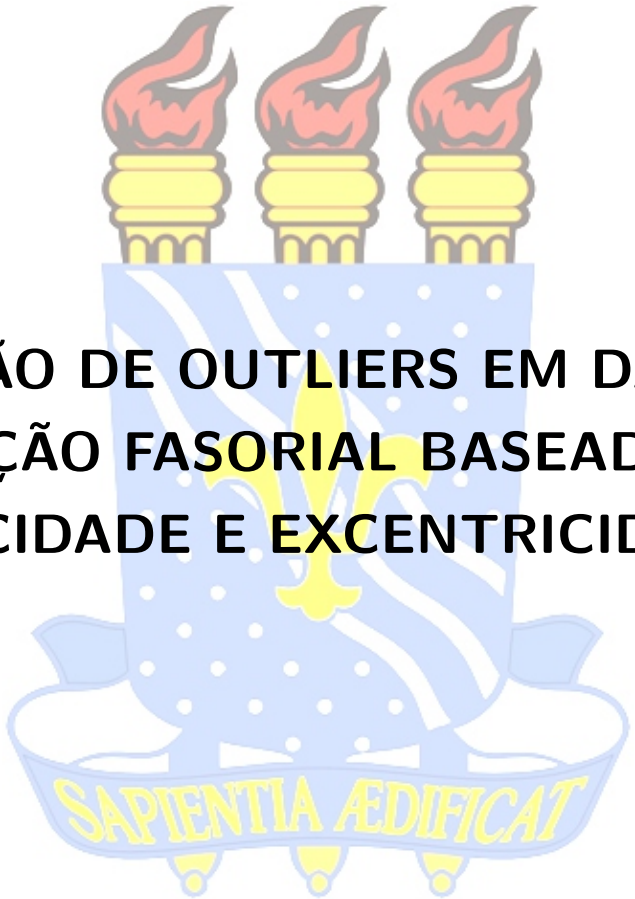
UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE ENERGIAS ALTERNATIVAS E RENOVÁVEIS
PROGRAMA DE PÓS GRADUAÇÃO EM ENGENHARIA ELÉTRICA

Diego Henrique da Silva Cavalcanti



DETECÇÃO DE OUTLIERS EM DADOS DE MEDIÇÃO FASORIAL BASEADO EM TIPICIDADE E EXCENTRICIDADE

João Pessoa

2024

Diego Henrique da Silva Cavalcanti

DETECÇÃO DE OUTLIERS EM DADOS DE MEDIÇÃO FASORIAL BASEADO EM TIPICIDADE E EXCENTRICIDADE

Trabalho apresentado ao Programa de Pós Graduação em Engenharia Elétrica da Universidade Federal da Paraíba como requisito para obtenção do título de Mestre em Engenharia Elétrica.

Universidade Federal da Paraíba

Centro de Energias Alternativas e Renováveis

Programa de Pós Graduação em Engenharia Elétrica

Orientador: Prof. Dr. Juan Moises Maurício Villanueva

João Pessoa

2024

Catalogação na publicação
Seção de Catalogação e Classificação

C376d Cavalcanti, Diego Henrique da Silva.
Detecção de outliers em dados de medição fasorial baseado em tipicidade e excentricidade / Diego Henrique da Silva Cavalcanti. - João Pessoa, 2024.
81 f. : il.

Orientação: Juan Moises Maurício Villanueva.
Dissertação (Mestrado) - UFPB/CEAR.

1. Outliers. 2. Medição fasorial. 3. PMU. 4. Tipicidade e excentricidade. I. Villanueva, Juan Moises Maurício. II. Título.

UFPB/BC CDU 621.315(043)

Elaborado por CHRISTIANE CASTRO LIMA DA SILVA - CRB-15/865

UNIVERSIDADE FEDERAL DA PARAÍBA – UFPB

CENTRO DE ENERGIAS ALTERNATIVAS E RENOVÁVEIS – CEAR

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA – PPGEE

A Comissão Examinadora, abaixo assinada, aprova a Dissertação

DETECÇÃO DE OUTLIERS EM DADOS DE MEDIÇÃO FASORIAL BASEADO EM TIPICIDADE E EXCENTRICIDADE

Elaborada por

DIEGO HENRIQUE DA SILVA CAVALCANTI

como requisito parcial para obtenção do grau de

Mestre em Engenharia Elétrica.

COMISSÃO EXAMINADORA

gov.br Documento assinado digitalmente
JUAN MOISES MAURICIO VILLANUEVA
Data: 04/09/2024 23:16:16-0300
Verifique em https://validar.iti.gov.br

PROF. DR. JUAN MOISES MAURICIO VILLANUEVA
Orientador – UFPB

gov.br Documento assinado digitalmente
EULER CASSIO TAVARES DE MACEDO
Data: 12/09/2024 14:42:02-0300
Verifique em https://validar.iti.gov.br

PROF. DR. EULER CÁSSIO TAVARES DE MACEDO
Examinador Interno – UFPB

gov.br Documento assinado digitalmente
LUIZ AFFONSO HENDERSON GUEDES DE OLIVEI
Data: 05/09/2024 08:13:29-0300
Verifique em https://validar.iti.gov.br

PROF. DR. LUIZ AFFONSO HENDERSON GUEDES DE OLIVEIRA
Examinador Externo – UFRN

João Pessoa/PB, 30 de agosto de 2024

AGRADECIMENTOS

A Deus, por ter me dado a vida e todas as oportunidades que a acompanharam.

À minha família, em especial meus pais, Luiza Cristiane e João Calistrato, ao meu pai Luciano (in memoriam) e meus irmãos, Ana Alice e João Rafael, pelo apoio e incentivo durante toda essa jornada.

A minha noiva, Anna Mércia pelo o apoio incondicional, por ter me acalmado e me incetivado nos momentos de dificuldade, por me motivar sempre a buscar o melhor de mim e por toda paciência ao longo desses anos.

Aos professores Juan, Euler, Yuri e Helon, pela oportunidade de participar do Grupo de Inteligência Computacional (GICA) e pela confiança que depositaram em mim ao longo da minha permanência no grupo de pesquisa.

Ao meu orientador, Prof. Juan, por todos ensinamentos ao longo destes anos de convívio, paciência e incentivo constante.

Aos meus amigos, Ricardo Guerra, Kaike Albuquerque e Joel Adelaide, você foram indispensáveis durante toda a minha vida acadêmica.

À Maria Iná Correia Guerra, agradeço por todas as orações e palavras de incentivo.

Aos demais amigos do GICA, agradeço por todo o conhecimento compartilhado ao longo desses períodos que estivemos juntos, pela amizade construída e pelas valorosas discussões.

Aos colega do Operador Nacional do Sistema Elétrico-ONS, por todo apoio dado durante esse período de mestrado, em especial aos colegas da Gerência de Configuração de Redes, Diego Vila Nova, Jefferson, Ives, Nicole e Letícia. E ao meu gerente Alexandre De Marco pelo incentivo que sempre foi dado.

À CAPES pela concessão da bolsa durante parte do período de realização deste mestrado.

E a todos que de alguma forma contribuíram para a realização deste trabalho.

"Basta ser sincero e desejar profundo,
você será capaz de sacudir o mundo."
Raul Seixas

RESUMO

Atualmente, as unidades de medição fasorial(PMU) são equipamentos essenciais para operação em tempo real do sistema elétrico. No entanto, devido a fatores complexos, esses dados podem ser facilmente comprometidos por interferência, falha de sincronização ou até mesmo uma falha de algum equipamento. Isso levará a vários níveis de problemas de qualidade de dados, podendo afetar diretamente as aplicações que se baseiam nesses dados ou até mesmo ameaçar a segurança dos sistemas de energia. Este trabalho visa detectar tais anomalias orientado apenas aos dados da PMU, sem a necessidade do conhecimento de parâmetros elétricos da rede. Para isso foi proposto um algoritmo on-line baseado na excentricidade e tipicidade da amostra em relação ao seu conjunto de amostras anteriores, o TEDA. Foi proposto uma modificação no TEDA, o TEDA Janelado, onde seu principal objetivo foi deixar o método mais sensível as variações de média e variância ao longo do tempo, maximizando a excentricidade de uma amostra e favorecendo a detecção de *outliers* mais difíceis. Uma terceira estratégia para melhorar a detecção do TEDA também foi apresentado nesse trabalho, o fator de esquecimento, cujo papel é estabelecer um peso mais justo entre as novas amostras e a média/variância do seu conjunto de dados, esse método foi denominado como TEDA Esquecimento. Ao final, esses três métodos foram avaliados de acordo com métricas de desempenho utilizadas na literatura. Foram utilizados dois cenários de testes considerando dados reais de medição fasorial, sendo utilizado a frequência, tensão de fase e corrente, totalizando 36000 amostras para cada grandeza elétrica. Para todos os cenários analisados, os algoritmos utilizados tiveram um desempenho satisfatório, com destaque para o TEDA Janelado e o TEDA Esquecimento que obteve os melhores resultados nos cenários mais desafiadores. Ainda foi avaliado a capacidade dos métodos detectarem dados anômalos durante distúrbios elétricos. Por fim, é esperado que essa pesquisa contribua para o tema, apresentando uma nova abordagem para detecção de *outliers* em dados de PMU e contribuindo para melhoraria da qualidade dos dados, garantindo que suas aplicações nos sistemas de energia sejam utilizadas de forma mais segura.

Palavras-chave: Outliers, Medição Fasorial, PMU, TEDA Janelado, TEDA, Tipicidade e Excentricidade, Fator de esquecimento.

ABSTRACT

Currently, phasor measurement units (PMU) are essential equipment for real-time operation of the electrical system. However, due to complex factors, these data can easily be compromised by interference, synchronization failure or even a failure of some equipment. This will raise several levels of data quality problems, potentially affecting applications that are based on data quality or at the same time threatening to secure two power systems. This work aims to detect these anomalies based only on the data of the PMU, without the need for knowledge of electrical parameters of the network. For this purpose, an online algorithm was proposed based on the eccentricity and typicality of the sample in relation to its set of previous samples, or TEDA. It was proposed a modification in TEDA, or TEDA Janelado, where its main objective was to deixate the method more sensitive to variations in media and variation over time, maximizing the eccentricity of a sample and favoring the detection of *outliers* more difficult . A third strategy to improve the detection of TEDA was also presented in the work, or sketch factor, whose role is to establish a more fair weight between the new samples and the average/variance of its set of data, this method was called TEDA Sketch. Finally, these three methods were validated according to performance metrics used in the literature. Two test scenarios were used considering given phasor measurement areas, being used at frequency, phase voltage and current, totaling 36000 samples for each electrical magnitude. For all the scenarios analyzed, the algorithms used have a satisfactory performance, with emphasis on the TEDA Janelado and the TEDA Esquecimento that obtain the best results in the most challenging scenarios. It has also been certified to be capable of two methods to detect abnormalities during electrical disturbances. Finally, it is hoped that this research will contribute to this topic, presenting a new approach for detecting outliers in PMU data and contributing to better quality of data, guaranteeing that its applications in energy systems are used more safely.

Keywords: Outliers, Phasor Measurement, PMU, Windowed TEDA, TEDA, Typicality and Eccentricity, Thinning factor.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CNN	*Convolutional Neural Network*
DBSCAN	*Density Based Spatial Clustering of Applications with Noise*
DFT	Transformada Discreta de Fourier
EMS	*Energy Management System*
FCM	*Fuzzy C-means*
FFT	Transformada rápida de Fourier
FN	Falsos Negativos
FP	Falsos Positivos
GPS	Sistema de Posicionamento Global
KM	*K-means*
LOF	*Local Outlier Factor*
LoOP	*Local outlier probability*
LSTM	*Long Short Term Memory*
MCC	Coeficiente de correlação de Matthews
MDA	Desvio absoluto da mediana
ONS	Operador Nacional do Sistema Elétrico
PCA	Análise de componentes principais
PDC	Concentrador de Dados Fasoriais
PMU	Unidade de Medição Fasorial
RNA	Rede Neural Artificial
RNN	Rede neural recorrente
SCADA	*Supervisory Control And Data Acquisition*
SVM	*Suport Vector Machine*

TEDA	*Typicality and Eccentricity Data Analytics*
UTC	*Universal Time Coordinated*
VAR	Vetor auto-regressivo estocástico
VN	Verdadeiros Negativos
VP	Verdadeiros Positivos

SUMÁRIO

1 INTRODUÇÃO

1.1 PERTINÊNCIA E MOTIVAÇÃO DO TRABALHO

Desde o surgimento dos sistemas elétricos de potência, independente de onde sejam utilizados, há uma preocupação com a qualidade e constância da energia elétrica transmitida. Ao longo do tempo, fatores contribuíram para o aumento da sua complexidade como a expansão da rede, adição de elementos de geração, seja ela tradicional ou distribuída, rigor da regulamentação, instalação de novos equipamentos e novas tecnologias.

Esses fatores conferem um caráter crítico a operação de um sistema de potência, surgindo a necessidade de realizar um monitoramento da rede com qualidade e com segurança, capaz de detectar anomalias e aplicar correções de maneira que não haja impacto na operação do sistema. O impacto de eventos negativos no sistema é facilmente sentido e pode trazer danos graves à diferentes setores da sociedade. Medição síncrona por meio de PMU são tipicamente utilizados para monitoramento e controle da rede, além de análise do seu comportamento. Por isso é necessário que os dados coletados em um PMU sejam de confiança e alta qualidade, que permita um monitoramento em tempo real da rede (MAHAPATRA; CHAUDHURI; KAVASSERI, 2016; ZHOU et al., 2018).

Segundo informação do Operador Nacional do Sistema (ONS) todo agente de operação que deseja se integrar à Rede Básica deve instalar equipamentos de medição fasorial, por isso que é relevante desenvolver estudos que abordem soluções para problemas que possam afetar o desempenho do equipamento. Entretanto, este equipamento é susceptível a algumas fontes de erro, com relevância para este trabalho, principalmente interferências estruturais ou de comunicação que prejudiquem a qualidade do dado de PMU aquisitado. Por exemplo, é possível que haja erros no sinal transmitido, equipamentos mal calibrados podem inserir *offsets* nos sinais de medição ou até mesmo fatores físicos da rede podem inserir dados errôneos (SUNDARARAJAN et al., 2019).

De maneira geral, a PMU é utilizado em conjunto com outros equipamentos de monitoramento da rede e é responsável por transmitir informações de frequência, fasor de corrente e fasor de tensão. Por serem variáveis interdependentes, o efeito de *outliers* em qualquer uma delas refletirá em uma baixa confiabilidade da informação recebida.

Dito isso, o objeto de estudo se concentra nos dados de má qualidade provenientes de falhas dos dados de uma PMU. Um bom algoritmo de detecção de *outliers* influenciará diretamente na qualidade dos dados aquistados. Nosso objetivo ao examinar este assunto é apresentar uma avaliação científica da viabilidade da abordagem proposta para lidar

com os *outliers*. No capítulo posterior é abordado o estudo do estado da arte sobre o tema, apontando os métodos clássicos presentes na literatura e principais aplicações de algoritmos de detecção de *outliers* em sincrofasores.

1.2 OBJETIVOS

Este trabalho tem como objetivos:

1.2.1 OBJETIVO GERAL

O objetivo deste trabalho consiste em utilizar de técnicas orientadas a fluxos de dados para realizar a detecção de *outliers* em dados de medições fasoriais sincronizadas obtidas através de uma PMU.

1.2.2 OBJETIVO ESPECÍFICO

Apresentar uma metodologia para detecção de *outliers* para dados de sincrofasores baseado no conceito de tipicidade e excentricidade, utilizando o algoritmo TEDA.

A aplicação será realizada considerando conjunto de dados de uma única PMU com inserção de *outliers* em diversos cenários. Assim, para atingir o objetivo do trabalho será realizada a construção de algoritmos com um método baseado no TEDA, seguindo as etapas abaixo:

- Criação dos cenários de avaliação dos métodos;
- Proposição de uma melhoria do TEDA Clássico, focado na observação de um subconjunto de amostras restrito;
- Implementação de um fator de esquecimento para o TEDA Clássico;
- Avaliação de desempenho entre os algoritmos utilizando métricas amplamente utilizadas na literatura;
- Discussão dos resultados obtidos;

1.3 ORGANIZAÇÃO DO TRABALHO

Este trabalho está dividido em cinco capítulos. No primeiro, é feita uma introdução do tema, abordando a problemática envolvida e a proposta de solução. Além disso, foram citados os objetivos gerais e específicos.

No segundo capítulo, será feita uma fundamentação teórica sobre o tema detecção de anomalias em medição fasorial. Serão abordados conceitos como sincrofasores e sistemas de medição fasorial, tipos de anomalias, algumas técnicas existentes serão discutidas e o algoritmo do qual deriva a solução proposta.

O terceiro capítulo abordará a metodologia utilizada, partindo de uma visão geral da técnica utilizada concluindo numa descrição de todas as técnicas utilizadas e o seu funcionamento, através do que foi implementado.

No quarto capítulo, os resultados serão apresentados em forma de tabela com o resumo da matriz de confusão, em seguida uma análise do desempenho baseado em métricas amplamente utilizadas na literatura e comentários sobre cada cenário apresentado.

O quinto capítulo constará a conclusão do trabalho, comparando os resultados obtidos com os esperados ao traçar os objetivos. Ademais, serão discutidas as futuras melhorias deste trabalho, assim como o cronograma para defesa final.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo, exploraremos conceitos teóricos essenciais para a compreensão deste estudo. Iniciaremos por uma explicação sobre as Unidades de Medição Fasorial (*Phasor Measurement Units* – PMU) e seus principais componentes e características. Posteriormente, serão apresentadas as principais técnicas encontradas na literatura para detecção de outliers em séries temporais e em dados PMU.

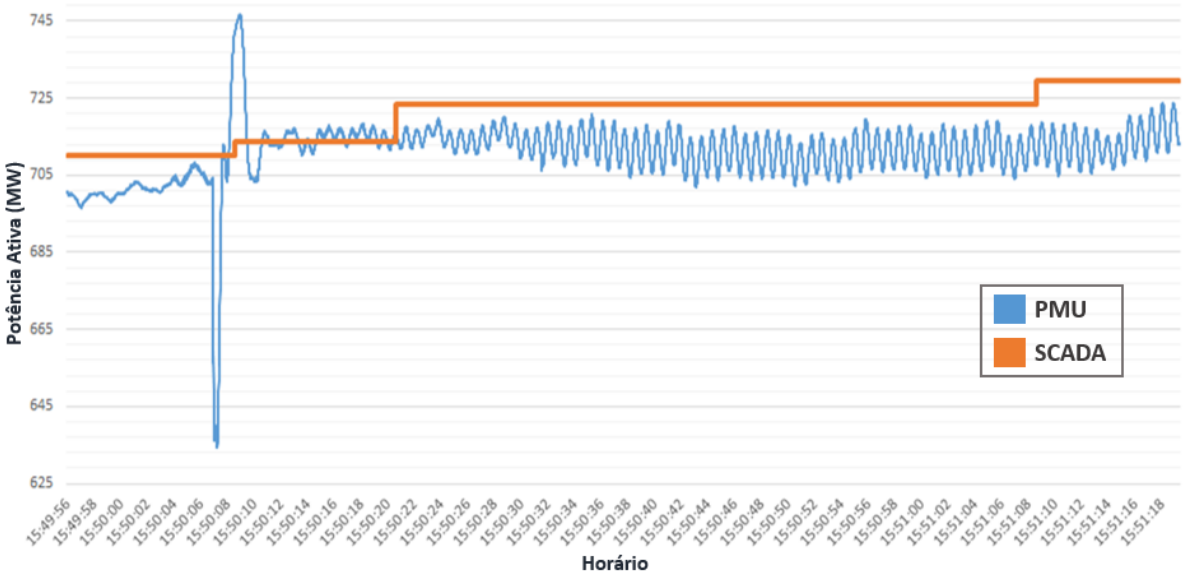
2.1 MEDIÇÃO FASORIAL SINCRONIZADA

A tecnologia de fasores sincronizados no tempo tem a sua história se iniciando a partir de um componente chave, o desenvolvimento da tecnologia de unidade de medição fasorial (PMU) na década de 1980. Este trabalho iniciou-se na American Electric Power Service Corporation, com a liderança principal de Arun Phadke e Jim Thorp, tendo posteriormente a equipe se transferido para a Virginia Tech. A tecnologia em si envolve o cálculo de medições fasoriais em tempo real sincronizadas com uma referência de tempo absoluta, fornecidas pelo Sistema de Posicionamento Global (GPS), para fornecer informações instantâneas e precisas sobre a condição do sistema elétrico devidamente monitorado, permitindo assim que ações corretivas imediatas ocorram (PHADKE, 2015).

Atualmente, as ferramentas SCADA(*Supervisory Control And Data Acquisition*) e EMS(*Energy Management System*) são empregadas para auxiliar o operador do sistema de potência em seus esforços para otimizar a operação do sistema elétrico. Entretanto, é comumente observado que os sistemas SCADA e EMS frequentemente carecem da agilidade necessária para capturar integralmente a dinâmica inerente ao sistema de potência (KARLSON; HEMMINGSSON; LINDAHL, 2004). Na Fig. 1 é apresentando gráficamente a difenrença entre os valores reais observados pela PMU e pelo SCADA durante uma oscilação transitória eletromagnética.

O controle de ângulo, por exemplo, manifesta maior nível de precisão quando baseado em PMU, dado que na ausência destas, a mensuração e regulação do ângulo se estabelecem por meio do fluxo de potência, um método de natureza indireta. Uma PMU é um dispositivo para medição sincronizada de tensões e correntes alternadas, com uma referência de tempo (ângulo) comum. A referência de tempo mais comum é o sinal do sistema de posicionamento global (GPS), que tem uma precisão melhor que 1 μs. Desta forma, as grandezas de corrente alternada podem ser medidas, convertidas em fasores (números complexos representados por sua magnitude e ângulo de fase) e com marcação de tempo (KARLSON; HEMMINGSSON; LINDAHL, 2004).

Figura 1 – Gráfico dos valores reais de uma PMU x SCADA durante oscilação transitória eletromagnética.



Fonte: Autoria Própria.

2.1.1 SINCROFASORES

Para entender o conceito de sincrofasores é necessário inicialmente conhecer a representação fasorial de uma grandeza elétrica. Um fasor consiste em uma expressão matemática polar que representa um número complexo, ou seja, podemos usar esse artifício para representar os valores de tensão e correntes senoidais na sua frequência fundamental, sendo composto por uma amplitude representada em valor eficaz (rms) e em um ângulo que coincide com a defasagem entre o pico da senoide e o eixo correlacionado à referência temporal. O conceito de fasor é apresentado na Fig. 2, onde a referência fasorial de tempo é definida em $t = 0$ (PHADKE; THORP, 2006).

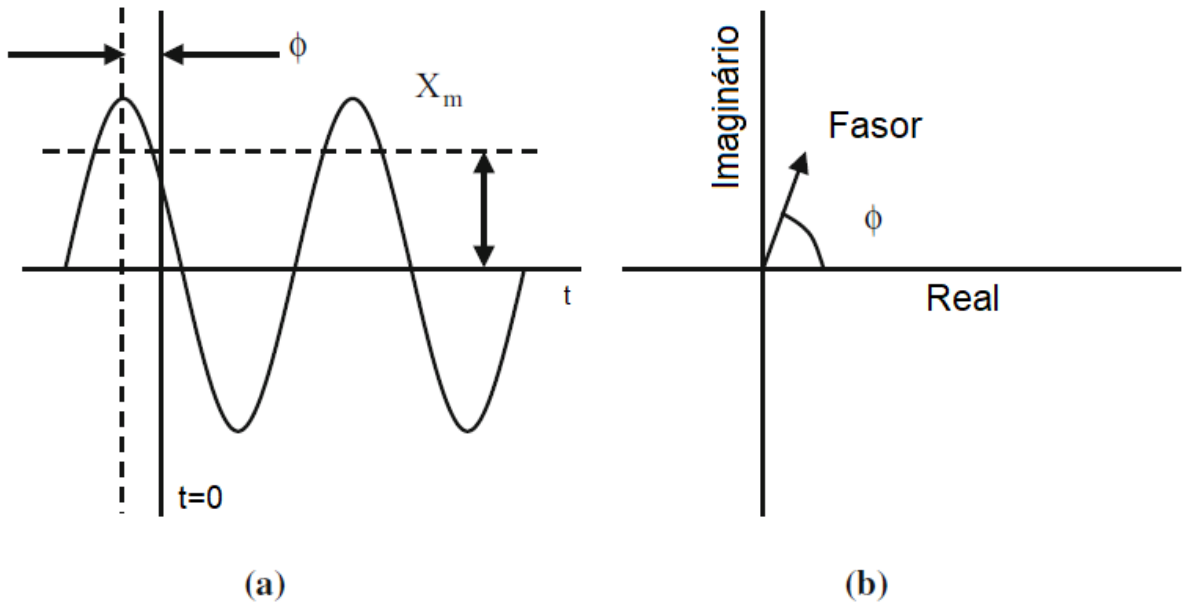
Segundo a norma IEC/IEEE 60255-118-1 (MARTIN et al., 2018), a representação fasorial de um sinal senoidal $x(t)$ é o valor complexo dado pela equação 2.1,

$$x(t) = \frac{X_m}{\sqrt{2}} e^{j\phi} = \frac{X_m}{\sqrt{2}} (cos\phi + jsen\phi) \tag{2.1}$$

onde, $\frac{X_m}{\sqrt{2}}$ é o valor rms do sinal $x(t)$ e ϕ é a sua fase. A fase é sincronizada com base no *Universal Time Coordinated* (UTC).

O ângulo de fase ϕ está relacionado diretamente à referência de tempo adotada. Logo, ao sincronizar o processo de amostragem de diferentes sinais, mesmo que fisicamente estejam distantes entre si, é plausível admitir que seus fasores podem ser representados em um único diagrama fasorial, desde que, atendam a necessidade de uma referência temporal única, ou seja, uma sincronização.

Figura 2 – Uma senoide (a) e sua representação como um fasor (b). O ângulo de fase do fasor é arbitrário, pois depende da escolha do eixo $t = 0$.



Fonte: (PHADKE; THORP, 2008).

Na Fig. 3 é ilustrado o processo de aquisição dos fasores sob uma mesma referência de tempo. A exatidão deste processo é devido a utilização do GPS, através do pulso por segundo transmitido por ele. Este pulso que é recebido por qualquer receptor no planeta coincide com todos os outros pulsos recebidos dentro de 1μs. Vale salientar que esses pulsos são fornecidos por relógios precisos que são mantidos pelo GPS e não consideram os fusos horários locais. Para que seja considerada uma hora universal em todas PMUs do sistema, os receptores também são responsáveis por compensar as diferenças de tempo em relação a rotação da Terra (PHADKE; BI, 2018).

