sempre o **TSSD-EWMA** é o método adequado. Nas análises individuais de cada série, foi possível perceber um bom valor de F1 para o **TSSD-EWMA**. Este método é baseado no SD-EWMA, diferindo em um teste estatístico que busca reduzir a quantidade de falsos positivos identificados no SD-EWMA. No entanto, ao reduzir os FP, ele reduz também a quantidade de acertos, e retorna detecções bem mais distantes do evento do que o SD-EWMA. Isso foi observado de maneira geral em todas as séries observadas nesta dissertação. A Tabela 6.15 mostra uma comparação isolada entre os 2 métodos, com uma série específica selecionada aleatoriamente do *Yahoo A*4.

Tabela 6.15: Comparação dos métodos SD-EWMA e TSSD-EWMA.

Método	**F1**	***bal_acc***	**FP**	**acertos**	***a posteriori***	***a priori***
SD-EWMA	0,9958	0,9958	14	11	0	-
TSSD-EWMA	0,9964	0,5448	2	1	854	-
GARCH	0,9612	0,9297	107	14	1	1
EWMA	0,9955	0,5667	0	2	0	-

Foi possível observar que o TSSD-EWMA apresentou melhor pontuação de F1 em relação ao SD-EWMA. Entretanto, nas demais métricas seu resultado ficou inferior ao esperado, reduzindo bem a quantidade de FP mas também os acertos do método. Consequentemente, o atraso ficou bem maior dado que em SD-EWMA a detecção ocorreu no ponto exato. De modo semelhante, EWMA mostrou um melhor F1, mas o GARCH teve uma melhor acurácia balanceada e uma maior quantidade de acertos na detecção. Com isso temos uma necessidade de avaliar bem cada métrica dentro do contexto específico, pois um bom resultado de uma métrica pode não ser suficiente para apresentar os bons resultados da detecção.

A Tabela 6.16 apresenta um resumo geral do desempenho de cada método por tipo de *dataset* (sintético ou real). Os valores apresentam quantas vezes o método obteve um melhor desempenho no tipo de *dataset* em questão.

Tabela 6.16: Análise geral por *dataset* e método.

Método	Sintético	Real
DE	3	10
NA	7	6
GARCH	2	4
SCP	5	1
CF	4	4
EWMA	6	8
PEWMA	1	0
SD-EWMA	4	2
TSSD-EWMA	2	6
KNN-CAD	1	10
KNN-LDCD	1	10

A partir da observação dos valores computados, é possível perceber quais métodos possuem melhor adequabilidade a cada caso em estudo. Para conjuntos de dados sintéticos, o método da normalização adaptativa (NA) apresentou melhor desempenho, enquanto que para dados que chegam em tempo real os métodos baseados em vizinhança e decomposição conseguiram resultados mais satisfatórios.

Capítulo 7

Conclusão

Esta dissertação abordou um estudo comparativo de métodos de detecção de eventos em séries temporais. Foi realizada uma revisão da literatura, onde 11 métodos foram selecionados para estudo e comparação. Os fundamentos encontrados mostram o crescente empenho dos autores na busca por uma melhor abordagem acerca do tema relacionado, onde diversificadas técnicas foram propostas envolvendo séries temporais univariadas, multivariadas, algoritmos *batch* e *online*, além de métodos envolvendo técnicas estatísticas, baseados em proximidade, vizinhança, decomposição, dentre outros disponíveis.

Frequentemente, especialistas precisam lidar com o problema de escolher o método de detecção de eventos mais adequado a uma série temporal e à aplicação conduzida. Essa escolha pode ser uma tarefa complexa, dado que existem diversos métodos de detecção na literatura, cada um especializando-se em séries temporais que apresentam propriedades estatísticas diversas ou fazendo suposições sobre a distribuição dos seus dados.

Os métodos abordados nesta pesquisa foram testados em séries sintéticas e reais, de tamanhos variados, e com diferentes quantidades de eventos presentes. O objetivo foi desenvolver uma metodologia de comparação para avaliar qual melhor método para determinado tipo de dado. O processo proposto envolve a aplicação de 7 métricas para avaliação do desempenho de cada método em estudo e ainda, o tempo. Então, nesta pesquisa foram computados: *(i)* tempo de execução, *(ii)* F1, *(iii)* acurácia balanceada, *(iv)* falsos positivos, *(v)* acertos, *(vi)* distância *a posteriori*, e *(vii)* distância *a priori*. De acordo com todo o estudo realizado nas referências encontradas, nenhum outro trabalho apresenta uma abordagem que envolve a comparação por diversas métricas. Poucos trabalhos apresentaram a detecção envolvendo mais de um tipo de evento.

