

Desenvolvimento de um modelo de clusterização de usinas eólicas para atendimento dos modelos de previsão do ONS

Paulo Sérgio de Castro Nascimento * Lucas De Souza Khenayfis **
Rogério José Menezes Alves *** André Luís Marques Marcato ****
Cláudia de Magalhães Santos Fonseca †
Patrícia de Sousa Oliveira Silva ‡

* *Gerência de Metodologias e Modelos Energéticos - ONS e Departamento de Engenharia Elétrica, Universidade Federal de Juiz de Fora, MG, (e-mail: castro.paulo@engenharia.ufjf.br).*

** *Gerência de Metodologias e Modelos Energéticos - ONS e Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro - PUC-Rio, RJ, (e-mail: lkhenayfis@ons.org.br).*

*** *Gerência de Metodologias e Modelos Energéticos - ONS e Departamento de Engenharia Elétrica, Universidade Federal do Espírito Santo, ES, (e-mail: rogerio.alves@ons.org.br).*

**** *Departamento de Engenharia Elétrica, Universidade Federal de Juiz de Fora, MG, (e-mail: andre.marcato@ufjf.edu.br).*

† *Departamento de Engenharia Elétrica, Universidade Federal de Juiz de Fora, MG, (e-mail: claudiabja@gmail.com).*

‡ *Departamento de Engenharia Elétrica, Universidade Federal de Juiz de Fora, MG, (e-mail: patricia.sousa@engenharia.ufjf.br).*

Abstract: The present work proposes to obtain a clustering of wind farms using the Gaussian Mixture Models (GMM) method, based on the wind profile and geographic location of wind farms operating in the South and Northeast regions from Brazil. The purpose of the generated grouping is to provide inputs for the wind generation forecasting models, helping the daily schedule of the operation of the National Electric System Operator (ONS). The success of this application allows the creation of generation and wind series more assertively within each cluster, thus increasing the reliability of the system operation. Thus, the main contribution of this work is to propose an alternative form of clustering, in relation to what is currently used by the ONS.

Resumo: O presente trabalho propõe a obtenção de uma clusterização de usinas eólicas com a utilização do método *Gaussian Mixture Models* (GMM), baseado no perfil de ventos e localização geográfica dos parques eólicos em operação nas regiões Sul e Nordeste do Brasil. O intuito do agrupamento gerado é fornecer insumos para os modelos de previsão de geração eólica, auxiliando a programação diária da operação do Operado Nacional do Sistema Elétrico (ONS). O sucesso dessa aplicação permite a criação de séries de geração e vento de forma mais assertiva dentro de cada *cluster*, assim aumentar a confiabilidade da operação do sistema. Dessa forma, a principal contribuição deste trabalho é propor uma forma alternativa de clusterização, em relação ao que atualmente é utilizado pelo ONS.

Keywords: Wind power; clustering; electrical systems planning.

Palavras-chaves: Geração eólica; clusterização; planejamento de sistemas elétricos.

1. INTRODUÇÃO

A crescente penetração da geração renovável, com natureza intermitente, tem apresentado novos desafios e complexidades ao ONS (Operador Nacional do Sistema Elétrico) na operação do SIN (Sistema Interligado Nacional). A grande transformação da matriz elétrica brasileira é decorrente da crescente expansão da capacidade instalada de usinas eólicas, superando o índice de outras fontes especificamente nas regiões Nordeste e Sul do Brasil. Tal situação impulsionou o ONS com necessidades de aperfeiçoamento

dos modelos de planejamento e operação para fazer face à geração de alta variabilidade.

Os maiores desafios da operação da fonte eólica estão relacionados à alta variabilidade da geração, aliada à complexidade de prever a condição futura, devido à dependência de variáveis climatológicas. Essas características implicam o acionamento de outras fontes de geração capazes de absorver essa intermitência. O que se acentua conforme o carregamento do ponto de conexão, ou seja, quanto maior é a capacidade de geração eólica conectada ao ponto.

Isto faz com que a acurácia da previsão dessa fonte seja imprescindível para a operação, especialmente em tempo real, devidos aos efeitos das rampas de geração (Eltohamy *et al.*, 2019).

Atualmente existe um número elevado de parques conectados ao SIN e, portanto, não é prático modelar estas centenas de usinas eólicas em uma única simulação para previsão. Em vista disso, geralmente os parques eólicos são divididos em grupos, podendo abranger áreas relativamente vastas. No entanto, é preciso cuidado ao se fazer esses agrupamentos, pois um grupo de usinas deve ser formado por usinas com características similares de velocidade de vento (Lima, 2017).

No modelo de previsão de geração eólica atualmente em uso, o ONS realiza esta previsão por agrupamentos definidos por critérios elétricos (ONS, 2018). De forma geral, esses agrupamentos são construídos considerando a conexão das usinas aos pontos de entrada na Rede Básica, os quais o ONS possui observabilidade. A Rede Básica é formada pelo conjunto de instalações do SIN, linhas de transmissão, subestações e outros equipamentos com tensão igual ou superior a 230 kV, porém alguns pontos de conexão monitorados pelo ONS podem estar em 69 kV. Este critério de agrupamento não possui nenhum embasamento técnico que o justifique, podendo, em alguns casos, agrupar usinas muito distantes entre si fisicamente, ou seja, com perfis de vento possivelmente muito distintos e estão em um mesmo *cluster*. Similarmente, há usinas muito próximas com comportamento de vento semelhante, mas que estão em grupos diferentes.

