

Análise de Componentes Principais para diagnóstico pós-falta em sistemas elétricos de grande porte

Victor F. Müller Jr.^{1,2}, Alexandre R. Aoki²
Ricardo Schumacher²

¹COPEL Geração e Transmissão, Curitiba, Paraná, Brasil
(e-mail: victor.muller@copel.com).

²Universidade Federal do Paraná, Curitiba, Paraná, Brasil
(e-mail: aoki@ufpr.br, ricardo.schumacher@ufpr.br)

Abstract: Synchronized phasor measurement systems are being widely used around the world, and have become essential elements in the evolution of the operation of large electrical systems. These systems are based on phasor measurement units, called PMUs, which are capable of recording data with a high sampling rate, thus generating a huge dataset. This work presents a methodology for selecting data for post-fault analysis, through dimensionality reduction, using principal component analysis. To validate the proposed methodology, real data related to a recent occurrence in the national interconnected system were used. The validation of the obtained results was done using an anomaly detection algorithm. With the application of the methodology, the possibility of using a data set smaller than the original was proved, while maintaining the characteristics of the attributes. This automatic data selection methodology benefits post-fault analysis due to the performance gain obtained by reducing the electrical attribute dataset, without significant loss in the accuracy of anomaly detection capability.

Resumo: Os sistemas de medição fasorial sincronizada estão sendo amplamente utilizados ao redor do mundo, e se tornaram elementos essenciais na evolução da operação dos sistemas elétricos de grande porte. Estes sistemas são baseados em unidades de medição fasorial, chamadas PMUs, as quais são capazes de gravar dados com alta taxa de amostragem, gerando assim um enorme conjunto de dados. Neste trabalho é apresentada uma metodologia para seleção de dados para análise pós-falta, através da redução da dimensionalidade, utilizando a análise de componentes principais. Para validar a metodologia proposta, foram usados dados reais relacionados a um recente caso de ocorrência no sistema interligado nacional. A validação dos resultados obtidos foi feita utilizando-se um algoritmo de detecção de anomalias. Com a aplicação da metodologia, comprovou-se a possibilidade de se utilizar um conjunto de dados menor do que o original, porém mantendo as características dos atributos. Esta metodologia automática de seleção de dados traz benefícios para análise pós-falta devido ao ganho de desempenho obtido com a redução do conjunto de dados dos atributos elétricos, sem perda significativa na precisão da capacidade de detecção de anomalia.

Keywords: Principal component analysis; Multivariate statistics; Wide area monitoring; Phasor measurement unit; Fault detection.

Palavras-chaves: Análise de componentes principais; Estatística multivariada; Monitoramento de área ampla; Unidade de medição fasorial; Detecção de faltas.

1. INTRODUÇÃO

Os sistemas de medição fasorial sincronizada estão sendo amplamente utilizados ao redor do mundo, com um crescimento exponencial nos últimos anos. Nos Estados Unidos desde a década passada já há uma rede continental com mais de 2.000 unidades de medição fasorial (PMUs) ajudando a melhorar a confiabilidade da rede de energia elétrica da América do Norte (OVERHOLT et al., 2015). Da mesma forma, também na China na década passada já havia uma rede composta por mais de 2.400 PMUs cobrindo todas as subestações de 500 kV no país e uma série de importantes

usinas de energia e subestações de 220/110 kV (LU et al., 2015). Na Europa continental atualmente existe um sistema de medição fasorial abrangendo quase toda a área territorial dos diversos países, com o qual apesar dos desafios de trocas de dados das PMUs entre os diferentes operadores dos sistemas de transmissão nacionais interconectados, as aplicações com os dados intercambiados permitem que os operadores gerenciem e operem todo o sistema de forma segura e eficiente (SATTINGER; GIANNUZZI, 2015).

Os sistemas de medição fasorial sincronizada (SMFS) são fundamentais para a operação do sistema elétrico de grande porte, que consiste na programação, coordenação, supervisão,

controle e análise, com a finalidade de garantir seu funcionamento de forma otimizada, confiável, segura e com qualidade (ONS, 2020).

Porém, o sistema elétrico não é imune a falhas, e quando elas acontecem devem ser analisadas detalhadamente para que o problema que a causou não se repita e nem se propague. A etapa de análise após uma falha corresponde a uma pesquisa bastante profunda sobre as causas da mesma. É uma tarefa bastante trabalhosa e que exige grande precisão, onde são utilizados todos os tipos de dados disponíveis (CARDOSO, ROLIM e ZÜRN, 2004). Com o crescimento exponencial dos dados de monitoramento, o diagnóstico de faltas enfrenta enormes desafios ao lidar com os grandes volumes de dados (XU et al., 2017).

As unidades de medição fasorial são instaladas em subestações do sistema de potência. Na maioria das aplicações, os dados fasoriais são usados em locais remotos das PMUs. Portanto, uma infraestrutura com arquitetura envolvendo PMUs, links de comunicação e concentradores de dados deve existir para obter todos os benefícios do sistema de medição de uma PMU (PHADKE; THORP, 2017).

A aplicação mais comum baseada em PMUs é o WAMS (*Wide Area Measurement System*), ou sistema de medição em área ampla. Sua finalidade é melhorar a estimação de estado, a análise pós-falha e as informações do agente operador. Nas aplicações de WAMS, várias PMUs são conectadas a um concentrador de dados fasoriais (PDC), com um armazenamento em massa acessível a partir da central de controle (KARLSSON et al., 2004).