Foram realizados experimentos com 3 *datasets* de diferentes áreas de aplicação: (i) Água, de dados sintéticos contendo informações sobre a qualidade da água, (ii) *Yahoo*, com dados sintéticos e reais do fluxo de dados do Yahoo, e (iii) 3W, com dados reais de exploração de petróleo. Os experimentos realizados sugerem que a análise comparativa dos métodos de detecção é capaz de nortear a escolha mais adequada para cada tipo de série. Foi constatado que essa análise auxilia na compreensão da natureza dos eventos detectados, provendo uma visão abrangente das detecções de

diferentes métodos. Com isso, busca-se evitar problemas como a desconsideração ou a identificação incorreta de eventos através de uma escolha inadequada do método de detecção, que podem trazer prejuízo às aplicações que dependem do monitoramento desses eventos.

Resultados experimentais apontam para a complexidade do problema de detecção de eventos, que está intimamente relacionado às propriedades estatísticas das séries temporais adotadas. No *dataset* 3W, coletado durante processos de exploração em poços de petróleo, foi observada a presença de eventos não desejáveis de diferentes tipos, apresentando diferentes semânticas e causando impactos de diferentes magnitudes no comportamento das séries temporais. Considerando-se que tais eventos podem ocorrer sequencialmente e concomitantemente através das variáveis coletadas, observa-se a complexidade da tarefa de seleção de métodos de detecção de eventos. A seleção deve ser adequada aos dados do domínio, que são mais complexos e desafiadores quando são analisados dados reais.

Os resultados apresentados sugerem a expansão da pesquisa para trabalhos futuros, onde estuda-se a possibilidade de realizar uma adaptação nos algoritmos para detectar eventos de forma *on-line*, em tempo real. Essa detecção sobre dados de *streaming* é muito importante, dado que sensores capturam informações a todo instante, e um processamento *on-line* fornece possíveis intervenções no sistema a ponto de evitar acidentes e falhas. Também considera-se incluir novos métodos, como os baseados em aprendizado de máquina e aprendizado profundo, onde se deixaria de ter uma referência nos dados para comparação da detecção e seria treinado um modelo baseado nos dados iniciais da série.

A extensão desta pesquisa também conduz a um problema de predição de eventos que se baseia não só em técnicas de detecção, mas também em soluções nas áreas de análise e predição de séries temporais e aprendizado de máquina. Este problema é particularmente importante, pois a predição de defeitos maquinários e instabilidades em operações e serviços possibilitam uma adoção de medidas preventivas ou rápidas correções, minimizando prejuízos aos processos e operações.

Bibliografia

Aggarwal, C. C. [2017]. *Outlier Analysis*, Springer International Publishing. 5, 20

Agrawal, S. and Agrawal, J. [2015]. Survey on anomaly detection using data mining techniques, *KES*, India, pp. 708–713. 31, 35

Alippi, C., Boracchi, G., Carrera, D. and Roveri, M. [2015]. Change detection in multivariate datastreams: Likelihood and detectability loss. 33, 35

Alkhamees, N. and Fasli, M. [2017]. Event detection from time-series streams using directional change and dynamic thresholds, *2017 IEEE International Conference on Big Data (Big Data)*, pp. 1882–1891. 1, 34, 35

Aminikhanghahi, S. and Cook, D. J. [2017]. A survey of methods for time series change point detection, *Knowledge and Information Systems* **51**(2): 339–367.
URL: *https://doi.org/10.1007/s10115-016-0987-z* 23, 31, 35

Anderson, K. D., Bergés, M. E., Ocneanu, A., Benitez, D. and Moura, J. M. F. [2012]. Event detection for non intrusive load monitoring, *IECON 2012 - 38th Annual Conference on IEEE Industrial Electronics Society*, pp. 3312–3317. 33, 35

Ba, A. and McKenna, S. [2015]. Water quality monitoring with online change-point detection methods, *Journal of Hydroinformatics* **17**: 7–19. 34, 35

Bach, K., Gundersen, O. E., Knappskog, C. and Ozturk, P. [2014]. Automatic case capturing for problematic drilling situations, pp. 48–62. 2