Em vista disso, a maior acurácia na representação espacial das usinas eólicas pode contribuir com a previsão de geração dessa fonte, o que possibilita um melhor uso dos recursos, de forma a reduzir a necessidade de reserva de potência operativa ou de geração termoeletrica adicional àquela definida por mérito econômico pelos modelos de otimização oficialmente em uso.

Muitos trabalhos da literatura apresentam métodos para clusterização aplicados à fonte eólica, todavia na grande maioria o interesse é encontrar agrupamentos ótimos de aerogeradores. O interesse de agrupamentos de parques eólicos está associado mais aos operadores de rede, por sua necessidade de operação macro dos sistemas elétricos de potência. Neste sentido, pode ser destacado o trabalho de Teng *et al.* (2019), sendo um algoritmo de clusterização *Support Vector Machine* (SVM) para agrupar turbinas eólicas em um parque. Além disso, os pesquisadores propõem uma melhoria no modelo com a aplicação de um método de detecção de pontos de fronteira. O trabalho utiliza dados históricos de velocidade e direção do vento. Já em Sun e Liu (2017) é abordado o agrupamento de parques eólicos para o planejamento do sistema, através da combinação de um *Bird Swarm Algorithm* com *K-means*. Neste trabalho os pesquisadores relatam que os parques eólicos podem ser divididos por equivalência dinâmica, mas citam outros trabalhos que aplicaram outros métodos, como exemplo: o *fuzzy C-means* proposto por Wu *et al.* (2016), que é um aprimoramento do *K-means*; Yang e Liu (2015), que utiliza o método *Gray Wolf Optimizer* para otimizar a clusterização.

Conforme mencionado anteriormente, uma das maiores dificuldades da operação, em relação às eólicas, é a obtenção de uma boa precisão na sua previsão. Em tempo real, essa previsão é essencial para que, na sala de controle, os operadores conheçam o comportamento da geração eólica para um futuro próximo, observando a injeção de potência em alguns pontos de conexão específicos da rede observada pelo ONS. Assim, podem realizar comandos para evitar possíveis problemas elétricos. O comportamento da clusterização por pontos de conexão é apresentado na Figura 1, onde podem ser observadas usinas muito próximas geograficamente agregadas em *clusters* diferentes e usinas mais distantes num mesmo *cluster*.

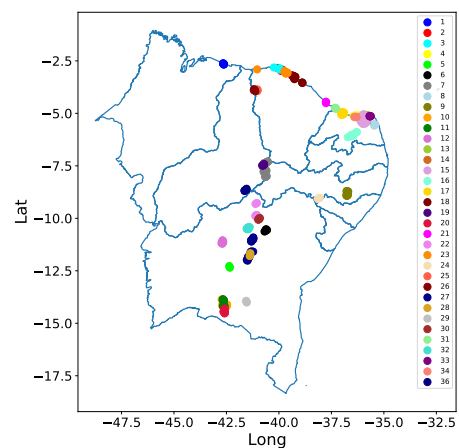


Figura 1. Clusterização por pontos de conexão.

Adicionalmente, as previsões dessa fonte de geração compõem o processo da programação diária da operação, que atualmente adota o modelo DESSEM (Modelo de Despacho Hidrotérmico de Curto Prazo), desenvolvido pelo CEPEL (Centro de Pesquisas de Energia Elétrica). Neste modelo, tal previsão é representada por agregações espaciais do montante de geração acumulado nas barras da rede elétrica, ou seja, *clusters* de usinas conforme a conexão à rede básica.

Em complemento, são esperados com os estudos realizados neste trabalho ganhos de qualidade nos resultados do planejamento da operação energética de médio e longo prazo, considerando o aperfeiçoamento na representação dessas fontes nos modelos DECOMP (Modelo de Determinação da Coordenação da Operação a Curto Prazo) e NEWAVE (Modelo de Planejamento da Operação de Sistemas Hidrotérmicos Interligados de Longo e Médio Prazo). Isto pode ser alcançado com o uso de agregações espaciais de acordo com o comportamento dos ventos no país. Desta forma, propor uma substituição à atual abordagem, que considera uma previsão baseada em fatores de capacidade apurados dos últimos cinco anos que precedem o PMO (Programa Mensal da Operação), computados por agregações espaciais em nível de subsistema. Conforme a Resolução Normativa nº 843/2019 ANEEL (2019).

Neste sentido, este trabalho busca trazer um novo método de clusterização, adaptado ao caso brasileiro, complementando o estado da arte em relação a este tema e contri-

buindo significativamente para o planejamento e operação do ONS no aspecto de geração eólica. Será aplicado o Modelo de Mistura de Gaussianas ou, em inglês, *Gaussian Mixture Models* (GMM) inicializado pelo *K-means* (*K*-médias). Todos os métodos e modelos foram implementados utilizando a linguagem Python, aplicando dados reais do ONS, das regiões Nordeste e Sul do Brasil.

2. REFERENCIAL TEÓRICO

Nesta seção são apresentados os métodos utilizados, neste artigo, para realizar a clusterização das usinas eólicas.

2.1 Método *k-means*

O *K-means* é um método de agrupamento que pode ser utilizado para a classificação em aplicações não-supervisionadas. O *k* neste método representa o número de *clusters*, ou grupos, em que deseja agregar o dado e é um hiperparâmetro do processo. O algoritmo compreende 4 passos principais, (Svensén e Bishop, 2007): (i) Determinar as posições iniciais dos centroides dos *k clusters*; (ii) Alocar cada elemento ao *cluster* com centroide mais próximo; (iii) Recalcular os centros dos *clusters* a partir da média dos elementos alocados; (iv) Repetir os passos de 2 a 4 até a convergência segundo algum critério.