A medição dos sincrofasores permite que o estado do Sistema Elétrico de Potência seja estimado de forma mais rápida, exata e confiável. Uma consequência direta desta facilidade é a possibilidade de se aferir a dinâmica do sistema elétrico.

2.1.2 SISTEMA DE MEDIÇÃO FASORIAL

Um sistema de medição fasorial constitui-se basicamente de três elementos: a Unidade de Medição Fasorial (PMU), o Concentrador de Dados Fasoriais (PDC) e os canais de comunicação (PHADKE; THORP, 2008). As medidas de tensão e de corrente são processadas pela PMU, convertidas em fasores e enviadas ao concentrador de dados a taxas que, normalmente, variam de 10 a 60 fasores por segundo. A Fig. 4 representa um sistema com N PMUs.

Figura 3 – Obtenção dos fasores sob uma referência comum temporal.

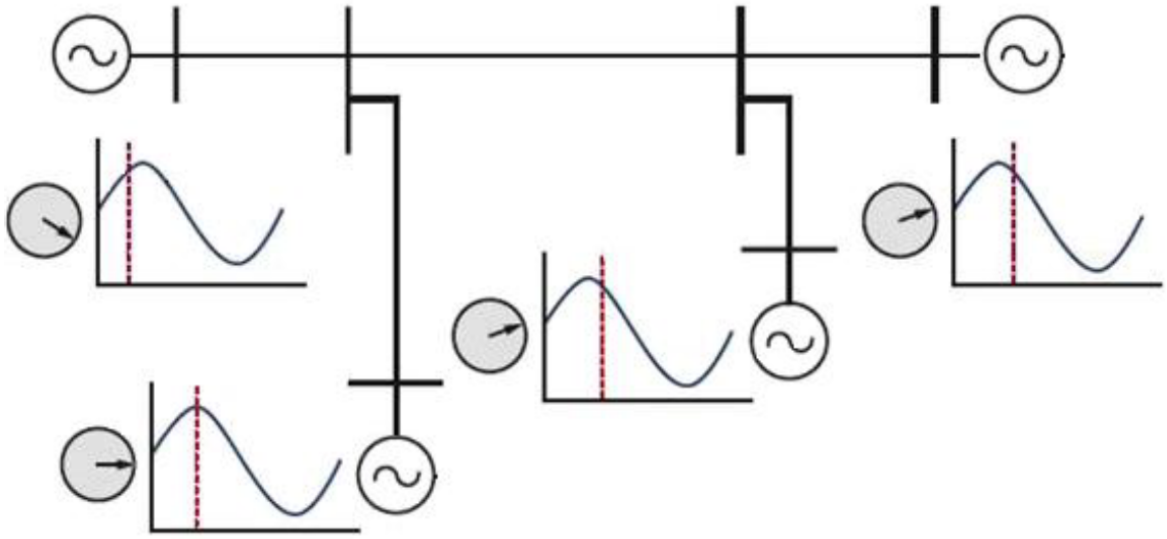
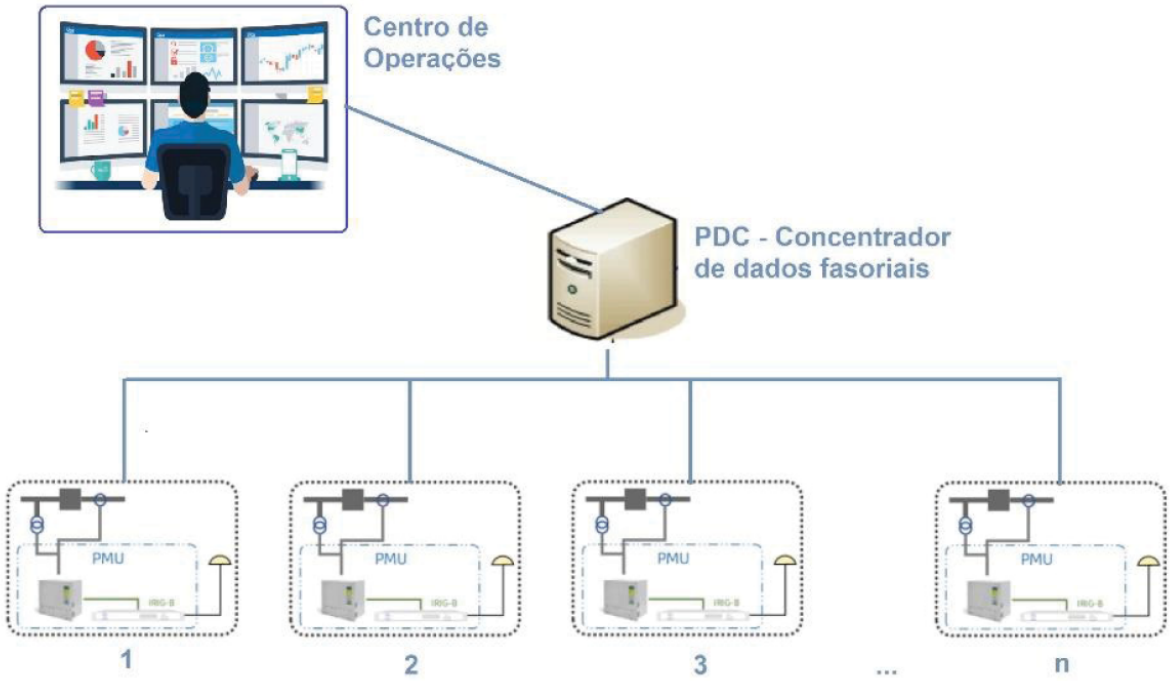


Figura 4 – Esquema de um sistema de medição fasorial de N PMUs.



2.1.2.1 CONCENTRADOR DE DADOS FASORIAIS - PDC

O PDC é um dispositivo (*hardware* e *software*) responsável por receber o fluxo de dados de PMU, calcular novos sinais, arquivar dados e enviar fluxos de dados para outros PDCs. A comunicação entre as PMUs e o PDC é assíncrona e, dessa forma, os dados enviados por estas unidades de medição chegam ao equipamento de forma desordenada,

podendo apresentar atrasos e até perdas de informações (KAMWA; SAMANTARAY; JOOS, 2012).

Sua função típica é reunir dados de várias PMUs, rejeitar dados inválidos, alinhar os registros de tempo e criar um registro coerente de dados gravados simultaneamente de todas os pontos de medição do sistema observado. Existem instalações de armazenamento locais nos PDCs, bem como funções de aplicativos que precisam dos dados de PMU, que devem ser disponibilizado pelos PDCs para os aplicativos locais em tempo real. Perceptivelmente, todo esse processo de comunicação e o gerenciamento de dados criam uma maior latência em tempo real, mas a experiência prática mostra que isso pode ser controlado (DERVIŠKADIĆ et al., 2016). O relógio interno do concentrador de dados é sincronizado com a referência de tempo UTC, da mesma forma que nas PMUs, visando o controle de tempo para monitoramento de desempenho do sistema e diagnóstico de problemas.

Devido a sua complexidade, o concentrador de dados deve atender alguns requisitos básicos de desempenho, para que suas funcionalidades não sejam afetadas durante a sua utilização, como:

- Alta capacidade de processamento;
- Disponibilidade e Confiabilidade;
- Sistema de comunicação;
- Capacidade de Armazenamento de dados;
- Compatibilidade aos diversos tipos de aplicações;
- Alta modularidade (Facilidade na integração e expansão).

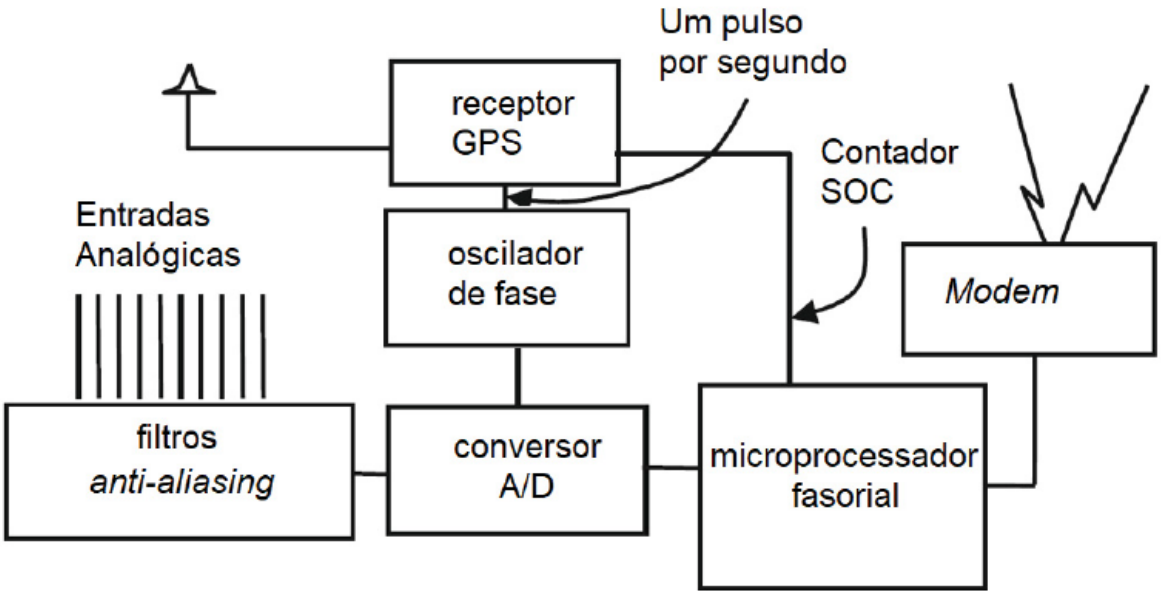
2.1.2.2 ESTRUTURA DA PMU

Os fabricantes de PMUs estão desenvolvendo seus produtos continuamente e por esse motivo esses equipamentos diferem uns dos outros em vários aspectos de *hardware* e *softwares*, o que torna difícil definir um modelo de PMU universal. Segundo a literatura, uma PMU genérica possui alguns componentes principais e essenciais para o seu funcionamento. Na Fig. 5 representa o que pode ser considerado um estrutura típica de uma PMU.

Inicialmente, o sinal analógico é obtido através dos transformadores de tensão e corrente. Em seguida, é necessário garantir a máxima frequência do sinal à metade da frequência de amostragem (critério de *Nyquist*), para tal são utilizados os filtros *antialiasing* (HUANG et al., 2008). Durante a conversão analógico-digital, o sinal passa por um processo de discretização e as amostras recebem uma etiqueta de tempo baseada

do sinal de sincronização, onde a referência de tempo é dado pelo sinal do GPS. No processamento, é feito a extração do fasor utilizando técnicas como a transformada discreta de Fourier(DFT) ou a transformada rápida de Fourier (FFT) (PHADKE; THORP, 2008). Por fim, a medição e sua estampa de tempo é enviada para o PDC por meio dos canais de comunicação.

Figura 5 – Modelo clássico de uma PMU



Fonte: Adaptado Phadke e Thorp (2008).

2.1.2.3 CANAIS DE COMUNICAÇÃO

Os canais de comunicação possuem um propósito bem definido, que é facilitar a transferência de dados entre a PMU e o PDC, entre diferentes PDCs, ou entre o PDC e as aplicações. Segundo Phadke e Bi (2018) as principais tecnologias utilizadas para essa comunicação são baseadas em fibra óptica e sistemas de comunicação sem fio.

Um dos principais fatores distintivos entre os sistemas de comunicação é o atraso na transferência de dados. Esses atrasos têm um impacto direto no desempenho do sistema de medição fasorial, especialmente em aplicações voltadas para controle e proteção. A seleção do sistema de comunicação a ser adotada está condicionada à aplicação específica em consideração. Tipicamente, aplicações de controle e proteção demandam uma velocidade de comunicação superior em comparação as aplicações de monitoramento.

A disponibilidade dos dados depende da combinação de performance das PMUS, PDCs e links de comunicação. A disponibilidade de dados mínima consiste no número mínimo de fasores completos para que a informação possa ser utilizada pela aplicação que os necessita. É aceitável que parte dos dados se perca, entretanto deve-se garantir que os

dados que restaram possam gerar a informação com qualidade necessária (DENG et al., 2019).

2.2 TIPOS DE OUTLIERS EM DADOS DE PMU

Os sincrofasores fornecem informações valiosas para a consciência situacional, operação e controle do sistema elétrico. Infelizmente, por se tratar de uma rede física, tais equipamentos estão suscetíveis a fatores externos como instabilidade de comunicação e/ou falha de *hardware*, com isso a qualidade de seus dados pode ser muito prejudicada por anomalias afetando diretamente as ferramentas que depende desses dados (BABA et al., 2022). Segundo Hill e Minsker (2010), em geral, as anomalias nesse tipo de sistema são ocasionadas por três motivos: falha do sensor, erro na transmissão de dados e comportamento pouco frequente do sistema.

A seguir, serão apresentados os principais tipos de anomalias nos dados de PMU que estão presentes na literatura: pico aleatório, pontos faltantes ou zeros, padrão errôneo e ruído de alta frequência.

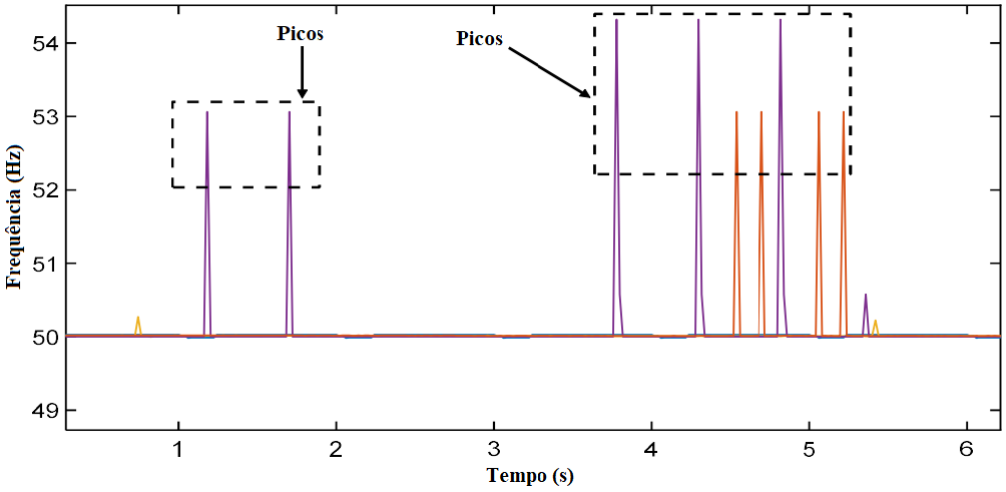
2.2.1 PICOS ALEATÓRIOS OU *SPIKES*

O pico aleatório ou *spike* é um típico problema presente nos dados de sincrofasores, que em geral é causado por problemas de *hardware* da PMU. A frequência e a magnitude dos picos variam de diferentes PMUs devido às suas características elétricas locais. Portanto, geralmente é difícil detectar um padrão para esse tipo de anomalia e aplicar um filtro uniforme para removê-los para todos os PMUs (BRAHMA et al., 2016). Conforme visto na Fig. 6, algumas medidas apresentam picos aleatórios no aspecto de amplitude e intervalo de tempo. Além disso, alguns picos continuam oscilando em torno da frequência do sistema. A tendência da oscilação indica que o valor médio da medição continua mudando ao longo do tempo.

2.2.2 DADOS FALTANTES OU ZEROS

Os dados faltantes em geral ocorrem devido a vários fatores, por exemplo, perda de sinal de GPS, falha de rede, falha de energia, etc. A detecção de dados ausentes é objetiva, pois cada medição de PMU recebe um estampa de tempo única, portanto, uma marcação de data/hora descontínuo implica na existência de dados ausentes. Todos os dados brutos são divididos em vários quadros não sobrepostos. Os dados ausentes resultam em descontinuidade e contornos das medições de PMU. Além disso, o número de quadros ausentes depende da janela da condição de falha. Em alguns casos, os dados ausentes do conjunto podem ser grandes, o que resulta em vales no conjunto de dados. Conforme

Figura 6 – Exemplo de alguns picos aleatórios nos dados de frequência.



Fonte: Adaptado de Deng et al. (2020).

mostrado na Fig. 7, dados ausentes coletados de uma rede elétrica podem ocorrer com frequência e resultar em uma queda repentina para 0 Hz na medição de frequência.

Figura 7 – Caso de alguns zeros nos dados de frequência de uma PMU real.

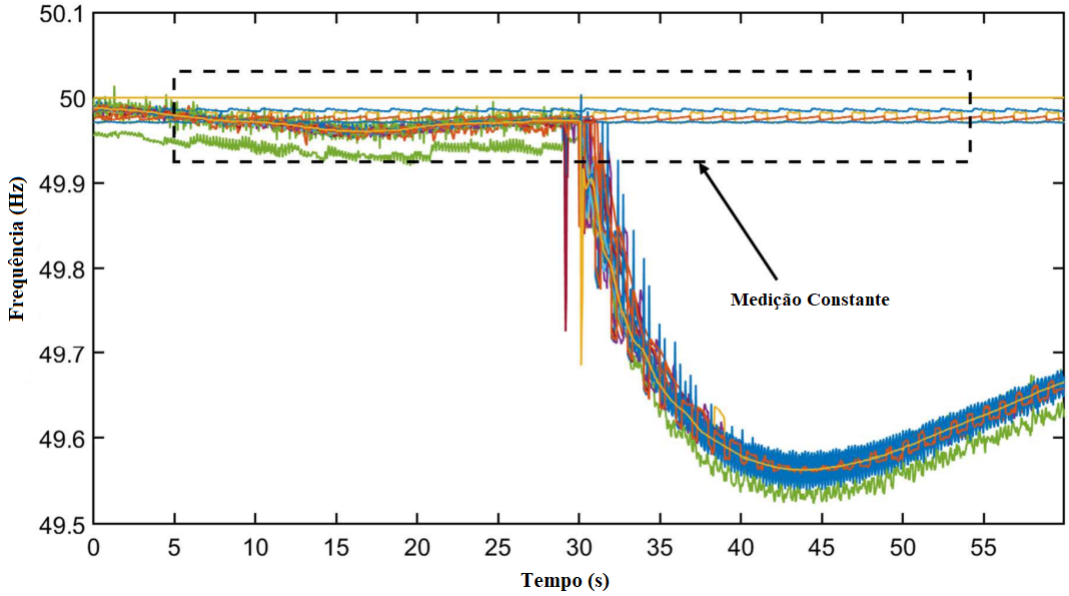
Fonte: Autoria Própria.

2.2.3 PADRÃO ERRÔNEO

Segundo Deng et al. (2020) define um tipo de anomalia denominada de Padrão Errôneo. Esse tipo de *outlier* são causados principalmente por problemas de *hardware* de PMU. Quando um erro de comunicação ou uma falha de *hardware* interrompe o fluxo de

dados normal e o concentrador de dados fasorial (PDC) pode repetir os últimos valores bons conhecidos e gravá-los em um banco de dados de série temporal. A Fig. 8 mostra um exemplo durante o desligamento de um gerador. Os aplicativos, ao considerar as informações desse conjunto de dados teriam um funcionamento ruim, pois alguns dados são preenchidos com valores constantes. Por outro lado, esse tipo de *outlier* pode auxiliar a detecção de perturbações (YU et al., 2019), incluindo o de ilhamento, devido a que os valores de frequência de várias PMUs divergem entre si quando ocorre o desligamento da máquina.

Figura 8 – Exemplo de um padrão errôneo de frequência.



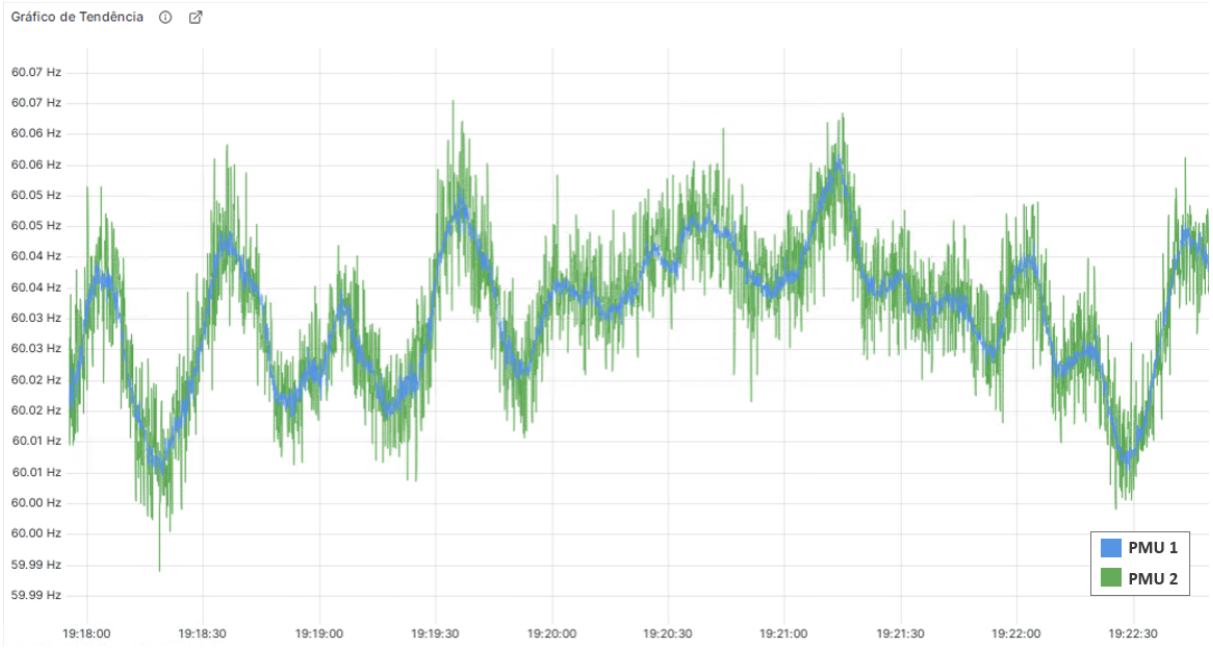
Fonte: Adaptado de Deng et al. (2020).

2.2.4 RUÍDOS DE ALTA FREQUÊNCIA

A medição com ruído de alta frequência é causada por um controle de intervalo de amostragem impreciso ou configuração inadequada do *hardware* da PMU. Esses ruídos podem continuar oscilando em torno da frequência do sistema (DENG et al., 2020). Com base nos dados de medição da rede elétrica, a amplitude do ruído varia em uma faixa relativamente ampla, de 0,01 a 0,2 Hz. Conforme mostrado na Fig. 9, o ruído de alta frequência para a PMU 02 tem pequenas diferenças na amplitude em relação a PMU 01. Assim, a aleatoriedade e a variedade do ruído de alta frequência dificultam sua remoção com um filtro de limite uniforme.

Conforme visto da Fig. 6 a Fig. 9, cada tipo de anomalia nas medições dos sincrofasores possui características únicas em comparação com os dados normais. De fato, convencionalmente, cada tipo de anomalia requer um método de detecção dedicado. Caso

Figura 9 – Caso de uma medição real de duas PMUs, sendo a PMU 02 com ruído de alta frequência.



Fonte: Autoria Própria.

contrário, alarmes falsos e problemas de queda de precisão em aplicações de sincrofasores podem surgir devido a detecções imprecisas e medições anormais.

2.3 ALGORITMOS DE DETECÇÃO DE OUTLIERS

A tecnologia de sincrofasores fornece medições fasoriais sincronizadas no tempo de um sistema de posicionamento global (GPS) de alta precisão. A utilização dessa ferramenta é de grande utilidade para o acompanhamento da dinâmica da rede elétrica e é amplamente utilizado em aplicações, incluindo detecção de distúrbios, localização de falhas, segurança cibernética, entre outros.

Entretanto a acurácia dessas aplicações têm relação direta com a qualidade das medições dos sincrofasores, e as anomalias nas medições podem afetar seu desempenho consideravelmente. Zhao et al. (2017) analisa o impacto de medições imprecisas de sincrofasores na aplicação de triangulação de distúrbios elétricos na rede. De acordo com este trabalho, o resultado de localização é sensível a medições de ângulo imprecisas quando as magnitudes dos distúrbios são pequenas. Medições anômalas também podem afetar a precisão da detecção de interrupção de linha (CHEN; WANG; ZHU, 2014). Além disso, dados fasoriais incompletos podem ter impactos de várias escalas nas localizações de faltas (BECEJAC; DEHGHANIAN; KEZUNOVIC, 2016) e na avaliação da estabilidade de tensão (TANG et al., 2011).

No entanto, a detecção de outliers em um conjunto de dados é sempre uma questão bastante complexa. De fato, a escolha de um método é direcionado a um tipo específico de anomalia, tornando tal método ineficiente para outros tipos de problema. Outro ponto, se dá nos ajustes dos parâmetros dos modelos, que é uma tarefa bastante complicada e que, em geral, depende de um conhecimento prévio e da sensibilidade do especialista. Por fim, a grande maioria dos modelos requer um alto esforço computacional o que os torna inviáveis para aplicações em tempo real.

Existem diversos trabalhos na literatura que propõe técnicas de detecção de *outliers*. Esses algoritmos podem ser classificados de acordo ao conhecimento prévio do conjunto de dados, sendo divididos em sua maioria, em dois grupos: supervisionados e não-supervisionados. Os algoritmos supervisionados necessitam de um conhecimento prévio dos dados, utilizando um conjunto de dados na fase de treinamento, que contém as amostras normais e os *outliers*. Todavia, os algoritmos não-supervisionados não possuem a necessidade do conhecimento prévio dos dados, ou seja, o algoritmo é capaz de identificar os *outliers* sem um treinamento anterior, criando uma facilidade para aplicação. Entretanto, os algoritmos não-supervisionados depende que os dados ditos "normais"sejam muito mais frequentes, caso contrário isso ocasionará em um elevado numero de falsos positivos ou falsos negativos.

