

Detecção de *Outliers* na Medição Fasorial baseado em Tipicidade e Excentricidade

Diego Henrique da Silva Cavalcanti*; Juan Moises Mauricio Villanueva*; Euler Cássio Tavares de Macêdo*; Helon David de Macêdo Braz*;

**Departamento de Engenharia Elétrica da Universidade Federal da Paraíba, João Pessoa, PB, Brasil
diego.cavalcanti@estudante.cear.ufpb.br, jmauricio@cear.ufpb.br,
euler@cear.ufpb.br, helon@cear.ufpb.br*

Abstract: 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, lack of synchronization or lack of equipment. This will raise several levels of data quality problems, being able to directly affect applications that are based on data quality. 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 based on the eccentricity and typicality of the sample in relation to its set of previous samples, or TEDA, was proposed, where its main objective was to define the most sensitive method as to the variations in media and variation over time, maximizing eccentricity and favoring the detection of more difficult outliers. To evaluate two methods, phasor frequency measurement data from a single PMU are used.

Resumo: 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 falha de equipamento. Isso levará a vários níveis de problemas de qualidade de dados, podendo afetar diretamente as aplicações que se baseiam nesses dados. 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, onde seu principal objetivo foi deixar o método mais sensível as variações de média e variância ao longo do tempo, maximizando a excentricidade e favorecendo a detecção de *outliers* mais difíceis. Para avaliação dos métodos, foram utilizados dados de medição fasorial de frequência uma única PMU.

Keywords: TEDA; Outliers; Phasor Measurement Units; Typicality; Eccentricity.

Palavras-chaves: TEDA; Anomalia; PMU; Tipicidade; Excentricidade.

1. INTRODUÇÃO

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, e instalação de novos equipamentos e 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 Unidades de Medição Fasorial (PMU, Phasor Measurement Unit) 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 et al., 2016).

Com a crescente expansão na utilização de PMU para operação do Sistema Elétrico, é relevante desenvolver estudos que abordem soluções para problemas que possam afetar o desempenho desse tipo de equipamento. Entretanto, este equipamento é susceptível a algumas fontes de erro, principalmente interferências estruturais ou de comunicação que prejudicam a qualidade do dado coletado. 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 utilizada em conjunto com outros equipamentos de monitoramento da rede e é responsável por transmitir informações de tensão, corrente e frequência. Por serem variáveis interdependentes, o efeito de *outliers* em qualquer uma delas compromete a confiabilidade da informação recebida.

O objetivo deste trabalho consiste em utilizar de técnicas orientadas a fluxos de dados para realizar a detecção de *outliers* em dados de medições sincronizadas obtidas por meio de uma PMU. Para este propósito, foi desenvolvido uma técnica baseada nos conceitos de tipicidade e excentricidade de uma amostra em relação ao seu conjunto de dados, denominada de TEDA (Typicality and Eccentricity Data Analytics) (Angelov, 2014).

Na seguinte seção será apresentada a fundamentação teórica para o entendimento do problema, na seção 3, será apresentada a metodologia proposta para detecção de *outliers*, e na seção 4, serão apresentados os resultados. Finalmente, serão discutidas as conclusões finais.

2. FUNDAMENTAÇÃO TEÓRICA

Neste capítulo, exploraremos conceitos teóricos essenciais para a compreensão deste estudo. Iniciaremos pela descrição dos principais tipos de anomalias presentes nos sinais de PMU. Logo em seguida, será apresentado o algoritmo TEDA, na sua versão clássica.

2.1 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. Segundo Hill and 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.

a. 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 ilustrado na Figura 1, algumas medidas apresentam picos aleatórios no aspecto de amplitude e intervalo de tempo.

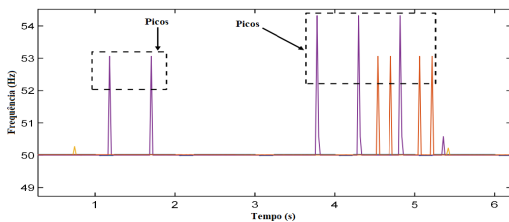


Figura 1. Exemplo de alguns picos aleatórios nos dados da frequência.

b. 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 mostrado na Figura 2, dados ausentes coletados de uma rede elétrica podem ocorrer com frequência e resultar em uma queda repentina para 0 Hz na medição de frequência.