Os dados da PMU se originam no ponto de medição e passam por um processo e caminho de comunicação até o seu ponto de uso. Problemas que surgem ao longo deste caminho podem afetar o desempenho de várias aplicações que fazem uso dos dados de fasores sincronizados e, portanto, comprometem a capacidade dessas aplicações de atender às necessidades e objetivos dos usuários de forma eficaz (NASPI, 2017).

A definição de quais são os atributos de dados mais importantes depende do problema de interesse a ser resolvido e da aplicação projetada para resolvê-lo. Na maioria dos casos, o analista deseja realizar uma análise específica, e para isto identifica os dados necessários para realizá-la e determina se os dados disponíveis são adequados para esse fim. Em alguns casos, o analista recebe um novo conjunto de dados e busca quais conhecimentos podem ser adquiridos, considerando a qualidade dos dados disponíveis (NASPI, 2017).

A técnica de análise de componentes principais foi inicialmente descrita por Karl Pearson (1901). Ele propôs um método prático de cálculo para duas ou três variáveis apenas. Uma descrição de métodos computacionais práticos veio muito mais tarde, feita por Hotelling (1933). A ideia central da análise baseia-se na redução do conjunto de dados a ser analisado, principalmente quando os dados são constituídos de um grande número de variáveis inter-relacionadas (MANLY; ALBERTO, 2019).

XIE et al. (2014) apresentaram um estudo que explora a dimensionalidade implícita nos dados de PMU, propondo um algoritmo de detecção de eventos, baseado em PCA (Análise de Componentes Principais). Neste estudo a redução de

dimensionalidade é realizada levando em consideração apenas os dados de frequência e tensão separadamente. Liu et al. (2015), apresentaram a implementação de um método de detecção de ilhamento baseado em PCA, usando os dados de frequência e tensão, cuja proposta envolve duas etapas: 1) modelagem de PCA offline usando dados históricos para obter os componentes principais e limites de controle e 2) monitoramento online para determinar se um evento de ilhamento ocorre. Cai et al. (2018) propuseram a integração do método kNN (vizinho mais próximo) com o método de monitoramento estatístico baseado em PCA para que um grande número de variáveis possa ser analisado em tempo real para monitoramento de área ampla. Dimitrovska et al. (2019) apresentaram uma proposta de um método aplicável em tempo real para triagem de contingências, através do reconhecimento de condições recorrentes e perigosas, da operação do sistema de potência. O método proposto utiliza o PCA para reconhecimento de padrões em uma base de dados, com amostras anteriores à falta. Chatterjee; Chaudhuri (2019) apresentaram uma proposta de adaptação do método PCA para tratar dados de baixa qualidade obtidos em PMUs, que podem estar induzindo à interpretação de que se trata de um evento altamente não linear, prejudicando a confiabilidade de detecção de eventos. Chu et al. (2017) apresentaram um método orientado a dados em tempo real para indicar a avaliação do estado de dados de PMU. Zhang et al. (2017) propuseram um modelo de análise de dados de conjunto para avaliação de estabilidade pré-contingência baseada em PMU, considerando medições de dados incompletos. Le; Benjapolakul (2018) demonstraram preocupação com a dependência que o uso de dados de PMU tem da rede de comunicação. De acordo com Ahmad e Senroy (2020), a qualidade das medições afeta muito o monitoramento de segurança do sistema de energia e outros aplicativos de análise de dados de missão crítica que dependem de dados de PMU.

A principal contribuição deste trabalho visa a análise sistêmica pós-falta em sistemas elétricos de grande porte, através da coleta de dados de um sistema de medição fasorial sincronizado e do desenvolvimento de uma metodologia de seleção de dados relevantes baseada em Análise de Componentes Principais.

O restante deste artigo está estruturado da seguinte forma. Na seção 2 é apresentado o sistema de medição fasorial sincronizado utilizado e é descrita a estrutura dos respectivos dados e suas características principais. Na seção 3 o método de execução é exposto através de um fluxo e descrito. Na seção 4 os resultados obtidos pela execução do método de seleção de dados são apresentados e analisados. O caso real de falta que forneceu os dados utilizados para realização das simulações é descrito, bem como os cenários de simulações avaliados são apresentados, assim como os resultados obtidos para todas as simulações. Além disso, é realizada a análise destes resultados e apresentada uma discussão baseada nos resultados obtidos. Por fim, a seção 5 apresenta as conclusões do trabalho desenvolvido.

2. ESTRUTURA DOS DADOS DAS PMUs

Os dados utilizados foram obtidos do sistema de medição fasorial sincronizado, da empresa COPEL Geração e

Transmissão S.A. A seleção das PMUs que estão sendo analisadas levou em consideração a existência de lacunas na coleta de dados de qualquer grandeza, sendo privilegiado o uso de PMUs que apresentam o conjunto completo de dados.

A arquitetura do SMFS utilizado para o fornecimento de dados é composta por diversas PMUs, instaladas em cada uma das subestações da COPEL Geração e Transmissão. Estas PMUs se comunicam pela rede com dois servidores, um principal e um de backup, os quais concentram os dados de todas as PMUs. A partir deste servidor, é possível acessar os dados de todas as PMUs. Atualmente o sistema de Medição Fasorial Sincronizada da COPEL Geração e Transmissão é composto de 56 PMUs.