Bai, X., Xiong, Y., Zhu, Y., Liu, Q. and Chen, Z. [2013]. Co-anomaly event detection in multiple temperature series, Vol. 8041, pp. 1–14. 33, 35

Batal, I., Cooper, G. F., Fradkin, D., Harrison, J. H., Mörchen, F. and Hauskrecht, M. [2015]. An efficient pattern mining approach for event detection in multivariate temporal data, *Knowledge and Information Systems* **46**: 115–150. 1, 33, 35

Batal, I., Fradkin, D., Harrison, J., Moerchen, F. and Hauskrecht, M. [2012]. Mining recent temporal patterns for event detection in multivariate time series data, Vol. 2012. 33, 35

Ben-Gal, I. [2005]. Outlier detection, *Data mining and knowledge discovery handbook*, Springer, pp. 131–146. 1

Berlinger, E., Illes, F., Badics, M., Banai, A. and Daroczi, G. [2015]. *Mastering R for Quantitative Finance*, Packt Publishing. 21

Boriah, S. [2010]. *Time series change detection: Algorithms for land cover change*, PhD thesis, University of Minnesota. 15

Braei, M. and Wagner, S. [2020]. Anomaly detection in univariate time-series: A survey on the state-of-the-art. 30, 35

Bulunga, M. L. [2012]. *Change-point detection in dynamical systems using auto-associative neural networks*, Mestrado, Faculty of Engineering at Stellenbosch University, África do Sul. 23

Calikus, E., Nowaczyk, S., Sant'Anna, A. and Dikmen, O. [2020]. No free lunch but a cheaper supper: A general framework for streaming anomaly detection, *Expert Systems with Applications* **155**: 113453.
URL: *http://dx.doi.org/10.1016/j.eswa.2020.113453* 32, 35

Campisano, R., Borges, H., Porto, F., Perosi, F., Pacitti, E., Masseglia, F. and Ogasawara, E. [2018]. Discovering tight space-time sequences, *DaWaK: Data Warehousing and Knowledge Discovery* pp. 247–257. 27

Cappers, B. C. M. and van Wijk, J. J. [2018]. Exploring multivariate event sequences using rules, aggregations, and selections, *IEEE Transactions on Visualization and Computer Graphics* **24**(1): 532–541. 2, 33, 35

Carmona, R. [2014]. *Statistical Analysis of Financial Data in R*, Springer-Verlag New York. 21

Carreño, A., Inza, I. and Lozano, J. [2019]. Analyzing rare event, anomaly, novelty and outlier detection terms under the supervised classification framework, *Artificial Intelligence Review* . 35

Carter, K. M. and Streilein, W. W. [2012]. Probabilistic reasoning for streaming anomaly detection, *2012 IEEE Statistical Signal Processing Workshop (SSP)*, pp. 377–380. 26

Chakravarthy, S., Krishnaprasad, V., Anwar, E. and Kim, S.-K. [1994]. Composite Events for Active Databases: Semantics, Contexts and Detection, *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 606–617.
URL: *http://dl.acm.org/citation.cfm?id=645920.672994* 1

Chandola, V., Banerjee, A. and Kumar, V. [2009]. Anomaly detection: A survey, *ACM Comput. Surv.* **41**. , 17, 19, 30, 35

Chen, H. and Zhang, N. [2015]. Graph-based change-point detection, *The Annals of Statistics* **43**(1): 139–176.
URL: *http://dx.doi.org/10.1214/14-AOS1269* 23

Chiu, B., Keogh, E. and Lonardi, S. [2003]. Probabilistic discovery of time series motifs, *the ninth ACM SIGKDD international conference, ACM Press* p. 493. 13, 27

Dancho, M. and Vaughan, D. [2019]. anomalize: Tidy Anomaly Detection.
URL: *https://CRAN.R-project.org/package=anomalize* 19

Dao, M., Zettsu, K., Pongpaichet, S., Jalali, L. and Jain, R. [2015]. Exploring spatio-temporal-theme correlation between physical and social streaming data for event detection and pattern interpretation from heterogeneous sensors, *2015 IEEE International Conference on Big Data (Big Data)*, pp. 2690–2699. 1

De Paepe, D., Hautte, S. V., Steenwinckel, B., De Turck, F., Ongenae, F., Janssens, O. and Hoecke, S. V. [2020]. A generalized matrix profile framework with support for contextual series analysis, *Engineering Applications of Artificial Intelligence* **90**. 32, 35

