Visualização de dados de turbinas eólicas baseado na Análise de Componentes Principais.

Danielle R. Pinna[1] **, Rodrigo Hamacher**[1] **, Fernando de Sá**[1]**,**
Rodrigo Toso[2] **, Felipe Henriques** [1] **, Diego Brandão**[1]

[1]Programa de Pós-graduação em Ciência da Computação (PPCIC)
Centro Federal de Educação Tecnológica Celso Suckow da Fonseca - CEFET/RJ

[2]Microsoft AI & Research.

danielle.pinna@aluno.cefet-rj.br, diego.brandao@cefet-rj.br

Resumo. *A energia eólica tornou-se uma das fontes energéticas mais importantes em todo o mundo. Sistemas de monitoramento são indispensáveis para manter o bom funcionamento dos componentes das turbinas eólicas. Em geral, a grande quantidade de dados geradas por estes sistemas não são usados de forma eficaz, desta maneira, se faz necessário o uso de técnicas capazes de reduzir a dimensão dos dados a fim de facilitar a interpretação das análises realizadas e melhor entender o comportamento dos dados. Neste trabalho, discutimos as visualizações dos dados reais de uma turbina eólica produzidos pelo sistema SCADA e aplicamos o método de Análise de Componentes Principais para reduzir a dimensão dos dados de alta dimensão.*

1. Introdução

A energia eólica é um importante recurso de energia limpa e renovável disponível na natureza.A turbina eólica (ou aerogerador) é responsável pela transformação da energia eólica em energia elétrica por meio do vento que movimenta as pás e faz girar o rotor, que transmite a rotação ao gerador.

De acordo com et al. [2017], a energia eólica é a fonte de energia renovável que mais cresce, porém a operação e manutenção das turbinas são responsáveis por cerca de 25% a 35% dos custos de geração.

Os problemas relacionados a manutenção da turbina eólica normalmente são as falhas do sistema elétrico e os provenientes das condições climáticas extremas. A falha do componente ocasiona redução da produtividade ou até mesmo o desligamento da turbina. Por essa razão, a maneira mais eficaz de reduzir os custos de manutenção é monitorar o status dos geradores e prever o seu mau funcionamento antes que o sistema falhe [et al.]. Assim, o diagnóstico precoce de falhas é um fator chave para reduzir significativamente os custos de manutenção.

Atualmente, os aerogeradores modernos já possuem um sistema de coleta e de armazenamento de dados, conhecido como sistema de controle e aquisição de dados (SCADA). Como o nome sugere, ele é um sistema de monitoramento alimentado por vários sensores que medem variáveis críticas do processo. Essas variáveis podem ser, por exemplo, do sistema hidráulico, do rotor e também variáveis meteorológicas, como temperatura, pressão atmosférica, velocidade do vento, umidade, etc.

A maioria dos trabalhos sobre detecção de falhas em turbinas eólicas são baseados em conjuntos de dados operacionais e de eventos, como os fornecidos pelo SCADA [et al., 2019].

O objetivo deste trabalho é apresentar uma análise exploratória inicial acerca de um conjunto de dados reais de turbinas eólicas, extraídos a partir do SCADA. Apresentamos visualizações gráficas e aplicamos a análise de componentes principais para reduzir a dimensionalidade dos dados.

O presente artigo está organizado em mais 5 seções. Na seção 2 é apresentado o referencial teórico, a seção 3 apresenta a metodologia. A base de dados é apresentada na seção 4, já os resultados são discutidos na seção 5 e, por fim, a seção 6 apresenta as considerações finais.

2. Fundamentação Teórica

Com os avanços da capacidade de armazenamento de dados, ocorre um fenômeno denominado de sobrecarga de informação. Isso significa que o número de variáveis disponíveis excede aquele necessário para entender o fenômeno estudado [Fodor, 2002]. No contexto de identificação de anomalias, é normal que haja uma redução do número de variáveis selecionadas do conjunto de dados para aplicação dos modelos.

A importância da aplicação de técnicas de redução de dimensionalidade é essencial quando se manipula um conjunto de dados de elevada dimensionalidade, cujo processamento demanda a retirada de características irrelevantes e redundantes para o modelo. Existem duas principais abordagens de técnicas de redução de dimensionalidade: extração de características e seleção de características. A primeira aplica uma transformação algébrica no espaço do conjunto de dados de entrada, resultando em um subespaço de dimensão inferior [et al., 2014]. A segunda técnica, de seleção de características, constrói um subconjunto de características a partir do conjunto de entrada original sem aplicar nenhuma transformação, no qual o objetivo é criar um subconjunto ótimo de variáveis do problema [Guyon and Elisseeff, 2003].