Cada PMU gera o seu respectivo conjunto de dados, contendo as grandezas relacionadas às medições do local onde está instalada. As medições que são coletadas são relativas às grandezas de tensão, frequência e taxa de variação da frequência (df/dt) do barramento e das correntes em todos os pontos de conexão com outras linhas de transmissão. Além disso, para a tensão e corrente é feita a medição da magnitude e do ângulo para cada uma das três fases.

Tabela 1 – Relação de atributos de uma PMU

Item	Atributo	Unidade
1	Estampa de Tempo	Data e Hora
2	Status	Texto
3	Frequência	Hertz
4	df/dt – taxa de variação da frequência	
5	Tensão Barra magnitude Fase A	Volt
6	Tensão Barra ângulo Fase A	Graus
7	Tensão Barra magnitude Fase B	Volt
8	Tensão Barra ângulo Fase B	Graus
9	Tensão Barra magnitude Fase C	Volt
10	Tensão Barra ângulo Fase C	Graus
11	Corrente Linha magnitude Fase A	Ampere
12	Corrente Linha ângulo Fase A	Graus
13	Corrente Linha magnitude Fase B	Ampere
14	Corrente Linha ângulo Fase B	Graus
15	Corrente Linha magnitude Fase C	Ampere
16	Corrente Linha ângulo Fase C	Graus

A Tabela 1 contém o padrão da lista de atributos que fazem parte do conjunto de dados de uma PMU. Os atributos relacionados com a corrente (11 a 16) se repetem para cada uma das linhas que estiverem conectadas no respectivo barramento.

3. METODOLOGIA

Para o desenvolvimento deste trabalho foi utilizado o software MATLAB, da MathWorks, Inc. Além das bibliotecas padrão da ferramenta, foi utilizado também a *toolbox* “Statistics and Machine Learning”, que disponibiliza funções baseadas em análise de componentes principais.

A Fig. 1 mostra a representação do fluxo de trabalho executado para a metodologia de análise dos dados das PMUs.

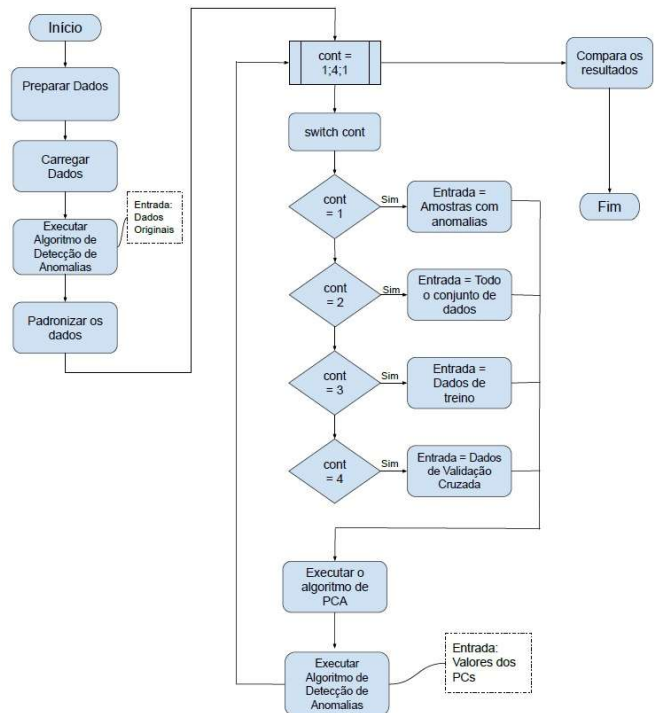


Fig 1 – Fluxograma da Metodologia

A seguir, é apresentado um detalhamento das etapas do processo representado na Fig. 1.

Preparação e carga dos dados: Este procedimento consiste basicamente na leitura dos arquivos de dados das PMUs e sua respectiva preparação. O procedimento permite a leitura de mais de um arquivo, possibilitando assim fazer a composição através da concatenação dos dados de várias PMUs. No caso de concatenação de vários arquivos, a premissa é de que todos referem-se ao mesmo intervalo de tempo. Assim a quantidade de registros será igual em todos eles. Os atributos são concatenados lado a lado, o que significa que, analisando sob o aspecto de uma tabela esta concatenação de dados aumenta a quantidade de colunas e mantém a quantidade de linhas original.

Após concluir a montagem do conjunto de dados, o procedimento separa o conjunto em três partes: dados de treino; dados de teste; dados de validação cruzada.

Algoritmo de detecção de anomalias: Neste processo utiliza-se um modelo gaussiano para detectar exemplos anômalos no conjunto de dados. Nesse conjunto de dados, encontra-se uma distribuição gaussiana e, em seguida, encontra-se valores que tenham probabilidade muito baixa e, portanto, podem ser considerados anomalias. Após estimarem-se os parâmetros gaussianos, pode-se investigar quais amostras têm uma probabilidade muito alta dada essa distribuição e quais amostras têm uma probabilidade muito baixa. As amostras de baixa probabilidade são mais prováveis de serem as anomalias no conjunto de dados. Neste caso, a maneira de determinar quais amostras são anomalias é selecionar um limite com base em um conjunto de validação cruzada. Para isto, utiliza-se um algoritmo para selecionar o limite ϵ usando o indicador *F1 Score* no conjunto de validação cruzada.