Dentre os critérios de convergência, os mais comuns são executar o algoritmo até que se atinja um número máximo de iterações ou os centroides deixem de se alterar significativamente. Esta significância é determinada previamente, representando um erro suficientemente pequeno que seja aceitável (Svensén e Bishop, 2007).

2.2 Método *GMM*

O Modelo de Misturas Gaussianas é outro método de aprendizagem não supervisionado, isto é, opera em dados não rotulados e sem conhecimento prévio das categorias presentes. Esse tipo de aprendizado é adequado a situações nas quais, coletar e rotular um grande conjunto de exemplos, pode se caracterizar como uma tarefa árdua ou mesmo inviável. Todavia, é necessário entender como aqueles exemplos se relacionam de modo a extrair informações úteis a seu respeito.

Ao contrário do *K-means*, a clusterização por *GMM* não visa encontrar apenas os centroides dos agrupamentos, mas sim representá-los segundo um arcabouço probabilístico mais complexo. Ao entender o dado como uma amostra advinda de um conjunto de distribuições normais multivariadas, este método é capaz de estimar agrupamentos com formatos distintos, não somente bolas.

Uma mistura gaussiana é definida por meio de uma função de densidade de probabilidade (Svensén e Bishop, 2007)

$$p(x) = \sum_{k=1}^K \pi_k \mathcal{N}(x | \mu_k, \Sigma_k). \quad (1)$$

A partir de uma amostra aleatória x , é assumido que elas são geradas por uma função densidade na qual: π_k são os coeficientes de mistura (necessariamente positivos

e somando 1); K representa o número de componentes da mistura; \mathcal{N} a distribuição de probabilidade gaussiana multivariada com média μ_k e matriz de variâncias e covariâncias Σ_k (Svensén e Bishop, 2007).

A estimação dos hiperparâmetros pode ser realizada por intermédio do algoritmo *Expectation Maximization* (EM), que visa a maximização de verossimilhança. Em suma, o EM é utilizado para encontrar estimadores consistentes e eficientes de parâmetros de modelos estatísticos nos casos em que as equações não podem ser resolvidas analiticamente.

O algoritmo EM necessita de uma inicialização dos parâmetros de cada normal, isto é, médias μ_k e matrizes de covariância Σ_k , bem como dos pesos de mistura. Uma possível inicialização dos parâmetros das normais consiste em uma execução preliminar do *K-means* e, subsequente, estimação de μ_k e Σ_k para cada conjunto, enquanto os pesos de mistura podem ser simplesmente assumidos iguais entre si.

A partir desta inicialização as iterações do EM são executadas até que se atinja um critério de convergência estabelecido, similar ao demonstrado por Silva e Dourado (2018).

2.3 Análise de Componentes Principais

É raro que aplicações de clusterização e reconhecimento de padrões operem diretamente sobre uma base de dados bruta, devido à possível alta dimensão e elevada redundância entre variáveis, características prejudiciais ao resultado dos algoritmos. Desta forma, são aplicados métodos para extração de características (do inglês, *feature extraction*) visando compactar o dado num conjunto de menor dimensão onde cada variável contém uma informação única.

O método mais tradicional para este fim é a Análise de Componentes Principais (*Principal Component Analysis* - *PCA* em inglês) ou transformação de Karhunen-Loève (Theodoridis, 2009). Em sua natureza, o *PCA* não é um método de compactação do dado, mas sim uma transformação ortogonal do mesmo para um sistema de coordenadas onde a projeção sobre eixos subsequentes apresenta cada vez menor variância, (Zhao *et al.*, 2019). Assim, ao projetar o dado sobre as primeiras coordenadas obtém-se uma representação de menor dimensionalidade na qual cada variável retém uma informação única atrelada a si.

A compactação através do *PCA* é expressa por:

$$X = F \times \Lambda' + \epsilon, \quad (2)$$

X é a matriz de dados $n \times p$ centralizada, sendo n o número de observações e p o número de variáveis originais; F é uma matriz $n \times k$ de componentes principais, contendo o dado projetado sobre as k primeiras dimensões da transformação ortonormal; Λ é uma matriz $k \times p$ denominada matriz de carregamentos e ϵ representa um erro aleatório de reconstrução do dado.

A estimação da matriz Λ , e subsequentemente F , pode ser executada por diversos métodos. O mais comum é baseado na decomposição espectral da matriz de variâncias e covariâncias amostral S , tal como:

$$S = \frac{1}{n} X'X = PDP', \quad (3)$$

sendo P uma matriz $p \times p$ contendo os autovetores de S em cada coluna e D uma matriz diagonal com os autovalores correspondentes. A matriz Λ corresponde às k primeiras colunas de P .

Por fim, é necessário estipular algum critério para eliminação de componentes. Um produto interessante da estimação do PCA da decomposição espectral é o cálculo dos autovalores da matriz S , correspondentes à variabilidade em cada eixo da transformação ortogonal e cuja soma representa a *variação total* do dado. Com isso é possível mensurar diretamente o quanto da variação total se perde com cada componente removida, buscando um compromisso entre compactação do dado e perda de informação.

3. DESENVOLVIMENTO METODOLÓGICO

O ONS tem o papel de coordenar a geração de energia em todo país e, em vista disso, o objetivo da aplicação demonstrada nesta seção é obter uma clusterização dos parques eólicos, de forma que propicie uma previsão de geração de energia eólica mais assertiva. Para tal, serão realizados agrupamentos considerando perfis de velocidade do vento em cada usina.