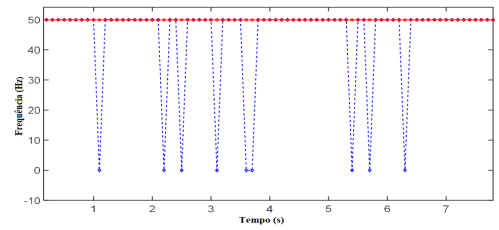


Figura 2. Exemplo de alguns zeros nos dados da frequência.

c. Padrão errôneo

Deng et al. (2020) define um tipo de anomalia denominada de Padrão Errôneo. Esse tipo de *outlier* ocorre 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 séries temporais. Por outro lado, esse tipo de *outlier* pode auxiliar a detecção de perturbações, porque os valores de frequência de várias PMUs divergem entre si quando ocorre o desligamento da máquina.

d. 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. Assim, a aleatoriedade e a variedade do ruído de alta frequência dificultam sua remoção com um filtro de limite uniforme.

2.2 Typicality and Eccentricity Data Analytics (TEDA)

O 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. 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, necessário um conjunto com apenas três amostras de dados, independência das amostras de dados individuais e é inteiramente baseado nos dados e sua distribuição mútua no espaço de dados;

O TEDA 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). A proximidade acumulada representa o quão similar é aquela amostra para o conjunto de dados, 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, e por fim, a excentricidade é o oposto da tipicidade, e representa o quão excêntrica é aquela amostra com relação ao conjunto.

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 (1) e (2).

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

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

em que: 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 (1) e (2) para o cálculo da excentricidade e tipicidade por (3) e (4), respectivamente:

$$\xi_k(x) = \frac{1}{k} + \frac{(u_k - x_k)(u_k - x_k)}{k\sigma_k^2} \quad (3)$$

$$\tau_k(x) = 1 - \xi_k(x) \quad (4)$$

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 et al. (1984). A desigualdade de Chebyshev é um princípio 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. Assim, a partir da excentricidade normalizada podemos definir um limiar para detecção de *outlier*:

$$\zeta_k(x) > \frac{m^2 + 1}{2k}, m > 0 \quad (5)$$

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.

3. METODOLOGIA PROPOSTA

Nesta seção, inicialmente é apresentado o conjunto de dados da PMU que será utilizado ao longo desse trabalho. Em seguida, é apresentado o algoritmo proposto para detecção de *outliers*, o TEDA Janelado.

3.1 Conjunto de Dados de PMU

Os dados de PMU utilizados nesse trabalho foram medidas da fase A no terminal de uma linha de 500 kV, localizado na região Nordeste do Brasil, no dia 05 de maio de 2023. Ao todo foram extraídos inicialmente três minutos de dados coletados pelas PMUs, com uma amostragem de 60 medições por segundo, o que corresponde a um volume de 10 mil amostras de frequência.

3.2 Algoritmos de detecção: TEDA Janelado

Para facilitar o entendimento, será utilizado um recorte com 1000 amostras do conjunto de dados de frequência elétrica, apresentado na Figura 3(a). Foram inseridos 10 de *outliers* do tipo vales e picos, conforme apresentado na Figura 3(b).

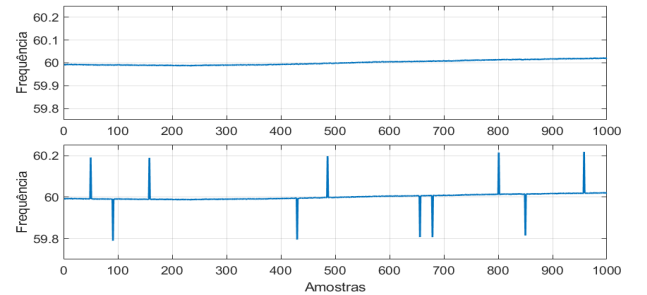


Figura 3. Recorte dos dados de frequência: a) sem *outliers*, b) com *outliers*.

O algoritmo proposto caracteriza-se por uma estrutura parecida com a do algoritmo TEDA Clássico. 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 análise é feita, conforme apresentado na Figura 4.