Neste trabalho, será utilizada a primeira técnica, de extração de características, aplicada ao conjunto de dados do SCADA por meio da análise de componentes principais, que consiste em transformar um conjunto de variáveis originais em um conjunto menor de variáveis (componentes) com uma perda mínima de informação.

2.1. Análise de Componentes Principais

Segundo Mingoti [2007], a técnica denominada de análise de componentes principais (ACP) tem como objetivo explicar a estrutura de variância e covariância de um vetor aleatório através da construção de combinações lineares das variáveis originais. Estas combinações são chamadas de componentes principais e são não correlacionadas entre si.

A j-ésima componente principal da matriz de covariância Σ_{pxp} é definida por:

$$Y_j = e_j' X = e_{j1}X_1 + e_{j2}X_2 + \cdots + e_{jp}X_p \qquad j = 1, \cdots, p$$

Se temos p-variáveis originais é possível obter-se a mesma quantidade p de componentes principais.

Em geral, o objetivo dessa análise é reduzir a quantidade de dados e facilitar a interpretação das análises realizadas. Assim, a informação contida nas p variáveis originais é substituída pela informação contida em k ($k < p$) componentes principais não correlacionadas. As componentes principais só dependem da matriz de covariância Σ ou da matriz de correlação P.

A variância das componentes principais é igual ao autovalor correspondente. A primeira componente principal é a de maior variância e as outras aparecem em ordem decrescente de variância.

$$Var(Y_i) = \lambda_i \;\; e \;\; \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$$

A variação total (soma das variâncias) de Y é a mesma de X. Isto é,

$$\lambda_1 + \lambda_2 + \cdots + \lambda_p = \sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp}$$

A proporção da variância total explicada ou devida ao k-ésimo componente é :

$$\frac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}, \;\; k = 1, 2, \cdots, p$$

Quando há uma discrepância muito acentuada entre as variâncias das variáveis originais, cada componente passa a ser extremamente dominada por uma variável em particular, o que torna as componentes sem muita utilidade prática. Esta discrepância é muitas vezes causada pela diferença das unidades de medidas das variáveis sendo assim necessário que seja realizado alguma transformação nos dados de modo a equilibrar melhor as variâncias.

Considere as variáveis padronizadas:

$$Z_i = \frac{X_i - \mu_i}{\sqrt{\sigma_{ii}}} \;\; i = 1, \cdots, p$$

A j-ésima componente principal da matriz de correlação P_{pxp} é definida por:

$$Y_j = e_j' Z = e_{j1} Z_1 + e_{j2} Z_2 + \cdots + e_{jp} Z_p \;\; j = 1, \cdots, p$$

3. Base de Dados

A base de dados utilizada é fornecida pela empresa de Energias de Portugal (EDP)[1]. De acordo com et al. [2020], os dados da EDP são um dos conjuntos de dados gratuitos disponíveis mais completos para análise de recursos eólicos e pesquisa do desempenho de turbinas eólicas.

A disponibilização desses dados se deu a partir de um desafio proposto pela empresa no qual o objetivo era a detecção de falhas de turbinas eólicas. Os registros foram extraídos do sistema de controle e aquisição de dados - SCADA, de 5 turbinas eólicas medidos nos anos de 2016 e 2017.

As informações disponíveis pela EDP são:

- *Metmast*: *Dataset* das variáveis meteorológicas, medido a cada 10 minutos. Os dados são extraídos de uma única torre. Exemplo de variáveis: Velocidade e direção do vento (2 sensores anemométricos), temperatura, pressão atmosférica, umidade, precipitação.
- *Failures*: *Dataset* com o registro das ocorrências de falhas em cada componente da turbina eólica, medido no tempo de cada ocorrência.
- *Logs*: *Dataset* do histórico dos eventos normais e anormais que ocorreram em cada turbina.
- *Signals*: *Dataset* das variáveis do sistema SCADA para os componentes e valores de produção mais importantes de cada turbina, leitura a cada 10 minutos.
- *Locations*: Localização das turbinas, contendo a latitude e longitude.