A seguir será exposta a metodologia desenvolvida para obtenção de agrupamentos alternativos àqueles utilizados atualmente pelo Operador, considerando semelhanças de comportamento da velocidade do vento. Para facilitar a compreensão do algoritmo proposto, a Figura 2 representa o processo iterativo de clusterização desenvolvido para solucionar o problema do caso brasileiro.

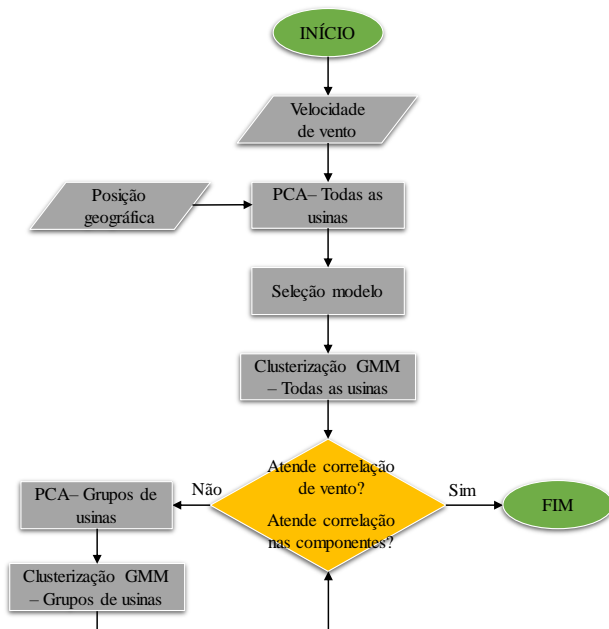


Figura 2. Fluxograma do Algoritmo Proposto.

O algoritmo deve ser inicializado com a entrada dos dados de velocidade de vento no formato descrito na Seção 4.1, isto é, um dado de séries temporais no qual

cada coluna representa uma usina e cada linha uma observação de velocidade do vento no tempo. Estes dados são processados preliminarmente através da Análise de Componentes Principais (PCA).

Tradicionalmente, estudos de clusterização buscam encontrar conjuntos de observações similares entre si. No estudo que será apresentado na seção 4, isto corresponde a encontrar grupos de registros históricos (linhas do dado) similares, enquanto nosso verdadeiro objetivo é identificar similaridade entre as colunas. Analisar o dado transposto seria pouco tratável, uma vez que este teria 33600 colunas e apenas 407 linhas, ver seção 4.1.

Propõe-se assim um uso menos ortodoxo do PCA: a classificação será realizada sobre a matriz de carregamentos Λ e não sobre as componentes principais extraídas. Recordando a exposição na Seção 2.3, as observações de uma dada usina são reconstruídas a partir de combinações lineares da linha correspondente em Λ . Assim, é possível interpretar as linhas como vetores k -dimensionais compactando o perfil de vento das usinas. Ainda, pela natureza da decomposição espectral, estes perfis carregam também informação multivariada da relação entre usinas.

Após a aplicação do PCA nas variáveis de velocidade de vento são adicionadas, artificialmente, mais duas componentes, representando as coordenadas geográficas das usinas eólicas. Desta forma, o modelo não só realizará a clusterização pelo comportamento do vento, mas também levando em conta a distância física entre as usinas.

Dado que os valores de coordenadas são, em magnitude, muito diferentes daqueles dos carregamentos na matriz Λ , é necessário escaloná-las para que seus limites se encaixem naqueles das componentes principais a partir da seguinte transformação:

$$CoE = \min[\Lambda] + \frac{\max[\Lambda] - \min[\Lambda]}{\max[Co] - \min[Co]} \cdot (Co - \min[Co]), \quad (4)$$

CoE representa os valores das coordenadas escalonadas; Λ é a matriz de carregamentos selecionada; Co são as coordenadas originais. As funções \max e \min destacam funções que capturam os valores máximos e mínimos dos vetores.

O processo pode ser descrito em duas etapas:

1ª etapa Esta etapa é responsável por preparação dos dados de vento, com a aplicação do PCA, inserção das componentes geográficas e agregação inicial das usinas. Esta primeira camada de agrupamento consiste na aplicação do GMM aos dados de perfil de vento e posição geográfica. O número de *clusters* é selecionado através de um processo de varredura no qual são testados valores em uma faixa, computando-se o BIC (*Bayesian Information Criterion*) em cada caso e selecionando ao final o número de *clusters* correspondente ao menor valor do critério.

O *Bayesian Information Criterion* (BIC) é um índice utilizado pela estatística bayesiana para seleção de modelos e também é conhecido como critério de informação Schwarz, sendo intimamente relacionado ao critério de informação de Akaike (AIC) (Burnham e Anderson, 2004).

2ª etapa Já nesta etapa, os *clusters* que não atenderem determinados critérios de qualidade são fragmentados novamente. Três critérios são utilizados e devem ser atendidos em conjunto para que o *cluster* seja considerado apto. Os dois primeiros são referentes a comportamentos esperados da matriz de correlação entre históricos de vento das usinas alocadas a um mesmo *cluster*: todos os valores na matriz devem ser superiores a 0,6 e ao menos 20% deles superiores a 0,7. Os cálculos dos dois primeiros critérios são expressos por:

$$\begin{aligned} \text{quantil} \left[C_k^{[NU_k \times NU_k]}, V_Q \right] &\geq V_c, \\ \text{e} \\ \min \left[C_k^{[NU_k \times NU_k]} \right] &\geq V_{cm}, \end{aligned} \quad (5)$$

C_k representa a matriz de correlação do *cluster* k ; NU_k é o número de usinas do *cluster* k ; V_Q é o valor do quantil de correlação aceitável, hiperparâmetro de entrada com valor igual a 0,2; V_c é o valor de limiar de correlação aceitável, hiperparâmetro de entrada com valor igual a 0,7; V_{cm} é o valor de de correlação mínima aceitável, hiperparâmetro de entrada com valor igual a 0,6.