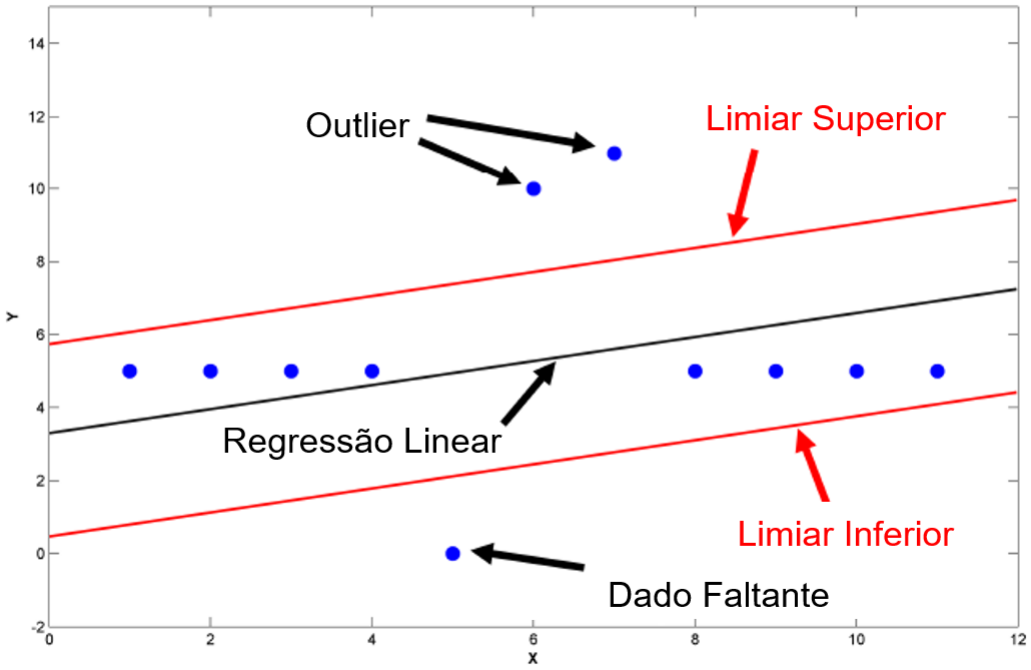
Logo, os algoritmos online são uma alternativa para contornar tais complexidades, pois possuem a vantagem de processar grande quantidade de dados em tempo real de maneira eficiente e eficaz (LIN et al., 2022). Essa característica demonstra sua relevância para aplicações em dados sincrofasoriais, onde os dados são contínuos e existe a necessidade de uma resposta rápida para as ferramentas que utilizam esse fluxo de dados.

Ren, Ye e Li (2017) apresentou uma abordagem de detecção de anomalias baseada em um modelo dinâmico de Markov para *streaming* de dados. Considerando a propriedade de memória curta de um modelo clássico de Markov e a propriedade de memória longa de um modelo de Markov de ordem superior, o modelo do autor propõe uma janela deslizante dos dados, que se adapta melhor a necessidade de cada conjunto. Especifica-se que os estudos experimentais utilizando conjuntos de dados simulados e conjuntos de dados de sistemas reais mostraram que o método proposto melhora a adaptabilidade e estabilidade da detecção de anomalias na serie temporal.

A regressão linear visa minimizar as distâncias verticais totais entre os pontos do conjunto de dados e a uma linha ajustada, conforme Fig. 10. O limiar superior e inferior é estabelecido pelo especialista, onde é definido um valor de k que é acrescido e decrescido da reta calculada (LAZAREVIC; KUMAR, 2005).

O teorema de Chebyshev determina um limite inferior da porcentagem de dados que existe dentro do número k de desvios padrão da média. No caso de distribuição normal,

Figura 10 – Detecção de outlier utilizando regressão linear.



Fonte: Adaptado de Zhou et al. (2018).

95% dos dados estão dentro de dois desvios padrão da média. Isto significa que se espera que 5% dos dados estejam fora de dois desvios padrão da média. O método proposto por (AMIDAN; FERRYMAN; COOLEY, 2005) utiliza a desigualdade de Chebyshev para identificar *outliers*, utilizando como limiar a Eq. 2.2.

$$P(|X - \mu| \leq k\sigma) \geq \left(1 - \frac{1}{k^2}\right) \tag{2.2}$$

onde: X representa os dados de entrada da PMU, μ é a média dos dados, σ é o desvio padrão dos dados e k representa o número de desvios padrão da média. Note que o parâmetro k é otimizado para detecção de valores discrepantes de PMU e os limites superior e inferior correspondentes H e L são dados como pelas Eq. 2.3 e 2.4. Os dados fora do intervalo determinado pelos limites superior e inferior são considerados uma anomalia.

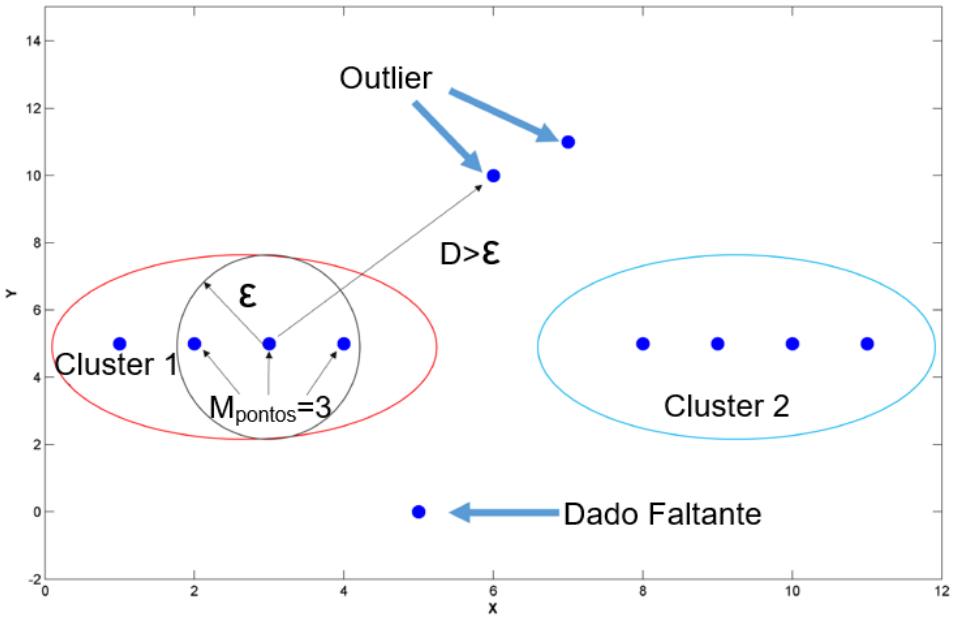
$$H = \mu + k * \sigma \tag{2.3}$$

$$L = \mu - k * \sigma \tag{2.4}$$

O DBSCAN(*Density Based Spatial Clustering of Applications with Noise*)(ESTER et al., 1996) é aplicado para detectar valores discrepantes e dados ausentes com ruído. Este método possuí vantagens para encontrar anomalias entre clusters de formato irregular e, devido a sua irregularidade não seriam detectados pelos modelos baseados em regressão linear e o na desigualdade de Chebyshev. É necessário definir dois parâmetros, o raio (ϵ) e

o número mínimo de pontos dentro do raio ϵ (M_{pontos}). A ideia central do DBSCAN é que um cluster é uma região densa de pontos que é separada por regiões de menor densidade. O algoritmo começa com um ponto de dados aleatório e verifica se há pontos próximos a uma distância ϵ. Se houver pontos suficientes dentro desse raio, ele forma um cluster e explora recursivamente os pontos próximos para expandir o cluster. Isso continua até que todos os pontos dentro do cluster tenham sido identificados. Pontos que não podem ser alcançados a partir de nenhum cluster são rotulados como *outlier*, Fig. 11.

Figura 11 – Detecção de outlier baseado em DBSCAN.



Fonte: Adaptado de Zhou et al. (2018).

Um algoritmo de aprendizagem baseado em modelo *ensemble* é proposto por Zhou et al. (2018), ele combina múltiplos algoritmos de detecção para criar um mecanismo de detecção mais robusto em um conjunto de dados de tensão de PMU. Para fase de treinamento, o autor utilizou três métodos como bases para detecção de *outliers*: Regressão Linear, Chebyshev, e DBSCAN, todos os métodos utilizam uma mesma janela de dados com comprimento N. A principal vantagem do modelo *ensemble* é que ele foi treinado de forma eficiente, alcançou alta precisão na detecção de erros diversificados e superou seus equivalentes que usam um único método de detecção.

Jones, Pal e Thorp (2014) apresenta um modelo de previsão de séries temporais combinado com filtro de Kalman e um algoritmo de suavização para limpar os dados ruins em dados de sincrofasores. Seu principal objetivo é evitar que dados ruins sejam propagados para as ferramentas, fornecendo um fluxo continuo de dados em boas condições. As simulações indicam que esses algoritmos podem compensar valores de dados incorretos de até 20% do conjunto de dados com relativa facilidade.

Outro método de detecção de *outliers* é o *Z-score*, que é uma medida numérica estatística que informa a que distância um dado está da média. Mas tecnicamente é uma medida do desvio padrão abaixo ou acima do desvio padrão da população onde os dados são encontrados e convertidos numa pontuação bruta. Shiffler (1988) usa *Z-score* como uma solução simples para detectar *outliers*. Basicamente precisamos saber a média e o desvio padrão do conjunto, então pode ser calculada o *Z-score* para cada ponto do conjunto, atribuindo uma pontuação de acordo com a Eq. (2.5).

$$z = \frac{(X - \mu)}{\sigma} \tag{2.5}$$

em que: z é o *z-score*, X é o valor do elemento, μ é média da população e σ é o desvio padrão da população.

A eficácia do *Z-Score* é questionável pois ao usar a média amostral e o desvio padrão amostral porque a média amostral pode ser muito afetada por certos valores extremos (*outliers*) ou mesmo valores extremos de um único valor. Para resolver este problema, a mediana e o desvio absoluto da mediana (MDA) são usados nos *Z-Score* modificados para substituir a média amostral e o desvio padrão, respectivamente (IGLEWICZ; HOAGLIN, 1993).

Iglewicz e Hoaglin (1993) recomenda marcar observações como *outliers* quando $|M_i|$>3,5. A pontuação M_i é válida para dados normais e é igual à *pontuação Z*.

Já o K-means (KM), é uma técnica de agrupamento criada por MacQueen et al. (1967) e é um dos algoritmos de aprendizado não supervisionado mais simples para resolver este problema bem conhecido. O procedimento segue uma abordagem simples e fácil para classificar um determinado conjunto de dados por um certo número de *clusters* (digamos k *clusters*). Existem maneiras de usar K-means como uma técnica de detecção de *outliers*. Cada *cluster* possui um centróide, e os pontos pertencentes a este *cluster* são definidos pela distância a este centróide. Você pode então classificar *outliers* com base em sua distância absoluta do centróide, bem como na distância relativa (ou seja, a distância absoluta comparada com a distância média dos pontos médios dos centróides do *cluster*). Assim visando minimizar uma função do erro quadrático:

$$J = \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2 \tag{2.6}$$

em que: $\|x - \mu_i\|^2$ é uma medida de distância escolhida entre um ponto de dados x e o centro de *cluster* μ_i, sendo um indicador da distância dos n pontos de dados de seus respectivos centroides.

Do K-Means, surge o Fuzzy C-means (FCM), desenvolvido por Dunn (1973) e melhorado por Bezdek (1981), é uma técnica de agrupamento em que cada ponto de dados

pode pertencer a vários *cluster*. Tal como acontece com K-means, o número necessário de *clusters* deve ser predefinido e localizado para realizar c-means fuzzy. Ao contrário dos métodos de *Clustering* global, os FCMs são mais flexíveis, pois atribuem cada objeto a diferentes *clusters* com diferentes graus de associação, conforme mencionado acima. Em outras palavras, enquanto no KM a adesão ao *cluster* é exatamente 0 ou 1, no FCM ela varia entre 0 e 1. FCM minimiza a função objetiva definida como:

$$J = \sum_{k=1}^{c} \sum_{i=1}^{N} (u_{ki})^2 \left| x_i - c_k \right|^2 ; \tag{2.7}$$

Portanto, o FCM pode ser melhor que o KM nos casos em que não podemos determinar facilmente que um objeto pertence a apenas um cluster, especialmente no caso de conjuntos de dados ruidosos ou *outliers*. Por esta razão, quando usado como técnica de detecção de *outlier*, o FCM funciona da mesma forma que o KM, com a adição de um parâmetro fuzzer. O parâmetro fuzzer m, à medida que se aproxima de 1, tende a tornar os *clusters* mais densos, uma vez que os centróides tendem a estar mais próximos das localizações concentradas dos objetos, mas à medida que m se aproxima do infinito, inversamente, os centróides tendem a estar mais próximos das localizações concentradas dos objetos. Posicionamento de forma descentralizada (IZAKIAN; PEDRYCZ, 2013). Desta forma, controlando melhor a localização do centróide, é possível isolar *outliers*, que é uma opção não controlada em k-means.

Se tratando de redes neurais, temos ainda o autoencoder. Inicialmente, pode-se dizer que um autoencoder é uma rede neural artificial usada para aprender codificação eficiente de dados de maneira supervisionada (KRAMER, 1991), com suas primeiras aplicações datando da década de 1980 (SCHMIDHUBER, 2015).

Um autoencoder consiste em um codificador que aprende uma representação latente z dos dados e um decodificador que reconstrói os dados $d(z)$ com base em ideias de estrutura de dados (ou seja, o propósito do autoencoder). Quando usados para detectar *outliers*, considerando que a camada oculta de um autoencoder geralmente consiste em menos neurônios, para reconstruir a entrada da forma mais realista possível, os pesos ocultos capturam os parâmetros que melhor representam os dados de entrada originais e ignoram os detalhes de entrada dos dados, como *outliers*. Em outras palavras, dados normais são mais fáceis de reconstruir do que *outliers*. Resumindo, quanto maior a diferença entre a saída (ou seja, a entrada reconstruída) e a entrada original, maior a probabilidade de que os dados de entrada correspondentes sejam uma anomalia (XIA et al., 2015).

Uma abordagem para detecção de *outliers* durante distúrbios da rede foi proposto por Mahapatra, Chaudhuri e Kavasseri (2016). Considerando que dados ruins na forma de valores discrepantes podem ter aparência muito semelhante à da resposta do sistema após eventos elétricos como falhas. Seu método é baseado na análise de componentes principais,

Principal Component Analysis(PCA), e tem como proposta distinguir entre os valores discrepantes causados por dados ruins daqueles causados por perturbações.

Os componentes principais no subespaço de dimensão inferior capturam as propriedades dinâmicas do sistema e esses componentes principais no subespaço de dimensão superior representam informações com ruídos. Usando esta propriedade, supõe-se que valores anômalos devido aos dados incorretos resultarão em maior atividade no subespaço de alta dimensão (segunda, terceira e quarta dimensão), uma vez que, valores discrepantes devido a falhas elétricas têm menos probabilidade de produzir tais alterações em dimensões superiores. Para validar a eficácia do método proposto, o autor realizou uma simulação de Monte Carlo de 100 execuções considerando magnitudes variadas de valores ruins e ruído gaussiano branco aditivo introduzido nas medições de tensões obtidas no sistema de teste de 11 barras apresentado por Kundur (1994).

Devido à sua simplicidade, o PCA é reconhecido como uma ferramenta eficiente no processamento de uma grande quantidade de dados em diversos campos da ciência e tecnologia. É amplamente reconhecido como uma técnica de redução de dimensionalidade para melhor compreender o comportamento oculto das propriedades de um sistema. O PCA tem sido utilizado para pós-processamento de dados da PMU, que contém medições de diversos barramentos da rede. Esses dados após a aplicação do PCA podem ser usados para detecção de falhas em sistemas de geração de energia eólica (TIANZHEN et al., 2013), detecção de ilhamento (LIU et al., 2015), identificação de coerência de geradores e barramentos (ANAPARTHI et al., 2005) e avaliação de vulnerabilidade dinâmica (CEPEDA; COLOMÉ; CASTRILLÓN, 2011).

Wu e Xie (2016) propôs um algoritmo online baseado em dados para identificar medições ruins em dados de PMU, utilizando o método *Local Outlier Factor* (LOF) e aproveita a relação espaço-temporal entre as medições dos sincrofasores. O LOF é um algoritmo de detecção de anomalias baseado em densidade, foi apresentado por Breunig et al. (2000). A ideia por trás do LOF é que os *outliers* são pontos que têm uma densidade local significativamente menor do que a densidade dos pontos ao seu redor.

A proposta de Wu e Xie utiliza as medições de um grupo composto de vários PMUs, todas as séries temporais medidas pelos sincrofasores formam um banco de dados D, que pode ser representando como uma matriz $X = [X_1, X_2, ..., X_m]$ de ordem nxm, onde X_i é um vetor $nx1$, m é a quantidade de PMU presentes na rede elétrica e n é o número de pontos no tempo dentro do período observado. Em uma determinada janela n, a "distância"entre os dados da PMUs X_i e X_j(ou duas séries temporais) é definida por:

$$d(X_i, X_j) = \frac{1}{n-1} \sum_{k=1}^{n} (x_{ik} - \hat{x_{ik}})(x_{jk} - \hat{x_{jk}}) \tag{2.8}$$

onde: x_{ik} e x_{jk} representam o elemento no instante k dos vetores X_i e X_j, respectivamente;

$\hat{x_{ik}}$ e $\hat{x_{jk}}$ é o valor médio amostral dos vetores X_i e X_j, respectivamente. Intuitivamente, $d(X_i, X_j)$ denota a covariância amostral entre X_i e X_j . Esta definição de distância quantifica a similaridade da taxa de variação de duas séries temporais. Em seguida é calculado o LOFs de cada série temporal. Então, é considerado um limite predefinido para que os LOFs calculados sejam comparados afim de identificar as anomalias. Portanto, com uma determinação adequada do valor limite, dados ruins de PMUs podem ser identificados com sucesso. O limite LOF é um valor dependente do sistema e pode ser determinado por um especialista ou através de treinamento offline, utilizando dados históricos obtidos do mesmo sistema.

Vale destacar, que o LOF possui um desempenho computacional rápido, sendo capaz de detectar *outliers* em dados de PMU durante condições de operação normal e perturbadas. Além de ter sua estrutura puramente baseada em dados, não sendo necessário conhecimento sobre os parâmetro de rede e/ou topologia do sistema.

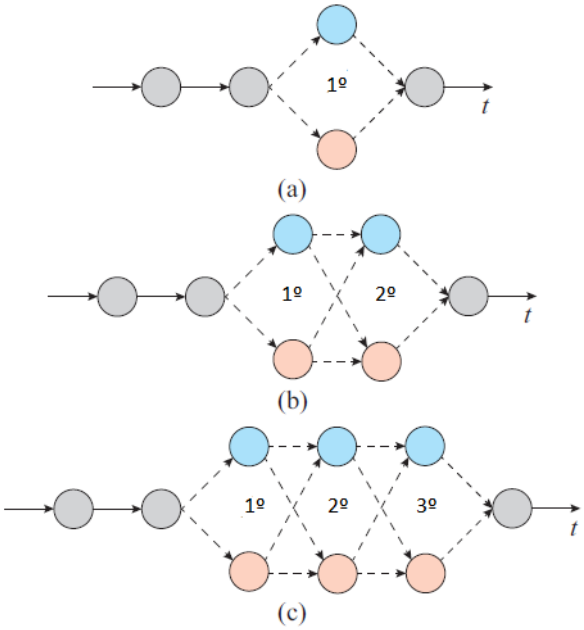
Matthews e Leger (2017) desenvolveram um *Map Reduce* para detectar anomalias nos dados e classifica-las como restritivas e temporais em grandes conjuntos de dados de PMU. A implementação é em sistema de 8 núcleos e os resultados são obtidos com bastante rapidez. Para fins de detecção de anomalias na rede elétrica, são analisadas medições de fasor de tensão, fasor de corrente e frequência. A principal vantagem deste trabalho é que apresentou melhorias significativas em comparação com trabalhos anteriores. Nesse caso, a detecção de anomalias é realizada online e com um funcionamento multivariado. Entretanto, uma desvantagem deste método é que requer enormes recursos computacionais.

Yang et al. (2020) a fim de melhorar a qualidade dos dados da PMU, propuseram um algoritmo de detecção outliers baseado em agrupamento espectral. O método possui a vantagem de não necessitar do conhecimento da topologia e dos parâmetros da rede elétrica. Inicialmente o autor propôs um método utilizando uma árvore de decisão para distinguir dados de eventos e dados ruins, usando o recurso de inclinação entre os os dados para caracterizar a variação entre as amostras.

Para montar a árvore de decisão, o autor considerou que os dados incorretos se desviam dos valores normais e em sua maioria estão isolados entre si. Na Fig. 12, os círculos azul representa os dados que assumiram valores superiores ao nominal, os círculos vermelhos representam os valores abaixo do nominal e o círculo cinza os valores nominais. Desse modo, considera-se que os dados ruins podem se apresentar até três vezes de maneira consecutivas, o que implica que as possibilidades máxima que a árvore pode assumir é igual a oito.

O autor ainda fez uma comparação considerando a regra 3σ (PUKELSHEIM, 1994), considerando que os dados ruins tende a ficar fora do intervalo $(\mu - 3\sigma, \mu + 3\sigma)$, onde μ representa o valor médio das amplitude e σ é o desvio padrão. Na Fig. 13, os dados

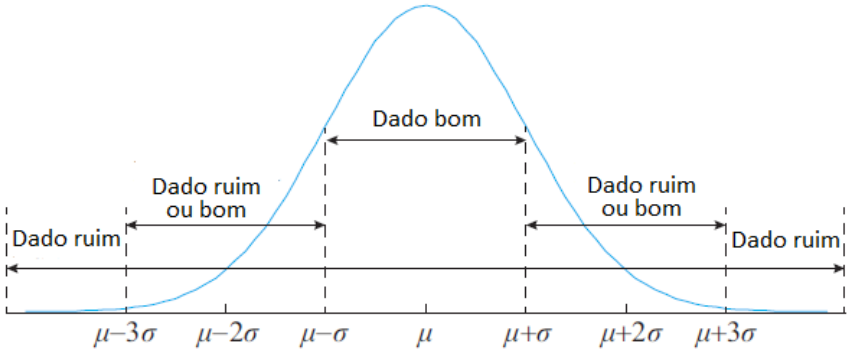
Figura 12 – Esquema de possíveis dados incorretos. (a) Um dado incorreto. (b) Dois dados incorretos. (c) Três dados incorretos.



Fonte: Adaptado de Yang et al. (2020).

distribuídos entre $(\mu - \sigma, \mu + \sigma)$ são considerados como dados normais. Os dados na extremidades da Gaussiana $(\mu - 3\sigma, \mu + 3\sigma)$ são considerados dados incorretos. Por fim, os dados entre $(\mu - 3\sigma, \mu - \sigma)$ e $(\mu + \sigma, \mu + 3\sigma)$ podem ser dados bons ou dados ruins, e não devem ser detectados pela regra e nem pela árvore de decisão.

Figura 13 – Diagrama de distribuição dos outliers.

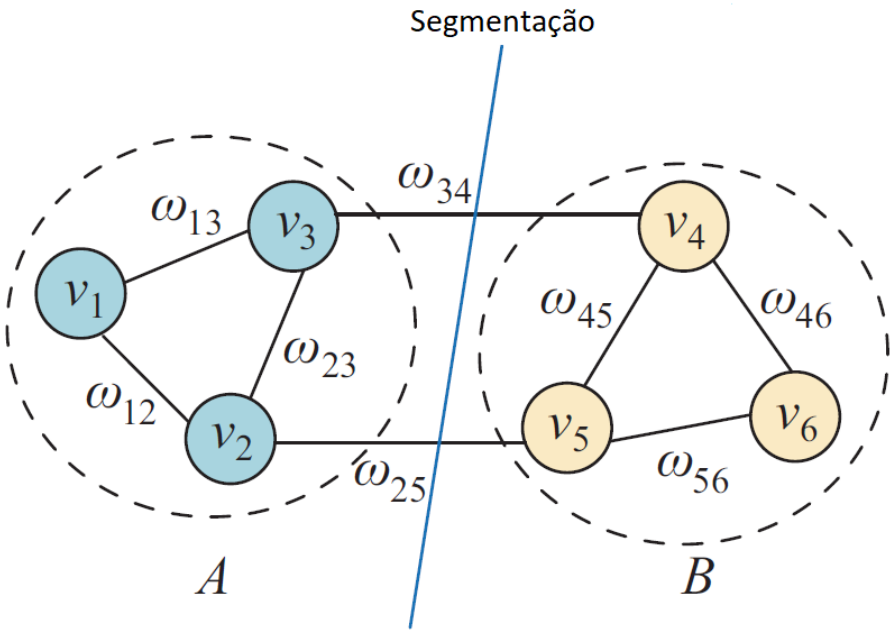


Fonte: Adaptado de Yang et al. (2020).

Nesse caso, quando a amplitude dos dados ruins estiverem próximo do valor médio do conjunto de dados sua detecção deve ser prejudicada, por isso o autor propôs uma melhoria com base no agrupamento espectral. Ao contrário do DBSCAN, o método de agrupamento espectral é exclusivamente baseado em grafos e transforma o problema de

agrupamento em um problema de segmentação de grafos, conforme apresentado na Fig. 14.

Figura 14 – Segmentação dos grafos em agrupamento espectral.



Fonte: Adaptado de Yang et al. (2020).

Na Fig. 14 os vértices v_i representam cada dado X_i da amostra, onde os vértices azuis representam os dados normais (subgrafo A), enquanto os vértices amarelos representam os dados ruins (subgrafo B). A relação entre essas amostras é o peso w_{ij} que indica o grau de similaridade entre v_i e v_j. Vale destacar que, por se tratar de um grafo não direcionado, $w_{ij} = w_{ji}$. O objetivo do *clustering* espectral é cortar o grafo para obter dois *clusters*: um com dados normais e outro com dados ruins. O que implica em uma maior similaridade dentro do subgrafo e menor similaridade entre os subgrafos. Essa proposta detectou *outliers* com pequenos valores de desvio, que em geral não são fáceis de detectar pelos métodos clássicos.

Uma abordagem baseada em dados utilizando os conceitos de aprendizagem profunda para detecção de anomalias em sincrofasores foi proposto por Deng et al. (2020). O método utilizado foi o *Convolutional Neural Network*(CNN), uma vez que, em comparação com outros métodos de aprendizagem tradicional, possui uma grande capacidade de processamento de big data, capacidade de extração automática de recursos e uma melhor capacidade de aprendizagem. A principal característica que diferencia as CNNs de outras redes neurais é a incorporação das chamadas camadas de convolução, que são projetadas especificamente para capturar padrões locais em um conjunto de dados.

Na prática as medições normais e algumas anomalias de dados estão em torno do

valor nominal, enquanto os picos ou dados faltantes podem ter um grande desvio do valor nominal. Isso causará uma compensação excessiva da CNN no processo de treinamento para um tipo específico de anomalia e reduzirá a sensibilidade para outras anomalias. Então se torna essencial uma etapa de normalização das entradas, reduzindo o desvio dos dados de entrada e, consequentemente, melhorar a convergência e a velocidade da CNN. Para o pré-processamento das medições brutas foram realizadas as seguintes etapas:

1. Para um conjunto de dados X é extraído o valor máximo (f^i_{max}) e o valor mínimo (f^i_{min});
2. Normalizar as amostras usando a Equação 2.9:

$$f^i_{j,norm} = \frac{f^i_j - f^i_{min}}{f^i_{max} - f^i_{min}} \tag{2.9}$$

3. Por fim, cada medição normalizada estará com valores na faixa de [0,1].

Para validar sua proposta, o autor comparou o desempenho do seu algoritmo com outros métodos de classificação de *outliers* amplamente utilizados na literatura: Support Vector Machine (SVM), Rede Neural Artificial (ANN), Análise de Componentes Principais-SVM (PCA-SVM) e Long Short Term Memory (LSTM). Ambos métodos foram implementados e testados com o mesmo conjunto de dados normalizados. Para RNA, foram utilizada duas camadas. Para comparar com o método de compressão dimensional, o PCA e o SVM são combinados para obter recursos de extração e classificação de recursos. Por fim, constatou-se que o SVM tem desempenho melhor que a RNA e o LSTM. Porém, o resultado do PCA-SVM é de 64,57% e não é tão bom quanto o SVM, o que indica que pode haver perda de informações durante o processo de PCA. No geral, a CNN proposta obtém a melhor precisão ($97,71\%$).