Padronização dos dados: Esta etapa é necessária para que as variáveis com grande escala não dominem o processo sobre as demais. No caso, os atributos como frequência, magnitude e ângulo de tensão e corrente, são representados por grandezas diferentes em suas respectivas escalas. Na análise comparativa dos atributos, é necessário que todos os atributos sejam transformados para que possam ser observados dentro do mesmo intervalo de escala. Isto é feito através da subtração da média e divisão pelo desvio padrão em cada atributo do conjunto de dados.

Algoritmo de análise dos componentes principais: Este processo executa a análise de componentes principais dos dados recebidos, efetua a seleção dos componentes principais com base na quantidade mínima e calcula e monta a lista de valores dos componentes principais (PCs).

Para execução da análise de componentes principais, propriamente dita, utiliza-se o algoritmo que retorna duas informações essenciais: a matriz de coeficientes e o vetor de explicação por componente principal.

Cada coluna da matriz de coeficientes contém os coeficientes para um componente principal e as colunas estão em ordem decrescente de variância do componente.

O vetor de explicação por componente principal contém a porcentagem da variância total explicada para cada um dos componentes principais identificados.

Uma vez que o objetivo neste processo é diminuir a dimensionalidade do conjunto de dados, é necessário definir quantos componentes principais serão utilizados. Para isto utiliza-se um critério de importância cumulativa, ou seja, seleciona-se os *n* primeiros componentes cuja soma cumulativa da porcentagem de variância excedam o limite de 80%. Este valor é obtido a partir do vetor de explicação por componente principal.

Tendo obtido a quantidade de componentes principais selecionados, é este número que define a nova dimensão do conjunto de dados que se tem como resultado, ou seja, o equivalente à quantidade de colunas da tabela. Para definir os novos valores de cada amostra do conjunto de dados, identifica-se a matriz de autovetores, que nada mais é que a matriz de coeficientes retornada pelo algoritmo de análise de componentes principais.

Para cada amostra, em cada um dos componentes principais selecionados, faz-se a soma cumulativa do autovetor, multiplicado pelo valor original. Ou seja, tem-se a seguinte equação:

$$PC_{z,j} = \sum_{i=1}^{atributos} autovetores_{ji} \cdot dados_{zi} \quad (1)$$

Com isto, este procedimento monta um novo conjunto de dados, mantendo a mesma quantidade de amostras do conjunto original, porém reduzindo a quantidade de atributos, para tantos quanto foram definidos na seleção dos componentes principais. Os valores do novo conjunto de dados são definidos através da aplicação das equações lineares de montagem de cada um dos componentes principais, que na realidade equivalem aos novos atributos deste novo conjunto de dados.

Deteção de anomalias com os componentes principais: Na sequência da montagem do novo conjunto de dados, com dimensões reduzidas, aplica-se novamente o mesmo algoritmo de deteção de anomalias. O conjunto de dados de treino, teste e validação cruzada é remontado com as PCs. Mas continua utilizando-se as mesmas proporções, e conseqüentemente a quantidade de registros que compõem cada um dos três conjuntos. Os critérios e a metodologia de deteção de anomalias continuam sendo exatamente os mesmos utilizados na etapa anterior, para que possa ser feita uma comparação fidedigna entre os resultados obtidos.

Métricas para comparação dos resultados: Na última etapa da metodologia, executa-se uma comparação dos resultados obtidos na deteção de anomalias a partir dos dados originais, com os resultados obtidos a partir dos dados com dimensionalidade reduzida. A abordagem adotada para fazer esta análise é a matriz de confusão, uma métrica que nada mais é do que uma tabela que mostra o número de previsões corretas e incorretas, sendo que neste caso, considera-se como uma previsão a deteção de anomalia baseada no conjunto de dados com dimensionalidade reduzida. A tabela tem uma estrutura como a demonstrada na Fig. 2.

		Resposta Prevista	
		yPrev = 1	yPrev = 0
Resposta Real	y = 1	Verdadeiro POSITIVO	Falso NEGATIVO
	y = 0	Falso POSITIVO	Verdadeiro NEGATIVO

Fig. 2 – Matriz de Confusão

A imagem representada na Fig. 2, tem por objetivo mostrar que em cada quadrante da tabela são calculadas as quantidades de amostras que atendem aos respectivos critérios (BRUCE; BRUCE, 2019), que são:

- verdadeiro POSITIVO: são computadas todas as amostras do conjunto de dimensionalidade reduzida que foram identificadas como anomalias e que também foram identificadas como anomalias no conjunto de dados original;
- falso POSITIVO: são computadas todas as amostras do conjunto de dimensionalidade reduzida que foram identificadas como anomalias, porém não foram identificadas como anomalias no conjunto de dados original;
- falso NEGATIVO: são computadas todas as amostras do conjunto de dimensionalidade reduzida que não foram identificadas como anomalias, porém foram identificadas como anomalias no conjunto de dados original;
- verdadeiro NEGATIVO: são computadas todas as amostras do conjunto de dimensionalidade reduzida que não foram identificadas como anomalias e que também não foram identificadas como anomalias no conjunto de dados original.