O terceiro critério é baseado no cálculo da média interna das componentes do PCA adicionadas as coordenadas, ou seja, os dados de entrada do GMM, de cada *cluster* gerado na primeira etapa. De posse da média interna dos *clusters*, é então realizado o cálculo da correlação das componentes de cada usina do *cluster* com a média. O *cluster* que não atende uma correlação mínima é considerado inapto e deverá ser fragmentado por mais uma vez, conforme:

$$\min \left[C_k^{[NU_k \times 1]} \right] \geq V_c, \quad (6)$$

onde C_k representa o vetor de correlação do *cluster* k ; NU_k é o número de usinas do *cluster* k ; V_c é o valor de correlação aceitável, com o mesmo valor do critério 1.

Para os *clusters* considerados inaptos o modelo GMM deverá ser aplicado novamente. Primeiramente, dividindo o *cluster* em dois *subclusters*. Caso algum dos *subclusters* não atenda o critério de parada, o número *subclusters* é novamente aumentado. Esse processo iterativo é continuado até que todos os *subclusters* são considerados aptos.

4. ESTUDO DE CASO E RESULTADOS

Os resultados apresentados para o estudo de caso, serão pautados no conjunto de usinas eólicas presentes no território brasileiro, especificamente nas regiões Sul e Nordeste, que apresentam a quase totalidade da capacidade instalada eólica do Brasil.

4.1 Base de dados

Os dados observados/medidos da velocidade do vento utilizados neste trabalho são oriundos de uma base de dados real utilizada pelo ONS. Esses dados foram obtidos do Sistema de Supervisão e Controle do ONS e do Sistema de Acompanhamento das Medições Anemométricas da Empresa de Pesquisa Energética (EPE). Em geral, esses dados possuem um número elevado de falhas, que nesta situação, são aplicados vários processos de tratamento.

Maiores detalhes poderão ser obtidos em ONS (2018). O conjunto de dados corresponde às séries de vento semi horárias das 487 usinas eólicas brasileiras, das quais 407 se encontram no Nordeste e 80 no Sul do Brasil. Esta base de dados foi coletada no período entre 01/01/2019 e 11/12/2020 pelo ONS, contendo assim um total de aproximadamente 33600 observações por usina.

A Figura 3 apresenta um gráfico da correlação das séries de vento entre cada par de usinas do Nordeste, seccionado por estados. Para facilitar a visualização, apenas correlações superiores a 0,7 são representadas em uma escala de cores indicando módulo de correlação. É notável que mesmo dentro de um estado em particular a correlação entre usinas pode ser baixa.

É possível, ainda, observar a existência de correlação entre usinas de estados diferentes, por exemplo, usinas do estado da Bahia possuem correlação com usinas do Piauí em determinada intensidade, no entanto, não possuem correlação com os estados do Ceará, Maranhão e Pernambuco.

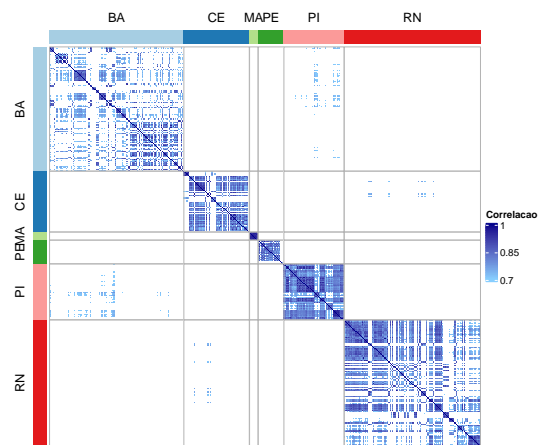


Figura 3. Correlação entre usinas eólicas do Nordeste do Brasil.

Na Figura 4 encontra-se o gráfico da dispersão geográfica das usinas eólicas no Nordeste do Brasil. Por este gráfico nota-se que a maior parte das usinas eólicas se encontram no interior do estado da Bahia e no estado do Rio Grande do Norte.

4.2 Aplicando Análise de Componentes Principais

A primeira etapa do modelo consiste na decomposição em componentes principais dos perfis de vento. As regiões Sul e Nordeste são tratadas separadamente, em função da baixa correlação entre elas frente às correlações internas de cada uma. A Figura 5 ilustra o comportamento da variância acumulada para os dados das 407 usinas eólicas do subsistema Nordeste, aplicando o método de componentes principais para redução do número de componentes.

O número de componentes foi escolhido de forma que explique ao menos 70% da variância total. Para o Nordeste 5 componentes são suficientes para explicar 70% da variância dos dados. Assim passamos a ter uma matriz de dados de 407×5 . Realizando o escalonamento das coordenadas e adicionando-as à matriz de componentes original, temos uma matriz de 407×7 para o classificador.