O trabalho proposto por Zhao et al. (2020) analisa as características de dados ruins de pequena magnitude e propõe um método para extrair as suas características de amplitude fasorial utilizando características de aprendizagem e memória de uma rede *Long Short-Term Memory*(LSTM). Uma LSTM é um tipo de rede neural recorrente(RNN) especializada em processar e modelar sequências de dados. Ela é projetada para superar as limitações das RNNs tradicionais, que sofrem de problemas como o desvanecimento do gradiente, que dificulta o aprendizado de dependências de longo prazo em sequências temporais. A principal característica distintiva das LSTMs é a capacidade de aprender e manter informações por longos períodos de tempo. Isso as torna muito eficazes em tarefas que envolvem dependências de longo prazo (BENGIO; SIMARD; FRASCONI, 1994).

Em seguida, o autor avalia as características dos dados de ângulo de fase das PMUs, o que denominou de método de utilização da inclinação do ângulo de fase. Por fim, utilizando as características da amplitude e do ângulo de fase dos sincrofasores, é utilizado

o DBSCAN para fazer uma análise das informações extraídas anteriormente e fazer uma detecção com maior eficácia dos *outliers*. O método proposto baseia-se unicamente nos dados de uma PMU e não requer o conhecimento na topologia do sistema de potência. Ele expande o alcance de detecção de dados ruins de uma PMU e também pode detectar dados ruins quando a amplitude é pequena e relativamente oculta.

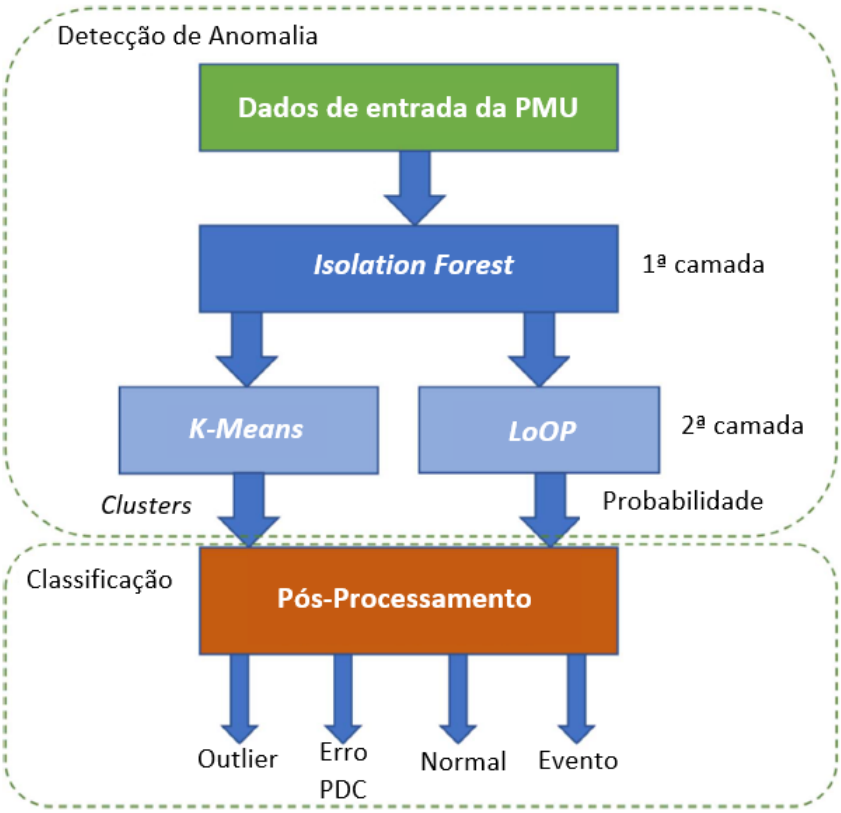
Em uma outra abordagem, Khaledian et al. (2020) apresenta uma proposta para detecção em tempo real de anomalias de dados nas medições de PMU e uma classificação em dados ruins, dados de eventos e erros no PDC. A ferramenta proposta, detecção e classificação de anomalias de sincrofasores(*synchrophasor anomaly detection and classification* - SyADC), utiliza uma técnica de conjunto. É utilizado um aprendizado de máquina não supervisionado, que combina algoritmos de detecção de valores discrepantes rápidos e precisos, como *Isolation Forest* e *local outlier probability* (LoOP) com o KMeans . O método proposto detecta as anomalias com uma acertividade de 99%. Além disso, a correlação entre as PMUs é calculada para detectar os eventos e erros do PDC, aumentando assim a precisão do algoritmo.

O método *Isolation Forest* aproveita a natureza das anomalias que são menos frequentes que as observações regulares e diferentes destas em termos de valores para isolá-las das demais amostras. Para calcular a probabilidade de um determinado ponto de dados em uma janela ser um valor discrepante de densidade local, o algoritmo de LoOP é aplicado. A saída obtida do *Isolation Forest* é alimentada no algoritmo LoOP, que se assemelha cuidadosamente a outro método, o LOF. Em paralelo, a saída do *Isolation Forest* também alimenta um K-Means que tem como objetivo classificar os resultados em clusters binários de classes normais/anomalias. No entanto, devido a sua sensibilidade dados normais com pequenas alterações podem ser considerados como um valor discrepante. Por fim, uma combinação dos resultados de K-Means e LoOP é usada para detectar as anomalias e classifica-las, conforme Fig.15.

Hannon et al. (2021) tem como proposta uma teoria de aprendizado de máquina estatístico. O método utilizado é o vetor auto-regressivo estocástico (VAR) para modelar a relação efetiva entre as amostras coletadas. Está abordagem multivariada, é atualizada em intervalos regulares e faz previsões de curto prazo das trajetórias futuras dos dados. Desse modo, as anomalias são então identificadas com base em desvios estatisticamente significativos após o recebimento dos dados. Por fim, o autor projeta uma ferramenta de classificação baseada em árvore de decisão para distinguir os *outliers* identificados em diferentes categorias interpretáveis.

Sua principal vantagem em relação a outras técnicas de aprendizado de máquina está na abordagem multivariada e com a classificação baseada em árvore de decisão, tornando o desempenho superior quando comparado com outras técnicas de classificação. Outro ponto importante na abordagem é o fato de não envolver ajuste de limites e

Figura 15 – Estrutura do algoritmo SyADC para detecção e classificação de anomalias.



Fonte: Adaptado de Khaledian et al. (2020).

parâmetros específicos de um sistema, tornando um algoritmo mais escalável. Entretanto, se faz necessário uma etapa de treinamento.

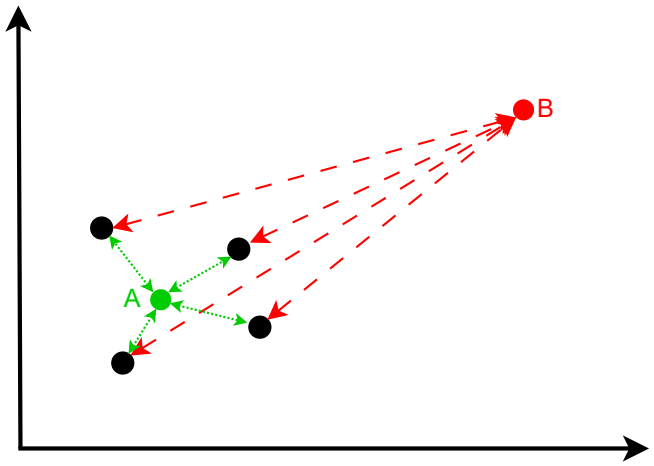
2.4 TYPICALITY AND ECCENTRICITY DATA ANALYTICS - TEDA

O *Typicality and Eccentricity Data Analytics* (TEDA) é um algoritmo estatístico para análise de dados proposto por Angelov (2014), utilizando conceitos de tipicidade e excentricidade de uma amostra em relação ao seu conjunto de dados. O conceito de tipicidade refere-se à semelhança de uma determinada amostra de dados com outras amostras dentro do mesmo conjunto de dados. Esta similaridade é geralmente calculada medindo a proximidade espacial entre as amostras em um espaço n-dimensional. A excentricidade, por outro lado, é o oposto da tipicidade, pois quantifica o grau de diferença entre uma amostra de dados específica e as outras amostras do conjunto de dados. Na Fig. 16 ilustra tal conceito, onde a amostra A possui uma maior tipicidade com relação ao dados do conjunto, enquanto a amostra B é mais excêntrica dos demais dados. Além disso, a estrutura TEDA

apresenta uma abordagem diferenciada que elimina algumas necessidades, são elas:

- Parâmetros iniciais específicos do problema;
- Presunções quanto à distribuição de dados;
- Um número infinito de observações, pois é necessário um conjunto com apenas três amostras de dados;
- Independência das amostras de dados individuais.
- É inteiramente baseado nos dados e sua distribuição mútua no espaço de dados;

Figura 16 – Exemplo do conceito de tipicidade e excentricidade entre amostras.



Fonte: Autoria Própria.

O TEDA pode ser visto como uma estrutura empírica que pode trabalhar de maneira eficiente com qualquer conjunto de dados, exceto processos puramente aleatórios nos quais amostras de dados individuais são completamente independentes umas das outras. Para tais casos (por exemplo, jogos de azar, jogos, ruído branco, etc.), a teoria tradicional da probabilidade é a melhor ferramenta a ser usada, pois foi desenvolvida com esses processos em mente e amadureceu ao longo de três ou mais séculos (COSTA et al., 2015).

O método fundamenta-se na ideia de avaliar a proximidade entre as amostras, e é importante notar que essas medidas não são idênticas à densidade utilizada em estatística e campos relacionados. Dentre os conceitos apresentados pelo TEDA, vale destacar três deles: a proximidade acumulada (π), a tipicidade (τ) e a excentricidade (ξ).

Defini-se o espaço de dados $x \in R^n$, onde x representa o conjunto de dados em um espaço n-dimensional, no qual, pode ser representado pela sequência:

$$\{x_1, x_2, ..., x_k, ...\}, x_k \in R^n, k \in N \tag{2.10}$$

onde: x_k representa um vetor de n dimensões e o índice k representa um sistema em um instante discreto, podendo, por exemplo, ter o significado físico de instante de tempo quando a amostra de dados chegou.

Quando $k > 1$ pode-se definir métricas de distâncias ($d(x, x_i)$) entre as amostras presentes no conjunto, essa distância pode ser calculada como Euclidiana, Mahabolis, Cosseno, etc. Logo, podemos definir a proximidade acumulada (π) de um ponto X como o somatório das distâncias entre o ponto X e todos os k pontos do conjunto, conforme enunciado na Eq. (2.11) (ANGELOV, 2014).

$$\pi_k(x) = \sum_{i=1}^{k} d(x, x_1), k > 1 \tag{2.11}$$

Ainda segundo Angelov (2014) a tipicidade (ξ) e excentricidade (ϵ) podem ser calculados a partir do momento que três ou mais amostras distintas tenham sido obtidas ($k > 2$), dado que considerando apenas duas amostras não idênticas elas serão igualmente atípicas ou excêntricas. Desta forma, a excentricidade de uma amostra x em um instante k é definida pela razão entre a sua proximidade acumulada e soma das proximidades acumuladas de todas as outras amostras, como pode ser visto na Eq. 2.12.

$$\xi_k(x) = \frac{2\pi_k(x)}{\sum_{i=1}^{k} \pi_k(x_i)}, k > 2, \sum_{i=1}^{k} \pi_k(x_i) > 0 \tag{2.12}$$

Por fim, a tipicidade (τ), é definida como o complemento da excentricidade:

$$\tau_k(x) = 1 - \xi_k(x), k > 2, \sum_{i=1}^{k} \pi_k(x_i) > 0 \tag{2.13}$$

Assim, tanto a excentricidade ξ quanto a tipicidade τ estão entre 0 e 2 e suas contrapartes normalizadas podem ser representadas, respectivamente, por:

$$\zeta_k(x) = \frac{\xi_k(x)}{2} \tag{2.14}$$

$$T_k(x) = \frac{\tau(x)}{k-2} \tag{2.15}$$

Ou seja, a tipicidade representa tanto o padrão de distribuição espacial quanto a frequência de ocorrência de uma amostra de dados simultaneamente e por amostra de dados.

Além disso, ainda de acordo com Angelov (2014) em casos que seja preciso uma aplicação online, as variáveis podem ser calculadas recursivamente. Assim, não é necessário

manter as amostras anteriores armazenadas localmente; em contrapartida, o cálculo das métricas podem ser realizados utilizando apenas a última amostra recebida e os valores agregados que representem o estado do sistema no instante anterior, conforme Eq. 2.16 e Eq. 2.17.

$$\mu_k(x) = \frac{k-1}{k}\mu_{k-1} + \frac{1}{k}x_k, \mu_1 = x_1 \tag{2.16}$$

$$\sigma_k^2(x) = \frac{k-1}{k}\sigma_{k-1}^2 + \frac{1}{k-1}\left\|x_k - \mu_k\right\|^2, \sigma_1^2 = 0 \tag{2.17}$$

onde: μ_k é a média no instante k e σ_k^2 é a variância no instante k.

Desse modo, é possível utilizar o formato recursivo das equações para o cálculo da excentricidade e tipicidade, respectivamente:

$$\xi_k(x) = \frac{1}{k} + \frac{(u_k - x_k)^T(u_k - x_k)}{k\sigma_k^2} \tag{2.18}$$

$$\tau_k(x) = 1 - \xi_k(x) = \frac{k-1}{k} - \frac{(u_k - x_k)^T(u_k - x_k)}{k\sigma_k^2} \tag{2.19}$$

De fato, a recursividade torna o TEDA um algoritmo vantajoso, pois consegue lidar com grandes quantidades de dados, de forma rápida, *online* e com baixo custo computacional. Outra característica que torna o método adequado aos *streams* de dados, uma vez que monitora e se ajusta de forma autônoma e adaptativa às mudanças no comportamento dos dados ao longo do tempo, sem a necessidade de treinamento prévio.

Para detectar valores discrepantes em qualquer distribuição de dados, mas assumindo um grande número de amostras de dados independentes, pode-se utilizar a conhecida desigualdade de Chebyshev, proposta por Saw, Yang e Mo (1984). A desigualdade de Chebyshelv é um principio estatístico amplamente utilizado para detecção de anomalias. Onde afirma que não mais que $1/m^2$ de observações de dados são maiores que a média. A distância entre os valores é maior que $m\sigma$, onde σ representa o desvio padrão dos dados. Assim, a partir da excentricidade normalizada podemos definir um limiar para detecção de *outlier*, denominada desigualdade de excentricidade (NADA, a):

$$\zeta_k(x) > \frac{m^2+1}{2k}, m > 0 \tag{2.20}$$

sendo m o número de desvios padrão que se deseja utilizar e k a quantidade atual de amostras do conjunto. Vale salientar que, o valor de m indica a sensibilidade utilizada no limiar do algoritmo, quanto maior valor m menor será sua sensibilidade para detecção de *outliers*. Logo, a Eq. 2.20 nos permite realizar análise local e individual para cada amostra obtida em um fluxo de dados.

Uma outra abordagem, proposta por Nunes e Guedes (2024), propõe a aplicação de um fator de esquecimento ao algoritmo do TEDA. Note que, para um grande conjunto de dados os fatores $k-1/k$ e $1/k$ das equações 2.16 e 2.17, tendem a 1 e 0, respectivamente. Isso significa que as amostras de dados mais antigas possuem uma maior relevância para o conjunto de dados, enquanto as novas amostras devem influência cade vez menor para o conjunto de dados.

Deste modo, levando em consideração o fator de esquecimento, podemos reescrever as equações 2.16 e 2.17 como:

$$\mu_k(x) = \alpha\mu_{k-1} + (1-\alpha)x_k \tag{2.21}$$

$$\sigma_k^2(x) = \alpha\sigma_{k-1}^2 + (1-\alpha)\left\|x_k - \mu_k\right\|^2, \sigma_1^2 = 0 \tag{2.22}$$

sendo, α o fator de esquecimento, μ_k a média no instante k e σ_k^2 a variância no instante k.

O fator de esquecimento pode ser interpretado como um fator de ajuste da sensibilidade para o TEDA. Os valores aceitáveis para α estão entre 0 e 1, mais próximo de 1 (NUNES; GUEDES, 2024). Quanto maior o valor de α, menos relevância a amostra x_k tem para o conjunto de dado e consequentemente, menor sensibilidade a valores ruidosos. Por outro lado, para valores de α menores, maior a importância da amostra para o conjunto de dados.

De fato, o fator de esquecimento pode favorecer o TEDA para as k-ésimas amostras do conjunto, ao equilibrar de forma mais equânime os novos valores aquisitados. Porém essa afirmação não é verdade para as amostras iniciais do conjunto, uma vez que elas terão um ponderação melhor. Para evitar esta ponderação anômala, as equações 2.21 e 2.22, só devem ser considerada após atendida a condição da equação 2.23.

$$1-\alpha \geq \frac{1}{k} \tag{2.23}$$

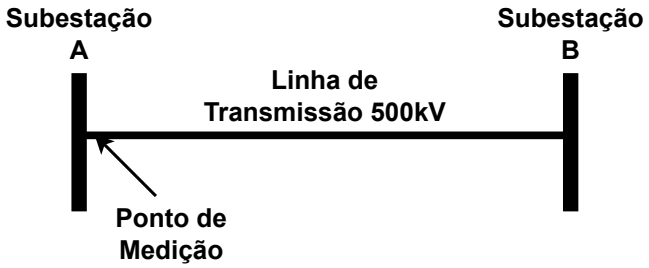
3 METODOLOGIA

Neste capítulo, inicialmente é apresentado o conjunto de dados de PMU que será utilizado ao longo desse trabalho. Em seguida, é apresentado o algoritmo proposto para detecção de *outliers*, os casos de estudos que serão analisados nesse trabalho e as métricas para avaliação dos resultados obtidos.

3.1 CONJUNTO DE DADOS DE PMU

Os dados de medição fasorial utilizados nesse trabalho são dados reais, oriundos de PMUs conectadas em terminais de conexão de linha de transmissão de 500kV, conforme Fig. 17, sendo utilizadas as medidas de frequência, tensão de fase e corrente de fase. As PMUs estão localizado na região Nordeste do Brasil. Neste trabalho será realizada uma análise observando dois cenários. O primeiro cenário foi feito um recorte considerando dados de uma única PMU, com o sistema em condições normais. O segundo cenário, foram coletados dados de três PMU em posições geograficamente diferentes, para esse caso buscou-se um intervalo com alguma ocorrência elétrica da rede. Cada período extraído foi considerado um intervalo de 10 minutos, com uma amostragem de 60 medições por segundo, o que corresponde a um volume de 36 mil amostras para cada grandeza observada.

Figura 17 – Modelo simplificado de uma linha de transmissão que conecta duas instalações.

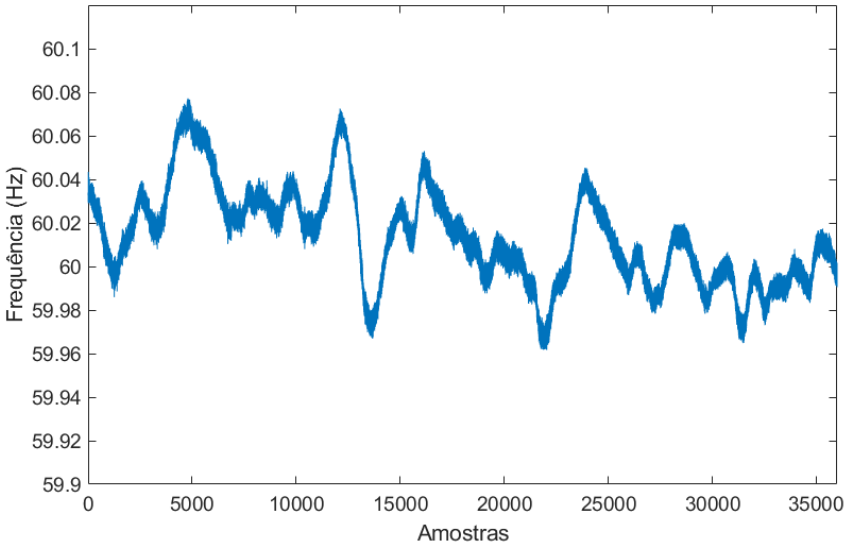


Fonte: Autoria própria.

Para o primeiro cenário, foi feito a coleta de dados de uma única PMU no dia 10 de Abril de 2024, entre as 08:05 e 08:10. Na Fig. 18 é apresentada a curva de frequência obtida, note que a frequência oscila em torno de 60 Hz, tendo uma variação de aproximadamente de 0,1 Hz no período observado. Na Fig.19 é apresentado a curva de corrente de fase, note que existe variações em torno de 1300 A, sendo o valor máximo observado de 1316 A e o mínimo 1286 A. Tais variações ao longo do tempo são normais devido a sua sensibilidade de

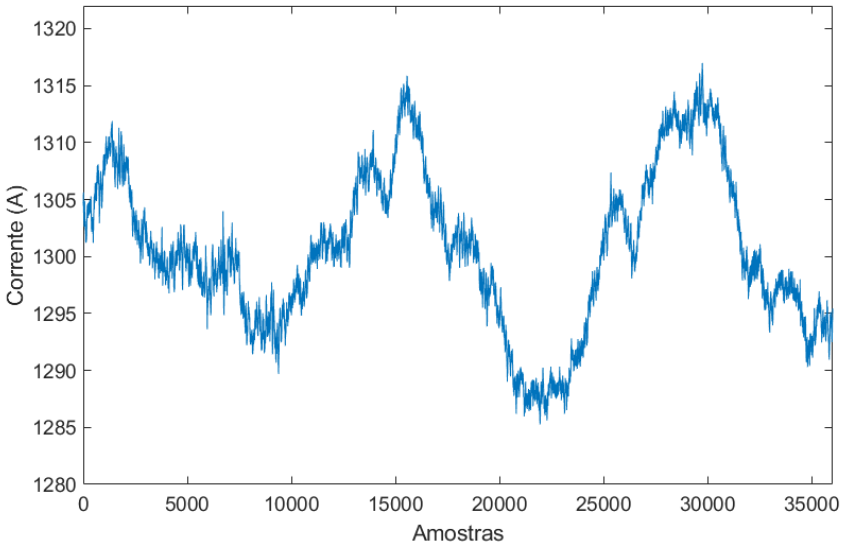
natureza elétrica. Por fim, a Fig. 20 apresenta a curva de tensão de fase, nela percebemos que os valores se alteram pouco ao longo do período.

Figura 18 – Dados de frequência extraídos.



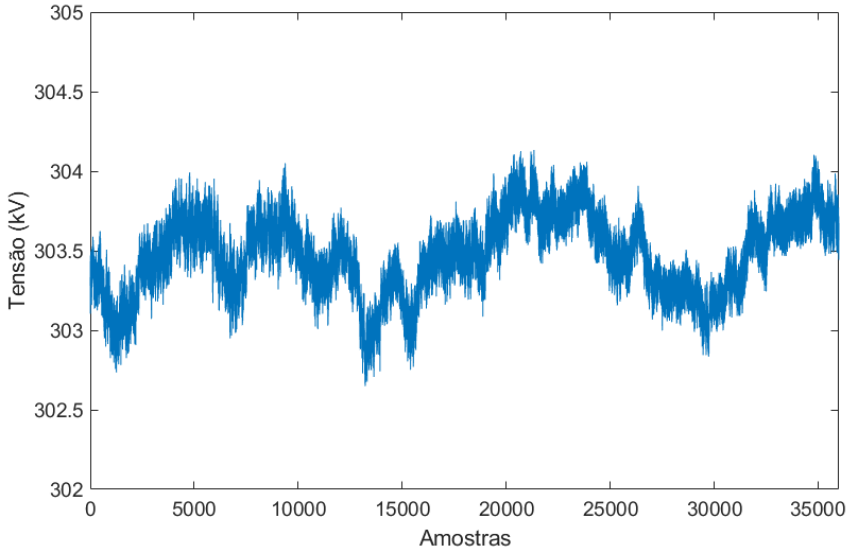
Fonte: Autoria própria.

Figura 19 – Dados de corrente de fase extraídos.



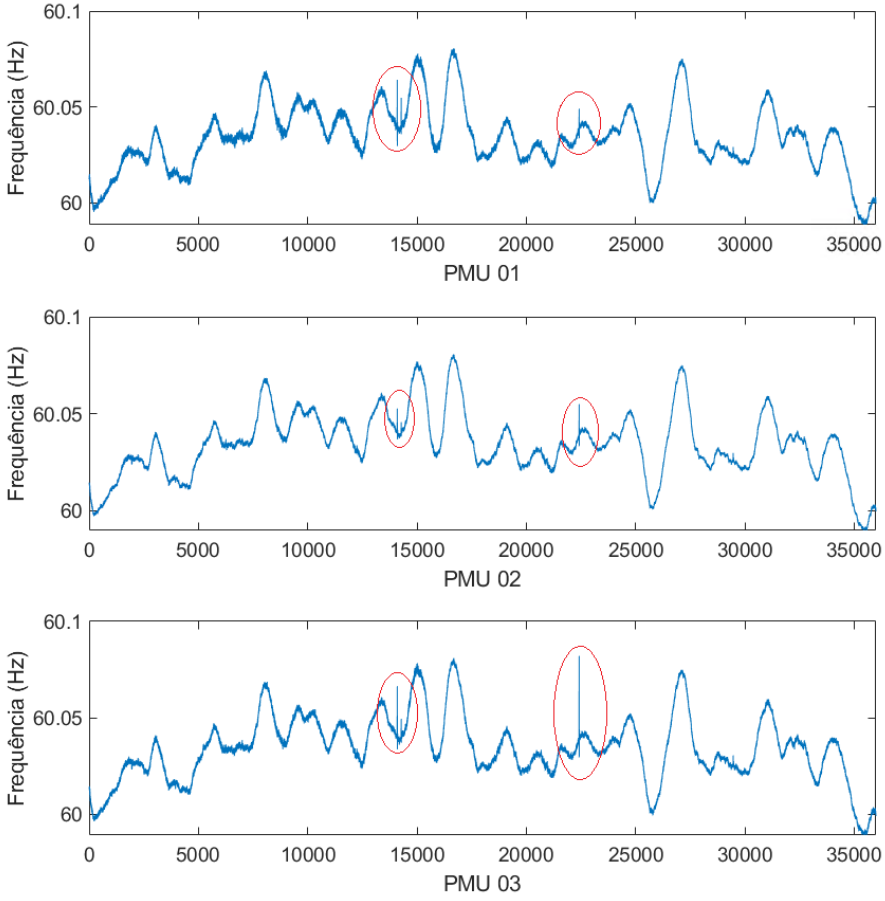
Fonte: Autoria própria.