Além disso, a empresa já disponibilizou os dados separados em treinamento e teste (sendo 80% para treinamento e 20% para teste), onde os dados de teste representam os últimos 4 meses do ano de 2017. Apenas não foi disponibilizado o conjunto de teste do *dataset Failures* porque a empresa avaliou o desempenho dos modelos desenvolvidos pelos competidores por esse conjunto de dados. A Tabela 1 apresenta a quantidade de observações total e a quantidade de variáveis em cada conjunto de dados.

Tabela 1. Descrição dos conjuntos de dados.

Conjunto de Dados	Quantidade de observações	Quantidade de variáveis
Metmast	87.528	41
Failures	23	4
Logs	318.835	5
Signals	521.784	83
Locations	17	3

4. Metodologia

As variáveis utilizadas neste trabalho foram as do conjunto de dados *Metmast*, que compõem as variáveis meteorológicas medidas a cada 10 minutos extraídos de uma única torre. Foram utilizadas as informações de velocidade e direção do vento, temperatura ambiente, pressão, umidade, precipitação, detecção de chuva e do anemômetro. As variáveis que continham apenas um único valor ao longo do tempo, identificadas como "Offset", foram desconsideradas da base.

Com o intuito de avaliar o comportamento de variação conjunta dessas variáveis foi realizada uma análise multivariada de componentes principais (PCA) com base na matriz de correlação. Para efetuar essa análise, os dados foram normalizados a fim de eliminar a discrepância das unidades de medida entre as variáveis. A seleção do número de componentes principais foi baseada no critério da análise gráfica e de acordo com o percentual desejado de perda mínima de informação.

Todas as análises deste artigo foram feitas por meio de rotinas computacionais implementadas no software R 4.1.1 (R Core Team [2014]).

5. Resultados

A Tabela 2 apresenta uma amostra do conjunto de dados *Metmast*.

Tabela 2. Amostra das primeiras observações do dataset *Metmast*.

Timestamp	Windspeed1				Windspeed2			
	Min	Max	Avg	Var	Min	Max	Avg	Var
2016-01-01T00:00:00+00:00	3,70	6,00	5,10	0,21	3,80	6,00	5,10	0,22
2016-01-01T00:10:00+00:00	4,10	6,00	5,10	0,09	4,10	6,00	5,20	0,10
2016-01-01T00:20:00+00:00	4,50	6,70	5,70	0,26	4,40	6,80	5,80	0,30
2016-01-01T00:30:00+00:00	5,10	7,00	6,30	0,11	5,10	7,10	6,40	0,12
2016-01-01T00:40:00+00:00	4,70	7,30	6,20	0,27	4,90	7,40	6,30	0,27
2016-01-01T00:50:00+00:00	4,90	7,40	6,60	0,33	5,00	7,60	6,80	0,35
2016-01-01T01:00:00+00:00	3,90	7,10	5,30	0,33	4,00	7,10	5,40	0,34
2016-01-01T01:10:00+00:00	4,30	6,70	5,70	0,26	4,50	6,90	5,80	0,26
2016-01-01T01:20:00+00:00	4,60	6,50	5,90	0,12	4,80	6,70	6,00	0,13
2016-01-01T01:30:00+00:00	4,40	6,50	5,60	0,16	4,60	6,70	5,70	0,15

Na Figura 1, temos o gráfico da análise temporal por variável meteorológica, sumarizado pela média diária das observações. A partir das visualizações pode-se notar que das 40 variáveis dos dados meteorológicos muitas possuem variação nula ao longo do tempo e também que há uma interrupção das medições de Fev/2017 a Abr/2017.

Figura 1. Análise temporal das variáveis.

Na Figura 2 é apresentado as 15 correlações cruzadas mais relevantes, ou seja, a classificação das variáveis com correlação mais altas obtidas em uma tabela cruzada.

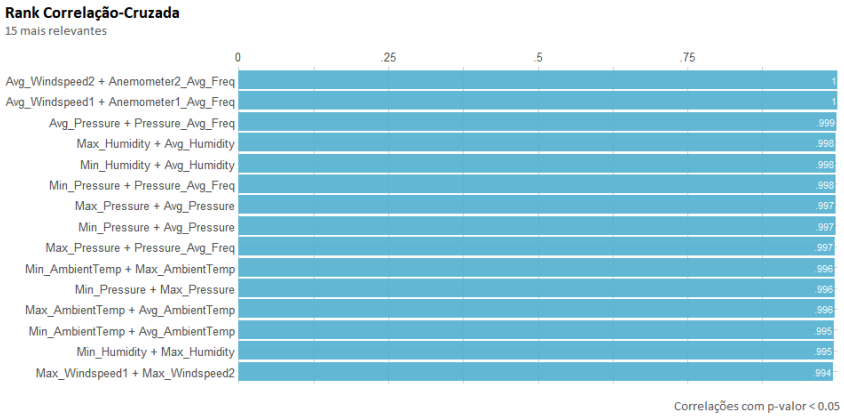

Figura 2. Rank das Correlações Cruzada do *Metmast*.