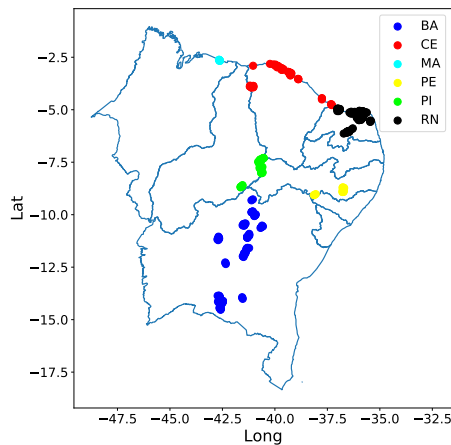


Figura 4. Dispersão eólicas do nordeste do Brasil.

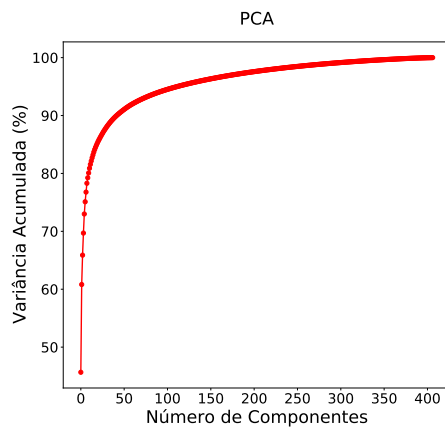


Figura 5. Variância acumulada por componentes do Nordeste.

4.3 Classificação da 1ª Etapa

Aplicando o processo de seleção do número de *clusters*, conforme descrito na seção da 1ª etapa, o número de 14 *clusters* retornou o melhor BIC. A Figura 6 apresenta os resultados da classificação de 1º nível com a definição dos 14 *clusters*.

O processo de clusterização da primeira etapa, ilustrado na Figura 6, captura os comportamentos físicos de vento similares, em níveis mais gerais, das regiões do subsistema Nordeste. Esta etapa gera *clusters* que carregam informações importantes e podem ser utilizados em modelos de previsão ou geração de cenários com foco nos modelos de planejamento, isto é, modelo de curto prazo, o DECOMP e médio-longo prazo, o NEWAVE, agregando evoluções a outros estudos em andamento no setor eletro-energético.

A Figura 7 apresenta a correlação dos dados de velocidade de vento, classificados e ordenados por *cluster*. Como pode ser observado, alguns *clusters* apresentam uma esparsidade considerável, indicando uma necessidade de reclusterização.

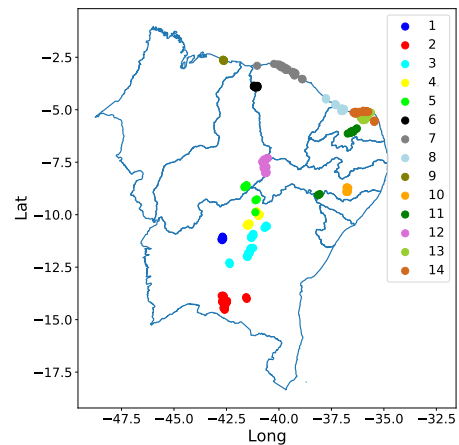


Figura 6. Clusterização eólicas do Nordeste - Etapa 1.

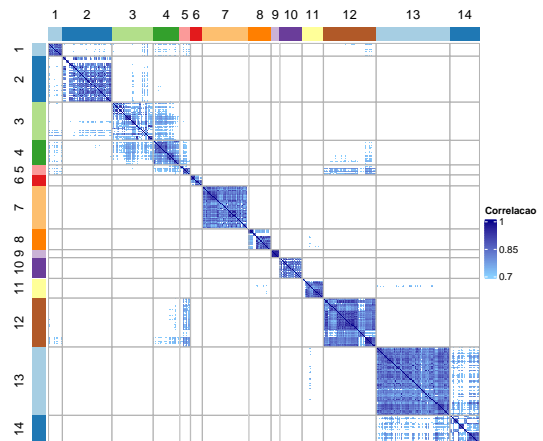


Figura 7. Correlação eólicas do Nordeste - Etapa 1.

A Figura 8 apresenta o comportamento, em duas dimensões, das componentes do *cluster* 11 após a primeira etapa de clusterização. Devido a característica dos dados das usinas desse *cluster*, as 5 componentes utilizadas apresentaram similaridade, sendo alocadas usinas pertencentes aos estados do Rio Grande do Norte e de Pernambuco, com distância considerável em um mesmo *cluster* (ver últimas duas componentes e Figura 6). Todavia, a adição das componentes de localização fará com que a correlação das componentes das usinas com a média interna seja menor que o critério exige e, portanto, esse *cluster* será particionado novamente.

4.4 Classificação da 2ª Etapa

De fato, nem todos os 14 *clusters* atendem aos critérios apresentados na seção metodológica. A próxima etapa consiste numa reclusterização na qual será aplicada novamente a redução em componentes principais para cada *cluster*. Então é refeita a clusterização até que todos os *clusters* atendam os critérios das Equações 5 e 6.

A Figura 9 apresenta os resultados gráficos da classificação após o atendimento de todos os critérios para cada *cluster*. Sendo a notação, $N.n$, com N clusterização da primeira etapa e n clusterização da segunda etapa.

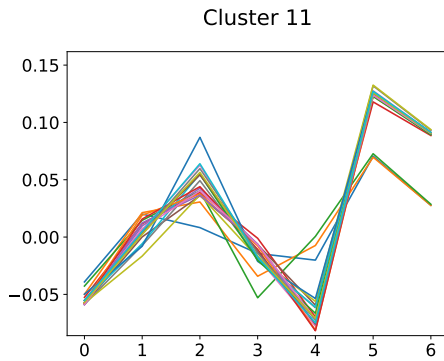


Figura 8. Comportamento das componentes *cluster 11*.