Figura 20 – Dados de tensão de fase extraídos.



Fonte: Autoria própria.

Figura 21 – Dados de frequência obtidos das três PMUs diferentes.

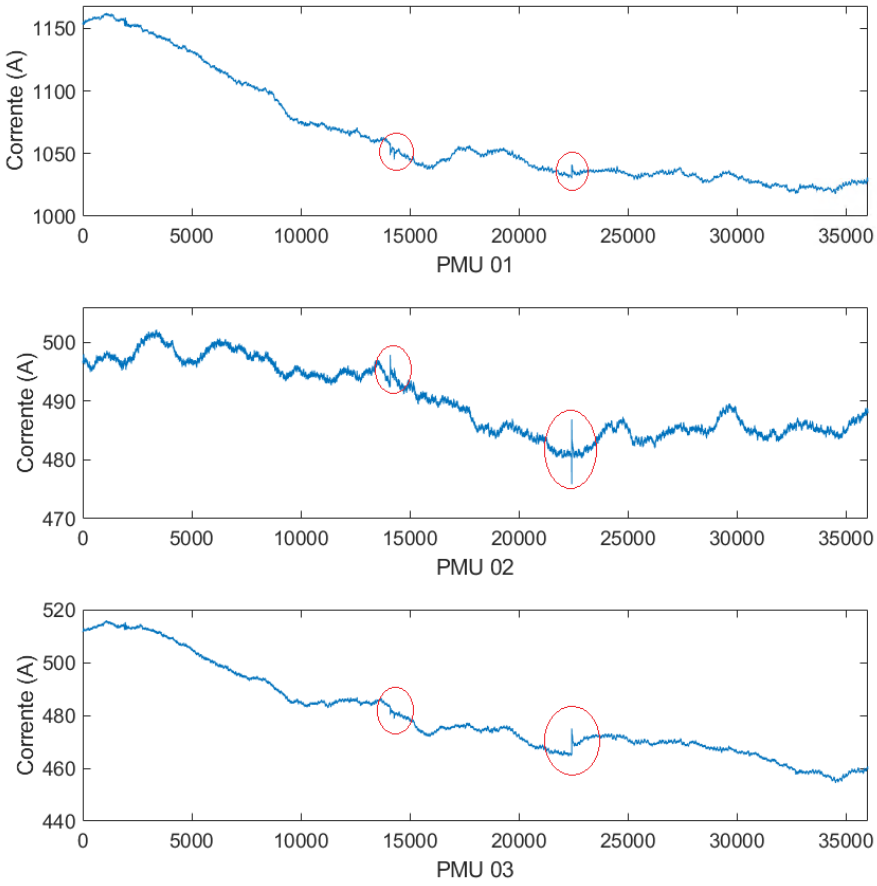


Fonte: Autoria própria.

O segundo cenário utilizado neste trabalho foram coletados o conjunto de dados de três PMUs distintas. Os dados são referentes ao dia 15 de junho de 2024, no período entre as 16:30 e 16:40. Durante o período observado ocorreu um desligamento de uma linha de transmissão de 500 kV próximo das PMUs observadas, em um primeiro momento a linha ficou energizada apenas por um terminal e em seguida, foi desligada em ambos os terminais. Na Fig. 21 são apresentadas as medições de frequências, na qual podem se observar alguns dados anômalos no momento da abertura do primeiro disjuntor e na abertura do segundo disjuntor. Perceba que todas as PMUs foram sensibilizadas pelo desligamento, entretanto com intensidades diferentes. Esse efeito ocorre devido as distâncias elétricas entre o equipamento contingenciado e a sua localização do sincrofasor.

Na Fig. 22 é apresentando as medições de corrente de fase obtida para as três PMUs. Aqui ocorre um comportamento parecido com a frequência, alguma medidas anômalas no instante da ocorrência.

Figura 22 – Dados de corrente de fase obtidos das três PMUs diferentes.

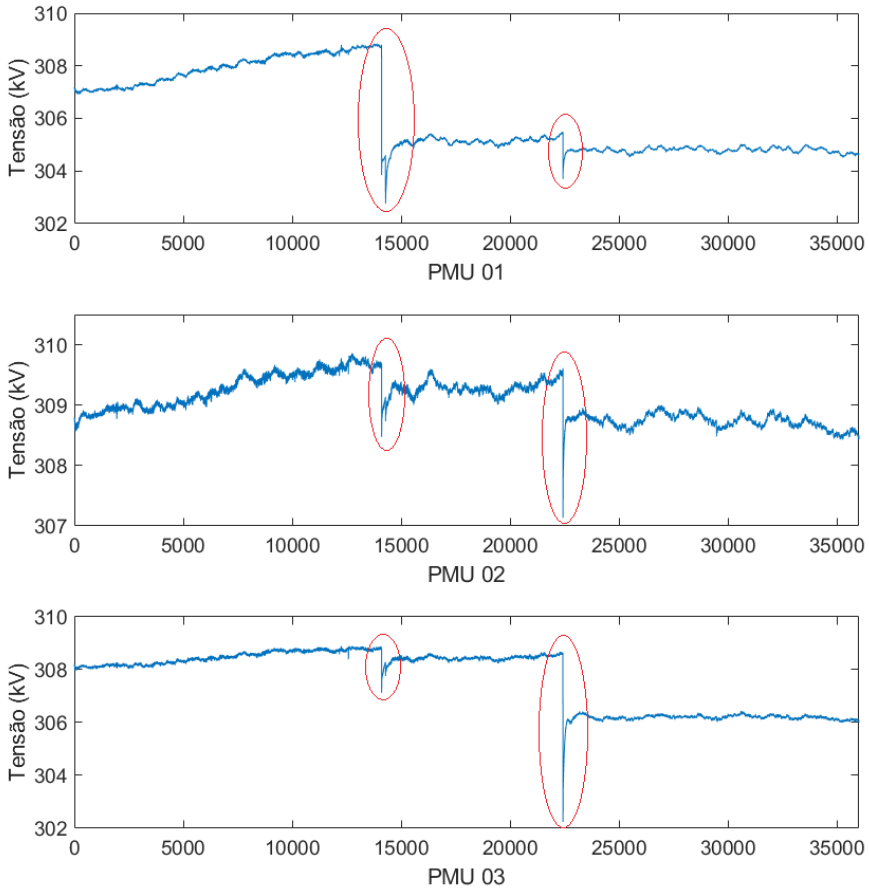


Fonte: Autoria própria.

Por fim, na Fig. 23 é apresentado as medições de tensão de fase das PMUs. No momento desligamento da linha de 500 kV, devido os efeitos eletromagnéticos do

equipamento, uma grande parcela de potência reativa é retirada abruptamente do sistema, causando uma redução nos valores de tensão. Nesta figura pode-se observar nitidamente os momentos de abertura dos dois terminais da linha.

Figura 23 – Dados de tensão de fase obtidos das três PMUs diferentes.



Fonte: Autoria própria.

Diante deste contexto, as curvas de PMU podem frequentemente encontrar-se deturpadas com *outliers*, os quais devem ser detectados e corrigidos para plena utilização nas atividades relacionadas como operação, manutenção e planejamento do sistema elétrico de potência.

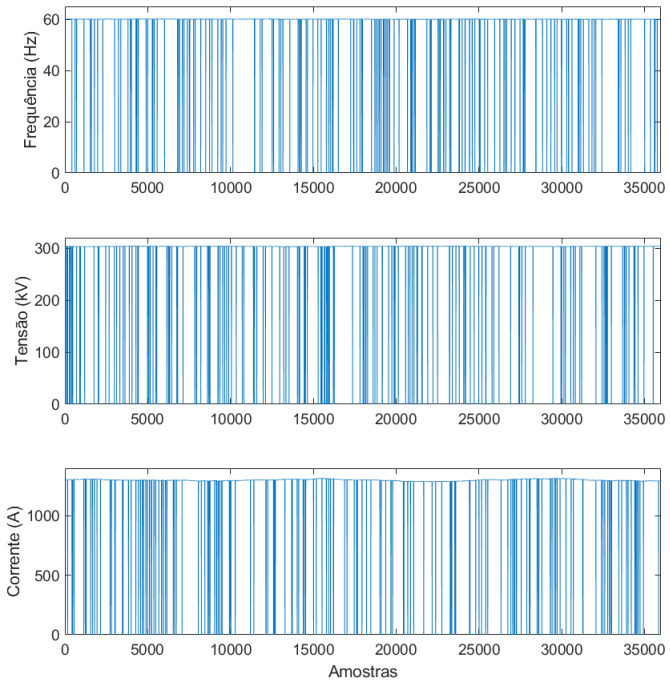
3.2 INSERÇÃO DE *OUTLIERS*

Conforme apresentado na seção anterior, todos os cenários utilizados para esse estudo possuem 36000 mil amostras para cada grandeza. A fim de analisarmos o desempenho dos algoritmos para detecção de picos e zeros. Nesse momento, será utilizados apenas o conjunto de dados do primeiro cenário. Foram inseridos algum tipo *outliers* de maneira

aleatória nos conjuntos de dados de frequência, corrente e tensão. No total utilizaremos 3 casos para o estudo, são eles:

- **Caso 1:** Inserção de 180 zeros totalizando 0,5% das amostras nos três conjuntos de dados, conforme ilustrado na Fig. 24.

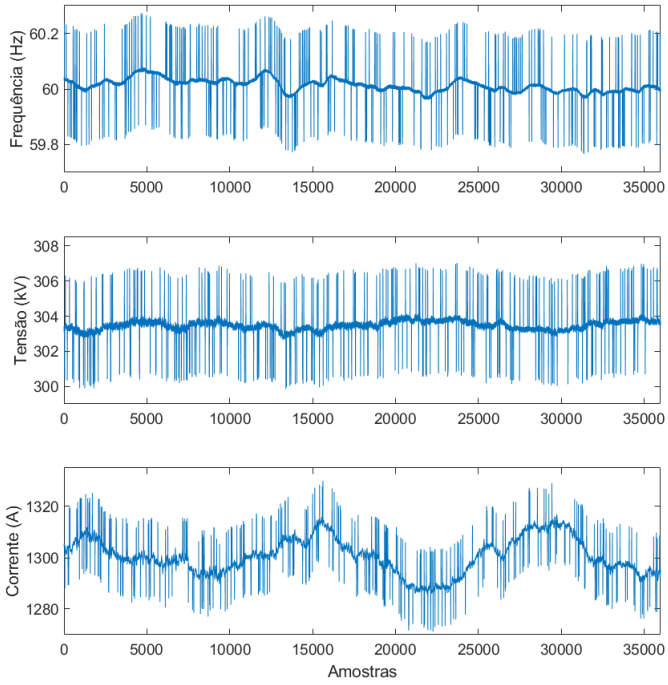
Figura 24 – Dados do Caso 01 após inserção de *outliers* tipo zeros.



Fonte: Autoria própria.

- **Caso 2:** Inserção de picos positivos(0,5%) e picos negativos(0,5%), totalizando 360 *outliers*, conforme ilustrado na Fig. 25. Para os valores de frequência foi considerado um acréscimo de ± 0,2 Hz. Para tensão foi considerado o acréscimo de ± 3 kV, e para corrente, um acréscimo de ± 15 A.

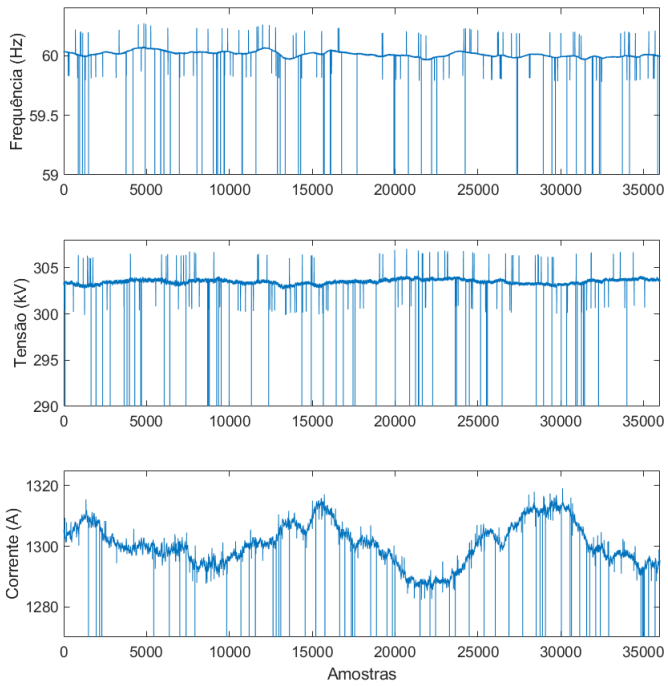
Figura 25 – Dados do Caso 02 após inserção de *outliers*.



Fonte: Autoria própria.

- **Caso 3:** Inserção de 60 picos positivos, 60 picos negativos e 60 zeros no conjunto de dados, seguindo o critério dos casos 01 e 02, conforme ilustrado na Fig. 26.

Figura 26 – Dados do Caso 03 após inserção de *outliers*-tipo picos.



Fonte: Autoria própria.

3.3 MÉTRICAS DE AVALIAÇÃO

Com o intuito de possibilitar uma análise da eficiência do método proposto, algumas métricas de avaliação foram levadas em consideração. Neste subtópico, exploraremos cada uma dessas métricas.

3.3.1 MATRIZ DE CONFUSÃO

As matrizes de confusão são uma ferramenta importante para avaliar modelos de classificação, fornecendo uma visão detalhada do desempenho do modelo. Para interpretar corretamente os resultados de uma matriz de confusão, é necessário compreender os diferentes elementos que a compõem, incluindo verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos. Neste subtópico, exploraremos como interpretar cada métrica e como usá-las para avaliar a precisão, sensibilidade, especificidade e outras métricas importantes de um modelo de classificação.

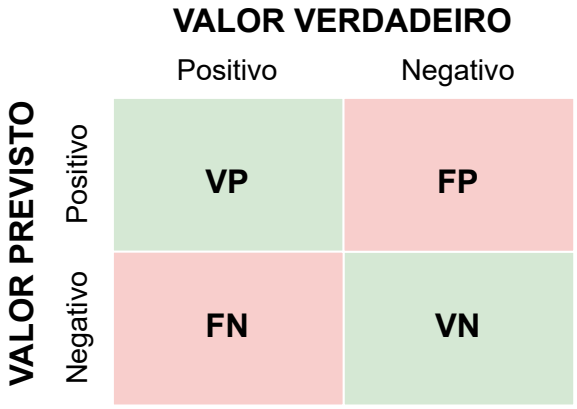
Dentro da matriz de confusão existem quadrantes que representam quatro parâmetros e, com esses parâmetros é possível realizar cálculos de alguns tipos de métricas de erro, permitindo uma análise detalhada do desempenho do modelo em relação aos dados de teste. A matriz de confusão é particularmente útil quando se trata de problemas de classificação binária, onde um modelo tenta categorizar os exemplos em duas classes, como positivo ou negativo, sim ou não, spam ou não spam, etc. Ela consiste em quatro componentes principais:

- Verdadeiros Positivos(VP): Esses são os casos em que o modelo previu corretamente uma instância como positiva quando ela realmente era positiva. Por exemplo, se o valor real do dado indicar que o dado não é um *outlier* e o resultado informa a mesma coisa. Em termos simples, é quando o modelo acertou.
- Verdadeiros Negativos(VN): Esses são os casos em que o modelo previu corretamente uma instância como negativa quando ela realmente era negativa. Por exemplo, se o valor real indicar que é um *outlier* e a classe prevista diz a mesma coisa.
- Falsos Positivos(FP): Esses são os casos em que o modelo previu erroneamente uma instância como positiva quando ela era, na verdade, negativa. Por exemplo, se o valor real diz que é um *outlier*, mas a classe prevista informa que não.
- Falsos Negativos(FN): Esses são os casos em que o modelo previu erroneamente uma instância como negativa quando ela era positiva. Por exemplo, se o valor real da classe indicar que o ponto não é um *outlier* e a classe prevista informa é um *outlier*.

É apresentado na Fig. 27 uma matriz de confusão, onde o quadrante superior esquerdo representa os valores Verdadeiros Positivos, o quadrante superior direito representa

os valores Falsos Negativos, o quadrante inferior esquerdo representa os valores Falsos Positivos e o por fim, o último quadrante representa os Verdadeiros Negativos.

Figura 27 – Matriz de Confusão.



Fonte: Autoria própria.

A interpretação desses elementos da matriz de confusão permite calcular várias métricas importantes de desempenho do modelo (STEHMAN, 1997), são eles:

- Precisão: A precisão mede a fração de previsões corretas feitas pelo modelo e é calculada pela Eq. 3.1. Indica a capacidade geral do modelo de fazer previsões corretas.

$$Precisao = \frac{VP + VN}{VP + VN + FP + FN} \tag{3.1}$$

- Sensibilidade ou *Recall*: A sensibilidade mede a capacidade do modelo de identificar corretamente todos os casos positivos e é calculada por Eq. 3.2). Ela é útil quando é importante evitar falsos negativos.

$$Recall = \frac{VP}{VP + FN} \tag{3.2}$$

- Especificidade: A especificidade mede a capacidade do modelo de identificar corretamente todos os casos negativos e é calculada pela Eq. 3.3 VN / (VN + FP). É relevante quando a minimização de falsos positivos é crítica.

$$Especificidade = \frac{VN}{VN + FP} \tag{3.3}$$

- Taxa de Falsos Positivos: Essa métrica é calculada pela Eq. 3.4 e fornece informações sobre a taxa de falsos positivos em relação ao total de negativos verdadeiros.

$$Taxa_{FP} = \frac{FP}{FP + VN} \tag{3.4}$$

A interpretação adequada da matriz de confusão ajuda a compreender o desempenho do modelo em diferentes aspectos, permitindo ajustes e melhorias quando necessário. Além disso, auxilia na seleção de limiares de classificação ideais e na identificação de áreas específicas que podem exigir atenção para otimizar o desempenho do modelo de classificação.

3.3.2 COEFICIENTE DE CORRELAÇÃO DE MATTHEWS

Uma outra métrica proposta por Matthews (1975) é intitulada de coeficiente de correlação de Matthews(MCC). É uma métrica de avaliação de desempenho que é frequentemente usada para medir a qualidade de modelos de classificação binária. O MCC leva em consideração todos os quatro elementos da matriz de confusão (Verdadeiros Positivos, Verdadeiros Negativos, Falsos Positivos e Falsos Negativos) para calcular uma única pontuação que varia de -1 a +1, onde -1 representa um modelo totalmente incorreto e +1 representa um modelo perfeito. O MCC pode ser calculado diretamente da matriz de confusão usando a Eq. 3.5.

$$MCC = \frac{VP.VN - FP.FN}{\sqrt{(VP + FP)(VP + FN)(VN + FP)(VN + FN)}} \tag{3.5}$$

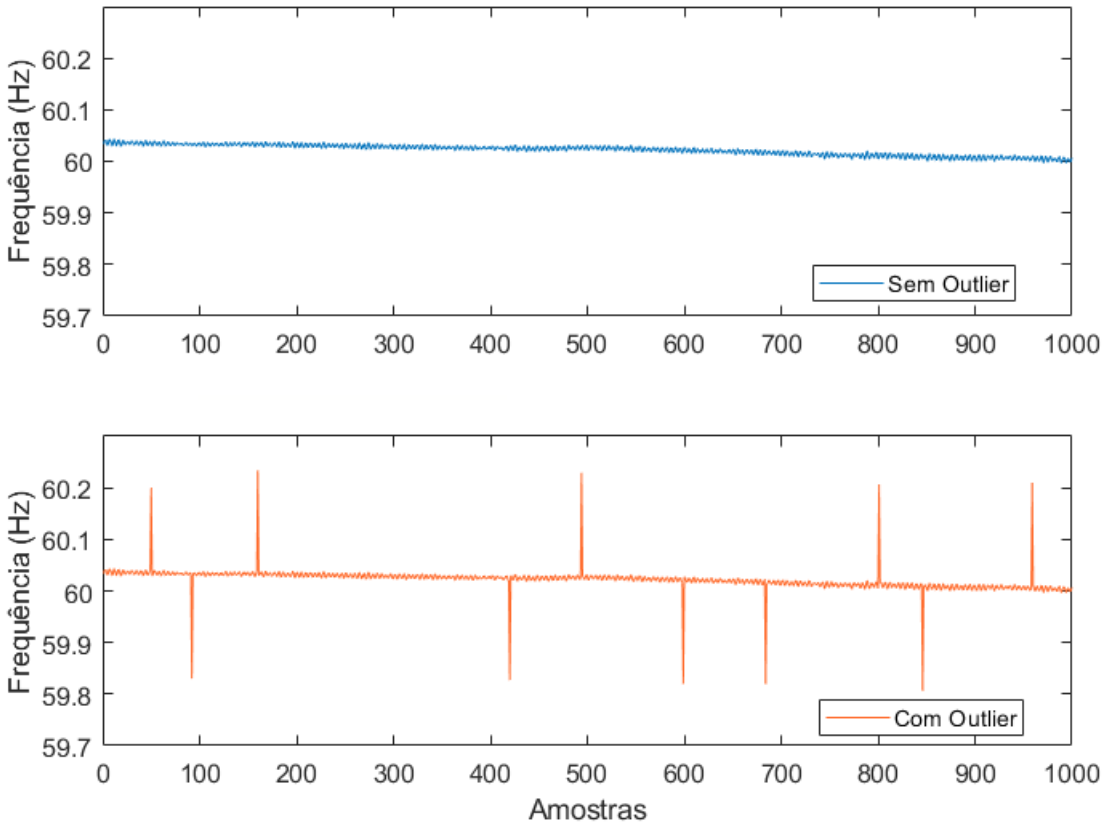
O MCC é uma métrica particularmente útil quando você deseja uma única medida que capture a qualidade geral do modelo, levando em consideração o equilíbrio entre sensibilidade e especificidade. Ele atribui maior pontuação a modelos que têm um bom equilíbrio entre prever corretamente as instâncias positivas e negativas, enquanto penaliza modelos que têm desempenho desequilibrado.

3.4 ALGORITMO DE DETECÇÃO UTILIZADOS

O método proposto neste trabalho é baseado nos conceitos de tipicidade e excentricidade proposto por Angelov (2014) que utilizaremos para detecção de anomalia em um conjunto de dados de PMU. Uma de suas vantagens para aplicação do método se dá pelo fato dos modelos serem orientados exclusivamente aos dados, ou seja, não há a necessidade de definição de parâmetros anteriores e/ou conhecimento prévio do comportamento da curva. Desta forma, não é necessário o treinamento prévio ou utilização de modelos matemáticos, onde com apenas a entrada de um dado se calcula a média, variância e excentricidade. A seguir será apresentado algumas especificidades do TEDA Clássico, o TEDA Janelado e o TEDA Esquecimento.

Para facilitar o entendimento, será utilizado um recorte com 1000 amostras do conjunto de dados de frequência elétrica, apresentado na Fig. 28(a). Foram inseridos 10 de *outliers* do tipo vales e picos, conforme apresentado na Fig. 28(b).

Figura 28 – Recorte dos dados de frequências. (a) Sem *outliers* (b) Com *outliers*.



Fonte: Autoria própria.

3.4.1 TEDA CLÁSSICO

A abordagem consiste em um algoritmo para detecção de anomalias que se caracteriza por modelar uma distribuição não paramétrica dos dados baseado apenas na proximidade de uma amostra particular para todo um conjunto de dados, considerando a desigualdade de Chebyshev para definir um limiar entre o que é *outlier* ou não. No Algoritmo 1 é apresentado o pseudocódigo do TEDA implementado neste trabalho.

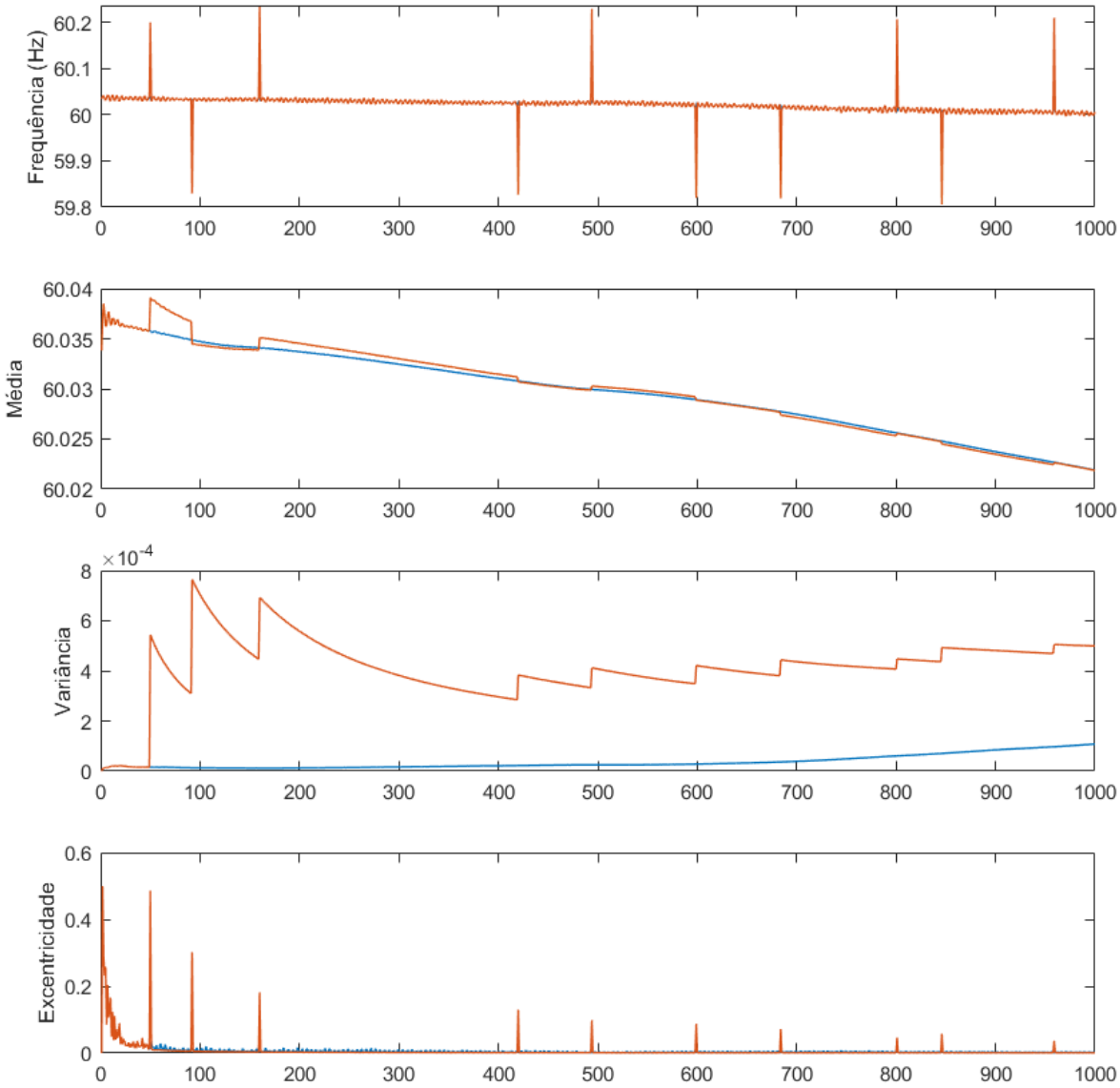
Note que a excentricidade (ξ_i) é calculada em função do dado de entrada X_i, da média(μ_i) e da variância(σ_i), e esses valores são calculados recursivamente pelas equações Eq. 2.16 e Eq. 2.17 a cada chegada de um novo dado. Logo, é factível concluir que esses valores tendem a manter um resíduo de todos os dados, bons ou ruins, da curva ao longo de todo período de tempo.