Nesta proposta, é inserido um novo parâmetro que determina o tamanho da janela de processamento de dados (Tam). Desse modo, considerando a quantidade de

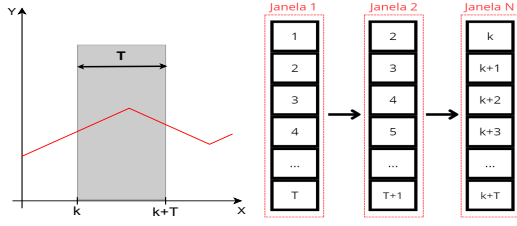


Figura 4. Modelo da janela deslizante adotado.

amostras (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 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 por (5) e (6), 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 em (7).

$$\mu_i(x) = \frac{1}{Tam} \sum_{k=Tam-i}^{Tam} x_k \quad (6)$$

$$\sigma_i^2(x) = \frac{1}{Tam} \sum_{k=Tam-i}^{Tam} (x_k - \mu_i(x))^2 \quad (7)$$

$$\xi_i(x) = \frac{1}{Tam} + \frac{(u_i - x_i)^T(u_i - x_i)}{Tam \cdot \sigma_i^2} \quad (8)$$

A Figura 5 apresenta uma análise da abordagem proposta para o conjunto de dados proposto no início dessa seção. A curva em azul representa os dados com *outliers* e a curva laranja sem *outliers*. Quando os *outliers* vão se apresentando, a média e a variância vão sendo sensibilizadas como ilustrado na Figura 5(b) e Figura 5(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 mantêm valores altos valores de excentricidade, Figura 5(d).

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 em (9).

$$\begin{cases} i \leq Tam, & \zeta_i(x) > \frac{m^2+1}{2 \cdot i} \\ i > Tam, & \zeta_i(x) > \frac{m^2+1}{2 \cdot Tam} \end{cases} \quad (9)$$

Por fim, é apresentado na Figura 6 o limiar com base em (9), considerando uma janela de 100 amostras ($Tam = 100$) 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 positivos, porém

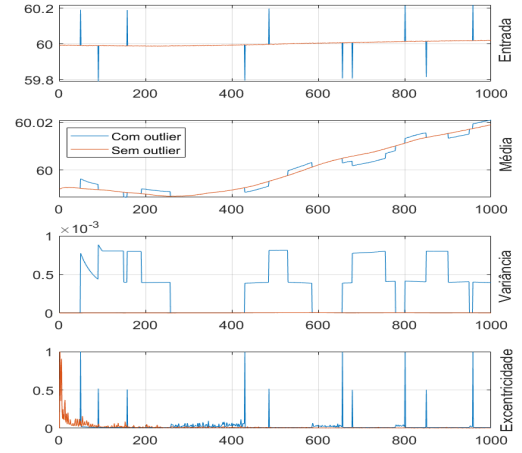


Figura 5. a) Conjunto de dados de entrada de frequência, b) Média, c) Variância, d) Excentricidade.

para o TEDA Janelado, valores de $m \geq 3$ 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*.

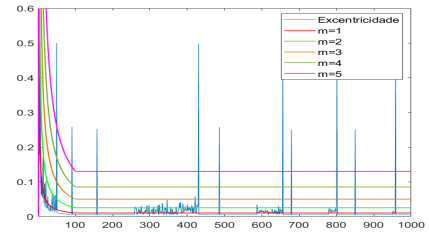


Figura 6. Excentricidade x Limiar de detecção.

4. RESULTADOS

Neste capítulo serão apresentados os resultados e as discussões para os métodos de detecção de *outliers* proposto neste trabalho, considerando medições fásoriais reais. Para extração dos resultados, foram analisados os dois métodos considerando quatro cenários contendo diferentes tipos de *outliers* nos dados. Também foi considerado variações dos valores do limiar de detecção (m) e do tamanho da janela (Tam). Os resultados são apresentados por meio da matriz de confusão e o coeficiente de correlação de Matthews (MCC), utilizada para medir a qualidade de modelos de classificação binária.