A Tabela 3 apresenta as dez primeiras componentes geradas pelo PCA com o sua respectivo percentual de variância explicada e acumulada. As 6 primeiras componentes possuem Autovalor maior que 1 e explicam aproximadamente 85% da variância dos dados. Ou seja, podemos efetivamente reduzir a dimensionalidade de 28 para 6 enquanto "perdemos" cerca de 15% da variância.

Tabela 3. Autovalores e Variância das 10 primeiras componentes.

Componentes	Autovalor	% Variância	% Variância Acumulada
1	9,30	33,20	33,20
2	4,74	16,94	50,14
3	3,99	14,24	64,38
4	2,56	9,16	73,54
5	2,08	7,41	80,95
6	1,21	4,33	85,28
7	1,02	3,64	88,92
8	1,00	3,57	92,48
9	0,93	3,34	95,82
10	0,45	1,60	97,42

Também observamos através da Figura 3 que podemos explicar mais de 50% da variância apenas com as duas primeiras componentes.

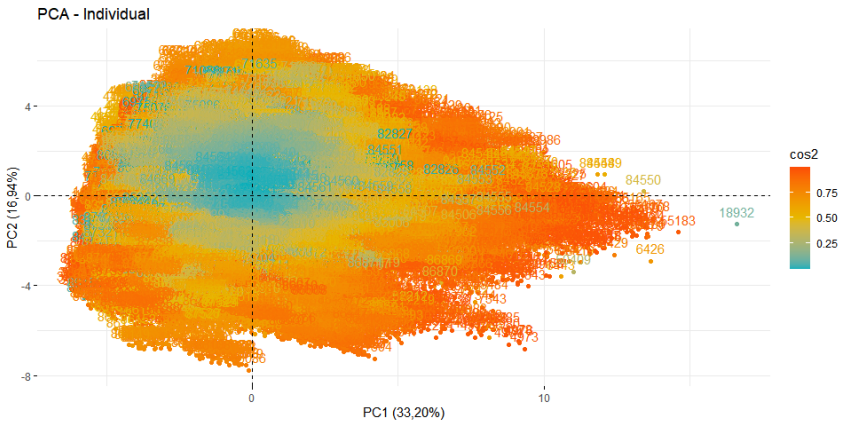

Figura 3. *Scatter Plot* das 2 primeiras Componentes Principais.

Com a seleção das seis componentes principais, a redução da dimensão de 28 variáveis originais para essas componentes é bastante razoável. Portanto decidiu-se utilizar unicamente os 6 primeiros componentes principais responsáveis por 85,28% da variância total no conjunto de dados.

Como intuito de entender a importância de cada variável na construção das componentes, são apresentados na Tabela 4 os coeficientes de ponderação de cada característica. Na primeira componente principal destacaram-se as variáveis de Velocidade do Vento e Anemômetro e neste caso pode-se chamá-la de componente de indicador do Vento. A segunda componente principal pode ser chamada de componente da Temperatura Ambiente, a terceira componente de Umidade, a quarta relacionada a Precipitação, a quinta componente ficou com a Direção do Vento e por último, na sexta componente destacaram-se conjuntamente as variáveis de Temperatura Ambiente e Umidade.

Tabela 4. Autovetores das 6 primeiras componentes selecionadas.