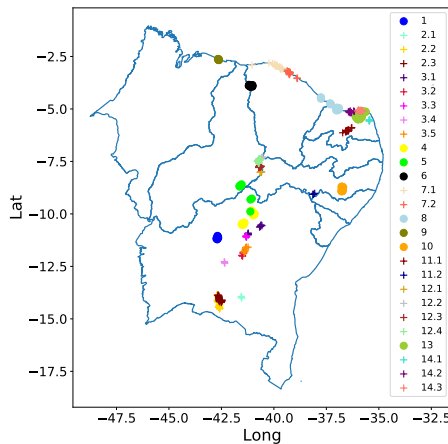


Figura 9. Clusterização eólica do Nordeste - Etapa 2.

O processo de clusterização da segunda etapa, ilustrado na Figura 9, apresenta refinamentos dos *clusters* da primeira etapa, trazendo informações mais localizadas das regiões em relação ao comportamentos dos ventos. Esta etapa gera informações que serão a base do processo de previsão da geração eólica utilizada na programação diária e no modelo DESSEM pelo ONS, podendo substituir a clusterização por pontos de conexão atualmente utilizada pelo modelo oficial de previsão eólica do ONS, conforme ONS (2018).

Para uma avaliação dos benefícios da metodologia, é feita uma análise aprofundada a respeito da clusterização por pontos de conexão, comparada aos resultados encontrados neste trabalho. Foi realizado uma ampliação de dois agrupamentos por ponto de conexão do estado da Bahia que se encontram muito próximos: os conjuntos 27 e 28, que representam os pontos de conexão à rede básica MORRO CHAPEU2 - 230 kV e MORRO CHAPEU2 - 69 kV, respectivamente. Estes *clusters* estão representados pelas cores azul e amarelo e são apresentados na Figura 10. Analisando a Figura 10, nota-se que as usinas representadas pelos pontos amarelos estão próximas de algumas usinas em azul ao passo que outras usinas em azul ao norte estão mais distantes deste grupo. Esta situação é um exemplo daquilo que impulsionou esse trabalho, motivando a busca por uma agregação diferente da utilizada atualmente.

Extraindo da Figura 10 três usinas eólicas, sendo duas usinas do *cluster* em azul denominadas usinas *A* e *C*, uma

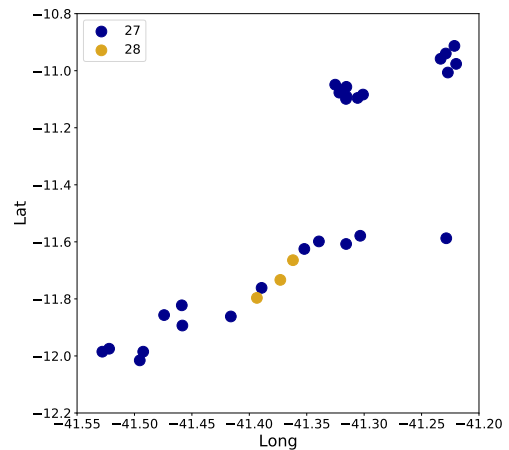


Figura 10. Pontos de conexão 27 e 28 - interior da Bahia.

mais ao leste e outra mais a oeste, e a usina central do *cluster* amarelo denominada *B*. Percebe-se que a usina pertencente ao *cluster* caracterizado pela cor amarela está entre duas usinas caracterizadas pela cor azul.

Pela Tabela 4.4, que apresenta a correlação das três usinas, é possível concluir que, a correlação das usinas *A* com a *B* e *B* com a *C*, de *clusters* diferentes, são maiores do que a correlação entre as usinas *A* e *C* de mesmo *cluster*.

Tabela 1. Correlação das três usinas escolhidas

Usinas	Correlação
A-B	0.687711
A-C	0.592313
B-C	0.721921

Tomando agora os resultados do modelo proposto de clusterização, o comportamento de usinas do *cluster 28* imersas dentro do *cluster 27* não ocorre. Na verdade, devido à similaridade de vento e distância entre as usinas, o modelo classificador encontrou 4 conjuntos para esse grupo de usinas conectadas aos pontos de conexão MORRO CHAPEU2 - 230 kV e MORRO CHAPEU2 - 69 kV. O resultado ampliado da Figura 6 para esse caso pode ser visualizado na Figura 11.

Diante do exposto, nota-se a maior aderência na representação em conjuntos dessas usinas. A agregação de forma a considerar o perfil de vento torna o processo de criação das curvas Vento x Potência por *cluster* mais eficiente, conforme necessidade do modelo de previsão adotado pelo ONS, descrito em detalhes em ONS (2018). Dessa forma, a metodologia desse artigo pode contribuir no processo de previsão de geração eólica atualmente utilizado pelo ONS.

4.5 Classificação das usinas do subsistema Sul

Os comportamentos apresentados neste trabalho para o subsistema Nordeste replicam-se ao subsistema Sul, entretanto com menos intensidade, uma vez que, o número de usinas eólicas nesse subsistema é bem mais baixo. Atualmente, para o subsistema Sul, o ONS observa 12 pontos de conexão que possuem usinas eólicas conectadas. A Figura 12 apresenta os resultados da classificação encontrada para o subsistema Sul.