4.1 Dados de Frequência

Nesta seção será realizada a comparação dos algoritmos entre TEDA Clássico e o TEDA Janelado em relação ao conjunto de dados de frequência. Para este conjunto de dados utilizados, foram propostos quatro cenários com diferentes tipos de *outliers*, como se ilustra na Figura 7, tais como:

- Cenário 01: 100 *outliers* do tipo pico positivo na curva de frequência (1% do conjunto de dados), considerando o incremento de 0,2 Hz;
- Cenário 02: 100 *outliers* do tipo zeros na curva de frequência;
- Cenário 03: 300 *outliers* do tipo pico positivo e pico negativo na curva de frequência (3% do conjunto de dados), considerando o incremento de $\pm 0,2$ Hz;

- Cenário 04: 125 *outliers*, sendo 100 do tipo pico positivo e pico negativo e 25 do tipo zero. Para esse caso, também foi considerado um incremento de $\pm 0,2$ Hz.

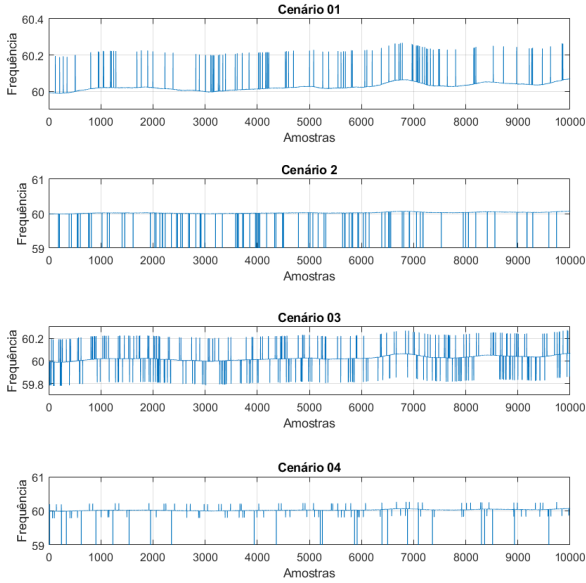


Figura 7. Dados de frequência após inserção de *outliers*.

a. Cenário 01

No primeiro cenário de teste, foi analisado a injeção de *outliers* aleatórios na curva de frequência do tipo pico positivo, totalizando 1% do dados. Na Tabela 1 é apresentada uma síntese da matriz de confusão obtida para os dois métodos com configurações distintas.

Tabela 1. Matriz de confusão para o cenário 1.

	m	TAM	Verdadeiro Positivo	Falso Positivo	Verdadeiro Negativo	Falso Negativo
TEDA	3	-	100	0	9900	0
Clássico	4	-	100	0	9900	0
TEDA JANELADO	3	300	100	1	9899	0
	4	300	100	0	9900	0
	3	500	100	0	9900	0
	4	500	100	0	9900	0
	3	1000	100	0	9900	0
	4	1000	100	0	9900	0

Tabela 2. Análise do cenário 1.

	m	TAM	Precisão	Recall	Especif.	Taxa FP	MCC
TEDA	3	-	1	1	1	0	1
Clássico	4	-	1	1	1	0	1
TEDA JANELADO	3	300	0,99	1	0,9998	0,001	0,995
	4	300	1	1	1	0	1
	3	500	1	1	1	0	1
	4	500	1	1	1	0	1
	3	1000	1	1	1	0	1
	4	1000	1	1	1	0	1

Perceba que resultados de ambos os algoritmos, TEDA Clássico e TEDA Janelado, tiveram um desempenho excelente, identificando todos os *outliers*. Porém o TEDA Janelado, quando Tam = 300 e m = 3 houve uma sinalização de falso positivo, que pode ser explicado pelo baixo limiar de detecção e pelo tamanho reduzido da janela, que torna o método ainda mais sensível.

Uma interpretação dos resultados da Tabela 1 é apresentada na Tabela 2. Apenas a configuração que sinalizou

um falso positivo sofreu uma redução do seu desempenho. Logo, para o Cenário 01 ambos os métodos podem ser considerados adequados para detecção desse tipo de *outliers*.

b. Cenário 02

No segundo cenário foram inseridos de maneira aleatória 100 zeros por todo o conjunto de dados. O desempenho dos dois métodos são apresentados por meio da sua matriz de confusão na Tabela 3.