Ding, D., Zhang, M., Pan, X., Yang, M. and He, X. [2019]. Modeling extreme events in time series prediction, *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, KDD '19, Association for Computing Machinery, New York, NY, USA, p. 1114–1122.
URL: *https://doi.org/10.1145/3292500.3330896* 20

Ding, J., Xiang, Y., Shen, L. and Tarokh, V. [2016]. Multiple change point analysis: Fast implementation and strong consistency, *IEEE Transactions on Signal Processing* **PP**. 23

Dutra, M. G. [2016]. *Discovering motifs in spatial-time series seismic datasets*, Mestrado em engenharia de produção e sistemas, Centro de Educação Tecnológica Celso Suckow da Fonseca - CEFET/RJ, Rio de Janeiro. , 4, 14

Eriksson, B., Barford, P., Bowden, R., Duffield, N., Sommers, J. and Roughan, M. [2010]. Basisdetect: A model-based network event detection framework, *Proceedings of the 10th ACM SIGCOMM*, Association for Computing Machinery, New York, NY, USA, p. 451–464. 32, 35

Esling, P. and Agon, C. [2012]. Time-series data mining, *ACM Computing Surveys* **45**(1): 1–34. 6

Fearnhead, P. and Rigaill, G. [2016]. Changepoint detection in the presence of outliers, *Journal of the American Statistical Association* . 31, 35

Freitas, I. W. S. [2019]. *Um estudo comparativo de técnicas de detecção de outliers no contexto de classificação de dados*, Mestrado, Universidade Federal Rural do Semi-Árido, Brasil. 31, 35

Gabarda, S. and Cristóbal, G. [2010]. Detection of events in seismic time series by time-frequency methods, *IET Signal Processing* **4**(4): 413–420. 34, 35

Gammerman, A. and Vovk, V. [2007]. Hedging predictions in machine learning, *The Computer Journal* **50**(2): 151–163. 19, 22

García, Y. G., Shadaydeh, M., Mahecha, M. D. and Denzler, J. [2018]. Extreme anomaly event detection in biosphere using linear regression and a spatiotemporal mrf model, *Natural Hazards* pp. 1–19. 34, 35

Gensler, A. and Sick, B. [2017]. Performing event detection in time series with swiftevent: an algorithm with supervised learning of detection criteria, *Pattern Analysis and Applications* . 32, 35

Ghalanos, A. [2014]. *rugarch: Univariate GARCH models.* R package version 1.4-0. 40

Gupta, M., Gao, J., Aggarwal, C. and Han, J. [2014]. Outlier Detection for Temporal Data: A Survey, *IEEE Transactions on Knowledge and Data Engineering* . 16

Guralnik, V. and Srivastava, J. [1999]. Event Detection from Time Series Data, *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '99, ACM, New York, NY, USA, pp. 33–42. event-place: San Diego, California, USA.
URL: *http://doi.acm.org/10.1145/312129.312190* , 1, 2, 24, 31, 35

Hahn, G., Fearnhead, P. and Eckley, I. [2020]. Bayesproject: Fast computation of a projection direction for multivariate changepoint detection, *Statistics and Computing* pp. 1573–1375. 32, 35

Han, J., Kamber, M. and Pei, J. [2011]. *Data Mining: Concepts and Techniques*, 3 edn, Morgan Kaufmann, Haryana, India; Burlington, MA. 00000. , 1, 16

Harchaoui, Z., Vallet, F., Lung-Yut-Fong, A. and Cappe, O. [2009]. A regularized kernel-based approach to unsupervised audio segmentation, *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP '09, IEEE Computer Society, USA, p. 1665–1668.
URL: *https://doi.org/10.1109/ICASSP.2009.4959921* 23

Hawkins, D. [1980]. *Identification of Outliers*, Springer Netherlands. 15, 30

Hoan, M. V. and Exbrayat, M. [2013]. Time series symbolization and search for frequent patterns, *ACM International Conference Proceeding Series* pp. 108–117. 14

Hunter, J. and McIntosh, N. [1999]. Knowledge-based event detection in complex time series data, *Proceedings of the Joint European Conference on Artificial Intelligence in Medicine and Medical Decision Making*, AIMDM '99, Springer-Verlag, Berlin, Heidelberg, p. 271–280. 33, 35

Hyndman, R. J. and Athanasopoulos, G. [2013]. *Forecasting: Principles and Practice*, 2 edn, Texts, Australia. 11