A partir dos números obtidos para preenchimento desta tabela, efetua-se o cálculo das seguintes métricas: acurácia, precisão,

revocação e *F1 Score*, cujas equações são (BRUCE; BRUCE, 2019):

$$acurácia = \frac{\sum VerdadeiroPositivo + \sum VerdadeitoNegativo}{\sum Amostras} \quad (2)$$

$$precisão = \frac{\sum VerdadeiroPositivo}{\sum VerdadeiroPositivo + \sum FalsoPositivo} \quad (3)$$

$$revocação = \frac{\sum VerdadeiroPositivo}{\sum VerdadeiroPositivo + \sum FalsoNegativo} \quad (4)$$

$$F1\ Score = 2 \cdot \frac{precisão \cdot revocação}{precisão + revocação} \quad (5)$$

4. TESTES E ANÁLISE DOS RESULTADOS

Para o desenvolvimento deste trabalho foi utilizado um caso real. Selecionou-se um evento ocorrido no dia 28/05/2021, às 11h26, em que ocorreu o desligamento automático do polo 2 do Elo CC 800 kV Xingu / Estreito (BMTE) e de sete unidades geradoras da UHE Belo Monte (Norte Energia S.A.), as quais estavam gerando no momento um montante de aproximadamente 4.050 MW.

Para o estudo do caso selecionou-se o intervalo de tempo compreendido entre 11h15 e 11h36. Considerando que os dados coletados pelas PMUs são fornecidos em uma taxa de 60 amostras por segundo, isso significa dizer que este conjunto de dados é composto por 75.600 registros de leitura. Neste intervalo de tempo o comportamento da frequência lida pelas PMUs selecionadas teve a variação conforme demonstrado na Fig. 3.

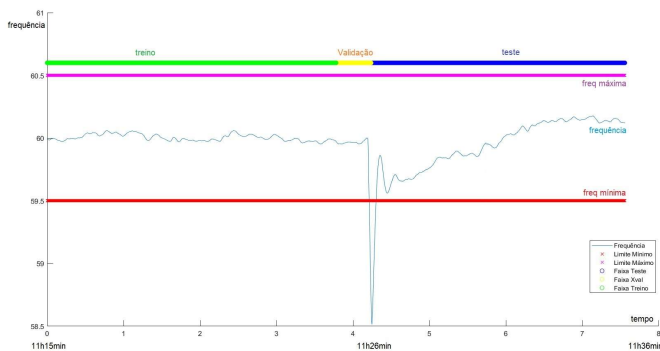


Fig. 3 – Gráfico de frequência

A Fig. 3 mostra um gráfico de variação da frequência ao longo do tempo. O eixo das ordenadas representa a variação do valor da frequência em Hz (Hertz) e o eixo das abscissas representa a variação das leituras ao longo do tempo, iniciando em 11h15 e indo até 11h36. No gráfico, seguindo uma sequência de observação das linhas de baixo para cima, tem-se a seguinte interpretação:

Linha verde, amarela e azul: representa a identificação da delimitação dos conjuntos de dados para treinamento, teste e validação cruzada. Sendo que a faixa em verde representa os dados para treinamento, a faixa em amarelo representa os dados para validação cruzada e a faixa em azul representa os dados para teste;

Linha magenta: representa o limite superior aceitável para a oscilação da frequência, 60,5 MHz, conforme submódulo 2.9 – Requisitos mínimos de qualidade de energia elétrica para acesso ou integração à Rede Básica do ONS;

Linha azul claro: representa a frequência.

Linha vermelha: representa o limite inferior aceitável para a oscilação da frequência, que é 59,5 MHz, conforme submódulo 2.9 – Requisitos mínimos de qualidade de energia elétrica para acesso ou integração à Rede Básica do ONS;

Assim, pode-se observar que no instante citado anteriormente, 11h26, é possível identificar uma variação na frequência que extrapola o limite mínimo regulamentado pelo ONS.

Para análise do caso, entre os 41 arquivos disponibilizados das PMUs, foram selecionadas oito PMUs localizadas em subestações da COPEL Geração e Transmissão, tendo como critério de seleção a qualidade dos dados. Foi feita uma verificação nos arquivos coletados para identificar as PMUs que não apresentassem nenhuma falha de leitura ou coleta no intervalo de tempo previamente selecionado. Quatro delas estão na região da cidade de Curitiba e quatro estão no interior do estado do Paraná, mais especificamente nas regiões das cidades de Jaguariaíva, Pinhão, Cascavel e Guaira.

Adotou-se duas abordagens distintas para realizar a análise da aplicação do PCA. Na primeira abordagem, relacionada à quantidade de PMUs utilizada, avalia-se a eficiência da solução aplicada individualmente por PMU ou aplicada em conjuntos de PMUs agrupadas. O objetivo é verificar e comparar os resultados obtidos com o olhar individual sobre cada PMU ou o olhar mais sistêmico, juntando dados do maior número possível de PMUs. Na segunda abordagem, relacionada à forma de alimentar o PCA, verifica-se qual conjunto de dados pode ser mais eficaz para executar o algoritmo de PCA. O objetivo é verificar os resultados obtidos utilizando-se todo o conjunto de dados e comparar com os resultados obtidos usando subconjuntos de dados para execução do PCA.

4.1 Abordagem 1 – Quantidade de PMUs

A primeira abordagem relaciona-se com a quantidade de PMUs sendo analisada, podendo ser tanto uma análise individual para cada PMU, quanto uma análise por agrupamentos de PMUs. No cenário de análise individual, o método é aplicado individualmente em cada uma das PMUs, e no cenário de análise agrupada, o método é aplicado em grupos de PMUs, montados aleatoriamente. A Tabela 2 apresenta a quantidade de atributos e a quantidade de anomalias em cada uma das 12 configurações utilizadas nesta abordagem.

