

Análise de Métodos para Agrupamento de Cargas Elétricas em Redes de Distribuição de Grande Porte usando Sistema de Informação Geográfica

Chumbi Wilson E* C. José Eduardo* Jaramillo-Leon Brian*
B. Leite Jônatas*

* Departamento de Engenharia Elétrica, Universidade Estadual Paulista, Ilha Solteira/SP, Brasil, (e-mails: enrique.quito@unesp.br, jose.chillogalli@unesp.br, brian.jaramillo@unesp.br, jb.leite@unesp.br).

:

Abstract: The growing integration of renewable energies and the new technologies adoption make the coupling between supply and demand a challenge. Furthermore, the allocation of distributed energy resources on the electrical grid allows islands configurations, as microgrids (MG). In a MG, when the consumption exceeds the generation capacity, it leads to an inefficient operation of the system, therefore, it is necessary to know the electrical demand patterns. However, electrical load grouping of a study area into categories is a technique that allows finding consumption relationships within and between groups. This article presents and compares two methods to identify electric load clusters in distribution networks using geographic information systems (GIS), DBSCAN and K-means methods are tested. Electric load density and electrical distance prove to be effective setting factors. The performance of the methods is evaluated in the coverage area of three distribution substations of an Ecuadorian electric utility. The results show clusters, that maximize electrical and geographical proximity, which represents a support tool in the MG planning process.

Resumo: A crescente integração de energias renováveis e a adoção de novas tecnologias tornam o balanço da oferta e demanda um desafio. Além disso, a alocação de recursos energéticos distribuídos na rede elétrica tem potencial de formar regiões ilhadas, como uma micro-rede (MG). Na MG, quando a carga excede a capacidade de geração, há uma operação ineficiente do sistema, logo conhecer os padrões de demanda necessários é fundamental. O agrupamento da carga elétrica de um área de estudo em categorias é uma técnica eficaz que permite encontrar relações de consumo dentro e entre os agrupamentos. Este artigo apresenta e compara a aplicação de dois métodos para identificar agrupamentos de cargas elétricas em redes de distribuição usando sistemas de informação geográfica: DBSCAN; e *k-means*. A densidade de carga e distância elétrica demonstram ser fatores eficazes de ajuste e seleção. O desempenho dos métodos é avaliado na área de cobertura de três subestações de distribuição de uma concessionária de energia Equatoriana. Os resultados apresentam aglomerados de cargas que maximizam a homogeneidade e proximidade tanto elétrica quanto geográfica, o que representa uma ferramenta de apoio no processo de planejamento de MGs.

Keywords: Electric load, DBSCAN method, k-means method, microgrid, distribution electrical network, geographic information system.

Palavras-chaves: Carga elétrica, método DBSCAN, método k-means, micro-rede, rede elétrica de distribuição, sistema de informação geográfica.

1. INTRODUÇÃO

Com o desenvolvimento da tecnologia, os sistemas de energia elétrica estão passando por uma fase de transformação para aproveitar os recursos energéticos distribuídos (DERs). A integração de fontes de geração distribuída (GD) nos sistemas de distribuição (SD) está em franca expansão (Blanco, 2009). A maior parte da capacidade instalada no mundo vem de fontes eólica e solar fotovoltaica. No Brasil, até Julho de 2017, o número de unidades GDs

instaladas foi superior a 15.500 e mais do 97% era do tipo fotovoltaico (de Castro and Dantas, 2017).

O paradigma de modelo centralizado está mudando para um modelo de gerenciamento híbrido centralizado-distribuído. A finalidade do modelo híbrido é aproveitar os GDs e satisfazer a demanda local, cobrindo a diferença de potência com a rede principal. Os GDs podem também melhorar as características da rede fornecendo reservas em condições críticas de falta de energia ou sobrecarga. Os agentes envolvidos na transição destes sistemas de energia descentralizados são apresentados em (Adil and Ko, 2016).

Da perspectiva da rede elétrica principal, um esquema de MG pode ser considerado como um agrupamento de GDs autônomo fechado, onde os recursos e as demandas estão complementados. As MGs podem apresentar vários esquemas de operação baseado no modelo de comercialização. Neste contexto, podem existir MGs isoladas ou conectadas à rede comercializando energia entre MGs e com a rede principal (Che et al., 2017).