Ishimtsev, V., Nazarov, I., Bernstein, A. and Burnaev, E. [2017]. Conformal k-nn anomaly detector for univariate data streams. 22

Keogh, E. and Kasetty, S. [2003]. On the need for time series data mining benchmarks: A survey and empirical demonstration, *Data Mining and Knowledge Discovery* pp. 349–371. 2

Keogh, E. and Lin, J. [2005]. Clustering of time-series subsequences is meaningless: implications for previous and future research, *Knowledge and Information Systems* **8**: 154–177. 13

Lampert, C., Blaschko, M. and Hofmann, T. [2008]. Beyond sliding windows: Object localization by efficient subwindow search, *CVPR 2008*, Max-Planck-Gesellschaft, IEEE Computer Society, Los Alamitos, CA, USA, pp. 1–8. Best paper award. 13

Lin, J., Keogh, E., Lonardi, S. and Patel, P. [2002]. Finding motifs in time series, *Conference on Knowledge Discovery and Data Mining* pp. 53–68. , 27, 28

Liu, S., Qin, Z., Gan, X. and Wang, Z. [2019]. Scod: A novel semi-supervised outlier detection framework, pp. 316–321. 34, 35

Liu, S., Smith, K. and Che, H. [2015]. A multivariate based event detection method and performance comparison with two baseline methods, *Water Research* **80**. , 7, 34, 35

Liu, S., Yamada, M., Collier, N. and Sugiyama, M. [2013]. Change-point detection in time-series data by relative density-ratio estimation, *Neural Networks* **43**: 72–83.
URL: *http://dx.doi.org/10.1016/j.neunet.2013.01.012* 34, 35

Lu, G., Zhou, Y., Lu, C. and Li, X. [2016]. A novel framework of change-point detection for machine monitoring, *Mechanical Systems and Signal Processing* **83**. 32, 35

Ma, J. and Perkins, S. [2003]. Online novelty detection on temporal sequences, *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD

'03, ACM, New York, NY, USA, pp. 613–618.
URL: *http://doi.acm.org/10.1145/956750.956828* 2

Malik, K., Sadawarti, H. and Kalra, G. [2014]. Comparative analysis of outlier detection techniques, *International Journal of Computer Applications* **97**: 12–21. 30, 35

Mao, Y., Qi, H., Chen, X. and Li, X. [2017]. Event detection with multivariate water parameters in the water monitoring applications, *2017 IEEE 4th International Conference on Cyber Security and Cloud Computing (CSCloud)*, pp. 321–326. 2

McLaren, C., Elliott, D. and Kirchner, R. [2018]. *Handbook on Seasonal Adjustment*, 1nd edition edn, Eurostat, Luxembourg, European Union. 8, 10, 12

Moritz, S. and Rehbach, F. [2020]. *EventDetectR*. R package version 0.3.4.
URL: *https://CRAN.R-project.org/package=EventDetectR* 50

Moskovitch, R., Walsh, C., Hripcsak, G. and Tatonetti, N. [2014]. Prediction of biomedical events via time intervals mining, *In: Proceedings of ACM SIGKDD workshop on connected health at big data Era (BigCHat2014), New York, US*. 2

Mueen, A., Keogh, E., Zhu, Q., Cash, S. and Westover, B. [2009]. Exact discovery of time series motifs, *Society for Industrial and Applied Mathematics - 9th SIAM International Conference on Data Mining* pp. 469–480. 4

Naoki, I. and Kurths, J. [2010a]. Change-point detection of climate time series by nonparametric method, *Lecture Notes in Engineering and Computer Science* **2186**. 33, 35

Naoki, I. and Kurths, J. [2010b]. Change-point detection of climate time series by nonparametric method, *Lecture Notes in Engineering and Computer Science* **2186**. 33, 35

Ogasawara, E., Martinez, L. C., Oliveira, D. d., Zimbrao, G., Pappa, G. L. and Mattoso, M. [2010]. Adaptive normalization: A novel data normalization approach for non-stationary time series, *The 2010 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. 19, 20

Papataxiarhis, V. and Hadjief., S. [2018]. Event correlation and forecasting over multivariate streaming sensor data, *ArXiv* **abs/1803.05636**. 33, 35

Perelman, L., Arad, J., Housh, M. and Ostfeld, A. [2012]. Event detection in water distribution systems from multivariate water quality time series, **46**: 82128219. 2, 34, 35

Persons, W. M. [1919]. *Indices of business conditions*, 1 edn, The Review of Economics and Statistics. 9