Tabela 2 – Lista de PMUs

PMU	Qtde. de atributos	Qtde. de anomalias
PIL	32	215
SMC	20	687
GRL	11	548
DJP	26	3982
JGI	32	290
SGD	26	965
CVN	14	537
GUA	14	2432
PIL+SMC+GRL	66	692
DJP+JGI+SGD	72	1735
PIL+SMC+GRL+DJP+JGI+SGD	144	962
Todas as oito PMUs	172	857

4.2 Abordagem 2 – Alimentação do PCA

A segunda abordagem relaciona-se com o conteúdo do conjunto de dados que é utilizado para executar o algoritmo de análise de componentes principais, onde são experimentadas quatro configurações. Na primeira configuração, o algoritmo de análise de componentes principais é alimentado somente com os dados das amostras que foram identificadas como anomalias, durante a etapa do processo em que é feita a detecção de anomalias no conjunto de dados completo (Dados 1). Na segunda configuração, o algoritmo de análise de componentes principais é alimentado com o conjunto completo de amostras (Dados 2). Na terceira configuração, o algoritmo de análise de componentes principais é alimentado somente com as amostras que foram selecionadas para compor o conjunto de dados de treino, para o algoritmo de detecção de anomalias (Dados 3). Na quarta configuração, o algoritmo de análise de componentes principais é alimentado somente com as amostras que foram selecionadas para compor o conjunto de dados de validação cruzada e testes, para o algoritmo de detecção de anomalias (Dados 4).

4.3 Resultados obtidos

A combinação destas duas abordagens resulta em 48 diferentes cenários para análise. A título de exemplo, na Tabela 3 são mostrados os resultados de quatro cenários, obtidos ao se executar o algoritmo de detecção de anomalias após alimentar o algoritmo de análise de componentes principais exclusivamente com os dados do conjunto composto pelas PMUs PIL, SMC, GRL, DJP, JGI e SGD, em suas quatro configurações.

Na Tabela 3, as quatro colunas de valores mostram os resultados obtidos com cada uma das quatro abordagens de alimentação do PCA. As linhas apresentam as seguintes informações:

- Acurácia: resultado da acurácia, conforme (2);
- Precisão: resultado da precisão, conforme (3);

- Revocação: resultado da revocação, conforme (4);
- *F1 Score*: resultado do *F1 Score*, conforme (5).
- PCs Selecionados: Mostra a relação dos componentes principais selecionados e as respectivas contribuições.

Tabela 3 – Resultados do conjunto PIL+SMC+GRL+DJP+JGI+SGD

Itens	Dados 1	Dados 2	Dados 3	Dados 4
Acurácia	13,03%	99,67%	13,78%	100,00%
Precisão	100,00%	100,00%	100,00%	100,00%
Revocação	1,04%	99,62%	1,89%	100,00%
<i>F1 Score</i>	0,021	0,998	0,037	1,000
PCs Selecionados				
PC 1	40,95%	26,07 %	40,24 %	24,07 %
PC 2	14,52%	14,81 %	14,39 %	17,40 %
PC 3	12,70%	13,99 %	12,77 %	13,92 %
PC 4	4,67%	12,45 %	4,84 %	12,92 %
PC 5	4,01%	4,87 %	4,06 %	4,24 %
PC 6	3,11%	3,74 %	3,17 %	3,19 %
PC 7	3,06%	3,04 %	3,14 %	3,15 %
PC 8		2,84 %		2,23%

Na Tabela 3 observa-se que tanto o melhor resultado de acurácia, quanto do *F1 Score*, são obtidos com os Dados 4 alimentando o algoritmo de análise de componentes principais. E neste caso, obtém-se o critério de seleção através da acumulação de 80% da contribuição com oito componentes principais.

Tabela 4 – Resultados da métrica *F1 SCORE*

	Dados 1	Dados 2	Dados 3	Dados 4
PIL	0,148	0,956	0,083	0,767
SMC	0,021	0,516	0,008	0,982
GRL	0,046	1	0,171	1
DJP	0,000	0,922	0,001	0,952
JGI	0,104	1	0,215	1
SGD	0,091	1	0,511	1
CVN	0,016	1	0,371	1
GUA	0,166	1	0,191	0,996
PIL+SMC+GRL	0,027	0,707	0,028	0,8
DJP+JGI+SGD	0,011	1	0,031	1
PIL+SMC+GRL+DJP+JGI+SGD	0,021	0,998	0,037	1
Todas as oito PMUs	0,034	1	0,035	1

Pelo fato do *F1 Score* ser o resultado combinado de outras métricas, este será adotado para fazer as análises subsequentes. A Tabela 4 mostra uma compilação dos valores obtidos pela métrica do *F1 Score* para todas as 48 configurações das duas abordagens. A Tabela 4 mostra em destaque na cor vermelha

os melhores resultados do *FI Score* obtidos com a configuração diferente na alimentação do PCA para cada agrupamento de PMUs.

4.4 Eixo de análise 1 – Quantidade de PMUs

A primeira linha de análise que será feita, tem por base a primeira abordagem, relacionada à quantidade de PMUs. Para cada PMU individualizada e para cada agrupamento de PMUs, serão analisados os resultados obtidos na execução do procedimento.

Foi feita uma comparação, linha a linha, dos resultados. Pode-se observar na Tabela 4 que para oito dos doze conjuntos, o melhor resultado obtido foi com a alimentação do PCA através do conjunto Dados 2. Isto significa que para 66% dos casos esta é a forma mais efetiva de alimentar o algoritmo do PCA. Porém para dez dos doze conjuntos, o melhor resultado obtido foi com a alimentação do PCA através dos Dados 4, ou seja, para 83% dos casos esta é a forma mais efetiva de alimentar o algoritmo do PCA.