Os recursos energéticos distribuídos podem melhorar a qualidade e confiabilidade da energia elétrica, reduzir as perdas técnicas e minimizar custos de operação (Borges, 2012; Kroposki et al., 2013; Song et al., 2017). Porém, uma alta inserção desses recursos em grande escala pode trazer problemas na confiabilidade e operação da rede elétrica. Alguns desses problemas são sobretensões, subtensões, aumento de perdas e sobrecarga nas linhas e transformadores, fluxo reverso de energia, falha de proteção, distorção harmônica, etc. (Zain ul Abideen et al., 2020). A maioria dos problemas estão relacionados à natureza estocástica dos DERs. Uma das principais desvantagens está associada a intermitência dos recursos causando o desbalanço entre geração e carga (Mulenga et al., 2020). Conhecer o comportamento e distribuição da demanda proporciona critérios para definir as MGs e estabelecer as novas topologias de rede no processo de planejamento de MGs. Portanto, um adequado estudo de carga permitirá que as futuras MGs aumentem seu desempenho.

Entre as técnicas não supervisionadas de mineração de dados, os algoritmos de agrupamento estão sendo amplamente utilizados na análise e detecção de padrões (Flach, 2012). Os dados de uso de energia junto com uma abordagem de agrupamento têm sido utilizados para encontrar padrões de consumo (Panapakidis, 2016). Uma metodologia híbrida, combinando o algoritmo *K-means* com uma rede neural foi utilizada para identificar comportamentos na estrutura de dados e prever a radiação solar (Benmouiza and Chekane, 2013). O agrupamento de carga elétrica permite extrair conhecimento e definir estratégias de planejamento da infraestrutura do SD. Uma síntese de algoritmos de agrupamento e as métricas de avaliação são discutidas sob a ótica de aplicações do agrupamento de carga elétrica na rede inteligente (Si et al., 2021). Uma extensão do algoritmo *K-means* é utilizada para resolver o problema de agrupamento de cargas de acordo com as suas características de tamanho e localização geográfica (Yu et al., 2018). Esse agrupamento baseado na distribuição de cargas (densidade), forma parte da estratégia de dimensionamento de cobertura das subestações de distribuição.

Nos últimos anos, os avanços das tecnologias da informação possibilitaram a digitalização de uma grande quantidade de dados e seu gerenciamento através dos sistemas baseados em computador (Lawanson et al., 2018). Estes sistemas de medição, gestão e controle utilizados pelas concessionárias de energia mantêm um registro de ativos da rede. Um componente importante dentro da gestão de ativos da rede é o sistema de informação geográfica (SIG), que permite gerenciar dados georreferenciados e aprimora o monitoramento, análise e visualização do sistema de distribuição (Meehan, 2013). Os novos tipos de dados estão sendo aproveitados tanto para extrair informação útil quanto para simular modelos de rede sofisticados.

Em (Gadelha Filho et al., 2021), foi proposta uma metodologia para encontrar topologias de rede utilizando uma técnica de agrupamento baseado na densidade, na qual se avalia cada agrupamento como uma possível comunidade de energia. Um método de agrupamento baseado em particionamento de grafos foi apresentado em (Zambrano-Asanza et al., 2021), os blocos de cargas determinados podem ser usados para planejar a expansão e topologia de rede dos sistemas de distribuição. A revisão da literatura mostra que poucas propostas adotam técnicas de agrupamento de carga elétrica para achar comunidades de energia. A densidade de carga elétrica é um indicador direto da intensidade e distribuição geográfica da demanda de uma localidade. O SIG possui tecnologias que permitem adicionar o conhecimento espacial como a densidade de carga na avaliação de projetos de energia (Hable, 2009; Gadelha Filho et al., 2021). Por tanto, este trabalho apresenta e compara a aplicação de dois métodos, DBSCAN e *K-means* para identificar agrupamentos de carga elétrica em redes de distribuição usando o SIG.