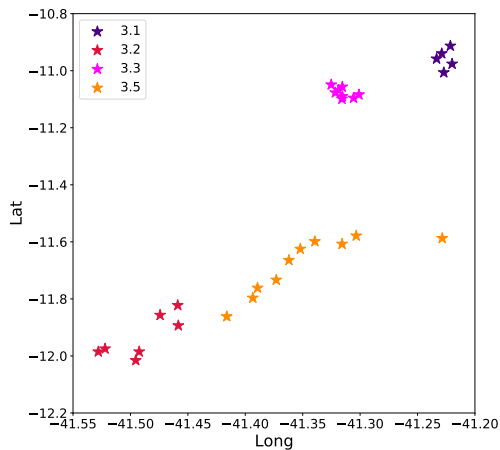


Figura 11. Clusters de ordem 3 - interior da Bahia.

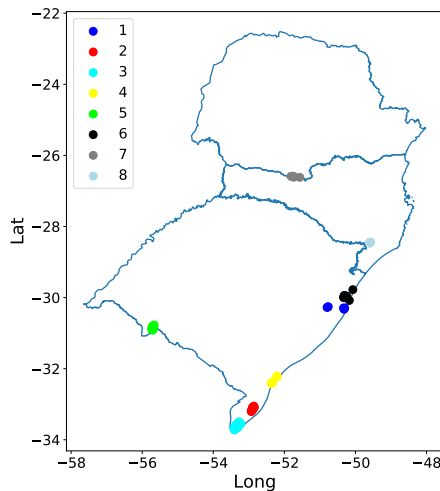


Figura 12. Clusterização eólica do Sul - Etapa 2.

O número de oito *clusters* foi selecionado pelo critério de informação BIC. Sendo que, esse número foi suficiente para que todos os *clusters* da primeira etapa atendessem os critérios estabelecidos nas Equações 5 e 6, conforme lustrado na Figura 12,. Isso ocorre, pois, o subsistema Sul apresenta um número de usinas menor que o Nordeste.

5. CONCLUSÃO

Neste trabalho o método de clusterização GMM foi aplicado com objetivo de agrupar usinas eólicas presentes no território brasileiro, de forma à auxiliar a decisão do planejamento diário de operações do ONS.

A aplicação do GMM, conforme a metodologia desenvolvida, permite obter uma nova forma de agrupamento das usinas eólicas, baseado no perfil de vento e localização das mesmas. Portanto, a principal contribuição deste estudo é a obtenção de uma melhor previsão de geração eólica, auxiliando o planejamento do Operador Nacional do Sistema Elétrico brasileiro.

Adicionalmente à contribuição direta aos modelos de previsão da geração eólica, esta metodologia surte efeito direto na cadeia de planejamento hidrotérmico, uma vez que esta previsão de geração é insumo para o modelo de curto prazo DESSEM. Não obstante, é possível vislumbrar aplicações deste mesmo método, em menor nível de detalhe exigido dos *clusters* finais, para previsão e/ou geração de cenários em médio ou longo prazo, alimentando os demais modelos da cadeia (DECOMP e NEWAVE).

REFERÊNCIAS

- ANEEL (2019). Resolução normativa nº 843. https://www.in.gov.br/materia/-/asset_publisher/Kujrw0TZC2Mb/content/id/70268458. Acesso em 14 de dezembro de 2021.
- Burnham, K.P. e Anderson, D.R. (2004). Multimodel inference: understanding aic and bic in model selection. *Sociological methods & research*, 33(2), 261–304.
- Eltohamy, M.S., Moteleb, M.S.A., Talaat, H., Mekhemer, S.F., e Omran, W. (2019). Analyzing wind power ramps for high penetration of variable renewable generation. In *2019 21st International Middle East Power Systems Conference (MEPCON)*, 768–775. IEEE.
- Lima, D.A. (2017). *Avaliação dinâmica de ilhamento de subsistemas elétricos com geração dispersa interligados à rede de transmissão*. Tese de doutorado, PUC-Rio.
- ONS (2018). Nt 0151-2018 - desenvolvimento metodológico para previsão de geração de fonte eólica. https://sintegre.ons.org.br/sites/6/27/48/paginas/servicos/produtos-pasta.aspx?RootFolder=/sites/6/27/48/Produtos/439/17-07-2019_152242. Acesso em 11 de julho de 2021.
- Silva, H.R.d. e Dourado, J.R. (2018). *Reconhecimento automático de locutor utilizando modelo de misturas gaussianas treinado pelo algoritmo de maximização da expectativa*. Trabalho de conclusão de curso, Universidade Tecnológica Federal do Paraná.
- Sun, Y. e Liu, S. (2017). Dynamic equivalence for wind farm based on ibsa-k-means algorithm. In *2017 2nd International Conference on Power and Renewable Energy (ICPRE)*, 443–446. IEEE.
- Svensén, M. e Bishop, C.M. (2007). *Pattern recognition and machine learning*.
- Teng, W., Wang, X., Meng, Y., e Shi, W. (2019). An improved support vector clustering approach to dynamic aggregation of large wind farms. *CSEE Journal of Power and Energy Systems*, 5(2), 215–223.
- Theodoridis, S. e Koutroumbas, K. (2009). *Pattern Recognition*. Elsevier Inc. All.
- Wu, Q., Zhong, Q., Wang, G., e Li, H. (2016). Multi-machine equivalent approach of wind farm based on fuzzy c-means clustering. *Modern Electric Power*, 33(6), 27–32.
- Yang, H. e Liu, J. (2015). A hybrid clustering algorithm based on grey wolf optimizer and k-means algorithm. *Journal of Jiangxi University of Science and Technology*, 5, 015.
- Zhao, H., Zheng, J., Xu, J., e Deng, W. (2019). Fault diagnosis method based on principal component analysis and broad learning system. *IEEE Access*, 7, 99263–99272.