4.5 Eixo de análise 2 – Alimentação do PCA

A segunda linha de análise feita, tem por base a abordagem relacionada à fonte de alimentação do algoritmo de análise de componentes principais, verificando para cada uma das quatro configurações, os resultados obtidos em cada um dos conjuntos de dados de PMUs (individualizados e agrupados).

Para isto, foi feita uma comparação, coluna a coluna, dos resultados. Neste caso, não foi facilmente identificado um padrão, como no caso da análise anterior. Para cada configuração diferente da fonte de entrada de dados, os resultados obtidos em cada conjunto de dados de PMUs foram bastante diversos.

A tabela 4 mostra que os conjuntos de Dados 2 e Dados 4 apresentam os melhores resultados de *FI Score*. Nestes dois casos a composição com todas as PMUs traz um ótimo desempenho, compensando desempenhos inferiores proporcionados pelas análises individuais. Estas observações apontam para resultados mais consistentes ao se adotar uma análise agrupada.

4.6 Redução da dimensionalidade

Outro aspecto importante a ser observado nos resultados obtidos é com a relação à redução de dimensionalidade obtida com a execução do algoritmo de análise de componentes principais. A tabela 5 mostra a quantidade original de atributos existente em cada configuração de conjuntos de PMUs e relaciona ao lado a quantidade de componentes principais obtida na execução de cada cenário.

Ao se observar a taxa de compressão na tabela 5 pelo aspecto do eixo de abordagem relacionado à quantidade de PMUs, observa-se que quanto maior a quantidade de atributos maior é a taxa de compressão obtida. Isto implica em dizer que uma quantidade maior de atributos originais obtida pela junção de várias PMUs pode ser representada quase que pela mesma quantidade de componentes principais obtida pela análise de uma única PMU.

Tabela 5 – Quantidade de PCs por cenário

Conjuntos de PMUs	Qtde Original	Quantidade de PCs			
		Dados 1	Dados 2	Dados 3	Dados 4
PIL	32	4	5	4	6
SMC	20	5	5	5	6
GRL	26	5	5	5	5
DJP	26	4	6	5	6
JGI	32	9	9	9	9
SGD	26	5	5	5	4
CVN	14	5	5	5	5
GUA	14	5	5	5	5
PIL+SMC+GRL	78	5	6	5	7
DJP+JGI+SGD	84	8	9	8	9
PIL+SMC+GRL+DJP+JGI+SGD	162	7	8	7	8
Todas as 8 PMUs	190	7	8	7	8

4.7 Discussão

A análise dos resultados no eixo de análise 1, que aborda a quantidade de PMUs, indica que independente da quantidade de PMUs utilizada, a estratégia mais eficaz para alimentação do algoritmo de análise de componentes principais é com o conjunto de Dados 4, que leva uma pequena vantagem sobre o conjunto de Dados 2. A análise dos resultados no eixo de análise 2, que aborda a perspectiva da alimentação do algoritmo de análise de componentes principais, indica que os melhores resultados são obtidos quando se utiliza dados da maior quantidade possível de PMUs de forma agrupada. A análise da redução de dimensionalidade, através da quantidade de componentes principais, indica que é possível obter taxas de compressão melhores utilizando-se mais atributos obtidos através do maior agrupamento possível de PMUs.

Ao combinar os resultados destas análises, infere-se que o melhor cenário é quando da utilização de dados de validação cruzada e teste (Dados 4), da maior quantidade possível de PMUs, para alimentação do algoritmo de análise de componentes principais.

Este cenário permite interpretar que pode ser adotada uma metodologia em que é possível realizar um pré-processamento *offline* dos dados, para obtenção dos parâmetros de execução do algoritmo de análise de componentes principais. Este pré-processamento vai gerar a base de treino para o algoritmo de análise de dados, no caso deste trabalho para detecção de anomalias, e por ser *offline* pode ser executado periodicamente, de acordo com as necessidades e possibilidades da infraestrutura existente. Após a geração desta base de treino, o algoritmo de análise de dados pode ser executado a qualquer momento que for necessário, bastando acrescentar ao conjunto de dados de treino o conjunto de dados que se deseja verificar.

5. CONCLUSÕES

Este trabalho propôs uma abordagem para a junção de dados coletados de um sistema de medição fasorial sincronizada para seleção e tratamento das informações mais relevantes, de forma a disponibilizá-las para aplicações e análises pós-falta.

Foi desenvolvida uma metodologia para seleção de dados para análise pós-falta, através da redução da dimensionalidade, utilizando a análise de componentes principais. Com a aplicação da metodologia, comprovou-se a possibilidade de se utilizar um conjunto de dados menor do que o original, porém mantendo as características dos atributos. Demonstrou-se um caso usando dados reais em que foi possível fazer a redução de 190 atributos, para um conjunto de oito componentes principais com uma precisão de 100% no *F1 Score*. Esta metodologia automática de seleção de dados traz benefícios para análise pós-falta devido ao ganho de desempenho obtido com a redução do conjunto de dados dos atributos elétricos, sem perda significativa na precisão da capacidade de detecção de anomalia.

O método desenvolvido permite a utilização com qualquer conjunto de dados de PMUs, independente de fabricante ou fornecedor destes equipamentos, ou da quantidade disponível. Foi demonstrada a aplicação da metodologia em cenários variando de apenas uma PMU a oito PMUs diferentes.