Plasse, J. and Adams, N. [2019]. Multiple changepoint detection in categorical data streams, *Statistics and Computing* pp. 1109–1125. 33, 35

Poonsirivong, K. and Jittawiriyanukoon, C. [2017]. A rapid anomaly detection technique for big data curation, pp. 1–6. 17

Prema, V. and Rao, K. U. [2015]. Time series decomposition model for accurate wind speed forecast, *Renewables: Wind, Water, and Solar* . 11

Raza, H., Prasad, G. and Li, Y. [2015]. Ewma model based shift-detection methods for detecting covariate shifts in non-stationary environments, *Pattern Recogn.* **48**(3): 659–669. 25, 26

Rehbach, F., Moritz, S., Chandrasekaran, S., Rebolledo, M., Friese, M. and Bartz-Beielstein, T. [2018]. GECCO 2018 Industrial Challenge: Monitoring of drinking-water quality, pp. 1–7. 49

Salles, R., Belloze, K., Porto, F., Gonzalez, P. and Ogasawara, E. [2019]. Nonstationary time series transformation methods: An experimental review, *Knowledge-Based Systems* **164**: 274–291. , 7, 8, 9, 20, 49

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. [2020]. Harbinger: Um framework para integração e análise de métodos de detecção de eventos em séries temporais, *SBBD - Simpósio Brasileiro de Banco de Dados* p. 106223.
URL: *http://www.sciencedirect.com/science/article/pii/S0920410519306357* 41

Shumway, R. H. and Stoffer, D. S. [2017]. *Time Series Analysis and Its Applications: With R Examples*, 4 edn, Springer, New York, NY. , 4, 6, 7, 8

Silva, H. J., Lucio, P. and Brown, F. [2016]. Analise mensal, sazonal e interanual da evapotranspiração potencial para o leste do estado do acre, brasil, *Ciência e Natura* **38**. , 12

Sreevidya, S. S. [2014]. A survey on outlier detection methods, Vol. 5, pp. 8153–8156. 32, 35

Takeuchi, J. and Yamanishi, K. [2006]. A unifying framework for detecting outliers and change points from time series, *IEEE Transactions on Knowledge and Data Engineering* **18**(4): 482–492. , 24, 25, 31, 35

Talagala, P. D., Hyndman, R. J., Smith-Miles, K., Kandanaarachchi, S. and Muñoz, M. [2020]. Anomaly Detection in Streaming Nonstationary Temporal Data, *Journal of Computational and Graphical Statistics* **29**(1): 13–27. 32, 35

Truong, C., Oudre, L. and Vayatis, N. [2020]. Selective review of offline change point detection methods, *Signal Processing* **167**: 107299.
URL: *http://dx.doi.org/10.1016/j.sigpro.2019.107299* 34, 35

Vargas, R. E. V., Munaro, C. J., Ciarelli, P. M., Medeiros, A. G., Amaral, B. G., Barrionuevo, D. C., Araújo, J. C. D., Ribeiro, J. L. and Magalhães, L. P. [2019]. A realistic and public dataset with rare undesirable real events in oil wells, *Journal of Petroleum Science and Engineering* **181**: 106223.
URL: *http://www.sciencedirect.com/science/article/pii/S0920410519306357* 51

Webscope, Y. [2015]. *Labeled anomaly detection dataset*. 50

Wu, Y., Lin, Y., Zhou, Z., Bolton, D. C., Liu, J. and Johnson, P. [2019]. Deepdetect: A cascaded region-based densely connected network for seismic event detection, *IEEE Transactions on Geoscience and Remote Sensing* **57**(1): 62–75. 2, 34, 35

Xie, F., Song, A. and Ciesielski, V. [2012]. Event detection in time series by genetic programming, *2012 IEEE Congress on Evolutionary Computation*, pp. 1–8. 6, 15, 33, 35

Xiong, L., Jiang, C., Xu, C., Yu, K. and Guo, S. [2015]. A framework of change-point detection for multivariate hydrological series, *Water Resources Research* **51**. 32, 35

Yadav, R., Pradhan, A. and Kamwa, I. [2018]. Real-time multiple event detection and classification in power system using signal energy transformations, *IEEE Transactions on Industrial Informatics* **PP**. 33, 35

Zhang, W., James, N. A. and Matteson, D. S. [2017]. Pruning and nonparametric multiple change point detection, *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 288–295. 32, 35