Variáveis	Coeficiente de Ponderação					
	PC1	PC2	PC3	PC4	PC5	PC6
Min_Windspeed1	**0,29**	-0,12	0,07	-0,06	0,03	-0,02
Max_Windspeed1	**0,32**	-0,09	0,04	-0,01	-0,02	0,01
Avg_Windspeed1	**0,31**	-0,11	0,05	-0,03	0,00	0,00
Var_Windspeed1	**0,25**	-0,04	0,02	0,05	-0,06	0,03
Min_Windspeed2	**0,29**	-0,12	0,03	-0,04	-0,01	0,00
Max_Windspeed2	**0,32**	-0,09	0,05	-0,02	0,00	0,00
Avg_Windspeed2	**0,32**	-0,10	0,04	-0,03	-0,01	0,00
Var_Windspeed2	**0,23**	-0,03	0,09	0,02	0,00	-0,02
Min_Winddirection2	-0,02	0,05	0,17	-0,20	**0,47**	-0,14
Max_Winddirection2	0,00	0,08	0,18	-0,22	**0,49**	0,10
Avg_Winddirection2	-0,01	0,07	0,20	-0,23	**0,52**	0,03
Var_Winddirection2	-0,02	0,01	-0,04	0,04	-0,04	-0,08
Min_AmbientTemp	0,13	**0,31**	-0,20	0,04	0,05	**0,41**
Max_AmbientTemp	0,13	**0,31**	-0,20	0,04	0,05	**0,41**
Avg_AmbientTemp	0,13	**0,31**	-0,20	0,04	0,05	**0,41**
Min_Pressure	-0,07	-0,36	-0,25	0,06	0,17	0,15
Max_Pressure	-0,07	-0,36	-0,25	0,06	0,16	0,14
Avg_Pressure	-0,07	-0,36	-0,25	0,06	0,17	0,15
Min_Humidity	-0,12	-0,17	**0,35**	-0,11	-0,16	**0,36**
Max_Humidity	-0,12	-0,16	**0,35**	-0,10	-0,16	**0,35**
Avg_Humidity	-0,12	-0,16	**0,35**	-0,10	-0,16	**0,35**
Min_Precipitation	0,01	0,01	0,20	**0,51**	0,15	0,02
Max_Precipitation	0,01	0,01	0,22	**0,51**	0,15	0,02
Avg_Precipitation	0,01	0,01	0,21	**0,52**	0,15	0,02
Max_Raindetection	0,00	0,00	0,00	0,00	-0,01	0,02
Anemometer1_Avg_Freq	**0,31**	-0,11	0,05	-0,03	0,00	0,00
Anemometer2_Avg_Freq	**0,32**	-0,10	0,04	-0,03	-0,01	0,00
Pressure_Avg_Freq	-0,07	-0,36	-0,25	0,06	0,17	0,15

Com base na interpretação das componentes, pode-se observar que as variáveis de Velocidade do Vento e Anemômetro associadas a primeira componente são as que representam a maior variância dos dados e demonstram ser variáveis com grande potencial de contribuição para o estudo de falhas em turbinas.

6. Conclusão

A partir das visualizações pode-se notar que das 40 variáveis dos dados meteorológicos muitas possuem variação nula ao longo do tempo e também que há uma interrupção das medições de Fev/2017 a Abr/2017. Utilizou-se a técnica de Análise de Componentes Principais, com o objetivo de reduzir a dimensão dos dados e facilitar a interpretação das análises a ser realizadas. Assim, a informação contida nas variáveis originais pode ser substituída pela informação contida nas 6 componentes principais não correlacionadas que correspondem a 85% da variância total explicada, resultando em uma economia de recursos para futuros trabalhos que utilizarão essa mesma base de dados, sem perda significativa de informação.

Referências

Edp - open data. https://opendata.edp.com/pages/homepage/, Acessado em 15/08/21.

A. Blanco-M. et al. Impact of target variable distribution type over the regression analysis in wind turbine data. In *2017 International Conference and Workshop on Bioinspired Intelligence (IWOBI)*, pages 1–7, 2017.

A. Stetco et al. Machine learning methods for wind turbine condition monitoring: A review. *Renewable Energy*, 133:620–635, 2019.

D. Menezes et al. Wind farm and resource datasets: A comprehensive survey and overview. *Energies*, 13(18), 2020.

S. Khalid et al. A survey of feature selection and feature extraction techniques in machine learning. In *2014 Science and Information Conference*, pages 372–378, 2014.

S. Qin et al.

I.K. Fodor. A survey of dimension reduction techniques. 5 2002. URL https://www.osti.gov/biblio/15002155.

I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182, mar 2003.

S. A. Mingoti. Análise de dados através de métodos estatística multivariada: uma abordagem aplicada. In *Análise de dados através de métodos estatística multivariada: uma abordagem aplicada*, page 295. Editora UFMG, 2007. ISBN 9788570414519.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.