Tabela 3. Matriz de confusão para o cenário 2.

	m	TAM	Verdadeiro Positivo	Falso Positivo	Verdadeiro Negativo	Falso Negativo
TEDA	3	-	100	0	9900	0
Clássico	4	-	100	0	9900	0
TEDA JANELADO	3	300	100	2	9898	0
	4	300	100	0	9900	0
	3	500	100	0	9900	0
	4	500	100	0	9900	0
	3	1000	100	0	9900	0
	4	1000	100	0	9900	0

De maneira similar ao Cenário 01 todos os métodos tiveram um bom desempenho, exceto o TEDA Janelado, que em uma das execuções sinalizou dois falsos positivos. Mais uma vez a presença do falso positivo pode ser justificada pelo limiar de detecção baixo e o número da janela menor, o que torna o algoritmo mais sensível.

Na Tabela 4 é apresentado uma análise dos desempenhos para cada caso estudado. Sendo apenas o método TEDA Janelado a apresentar um resultado diferente do ideal devido a presença de falsos positivos, porém ainda assim, obteve uma precisão de 99,98%. Já os demais métodos, obtiveram uma precisão de 100%.

Tabela 4. Análise do cenário 2.

	m	TAM	Precisão	Recall	Especif.	Taxa FP	MCC
TEDA	3	-	1	1	1	0	1
Clássico	4	-	1	1	1	0	1
TEDA JANELADO	3	300	0,9998	1	0,9998	0,0002	0,99
	4	300	1	1	1	0	1
	3	500	1	1	1	0	1
	4	500	1	1	1	0	1
	3	1000	1	1	1	0	1
	4	1000	1	1	1	0	1

c. Cenário 03

No Cenário 03 foram acrescentado aleatoriamente *outliers* em 300 amostras do tipo pico positivo e negativo. Este cenário apresenta um número três vezes a mais de anomalias, quando comparado com os anteriores. A Tabela 5 apresenta um extrato da matriz de confusão para algumas variações dos parâmetros e métodos.

Tabela 5. Matriz de confusão para o cenário 3.

	m	TAM	Verdadeiro Positivo	Falso Positivo	Verdadeiro Negativo	Falso Negativo
TEDA	3	-	300	0	9700	0
Clássico	4	-	293	0	9700	7
TEDA JANELADO	3	300	300	0	9700	0
	4	300	288	0	9700	12
	3	500	300	0	9700	0
	4	500	293	0	9700	7
	3	1000	300	0	9700	0
	4	1000	293	0	9700	7

Note que para esse conjunto de dados, os métodos que utilizaram o parâmetro m = 4 apresentaram a presença de falsos negativos, enquanto para m = 3 tiveram uma melhor

assertividade. De fato, o valor de m está diretamente relacionado ao limiar de detecção, ocasionando uma maior incidência de falso negativo.

Uma análise do desempenho é apresentado na Tabela 6. Note que a sensibilidade (recall) dos algoritmos que utilizaram o parâmetro $m=4$ foram inferiores ao que utilizaram $m=3$, 97%, 66% e 100% respectivamente. Nesse cenário podemos concluir que todos os métodos possuem um bom desempenho, desde que utilizando o parâmetro $m=3$.

Tabela 6. Análise do cenário 3.

	m	TAM	Precisão	Recall	Especif.	Taxa FP	MCC
TEDA	3	-	1	1	1	0	1
Clássico	4	-	0,993	0,9766	1	0	0,9879
	3	300	1	1	1	1	1
	4	300	0,998	0,96	1	0	0,9792
TEDA	3	500	1	1	1	0	1
JANELADO	4	500	0,9993	0,9766	1	0	0,9879
	3	1000	1	1	1	0	1
	4	1000	0,9993	0,9766	1	0	0,9879

d. Cenário 04

Por fim, no Cenário 04 foram acrescentados 100 *outliers* do tipo pico positivo e negativo e 25 zeros ao conjunto de dados, com o objetivo de verificar o desempenho dos algoritmos diante de variações de *outliers*. Um resumo da matriz de confusão é apresentado na Tabela 7 para as diferente configurações analisadas.

Tabela 7. Matriz de confusão para o cenário 4.

	m	TAM	Verdadeiro Positivo	Falso Positivo	Verdadeiro Negativo	Falso Negativo
TEDA	3	-	26	0	9875	99
Clássico	4	-	25	0	9875	100
	3	300	73	2	9873	52
	4	300	72	0	9875	53
TEDA	3	500	57	0	9875	68
JANELADO	4	500	56	0	9875	69
	3	1000	34	0	9875	91
	4	1000	33	0	9875	92

O TEDA Clássico teve uma maior dificuldade para detecção dos *outliers* do tipo pico positivo e negativo, provavelmente devido à forte influência das amostras do tipo zero na excentricidade dos dados. Por sua vez o TEDA Janelado teve um desempenho melhor, uma vez que as amostras ruins não afetam a excentricidade ao longo do tempo, afetando apenas durante a janela, uma vez que, a abordagem observar apenas um subconjunto de amostras. Os valores do tamanho da janela (Tam) e o limiar de detecção (m) quando calibrados corretamente, favorecem para uma maior assertividade do método.