Algorithm 1: TEDA Clássico

Entrada: Conjunto X de amostras.
Saída: É *Outlier*?
Amostra X_i;
while X_i *existe?* **do**
  ler a amostra $X_i \in X$;
  if i==*1* **then**
    $\mu_i = x_i$;
    $\sigma_i^2 = 0$;
  else
    $\mu_i = \frac{i-1}{i}\mu_{i-1} + \frac{1}{i}x_i$;
    $\sigma_i^2 = \frac{i-1}{i}\sigma_{i-1}^2 + \frac{1}{i-1}\left\|x_i - \mu_i\right\|^2$;
    $\xi_i(x) = \frac{1}{i} + \frac{(u_i - x_i)^T(u_i - x_i)}{i\sigma_i^2}$;
    $\zeta_i(x) = \frac{\xi_i}{2}$;
    if $\zeta_i(x) > \frac{m^2+1}{2i}$ **then**
      X_i é *outlier*;
    else
      X_i não é *outlier*;
    end
  end
  $i \leftarrow i + 1$;
end

A Fig. 29 apresenta uma análise nos dados do conjunto proposto, onde a curva em azul representa os dados sem *outliers* e a curva laranja com *outliers*. Perceba que a partir do momento em que os dados ruins vão sendo aquistados tanto a média, Fig. 29(b), e quanto a variância, Fig. 29(c), sofrem um desprendimento da curva sem *outliers* conforme esperado. Entretanto, ao longo do tempo, essas variações tendem a ser cada vez menor, já que os dados bons predominam nas amostras como um todo, tornando os novos *outliers* menos excêntricos em relação aos *outliers* iniciais, Fig. 29(d).

Figura 29 – (a)Conjunto de dados de entrada de frequência. (b) Média. (c) Variância. (d)Excentricidade.

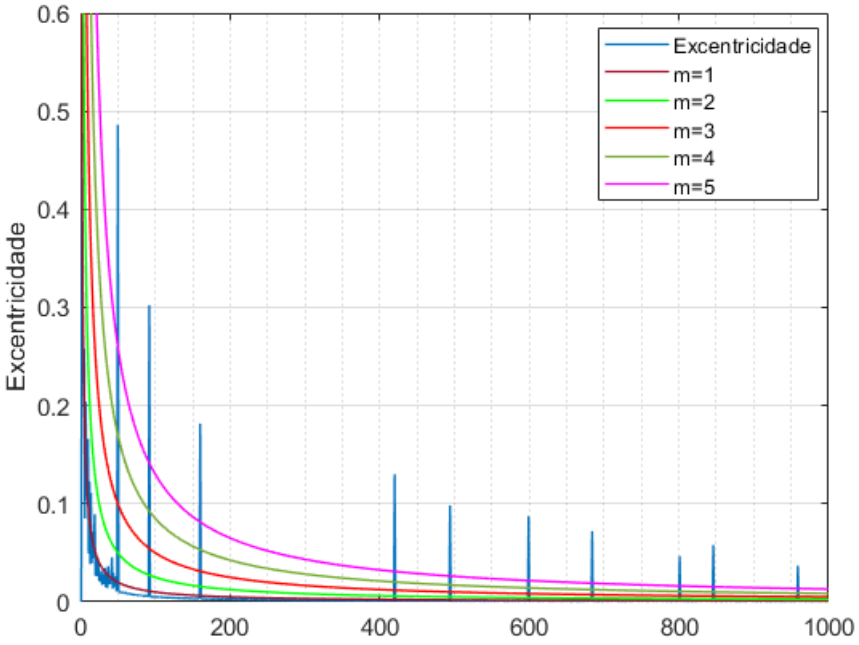


Fonte: Autoria própria.

Em seguida, a excentricidade é normalizada e utilizada com a definição de um limiar com base na desigualdade de Chebyshev(SAW; YANG; MO, 1984) para detecção de *outliers*. A condição expressa pela Eq. 2.20 define se a amostra atual X_i está "$m\sigma$" distante da média, onde o parâmetro m uma constante que define "quantos desvios-padrão" distante da média uma amostra deve estar para ser considerada um *outlier*. Na maioria dos casos, o valor de m é definido como 3 (NADA, a). Na Fig. 30 é apresentando uma análise do limiar de Chebyshev em função de diferente valores de m. Quando m=1, o limiar está mais sensível favorecendo a ocorrência de falsos positivos, enquanto para m=5 o limiar está mais restritivo o que irá acarretar numa maior incidência de falso negativos. Entretanto esse limiar é ponderado pela quantidade de amostras, ou seja, na i-ésima amostra o valor

do limiar tende a zero. Em outras palavras, quanto mais amostras são utilizadas no cálculo do TEDA, mais robustas são as estimativas da média e da variância e, consequentemente, menos restritivo será o limiar de *outlier*.

Figura 30 – Excentricidade normalizada x Desigualdade de Chebyshev.



Fonte: Autoria própria.

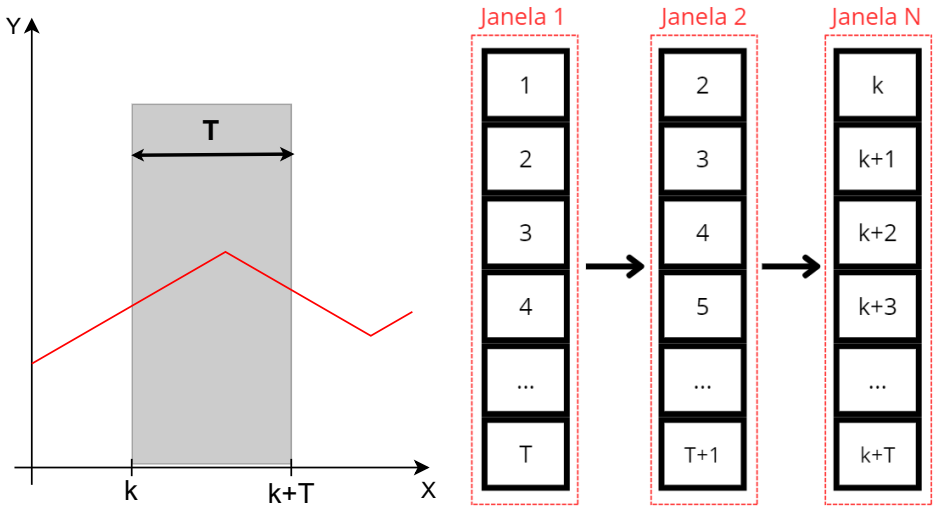
A seguir, será apresentado uma abordagem diferente para o TEDA Clássico. O novo conceito se baseia em uma janela móvel com o intuito que os valores não sofram tanta influência de valores passados.

3.4.2 TEDA JANELADO

O algoritmo proposto caracteriza-se por uma estrutura parecida com a do algoritmo anterior. A premissa principal do TEDA Janelado é executar uma análise de tipicidade e excentricidade com base em um conjunto de dados finitos, equivalente a um subconjunto das amostras. Considerando um fluxo de dados contínuo, esses dados são armazenados em um *buffer* semelhante a uma estrutura de uma fila de dados, a cada nova amostra, essa janela de tamanho fixo é deslocada e uma nova analise é feita, conforme apresentado na Fig. 31.

No Algoritmo 2 é apresentado o pseudocódigo do TEDA Janelado implementado para este trabalho. Essa abordagem possui um novo parâmetro que determina o tamanho da janela de processamento de dados (Tam). Desse modo, considerando a quantidade de dados aquistados (i), até que a condição $i > Tam$ seja atendida o método é similar ao TEDA Clássico. Quando $i < Tam$ os cálculos de média e variância serão afetados dado que

Figura 31 – Modelo da janela adotada.



Fonte: Autoria própria.

o cálculo já não é mais recursivo, a atualização desses valores deve ser calculada sempre após uma nova entrada e considerando apenas o subconjunto de dados que está sendo observado, ou seja, o novo método leva em consideração apenas os valores da janela atual. Os valores de média e variância são calculados pelas Eq. 3.6 e Eq. 3.7, respectivamente.

Devido a janela de dados ter um tamanho definido o cálculo da excentricidade também é modificado, pois o modelo não observa mais as *i-ésimas* amostras do conjunto, ele está restrito ao tamanho Tam do conjunto de dados, conforme enunciado na Eq. 3.8.

$$\mu_i(x) = \frac{1}{Tam} \sum_{k=Tam-i}^{Tam} x_k \tag{3.6}$$

$$\sigma_i^2(x) = \frac{1}{Tam} \sum_{k=Tam-i}^{Tam} (x_k - \mu_i(x))^2 \tag{3.7}$$

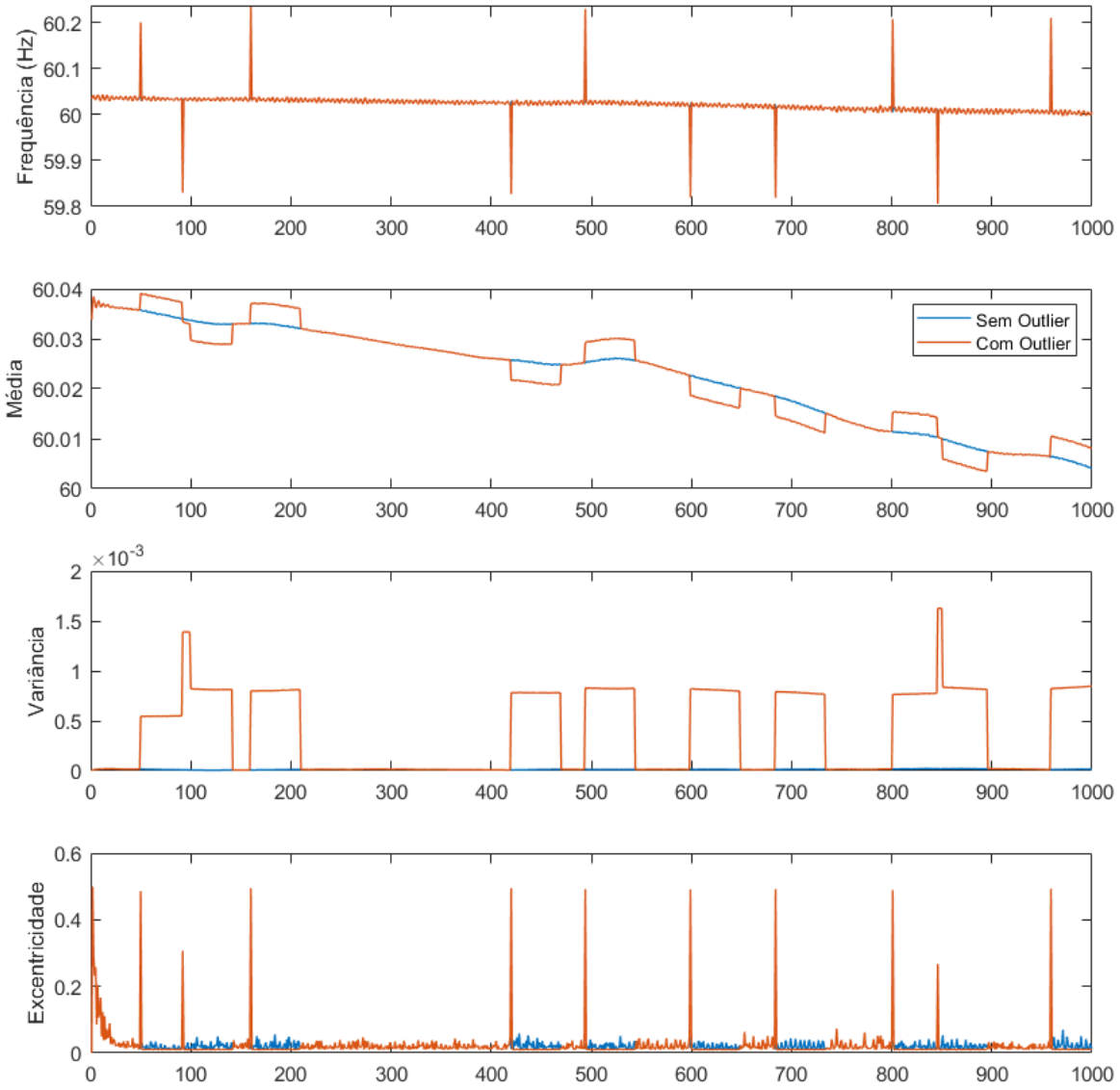
$$\xi_i(x) = \frac{1}{Tam} + \frac{(u_i - x_i)^T (u_i - x_i)}{Tam * \sigma_i^2} \tag{3.8}$$

Algorithm 2: TEDA Janelado

Entrada: Conjunto X de amostras.
Saída: É *Outlier*?
$Tam \leftarrow$ Tamanho da Janela ;
Amostra X_i;
while X_i *existe?* **do**
  ler a amostra $X_i \in X$;
  if i==*1* **then**
    $\mu_i = x_i$;
    $\sigma_i^2 = 0$;
  else if $i > Tam$ **then**
    $\mu_i = \frac{\sum_{k=i}^{Tam} X_k}{Tam}$;
    $\sigma_i^2 = \frac{\sum_{k=i}^{Tam}(X_k - \mu_i)^2}{Tam}$;
    $\xi_i(x) = \frac{1}{Tam} + \frac{(u_i - x_i)^T(u_i - x_i)}{Tam\sigma_i^2}$;
    $\zeta_i(x) = \frac{\xi_i}{2}$;
    if $\zeta_i(x) > \frac{m^2+1}{2*Tam}$ **then**
      X_i é *outlier*;
    else
      X_i não é *outlier*;
    end
  else
    $\mu_i = \frac{i-1}{i}\mu_{i-1} + \frac{1}{i}x_i$;
    $\sigma_i^2 = \frac{i-1}{i}\sigma_{i-1}^2 + \frac{1}{i-1}\|x_i - \mu_i\|^2$;
    $\xi_i(x) = \frac{1}{i} + \frac{(u_i - x_i)^T(u_i - x_i)}{i\sigma_i^2}$;
    $\zeta_i(x) = \frac{\xi_i}{2}$;
    if $\zeta_i(x) > \frac{m^2+1}{2i}$ **then**
      X_i é *outlier*;
    else
      X_i não é *outlier*;
    end
  end
  $i \leftarrow i + 1$;
end

A Fig. 32(a) apresenta uma análise da abordagem proposta para o conjunto de dados proposto no inicio dessa seção. A curva em azul representa os dados sem *outliers* e a curva laranja com *outliers*. Quando os *outliers* vão se apresentando, a média e a variância vão sendo sensibilizadas, porém ao contrário do TEDA Clássico essa sensibilidade é ainda maior 32(b)(c). Esse efeito se dá devido a janela utilizada, pois é devido a relação com aquele conjunto de amostras que a anomalia se torna mais impactante. Logo, em relação ao TEDA Clássico, os novos *outliers* são mais excêntricos e mantém valores altos durante todo ciclo da Fig. 32(d).

Figura 32 – (a)Conjunto de dados de entrada de frequência. (b) Média. (c) Variância. (d)Excentricidade.



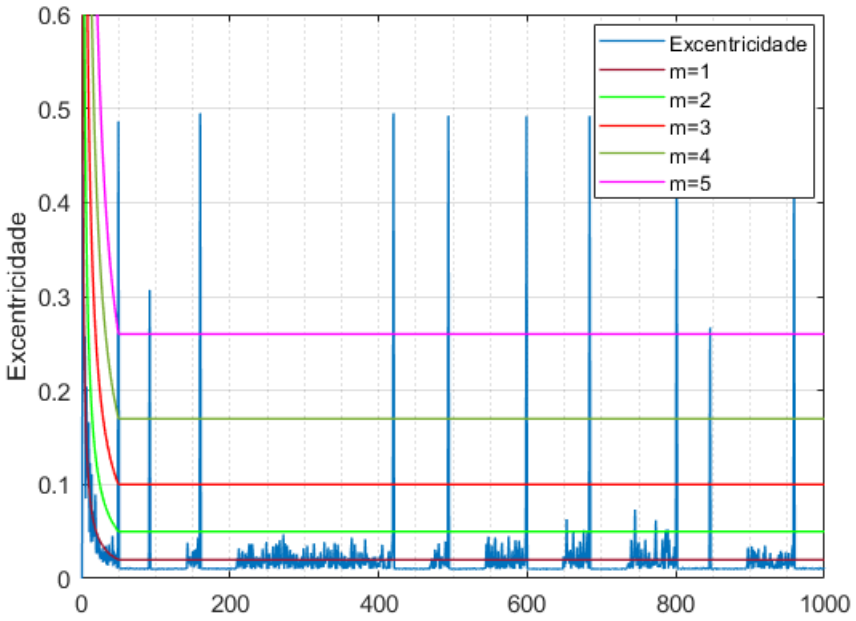
Fonte: Autoria própria.

Após realizado o cálculo de média, variância e excentricidade, a excentricidade é normalizada e comparada com relação ao limiar. Nesse momento, o cálculo do limiar é feito com base na desigualdade de Chebyshev. Entretanto há uma diferença com relação ao TEDA Clássico, quando o valor $i > Tam$ o limiar se tornará um valor constante, pois a excentricidade da amostra está sendo comparada apenas com a janela de dados específica, conforme enunciado na Eq. 3.9.

$$\begin{cases} i \leq Tam, & \zeta_i(x) > \frac{m^2+1}{2*i} \\ i > Tam, & \zeta_i(x) > \frac{m^2+1}{2*Tam} \end{cases} \tag{3.9}$$

Por fim, é apresentado na Fig. 33 o limiar com base na Eq. 3.9, considerando uma janela de 50 amostras ($Tam = 50$) e o valor de m variando de 1 a 5. De forma análoga ao TEDA Clássico, quando $m < 3$ o limiar é bastante sensível e pode favorecer a ocorrência de falsos negativos, porém para o TEDA Janelado, valores de $m \geqslant 3$ já apresenta um bom limiar de detecção para os dados utilizado. Perceba que o limiar continua constante quando $i > 100$, porém a excentricidade normalizada continua elevada, o que não prejudica a detecção de novos *outliers*.

Figura 33 – Excentricidade normalizada x Limiar de Detecção.



Fonte: Autoria própria.

Uma outra abordagem será apresentada a seguir, o TEDA Esquecimento. Assim como no TEDA Janelado, a premissa principal do modelo é a implementação de um fator de esquecimento, dando um peso diferente as novas amostras do conjunto de dados.

3.4.3 TEDA ESQUECIMENTO

O TEDA Esquecimento é baseado no TEDA Clássico, tendo como principal diferença a implantação de um fator de esquecimento nas equações recursivas de média e variância. Tal mecanismo visa uma maior capacidade de dar relevância aos dados recentes de um conjunto, ou seja, é possível ponderar o pesos das amostras antigas e novas para a definição do novo conceito. No Algoritmo 3 é apresentado o pseudocódigo do TEDA implementado neste trabalho com base na proposta de Nunes e Guedes (2024).

Algorithm 3: TEDA Esquecimento

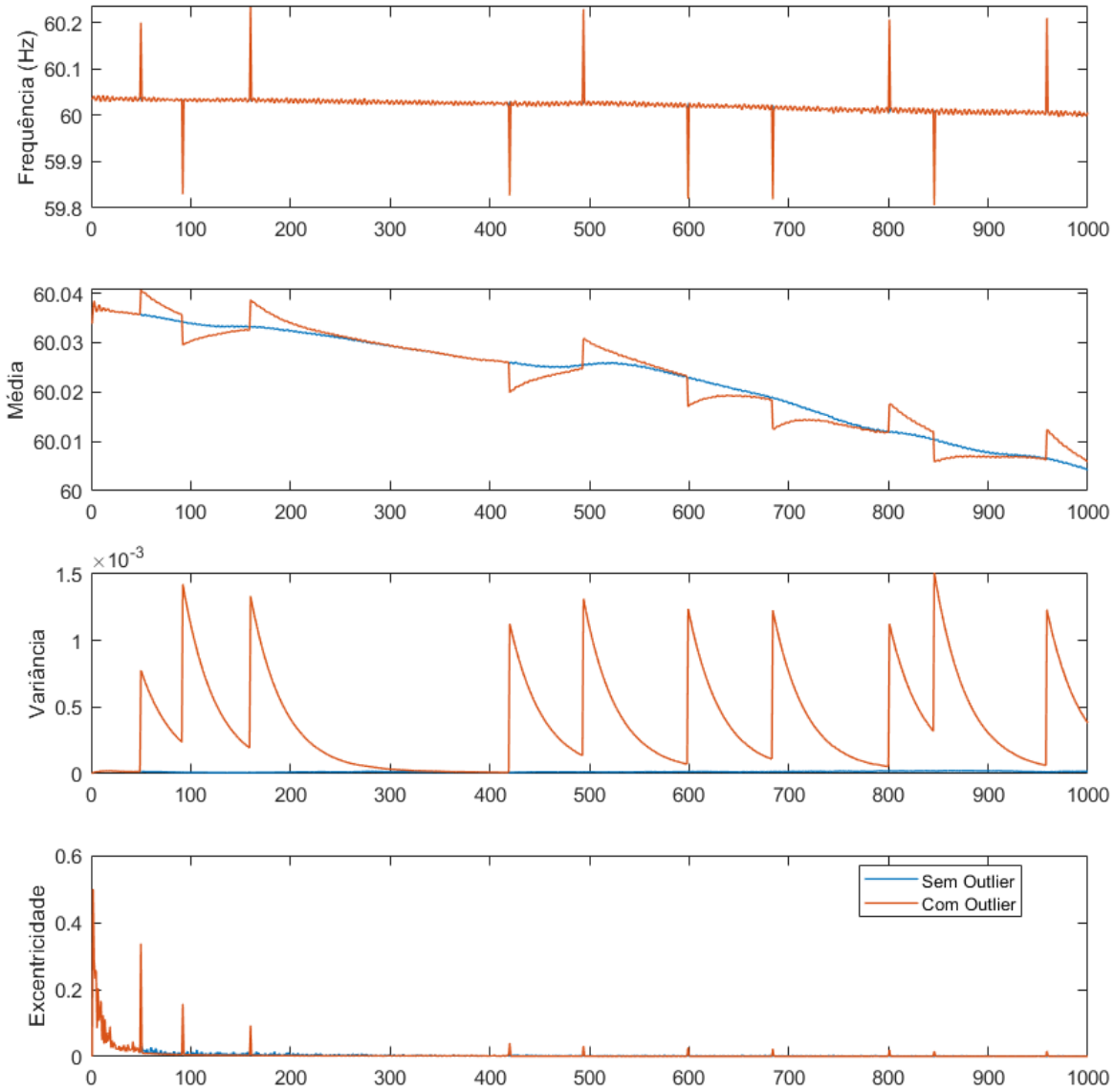
Entrada: Conjunto X de amostras.
Saída: É *Outlier*?
$\alpha \leftarrow$ Fator de esquecimento ;
Amostra X_i;
while X_i *existe?* **do**
  ler a amostra $X_i \in X$;
  if i==*1* **then**
    $\mu_i = x_i$;
    $\sigma_i^2 = 0$;
  else if $\left(\frac{k-1}{k}\right) \leq \alpha$ **then**
    $\mu_i = \frac{i-1}{i}\mu_{i-1} + \frac{1}{i}x_i$;
    $\sigma_i^2 = \frac{i-1}{i}\sigma_{i-1}^2 + \frac{1}{i-1}\|x_i - \mu_i\|^2$;
    $\xi_i(x) = \frac{1}{i} + \frac{(u_i - x_i)^T(u_i - x_i)}{i\sigma_i^2}$;
    $\zeta_i(x) = \frac{\xi_i}{2}$;
    if $\zeta_i(x) > \frac{m^2+1}{2i}$ **then**
      X_i *é outlier;*
    else
      X_i *não é outlier;*
    end
  else
    $\mu_i = \alpha\mu_{i-1} + (1-\alpha)x_i$;
    $\sigma_i^2 = \alpha\sigma_{i-1}^2 + (1-\alpha)\|x_i - \mu_i\|^2$;
    $\xi_i(x) = \frac{1}{i} + \frac{(u_i - x_i)^T(u_i - x_i)}{i\sigma_i^2}$;
    $\zeta_i(x) = \frac{\xi_i}{2}$;
    if $\zeta_i(x) > \frac{m^2+1}{2i}$ **then**
      X_i *é outlier;*
    else
      X_i *não é outlier;*
    end
  end
  $i \leftarrow i + 1$;
end

Perceba que enquanto a condição $\left(\frac{k-1}{k}\right) \leq \alpha$ não está atendida, os cálculos da média(μ_i) e da variância(σ_i) são calculados recursivamente pelas equações Eq. 2.16 e Eq. 2.17, igual ao TEDA Clássico. Entretanto, quando $\left(\frac{k-1}{k}\right) > \alpha$, os valores de média e variância passam a ser calculados levanto em consideração do fator de esquecimento(α), conforme as Eq. 2.21 e 2.22. Segundo Nunes e Guedes (2024), essa condição é necessária para que seja evitado ponderações anômalas nas amostras iniciais.

Na Fig. 34 é apresentando uma análise do algoritmo quanto a abordagem proposta para o conjunto de dados proposto no inicio dessa seção. A curva em azul representa os dados sem *outliers* e a curva em laranja os dados com *outliers*. Quando os dados ruins vão se apresentando, as curvas de média (Fig. 34(B)) e e a variância (Fig. 34(C)) vão se desprendendo da curva de referência. Ao contrário do TEDA Clássico, essa sensibilidade

é ainda maior e vai decaindo ao longo do tempo, chegando a se aproximar da curva de referência. Esse efeito pode ser justificado pela utilização do fator de esquecimento α, pois a partir de agora as novas amostras passam a ter uma maior relevância para o contexto atual.