Como trabalhos futuros, sugere-se: automatizar a análise da qualidade dos dados de entrada, utilizando-se técnicas analíticas de dados para este fim, para garantir maior precisão no processo e maior volume de dados; utilizar os dados selecionados para mais aplicações de análise pós-falta além de detecção de anomalias, tais como localização de falta e estabilidade de tensão.

AGRADECIMENTOS

Os autores agradecem a bolsa de desenvolvimento de tecnologia do CNPq – Conselho Nacional de Desenvolvimento Científico e Tecnológico do Ministério da Ciência, Tecnologia e Inovações. Os autores agradecem o apoio da CAPES – Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, fundação do Ministério da Educação (MEC). Os autores também agradecem à COPEL GeT e à ANEEL pelo financiamento do programa de pesquisa e desenvolvimento tecnológico (P&D), por meio do projeto PD-06491-0531/2019.

REFERÊNCIAS

- Ahmad, T.; Senroy, N. (2020). *Statistical Characterization of PMU Error for Robust WAMS Based Analytics*. IEEE Transactions on Power Systems, v. 35, n. 2, p. 920–928.
- Bruce, P.; Bruce, A. (2019). *Estatística prática para cientistas de dados: 50 conceitos essenciais*. Alta Books, Rio de Janeiro.
- Cai, L.; Thornhill, N. F.; Kuenzel, S. (2018). *Wide-Area Monitoring of Power Systems Using Principal Component Analysis and -Nearest Neighbor Analysis*. IEEE Transactions on Power Systems, v. 33, n. 5, p. 4913–4923.
- Cardoso Jr, G.; Rolim, J. G.; Zürn, H. H. (2004). *Diagnóstico de faltas em sistemas de potência: definição do problema e abordagens via inteligência artificial*. SBA: Controle & Automação Sociedade Brasileira de Automática, v. 15, n. 2, p. 215–229.
- Chatterjee, K.; Chaudhuri, N. R. (2019). *Corruption-Resilient Detection of Event-Induced Outliers in PMU Data: A Kernel PCA Approach*. 2019 IEEE Power & Energy Society General Meeting.
- Chu, L.; Qiu, R.; He, X.; Ling, Z.; Liu, Y. (2018). *Massive Streaming PMU Data Modelling and Analytics in Smart Grid State Evaluation based on Multiple High-Dimensional Covariance Test*. IEEE Transactions on Big Data, v. 4, n. 1, p. 55–64.
- Dimitrovska, T.; Rudez, U. (2019). *Real-Time Application of an Indirect Power-System Contingency Screening Method Based on Adaptive PCA*. IEEE Transactions on Power Systems, v. 34, n. 6, p. 4665–4673.
- Ensslin L., Ensslin S., Lacerda R., Tasca J. (2010). *Processo de seleção de portfólio bibliográfico*.
- Karlsson, D.; Hemmingsson, M.; Lindahl, S. (2004). *Wide Area System Monitoring and Control*. IEEE Power & Energy Magazine, p. 68–76.
- Le, N. T.; Benjapolakul, W. (2018). *A data imputation model in phasor measurement units based on bagged averaging of multiple linear regression*. IEEE Access, v. 6, p. 39324–39333.
- Liu, X.; Laverty, D. M.; Best, R. J.; Li, K. (2015). *Principal component analysis of wide-area phasor measurements for islanding detection—A geometric view*. IEEE Transactions on Power Delivery, v. 30, n. 2, p. 976 - 985.
- Lu, C.; Shi, B.; Wu, X.; Sun, H. (2015). *Advancing China's Smart Grid: Phasor Measurement Units in a Wide-Area Management System*. IEEE Power and Energy Magazine, v. 13, n. 5, p. 60–71.
- Manly, B. F. J.; Alberto, J. A. N. (2019). *Métodos estatísticos multivariados: uma introdução*. Bookman, Porto Alegre.
- NASPI. (2017). North American Synchrophasor Initiative. *PMU Data Quality: A framework for the attributes of PMU data quality and a methodology for examining data quality impacts to synchrophasor applications*.
- ONS. (2020). ONS. *O Sistema Interligado Nacional*. Disponível em: <<http://www.ons.org.br/paginas/sobre-o-sin/o-que-e-o-sin>>. Acesso em: novembro de 2020.
- Overholt, P.; Ortiz, D.; Silverstein, A. (2015). *Synchrophasor technology and the DOE: Exciting opportunities lie ahead in development and deployment*. IEEE Power and Energy, p. 14–17.
- Phadke, A. G.; Thorp, J. S. (2017). *Synchronized Phasor Measurements and Their Applications*. Springer.
- Sattinger, W.; Giannuzzi, G. (2015). *Monitoring Continental Europe: An Overview of WAM Systems Used in Italy and Switzerland*. IEEE Power and Energy Magazine, v. 13, n. 5, p. 41–48.
- Xu, Y.; Sun, Y.; Wan, J.; Liu, X.; Song, Z. (2017). *Industrial Big Data for Fault Diagnosis: Taxonomy, Review, and Applications*. IEEE Access, v. 5, p. 17368–17380.
- Zhang, Y.; Xu, Y.; Dong, Z. Y. (2017). *Robust Ensemble Data Analytics for Incomplete PMU Measurements-Based Power System Stability Assessment*. IEEE Transactions on Power Systems, v. 33, n. 1, p. 1124 - 126.