O desempenho dos métodos são apresentados na Tabela 8. Apesar do alto valor de precisão de todas as configurações (99%), quando analisando a sensibilidade, percebemos que a melhor configuração apresentou uma sensibilidade igual a 58,4%. Se considerado o MCC, que é a uma correlação entre o recall e a especificidade do método, o TEDA Janelado apresenta um melhor valor de MCC igual a 0,7518, contra um valor de MCC de 0,4538 para o TEDA Clássico. Por fim, para esse cenário o TEDA Janelado se mostrou superior em relação ao TEDA Clássico, principalmente quando utiliza janelas menores.

5. CONCLUSÕES

Neste trabalho foi apresentado uma metodologia para detecção de *outliers* em dados de medição fasorial baseado

Tabela 8. Análise do cenário 4.

	m	TAM	Precisão	Recall	Especif.	Taxa FP	MCC
TEDA	3	-	0,9901	0,208	1	0	0,4538
Clássico	4	-	0,993	0,9766	1	0	0,445
	3	300	0,9946	0,584	0,998	0,0002	0,7518
	4	300	0,9947	0,576	1	0	0,7569
TEDA	3	500	0,9932	0,456	1	0	0,673
JANELADO	4	500	0,9931	0,448	1	0	0,667
	3	1000	0,9909	0,272	1	0	0,5191
	4	1000	0,9908	0,264	1	0	0,5114

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 um modelo baseado no TEDA Clássico, onde a nova abordagem observaria apenas um subconjunto de amostras. Essa nova metodologia foi denominada como TEDA Janelado, uma vez que, ao contrário do TEDA Clássico, cada nova amostra é calculado a sua excentricidade e tipicidade com base apenas em uma janela de dados, ocasionando uma menor memória dos dados em relação ao TEDA Clássico.

Para trabalhos futuros, com a finalidade de avaliar a confiabilidade das metodologias propostas neste trabalho para a detecção de *outlier*: otimização dos parâmetros m e Tam; verificar a aplicação de um limiar adaptativo em relação a variância para os métodos; e, avaliar o desempenho dos métodos para os eventos elétricos;

REFERÊNCIAS

- Angelov, P. (2014). Outside the box: an alternative data analytics framework. *Journal of Automation Mobile Robotics and Intelligent Systems*, 8(2), 29–35.
- Brahma, S., Kavasseri, R., Cao, H., Chaudhuri, N., Alexopoulos, T., and Cui, Y. (2016). Real-time identification of dynamic events in power systems using pmu data, and potential applications models, promises, and challenges. *IEEE transactions on Power Delivery*, 32(1), 294–301.
- Deng, X., Bian, D., Wang, W., Jiang, Z., Yao, W., Qiu, W., Tong, N., Shi, D., and Liu, Y. (2020). Deep learning model to detect various synchrophasor data anomalies. *IET Generation, Transmission & Distribution*, 14(24), 5739–5745.
- Hill, D.J. and Minsker, B.S. (2010). Anomaly detection in streaming environmental sensor data: A data-driven modeling approach. *Environmental Modelling & Software*, 25(9), 1014–1022.
- Mahapatra, K., Chaudhuri, N.R., and Kavasseri, R. (2016). Bad data detection in pmu measurements using principal component analysis. In *2016 North American Power Symposium (NAPS)*, 1–6. IEEE.
- Saw, J.G., Yang, M.C., and Mo, T.C. (1984). Chebyshev inequality with estimated mean and variance. *The American Statistician*, 38(2), 130–132.
- Sundararajan, A., Khan, T., Moghadas, A., and Sarwat, A.I. (2019). Survey on synchrophasor data quality and cybersecurity challenges, and evaluation of their interdependencies. *Journal of Modern Power Systems and Clean Energy*, 7(3), 449–467.