Figura 34 – (a)Conjunto de dados de entrada de frequência. (b) Média. (c) Variância. (d)Excentricidade.

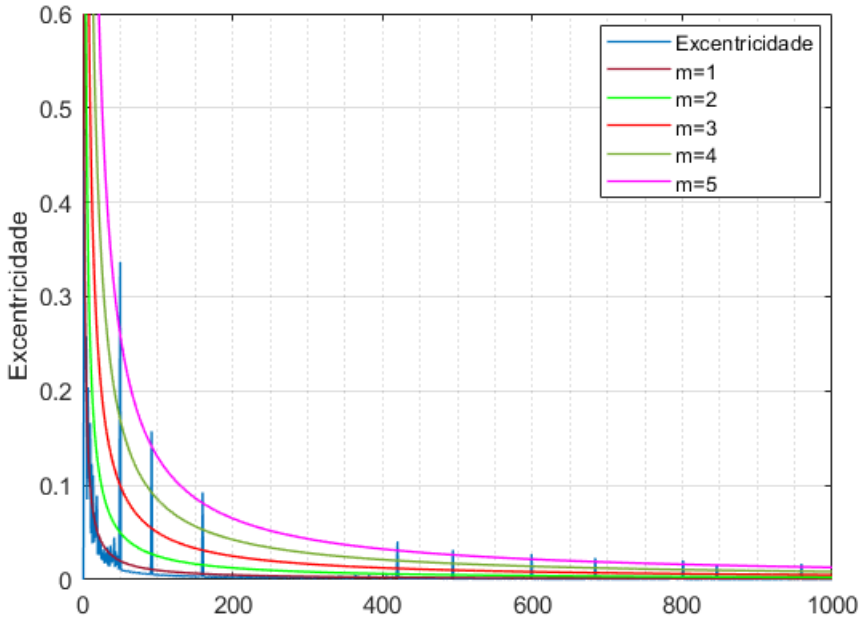


Fonte: Autoria própria.

Os limiares de detecção do modelo é similar ao TEDA Clássico com base na desigualdade de Chebyshev. O fator de esquecimento torna as novas amostras do conjunto de dados mais típicas, dado a existência de um peso maior para esses novos dados, ou seja, mesmo para as amostras mais excêntricas ao conjunto de dados, os valores de excentricidade tendem a diminuir, quando comparando com os outros modelos, sendo necessário a

utilização de um limiar de detecção m um pouco menor. Na Fig. 35 é apresentado uma análise do limiar de Chebyshev em função de diferente valores de m e a excentricidade do TEDA Esquecimento para o conjunto de dados apresentado anteriormente. Quando m=1 o limiar favorece a incidência de falsos positivos. Quando m=5 o limiar passa a ser mais restritivo acarretando em uma maior dificuldade de detecção, aumentando o número de falso negativos. Devido a natureza do TEDA Esquecimento, valores de excentricidade menor, é factível a adoção de valores de m um pouco menor, caso contrário, o algoritmo ficará cada vez mais restritivo.

Figura 35 – Excentricidade normalizada x Limiar de Detecção.



Fonte: Autoria própria.

4 RESULTADOS

Neste capítulo serão apresentados os resultados e as discussões para os métodos de detecção de *outliers* proposto neste trabalho, considerando medições fasoriais reais que foram coletadas de algumas instalações localizadas do Nordeste brasileiro, no ano de 2024.

Os três métodos apresentados nesse trabalho serão avaliados com base nos dois cenários de dados proposto. Inicialmente, utilizou-se o primeiro cenário, cujo dados tiveram a injeção aleatória de *outliers*. Para esse cenário, foram definidos três casos. Para cada método foi considerado variações dos valores do limiar de detecção (m), do tamanho da janela (Tam) e do fator de esquecimento (α). Os resultados serão avaliados conforme as métricas apresentadas no seção 3.3.

Por fim, será avaliado do desempenho dos modelos para detecção de anomalias nos dados de três PMUS de frequência, tensão e corrente. Nesse cenário, apesar dos dados serem anômalos, não podem ser considerados *outliers*. O cenário analisado é referente aos dados de um período que ocorreu um desligamento de uma linha de transmissão de 500 kV. A ideia é avaliar se todos os algoritmos são capazes de detectar esses fenômenos elétricos da rede.

4.1 CENÁRIO 01: DADOS COM INJEÇÃO ALEATÓRIA DE *OUTLIERS*

Nesta seção será realizada a apresentação do desempenho entre os TEDA Clássico, TEDA Janelado e o TEDA Esquecimento em relação aos conjuntos de dados apresentado na Seção 3.2 e utilizando as métricas de avaliação anteriormente descritas. A seguir, serão apresentado os resultados com base nos três casos definidos e para os diferentes ajustes de m, Tam e α.

4.1.1 CASO 01 - *OUTLIERS* TIPO ZEROS

Para o Caso 01 foi feita a injeção de dados errôneos do tipo zeros de maneira aleatória na curva de frequência, tensão e corrente, totalizando 180 *outliers*. Os resultados obtidos para frequência, corrente e tensão são apresentados nas Tabelas 1, 2 e 3, respectivamente.

É possível verificar dos resultados obtidos que todos os três algoritmos foram capazes de detectar todos os *outliers* do tipo zero em ao menos uma das configurações utilizadas. O TEDA Clássico se mostrou mais eficiente com relação aos demais, obtendo o MCC=1 para todas as configurações. O TEDA Janelado, quando utilizado os valores

Tabela 1 – Resumo da matriz de confusão e o MCC para os dados de frequência do Caso 01.

	m	*Tam*	α	VP	FP	VN	FN	MCC
TEDA	3	-	-	180	0	35820	0	1
Clássico	4	-	-	180	0	35820	0	1
	3	300	-	180	7	35813	0	0,9810
	3	500	-	180	14	35806	0	0,9630
TEDA	3	1000	-	180	4	35816	0	0,9890
Janelado	4	300	-	180	0	35820	0	1
	4	500	-	180	0	35820	0	1
	4	1000	-	180	0	35820	0	1
	3	-	0,98	180	0	35820	0	1
	3	-	0,97	180	1	35819	0	0,9972
TEDA	3	-	0,96	180	1	35819	0	0,9972
Esquecimento	4	-	0,98	180	0	35820	0	1
	4	-	0,97	177	0	35820	3	0,9915
	4	-	0,96	164	0	35820	16	0,9543

Tabela 2 – Resumo da matriz de confusão e o MCC para os dados de corrente do Caso 01.

	m	*Tam*	α	VP	FP	VN	FN	MCC
TEDA	3	-	-	180	0	35820	0	1
Clássico	4	-	-	180	0	35820	0	1
	3	300	-	180	61	35759	0	0,8634
	3	500	-	180	35	35685	0	0,9145
TEDA	3	1000	-	180	0	35820	0	1
Janelado	4	300	-	180	1	35819	0	0,9972
	4	500	-	180	0	35820	0	1
	4	1000	-	180	0	35820	0	1
	3	-	0,98	180	0	35820	0	1
	3	-	0,97	180	0	35820	0	1
TEDA	3	-	0,96	180	0	35820	0	1
Esquecimento	4	-	0,98	180	0	35820	0	1
	4	-	0,97	176	0	35820	4	0,9887
	4	-	0,96	166	0	35820	14	0,96014

de m=3 apresentou alguns falso positivos, que pode ser explicado pelo baixo limiar de detecção, tornando o método mais sensível e favorecendo a incidências de FPs. Por fim, o TEDA Esquecimento, se mostrou preciso para valores de $\alpha = 0,98$ e m=3, entretanto, para m=4 e valores de $\alpha = 0,97$ e $\alpha = 0,96$ não foi capaz de detectar todos as anomalias. Conforme Fig. 35 os valores de excentricidade para o TEDA Esquecimento tendem a ser menores que os demais métodos, tornando o algoritmo menos sensível.

Tabela 3 – Resumo da matriz de confusão e o MCC para os dados de tensão do Caso 01.

	m	*Tam*	α	VP	FP	VN	FN	MCC
TEDA	3	-	-	180	0	35820	0	1
Clássico	4	-	-	180	0	35820	0	1
	3	300	-	180	16	35804	0	0,9581
	3	500	-	180	3	35817	0	0,9917
TEDA	3	1000	-	180	4	35816	0	0,9890
Janelado	4	300	-	180	0	35820	0	1
	4	500	-	180	0	35820	0	1
	4	1000	-	180	0	35820	0	1
	3	-	0,98	180	0	35820	0	1
	3	-	0,97	180	0	35820	0	1
TEDA	3	-	0,96	180	0	35820	0	1
Esquecimento	4	-	0,98	180	0	35820	0	1
	4	-	0,97	178	0	35820	2	0,9887
	4	-	0,96	166	0	35820	14	0,96014

4.1.2 CASO 02 - *OUTLIERS* TIPO PICOS POSITIVOS E NEGATIVOS

No segundo caso de teste foram considerados a inserção de *outliers* do tipo pico positivo e negativo de forma aleatória nas três grandezas elétricas, totalizando 360 *outliers*. Na Tabela 4 é apresentado o resultado para os dados de frequência, na Tabela 5 é apresentado o resultado para os dados de corrente e na Tabela 6 é apresentado o resultado para os dados de tensão.

Tabela 4 – Resumo da matriz de confusão e o MCC para os dados de frequência do Caso 02.

	m	*Tam*	α	VP	FP	VN	FN	MCC
TEDA	3	-	-	360	0	35640	0	1
Clássico	4	-	-	360	0	35640	0	1
	3	300	-	360	0	35640	0	1
	3	500	-	360	0	35640	0	1
TEDA	3	1000	-	360	0	35640	0	1
Janelado	4	300	-	360	0	35640	0	1
	4	500	-	360	0	35640	0	1
	4	1000	-	360	0	35640	0	1
	3	-	0,98	360	0	35640	0	1
	3	-	0,97	360	0	35640	0	1
TEDA	3	-	0,96	358	0	35640	2	0,9971
Esquecimento	4	-	0,98	356	0	35640	4	0,99437
	4	-	0,97	336	0	35640	24	0,9657
	4	-	0,96	294	0	35640	66	0,9028

Para o conjunto de dados de frequência os três métodos tiveram um excelente desempenho, com destaque para o TEDA Clássico e o TEDA Janelado que detectou todos

os dados ruins nos diferente valores de m e *Tam*. Por outro lado, o TEDA Esquecimento até apresentou bons resultados para valores de m=3, porém, para valores de m=4 o algoritmo teve uma incidência maior de falso negativos, e sendo ainda pior quando $\alpha = 0.96$.

Tabela 5 – Resumo da matriz de confusão e o MCC para os dados de corrente do Caso 02.

	m	*Tam*	α	VP	FP	VN	FN	MCC
TEDA	3	-	-	167	0	35640	193	0,6792
Clássico	4	-	-	99	0	35640	261	0,5224
	3	300	-	360	13	35627	0	0,9822
	3	500	-	360	5	35635	0	0,9930
TEDA	3	1000	-	356	5	35635	4	0,9873
Janelado	4	300	-	360	0	35640	0	1
	4	500	-	357	0	35640	3	0,9957
	4	1000	-	329	0	35640	31	0,9555
	3	-	0,98	360	4	35636	0	0,9944
	3	-	0,97	360	3	35637	0	0,9955
TEDA	3	-	0,96	360	2	35638	0	0,9972
Esquecimento	4	-	0,98	353	1	35639	7	0,9887
	4	-	0,97	331	1	35639	29	0,9573
	4	-	0,96	294	0	35640	66	0,9028

Para os dados de corrente o TEDA Clássico apresentou uma baixa capacidade de detecção, tendo um péssimo resultado em relação as demais métodos. Tal fato é justificado pois o TEDA Clássico tem seu cálculo de excentricidade e tipicidades baseado nas médias e variâncias histórica de todo o conjunto de dados, e por isso, tende a ter uma menor sensibilidade ao longo do tempo. Por outro lado, o TEDA Janelado e o TEDA Esquecimento foram capazes de detectar todos os dados ruins em ao menos três das suas configurações. Note que para valores de m=4, o TEDA Esquecimento e o TEDA Janelado apresentou valores de falsos negativos. Vale salientar que o TEDA Janelado, quando m=4 e *Tam*=300, obteve o melhor MCC de todos as configurações avaliadas.

Tabela 6 – Resumo da matriz de confusão e o MCC para os dados de tensão do Caso 02.

	m	*Tam*	α	VP	FP	VN	FN	MCC
TEDA	3	-	-	360	0	35820	0	1
Clássico	4	-	-	360	0	35820	0	1
	3	300	-	360	1	35640	0	0,9986
	3	500	-	360	0	35640	0	1
TEDA	3	1000	-	360	0	35640	0	1
Janelado	4	300	-	360	0	35640	0	1
	4	500	-	360	0	35640	0	1
	4	1000	-	360	0	35640	0	1
	3	-	0,98	360	0	35640	0	1
	3	-	0,97	360	0	35640	0	1
TEDA	3	-	0,96	358	0	35640	4	0,9943
Esquecimento	4	-	0,98	356	0	35640	6	0,9915
	4	-	0,97	336	0	35640	23	0,9672
	4	-	0,96	294	0	35640	71	0,8951

Por fim, para o dados de tensão, os três métodos se mostraram eficientes. O TEDA Janelado apresentou alguns falsos positivos quando utilizado o valor de m=3, uma vez a presença do falso positivo é justificada pelo limiar de detecção baixo e o número da janela menor, o que torna o algoritmo mais sensível. Ao contrário do TEDA Janelado, o TEDA Esquecimento teve um pior desempenho quando utilizados valores de m=4, não sendo capaz de detectar todos os *outliers* para todos os valores de α.

4.1.3 CASO 03 - *OUTLIERS* TIPO ZEROS, PICOS POSITIVOS E PICO NEGATIVOS

Para o Caso 03 foram acrescentados 60 *outliers* do tipo pico positivo, 60 *outliers* do tipo pico negativo e 60 *outliers* do tipo zeros, na curvas de frequência, corrente e tensão. O objetivo é verificar o desempenho dos algoritmos diante das variações dos diferente tipos de *outliers*. Um resumo da matriz de confusão e MCC é apresentado nas Tabelas 7, 8 e 9 para as diferente configurações analisadas.

Os resultados do desempenho dos algoritmos para os valores de frequência são apresentados na Tabela 7. O TEDA Clássico encontrou mais dificuldades na detecção de outliers do tipo pico positivo e negativo, possivelmente devido à forte influência das amostras do tipo zero na excentricidade dos dados. O TEDA Janelado apresentou seu melhor desempenho com *Tam*=300, pois nesse caso as amostras ruins afetam a excentricidade por um intervalo menor. O TEDA Esquecimento teve o melhor coeficiente de correlação de Matthews quando comparado com os outros métodos, alcançando um MCC de 0,89. O melhor desempenho foi obtido com o fator de esquecimento igual a $\alpha = 0,96$, o que confere maior relevância às novas amostras do conjunto.

Tabela 7 – Resumo da matriz de confusão e o MCC para os dados de frequência do Caso 03.

	m	*Tam*	α	VP	FP	VN	FN	MCC
TEDA	3	-	-	64	0	35820	116	0,5953
Clássico	4	-	-	64	0	35820	116	0,5953
	3	300	-	139	5	35815	41	0,8627
	3	500	-	117	5	35815	63	0,7887
TEDA	3	1000	-	95	7	35813	85	0,7000
Janelado	4	300	-	139	0	35820	41	0,8782
	4	500	-	117	0	35820	63	0,7055
	4	1000	-	95	0	35820	85	0,7256
	3	-	0,98	116	0	35820	64	0,8020
	3	-	0,97	134	0	35820	46	0,8622
TEDA	3	-	0,96	144	0	35820	36	0,8939
Esquecimento	4	-	0,98	113	0	35820	67	0,7915
	4	-	0,97	129	0	35820	51	0,8459
	4	-	0,96	130	0	35820	50	0,8492

Para os dados de corrente, os resultados obtidos estão apresentados na Tabela 8. Todas as configurações analisadas enfrentaram dificuldades na detecção de outliers, devido às grandes variações nas amostras causadas por valores zero. Os três algoritmos são baseados na excentricidade, que é calculada usando a média e a variância dos dados. Valores muito dispersos tendem a reduzir essa excentricidade, dificultando a detecção de dados anômalos. Entre os três métodos, o TEDA Clássico teve o pior desempenho, geralmente detectando apenas os zeros. O TEDA Janelado apresentou seu melhor desempenho para valores de *Tam*=300, mas sofreu com os alto números de falsos positivos ao utilizar m=3. Por fim, os melhores resultados foram obtidos pelo TEDA Esquecimento, com valores do fator de esquecimento iguais a 0,97 e 0,96.

Na Tabela 9 são apresentados os resultados obtidos para os dados de tensão. Mais uma vez, os algoritmos tiveram dificuldade para detecção de todos os dados ruins. O TEDA Esquecimento teve o melhor desempenho com relação aos demais, apresentando uma detecção de 140 *outliers* de um total de 160 para $\alpha = 0.96$ e m=3. O TEDA Janelado foi capaz de detectar 130 *outliers* considerando o *Tam*=300, entretanto para m=3 houve a presença de 12 falso positivos. Por fim, o TEDA Clássico basicamente só foi capaz de detectar os *outliers* do tipo zero.

Tabela 8 – Resumo da matriz de confusão e o MCC para os dados de corrente do Caso 03.

	m	*Tam*	α	VP	FP	VN	FN	MCC
TEDA	3	-	-	64	0	35820	116	0,5953
Clássico	4	-	-	63	0	35820	117	0,5906
	3	300	-	118	97	35723	62	0,5976
	3	500	-	94	50	35770	83	0,58201
TEDA	3	1000	-	74	39	35781	106	0,4170
Janelado	4	300	-	107	5	35815	73	0,7527
	4	500	-	84	0	35820	96	0,6822
	4	1000	-	69	0	35820	111	0,6181
	3	-	0,98	99	3	35817	81	0,7297
	3	-	0,97	120	1	35819	60	0,8124
TEDA	3	-	0,96	124	3	35817	56	0,8194
Esquecimento	4	-	0,98	90	0	35820	90	0,7062
	4	-	0,97	103	0	35820	77	0,7556
	4	-	0,96	107	0	35820	73	0,7702

Tabela 9 – Resumo da matriz de confusão e o MCC para os dados de tensão do Caso 03.

	m	*Tam*	α	VP	FP	VN	FN	MCC
TEDA	3	-	-	61	0	35820	119	0,5811
Clássico	4	-	-	61	0	35820	119	0,5811
	3	300	-	130	12	35808	50	0,8123
	3	500	-	108	0	35820	72	0,7738
TEDA	3	1000	-	80	1	35819	100	0,6615
Janelado	4	300	-	130	0	35820	50	0,8492
	4	500	-	108	0	35820	72	0,7738
	4	1000	-	80	0	35820	100	0,6657
	3	-	0,98	117	1	35819	63	0,7297
	3	-	0,97	134	0	35820	46	0,8124
TEDA	3	-	0,96	140	0	35820	40	0,8894
Esquecimento	4	-	0,98	112	0	35820	68	0,7062
	4	-	0,97	127	0	35820	53	0,7556
	4	-	0,96	133	0	35820	47	0,7702

4.2 CENÁRIO 02: DADOS REAIS DE UM DISTÚRBIO ELÉTRICO

Apesar do escopo desse trabalho ser a detecção de dados ruins nos dados de medição fasorial sincronizada, avaliar os algoritmos utilizados nesse trabalho diante do cenários de eventos elétricos da rede pode ser de grande relevância para a literatura, uma vez que, diversos trabalhos do estado da arte buscam maneiras de capturar tais fenômenos elétricos a partir o conjunto de dados de forma autônoma. Entretanto, o TEDA Clássico não teve uma boa capacidade de detecção e por isso não será considerado nessa análise.

De fato, os dados da PMU durante uma contingência no sistema elétricos é uma anomalia de dados, destoando dos dados dito típicos para determinada grandeza elétrica. Uma premissa importante para a análise que vamos fazer a seguir é, como as PMUs tem suas etiquetagem de tempo com base no sinal de GPS, tal medida possui de fato uma sincronização com as demais PMUs, o que torna factível a possibilidade de analisarmos mais de uma PMU considerando o mesmo instante de tempo. Isso nós leva a uma certeza, se um dado anômalo for detectado por mais de uma PMU, ele deve ser é de fato um dado confiável. Nessa seção iremos dividir a análise em três abordagens: frequência, corrente e tensão.

4.2.1 DADOS DE FREQUÊNCIA

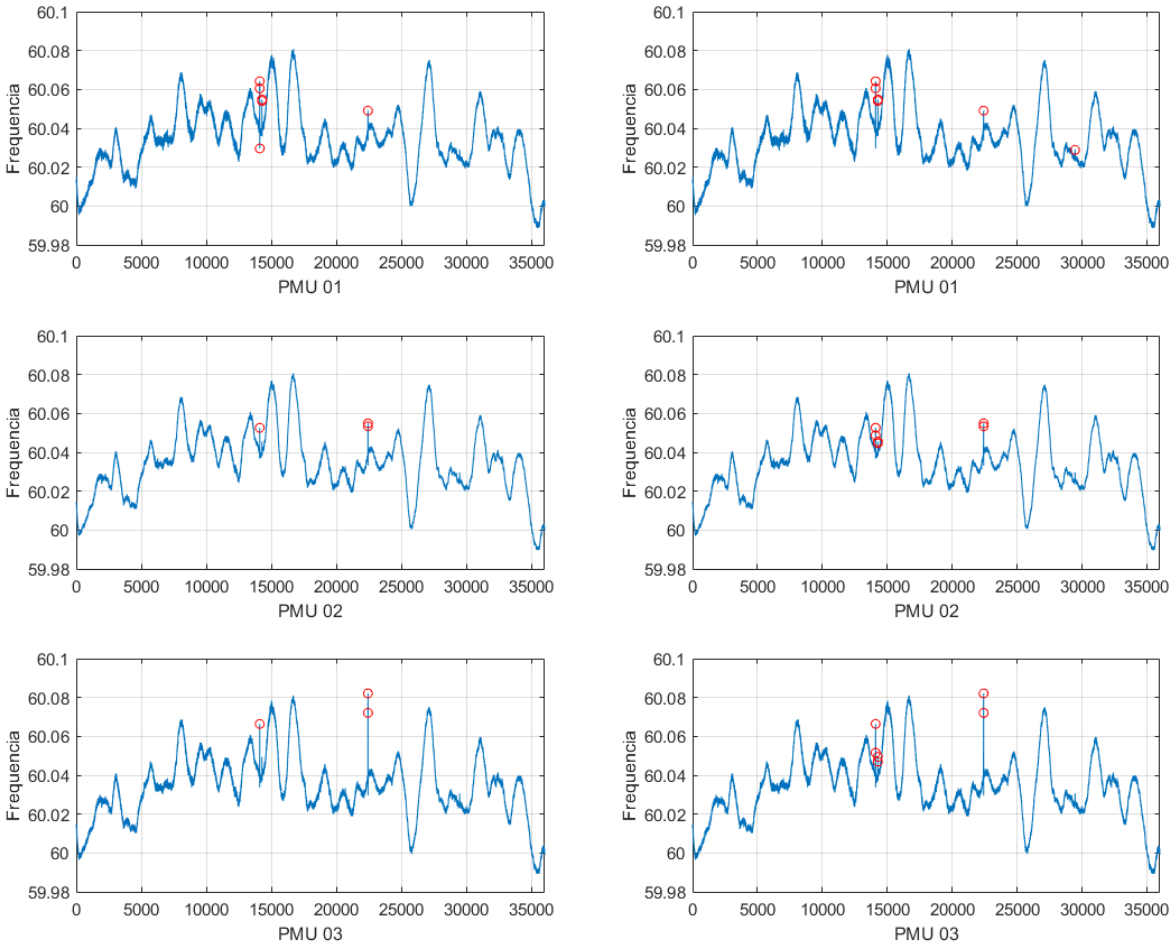
Para a análise dos dados de frequência, os parâmetro utilizados para cada método é apresentado na Tabela 10. Durante as alterações topológicas da rede, ou seja, energização e desenergização de algum equipamento causam picos instantâneo nos dados de frequência de uma PMU. Esse efeito é justificado pelo fato de que tais valores são cálculos por meio de estimativas considerando as amostas de tensão e corrente da PMU. Logo, para o caso analisado, no instante da abertura do primeiro terminal da linha de transmissão deve ter um pico e, em seguida, após a abertura do outro terminal, outro pico.

Tabela 10 – Parâmetros utilizados para os dados de frequência.

	m	Tam	α
TEDA Janelado	4	300	-
TEDA Esquecimento	4	-	0,98

Os resultado obtidos são apresentados na Fig. 36, sendo os círculos vermelhos os pontos que o algoritmo classificou como anômalos. Neste cenário, todas as PMUs foram sensibilizadas pelo o evento de desligamento de ambos os terminais da linha e os algoritmos também foram capazes de perceber essa mudança no contexto dos dados. Como em todo modelo de detecção é comum a incidência de falsos positivos, como por exemplo, a amostra i=29466 na curva da PMU 01 que foi considerada *outlier* em apenas apenas um dos métodos. Por outro lado, as amostras note que a vizinhança das amostras i=14100 e i=22420 foram detectadas nos dois modelos para os três dados de PMU. De fato, se faz necessário a existência de um mecanismo de inferência para fazer uma comparação direta na vizinhança das amostras detectadas entre as várias PMUs e expurgar o que não deve ser tratado como dado ruim. Logo, podemos atestar a capacidade de detecção da ocorrência de algum evento na rede elétrica para os dois métodos utilizados na curva de frequência.

Figura 36 – Resultado de detecção de anomalias nos dados de Frequência. A esquerda o TEDA Janelado. A direita o TEDA Esquecimento.



Fonte: Autoria própria.

4.2.2 DADOS DE CORRENTE

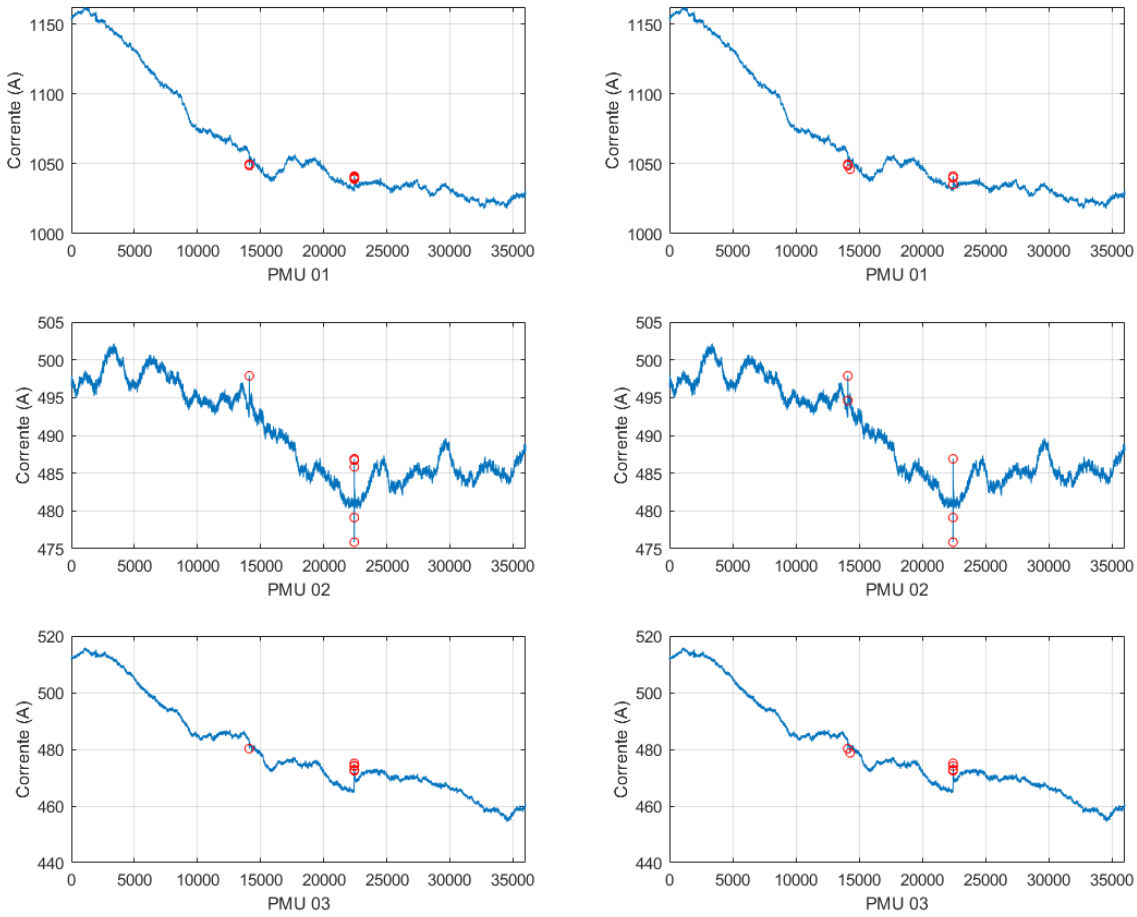
Para a análise dos dados de corrente, os parâmetros utilizados para cada método estão apresentados na Tabela 11. As medições de corrente variam ao longo do dia devido a mudanças topológicas e variações de carga. Esses valores são coletados diretamente dos equipamentos por meio de transformadores de corrente. Durante desligamentos intempestivos, o sistema como um todo exibe uma resposta eletrodinâmica, e, como esperado, as medições de corrente também são afetadas. No caso estudado, durante a abertura de cada terminal, as PMUs detectam essas variações com intensidades que variam conforme a localização geográfica.

Tabela 11 – Parâmetros utilizados para os dados de corrente.

	m	Tam	α
TEDA Janelado	4	300	-
TEDA Esquecimento	4	-	0,98

Na Figura 37, os resultados dos dados de corrente são apresentados graficamente. As curvas de corrente mostram diferenças significativas entre si, o que é esperado, já que estão em pontos elétricos distintos e têm carregamentos diferentes. No entanto, os métodos conseguiram detectar os dois eventos que ocorreram no sistema, nas proximidades das amostras i=141000 e i=22420, como previsto. Esse tipo de análise é mais complexo devido à diversidade de conceitos nos conjuntos de dados. Os dois modelos apresentaram a capacidade de detecção das duas ocorrências da rede elétrica.

Figura 37 – Resultado de detecção de anomalias nos dados de Corrente. A esquerda o TEDA Janelado. A direita o TEDA Esquecimento.



Fonte: Autoria própria.

4.2.3 DADOS DE TENSÃO

As medições de tensão de um equipamento são provenientes dos transformadores de potência e adquiridas pela PMU com uma frequência de amostragem de 60 medições por segundo. Essa alta frequência de amostragem permite a observação dos comportamentos transitórios da tensão durante distúrbios elétricos. As medições de tensão estão sempre correlacionadas com a quantidade de potência reativa no sistema. Por exemplo, a inserção de um banco de capacitores pode ajudar a elevar o perfil de tensão, enquanto o desligamento de uma linha de transmissão pode auxiliar no controle da tensão.

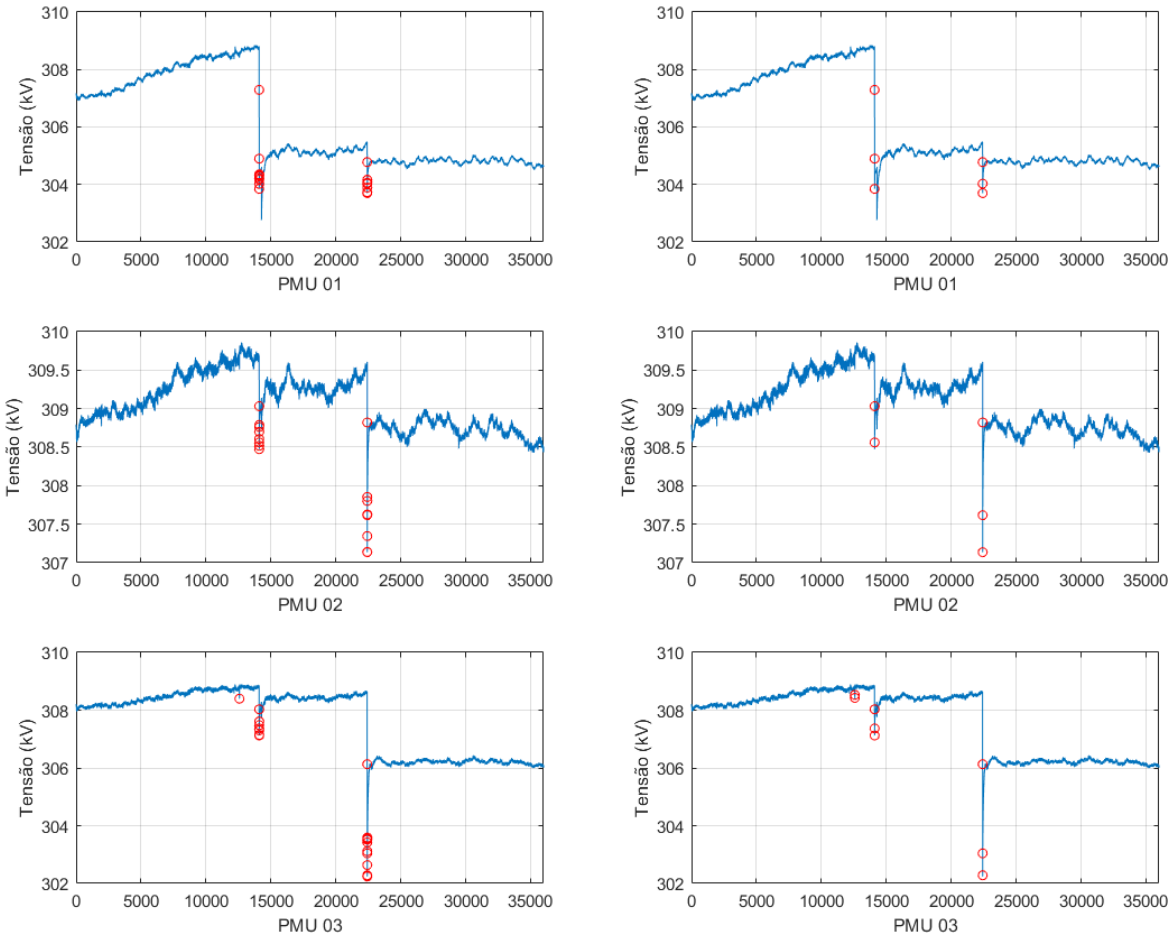
No caso do evento proposto para este estudo, espera-se um pequeno afundamento na tensão durante o desligamento do primeiro terminal da linha de transmissão. Após o desligamento do segundo terminal, outro afundamento de tensão também é esperado. Os parâmetros utilizados para cada método estão apresentados na Tabela 12.

Tabela 12 – Parâmetros utilizados para os dados de tensão.

	m	Tam	α
TEDA Janelado	4	300	-
TEDA Esquecimento	4	-	0,98

Os resultados obtidos para o caso de tensão são apresentandos na Fig. 38. Os dois modelos foram capazes de detectar com precisão os eventos que ocorrerão na vizinha dos instante $i = 14100$ e $i = 22420$. O TEDA Janelado apresentou um falso positivo nas medições da PMU 03, algo que poderia ser facilmente desconsiderado se for observados a existência de eventos nas outras PMU. Mais uma vez, os modelos utilizados foram capazes de detectar as duas ocorrências elétricas.

Figura 38 – Resultado de detecção de anomalias nos dados de tensão. A esquerda o TEDA Janelado. A direita o TEDA Esquecimento.



Fonte: Autoria própria.

5 CONCLUSÃO

Neste trabalho foi apresentado uma metodologia para detecção de *outliers* em dados de medição fasorial baseado no conceito de tipicidade e excentricidade. Essa abordagem é orientada apenas para os dados e depende de uma única PMU, não precisando do conhecimento da topologia ou dos parâmetros do sistema. Ele pode melhorar a qualidade dos dados da PMU e estabelecer uma base para uma melhor aplicação dos dados em softwares utilizados na operação do sistema elétrico.

A proposta de desenvolvimento foi a criação de dois modelos baseado no TEDA Clássico, onde a primeira abordagem observaria apenas um subconjunto de amostras e a segunda abordagem, um fator de esquecimento. A primeira metodologia foi denominada como TEDA Janelado, ao contrário do TEDA Clássico, cada nova amostra é calculado a sua excentricidade e tipicidade com base apenas em uma janela de dados móvel, ocasionando uma menor "mémoria" dos dados em relação ao TEDA Clássico. A segunda metodologia foi denominada como TEDA Esquecimento, foi implementando o fator de esquecimento no TEDA Clássico, criando uma ponderação fixa entre as novas amostras do conjunto e a média e variância do conjunto de dados.

Para avaliar o desempenho dos modelos foram avaliados em dois cenários, para as grandezas de frequência, corrente de fase e tensão fase. O primeiro cenário os métodos foram avaliado para a condição de alta incidência de dados ruins, onde foram criados três casos de acordo com a quantidade e os tipos de *outliers* que foram inseridos de maneira aleatória, conforme explanado na Seção 3.2. O segundo cenário estudado, os métodos foram avaliados em um cenário de dados reais durante um distúrbio elétricos, com o objetivo de avaliar a capacidade de detecção de eventos.

A eficácia dos algoritmos diante do cenário 01, visando a detecção de *outliers*, foi feita por meio de um estudo comparativo entre os métodos TEDA Clássico, TEDA Janelado e TEDA Esquecimento. Para todos os resultados foram apresentados sua matriz de confusão e a avaliados através do coeficiente de correlação de Matthews - MCC.

Considerando o Caso 01, todos métodos apresentaram um MCC=1, sendo capazes de detectar todos as anomalias. O TEDA Janelado para valores de m=3 apresentou uma alta incidência de falsos positivos e o TEDA Esquecimento apresentou alguns falsos negativos.

Para o Caso 02, quando comparado os dados de frequência e tensão, todos os métodos apresentaram ao menos uma solução com MCC=1. O pior desempenho foi o TEDA Clássico, com um MCC=0,5224. Tal dificuldade de detecção para os dados de

corrente, justificado pelo fato da menor amplitude dos *outliers* quando comparado com o desvio padrão dos dados do conjunto. Ainda no cenário de corrente, o TEDA Janelado teve o melhor desempenho, com um MCC=1 para as configurações de m=4 e Tam=300.

Quando os métodos foram avaliados no pior conjunto de dados, o Caso 03, o TEDA Clássico teve os piores valores de MCC quando comparado com os outros métodos. O TEDA Esquecimento apresentou os melhores valores de MCC, sendo o melhor MCC=0,8939. O TEDA Esquecimento apresentou uma alta capacidade de detecção de verdadeiros positivos e uma baixa incidência de falsos positivos. O TEDA Janelado apresentou uma boa capacidade de detecção para valores de janelas menores. Vale destacar, que para esse caso a característica principal do modelo janelado e o fator de esquecimento podem ser relacionado diretamente ao melhor desempenho dos modelos.

Em seguida, foi proposto um cenário 02, onde o TEDA Janelado e o TEDA Esquecimento foram aplicados para detecção de dados anômalos durante uma ocorrência elétrica nos dados de três PMUs distintas. Ambos os métodos foram capazes de detectar com precisão os instantes no qual ocorreu as contingências do sistema.

Em trabalhos futuros, com a finalidade de avaliar ainda mais a confiabilidade das metodologias propostas neste trabalho para a detecção de *outlier*, sugere-se:

- Avaliação da necessidade de otimização dos parâmetros: m, tamanho da janela e fator de esquecimento;
- Verificar a aplicação de um limiar adaptativo em relação a variância para os métodos;
- Utilizar os métodos em aplicações em tempo real;
- Avaliar o desempenho dos métodos para outros distúrbios elétricos;

REFERÊNCIAS

MATTHEWS, Brian W. In: . [S.l.: s.n.]. Citado 2 vezes nas páginas 42 e 56.

AMIDAN, B. G.; FERRYMAN, T. A.; COOLEY, S. K. Data Outlier Detection Using The Chebyshev Theorem. In: IEEE. *2005 IEEE Aerospace Conference*. [S.l.], 2005. p. 3814–3819. Citado na página 29.

ANAPARTHI, K. K. et al. Coherency Identification In Power Systems Through Principal Component Analysis. *IEEE Transactions on Power Systems*, IEEE, v. 20, n. 3, p. 1658–1660, 2005. Citado na página 33.

ANGELOV, P. Outside The Box: An Alternative Data Analytics Framework. *Journal of Automation Mobile Robotics and Intelligent Systems*, Sieć Badawcza Łukasiewicz-Przemysłowy Instytut Automatyki i Pomiarów, v. 8, n. 2, p. 29–35, 2014. Citado 3 vezes nas páginas 39, 41 e 53.

BABA, M. et al. A Review of The Importance of Synchrophasor Technology, Smart Grid, and Applications. *Bulletin of the Polish Academy of Sciences: Technical Sciences*, p. e143826–e143826, 2022. Citado na página 24.

BECEJAC, T.; DEHGHANIAN, P.; KEZUNOVIC, M. Impact of The Errors in The PMU Response on Synchrophasor-based Fault Location Algorithms. In: IEEE. *2016 North American Power Symposium (NAPS)*. [S.l.], 2016. p. 1–6. Citado na página 27.

BENGIO, Y.; SIMARD, P.; FRASCONI, P. Learning Long-term Dependencies With Gradient Descent is Difficult. *IEEE Transactions on Neural Networks*, IEEE, v. 5, n. 2, p. 157–166, 1994. Citado na página 37.

BEZDEK, J. C. Objective Function Clustering. In: *Pattern Recognition With Fuzzy Objective Function Algorithms*. [S.l.]: Springer, 1981. p. 43–93. Citado na página 31.

BRAHMA, S. et al. Real-time Identification of Dynamic Events In Power Systems Using PMU Data, and Potential Applications—models, Promises, and Challenges. *IEEE transactions on Power Delivery*, IEEE, v. 32, n. 1, p. 294–301, 2016. Citado na página 24.

BREUNIG, M. M. et al. LOF: Identifying Density-based Local Outliers. In: *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*. [S.l.: s.n.], 2000. p. 93–104. Citado na página 33.

CEPEDA, J.; COLOMÉ, D.; CASTRILLÓN, N. Dynamic Vulnerability Assessment Due to Transient Instability Based on Data Mining Analysis For Smart Grid Applications. In: IEEE. *2011 IEEE PES CONFERENCE ON INNOVATIVE SMART GRID TECHNOLOGIES LATIN AMERICA (ISGT LA)*. [S.l.], 2011. p. 1–7. Citado na página 33.

CHEN, C.; WANG, J.; ZHU, H. Effects of Phasor Measurement Uncertainty on Power Line Outage Detection. *IEEE Journal of Selected Topics in Signal Processing*, IEEE, v. 8, n. 6, p. 1127–1139, 2014. Citado na página 27.

COSTA, B. S. J. et al. Online Fault Detection Based on Typicality and Eccentricity Data Analytics. In: IEEE. *2015 International Joint Conference on Neural Networks (IJCNN)*. [S.l.], 2015. p. 1–6. Citado na página 40.

DENG, X. et al. Impact of Low Data Quality on Disturbance Triangulation Application Using High-density PMU Measurements. *IEEE Access*, IEEE, v. 7, p. 105054–105061, 2019. Citado na página 24.

DENG, X. et al. Deep Learning Model to Detect Various Synchrophasor Data Anomalies. *IET Generation, Transmission & Distribution*, Wiley Online Library, v. 14, n. 24, p. 5739–5745, 2020. Citado 3 vezes nas páginas 25, 26 e 36.

DERVIŠKADIĆ, A. et al. Architecture and Experimental Validation of a Low-latency Phasor Data Concentrator. *IEEE Transactions on Smart Grid*, IEEE, v. 9, n. 4, p. 2885–2893, 2016. Citado na página 22.

DUNN, J. C. A Fuzzy Relative of The ISODATA Orocess and Its Use in Detecting Compact Well-separated Clusters. Taylor & Francis, 1973. Citado na página 31.

ESTER, M. et al. A Density-based Algorithm for Discovering Clusters in Large Spatial Databases With Noise. In: *kdd*. [S.l.: s.n.], 1996. v. 96, n. 34, p. 226–231. Citado na página 29.

HANNON, C. et al. Real-time Anomaly Detection and Classification in Streaming PMU Data. In: IEEE. *2021 IEEE Madrid PowerTech*. [S.l.], 2021. p. 1–6. Citado na página 38.

HILL, D. J.; MINSKER, B. S. Anomaly Detection in Streaming Environmental Sensor Data: A Data-driven Modeling Approach. *Environmental Modelling & Software*, Elsevier, v. 25, n. 9, p. 1014–1022, 2010. Citado na página 24.

HUANG, Z. et al. Performance evaluation of phasor measurement systems. In: *2008 IEEE Power and Energy Society General Meeting - Conversion and Delivery of Electrical Energy in the 21st Century*. [S.l.: s.n.], 2008. p. 1–7. Citado na página 22.

IGLEWICZ, B.; HOAGLIN, D. Volume 16: How To Detect and Handle Outliers. *The ASQC Basic References in Quality Control: Statistical Techniques*, EF Mykytka, Ed, v. 16, 1993. Citado na página 31.

IZAKIAN, H.; PEDRYCZ, W. Anomaly Detection in Time Series Data Using a Fuzzy C-means Clustering. In: IEEE. *2013 Joint IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS)*. [S.l.], 2013. p. 1513–1518. Citado na página 32.

JONES, K. D.; PAL, A.; THORP, J. S. Methodology For Performing Synchrophasor Data Conditioning and Validation. *IEEE Transactions on Power Systems*, IEEE, v. 30, n. 3, p. 1121–1130, 2014. Citado na página 30.

KAMWA, I.; SAMANTARAY, S.; JOOS, G. Compliance Analysis of PMU Algorithms and Devices For Wide-area Stabilizing Control of Large Power Systems. *IEEE Transactions on Power Systems*, IEEE, v. 28, n. 2, p. 1766–1778, 2012. Citado na página 22.

KARLSON, D.; HEMMINGSSON, M.; LINDAHL, S. Wide Area System Monitoring and Control. *IEEE Power and Energy Magazine*, p. 69–76, 2004. Citado na página 18.

KHALEDIAN, E. et al. Real-time Synchrophasor Data Anomaly Detection and Classification Using Isolation Forest, Kmeans and Loop. *IEEE Transactions on Smart Grid*, IEEE, v. 12, n. 3, p. 2378–2388, 2020. Citado 2 vezes nas páginas 38 e 39.

KRAMER, M. A. Nonlinear Principal Component Analysis Using Autoassociative Neural Networks. *AIChE journal*, Wiley Online Library, v. 37, n. 2, p. 233–243, 1991. Citado na página 32.

KUNDUR, P. EPRI Power System Engineering Series. *Power System Stability and Control*, McGraw Hill, 1994. Citado na página 33.

LAZAREVIC, A.; KUMAR, V. Feature Bagging For Outlier Detection. In: *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*. [S.l.: s.n.], 2005. p. 157–166. Citado na página 28.

LIN, D. et al. PlatoGL: Effective and Scalable Deep Graph Learning System for Graph-enhanced Real-time Recommendation. In: *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. [S.l.: s.n.], 2022. p. 3302–3311. Citado na página 28.

LIU, X. et al. Principal Component Analysis of Wide-area Phasor Measurements For Islanding Detection—A geometric View. *IEEE Transactions on Power Delivery*, IEEE, v. 30, n. 2, p. 976–985, 2015. Citado na página 33.

MACQUEEN, J. et al. Some Methods For Classification and Analysis of Multivariate Observations. In: OAKLAND, CA, USA. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. [S.l.], 1967. v. 1, n. 14, p. 281–297. Citado na página 31.

MAHAPATRA, K.; CHAUDHURI, N. R.; KAVASSERI, R. Bad Data Detection in PMU Measurements Using Principal Component Analysis. In: IEEE. *2016 North American Power Symposium (NAPS)*. [S.l.], 2016. p. 1–6. Citado 2 vezes nas páginas 15 e 32.

MARTIN, K. et al. Synchrophasor Measurements for Power Systems under the Standard IEC/IEEE 60255-118-1. *IEEE, December*, 2018. Citado na página 19.

MATTHEWS, B. W. Comparison of The Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, Elsevier, v. 405, n. 2, p. 442–451, 1975. Citado na página 53.

MATTHEWS, S. J.; LEGER, A. S. Leveraging Mapreduce and Synchrophasors For Real-time Anomaly Detection in The Smart Grid. *IEEE Transactions on Emerging Topics in Computing*, IEEE, v. 7, n. 3, p. 392–403, 2017. Citado na página 34.

NUNES, Y. T.; GUEDES, L. A. Concept Drift Detection Based on Typicality and Eccentricity. *IEEE Access*, IEEE, 2024. Citado 3 vezes nas páginas 43, 61 e 62.

PHADKE, A. Pmu Memories: Looking Back Over 40 Years. *IEEE Power and Energy Magazine*, IEEE, v. 13, n. 5, p. 96–94, 2015. Citado na página 18.

PHADKE, A.; THORP, J. History and Applications of Phasor Measurements. In: IEEE. *2006 IEEE PES Power Systems Conference and Exposition*. [S.l.], 2006. p. 331–335. Citado na página 19.

PHADKE, A. G.; BI, T. Phasor Neasurement Units, WAMS, and Their Applications in Protection and Control of Power Systems. *Journal of Modern Power Systems and Clean Energy*, SGEPRI, v. 6, n. 4, p. 619–629, 2018. Citado 2 vezes nas páginas 20 e 23.

PHADKE, A. G.; THORP, J. S. *Synchronized Phasor Measurements and Their Applications*. [S.l.]: Springer, 2008. v. 1. Citado 2 vezes nas páginas 20 e 23.

PUKELSHEIM, F. The Three Sigma Rule. *The American Statistician*, Taylor & Francis, v. 48, n. 2, p. 88–91, 1994. Citado na página 34.

REN, H.; YE, Z.; LI, Z. Anomaly Detection Based on a Dynamic Markov Model. *Information Sciences*, Elsevier, v. 411, p. 52–65, 2017. Citado na página 28.

SAW, J. G.; YANG, M. C.; MO, T. C. Chebyshev Inequality With Estimated Mean and Variance. *The American Statistician*, Taylor & Francis, v. 38, n. 2, p. 130–132, 1984. Citado 2 vezes nas páginas 42 e 56.

SCHMIDHUBER, J. Deep Learning in Neural Networks: An Overview. *Neural Networks*, Elsevier, v. 61, p. 85–117, 2015. Citado na página 32.

SHIFFLER, R. E. Maximum Z Scores and Outliers. *The American Statistician*, Taylor & Francis Group, v. 42, n. 1, p. 79–80, 1988. Citado na página 31.

STEHMAN, S. V. Selecting and Interpreting mMeasures of Thematic Classification Accuracy. *Remote Sensing of Environment*, Elsevier, v. 62, n. 1, p. 77–89, 1997. Citado na página 52.

SUNDARARAJAN, A. et al. Survey on Synchrophasor Data Quality and Cybersecurity Challenges, and Evaluation of Their Interdependencies. *Journal of Modern Power Systems and Clean Energy*, SGEPRI, v. 7, n. 3, p. 449–467, 2019. Citado na página 15.

TANG, J. et al. Effects of PMU's Uncertainty on Voltage Stability Assessment in Power Systems. In: IEEE. *2011 IEEE International Instrumentation and Measurement Technology Conference*. [S.l.], 2011. p. 1–5. Citado na página 27.

TIANZHEN, W. et al. The Normalization PCA Model and Its Application in Fault Detection of Wind Power Generation System. In: IEEE. *2013 IEEE International Symposium on Industrial Electronics*. [S.l.], 2013. p. 1–6. Citado na página 33.

WU, M.; XIE, L. Online Identification of Bad Synchrophasor Measurements Via Spatio-temporal Correlations. In: IEEE. *2016 Power Systems Computation Conference (PSCC)*. [S.l.], 2016. p. 1–7. Citado na página 33.

XIA, Y. et al. Learning Discriminative Reconstructions For Unsupervised Outlier Removal. In: *Proceedings of the IEEE International Conference on Computer Vision*. [S.l.: s.n.], 2015. p. 1511–1519. Citado na página 32.

YANG, Z. et al. Bad Data Detection Algorithm For PMU based on Spectral Clustering. *Journal of Modern Power Systems and clean energy*, SGEPRI, v. 8, n. 3, p. 473–483, 2020. Citado 3 vezes nas páginas 34, 35 e 36.

YU, W. et al. Timestamp Shift Detection For Synchrophasor Data Based on Similarity Analysis Between Relative Phase Angle and Frequency. *IEEE Transactions on Power Delivery*, IEEE, v. 35, n. 3, p. 1588–1591, 2019. Citado na página 26.

ZHAO, H. et al. Small Magnitude PMU Bad Data Detection Based on Data Mining Technology. In: IEEE. *2020 IEEE Sustainable Power and Energy Conference (iSPEC)*. [S.l.], 2020. p. 2526–2532. Citado na página 37.

ZHAO, J. et al. Impact of Measurement Errors on Synchrophasor Applications. In: IEEE. *2017 IEEE Power & Energy Society General Meeting*. [S.l.], 2017. p. 1–5. Citado na página 27.

ZHOU, M. et al. Ensemble-based Algorithm for Synchrophasor Data Anomaly Detection. *IEEE Transactions on Smart Grid*, IEEE, v. 10, n. 3, p. 2979–2988, 2018. Citado 3 vezes nas páginas 15, 29 e 30.