Neste artigo propõe-se determinar um ajuste adequado de parâmetros para melhorar significativamente os resultados dos algoritmos de agrupamento de cargas elétricas em redes de distribuição reais. Além disso, avalia-se o desempenho das técnicas de agrupamento DBSCAN e *K-means* combinado com dados de carga elétrica usando o SIG.

O artigo está organizado da seguinte forma. Na seção 2, descreve-se a teoria dos métodos de agrupamento DBSCAN e *K-means*, respectivamente. A seção 3 apresenta o caso de estudo que é uma rede de distribuição no Equador. Na seção 4, apresenta-se uma abordagem da técnica de distância euclidiana, uma métrica utilizada para avaliar a qualidade do particionamento. Na seção 5 são apresentados os resultados obtidos. Por último, as conclusões e propostas de trabalhos futuros são apresentadas na seção 6.

2. MÉTODOS DE AGRUPAMENTO

Os métodos de agrupamento são técnicas de classificação que permitem organizar um conjunto de dados em categorias com base em um atributo específico. Pretende-se que dentro de cada grupo a semelhança de seus elementos seja maximizada.

2.1 DBSCAN

Os métodos de agrupamento baseado na densidade têm a funcionalidade de analisar dados espaciais. DBSCAN (*Density Based Spatial Clustering of Applications with Noise*), OPTICS (*Ordering Points to Identify Clustering Structure*), DENCLUE (*DENSITY CLUSTERING*) são alguns desses métodos. Neste trabalho o DBSCAN é o algoritmo utilizado para lidar com o agrupamento espacial. A ideia principal do algoritmo é identificar regiões de alta densidade de pontos que estão separadas das outras regiões que possuem uma baixa densidade. Algumas das vantagens do DBSCAN em relação a outras técnicas são:

- Não há necessidade de prever o número de agrupamentos, a própria noção de densidade apresenta uma guia de ajuste;

- É capaz de identificar agrupamentos de silhuetas arbitrárias, imediatamente valorizadas quando se trata de dados com referência geográfica;
- É robusto a *outliers* e na hora de priorizar a variabilidade em termos de densidade é possível mesclar com outros *clusters* se forem semelhantes;
- O ajuste de parâmetros é aberto para usuários com conhecimento sobre os dados considerados.

O DBSCAN exige dois parâmetros: o raio da vizinhança de um ponto (*eps*) e o número mínimo de pontos (*min_samples*) para que uma região seja considerada como densa. Na Fig. 1, apresentam-se alguns dos tipos de pontos identificados pelo algoritmo. Se a vizinhança de um objeto *p* contém ao menos um número mínimo de objetos igual a *min_samples*, então o objeto *p* é chamado de ponto central. Um ponto periférico tem menos objetos do que *min_samples* em sua vizinhança, mas fica na vizinhança de um ponto central. Um ponto periférico ainda faz parte do agrupamento. *Outliers* são pontos isolados que não são nem central nem ponto periférico (Ester et al., 1996).

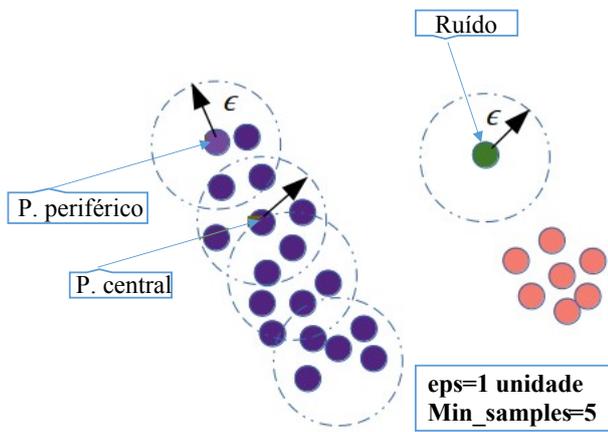


Figura 1. Abordagem dos parâmetros que formam parte do método DBSCAN.

2.2 K-Means

O algoritmo de agrupamento *K-means* procura uma configuração *C* ótima definida em (1), dividindo um conjunto de dados em *k* grupos não sobrepostos.

Seja $X = \{x_i\}$, $i = 1, 2, 3, \dots, n$ um conjunto de dados *n*-dimensional e o parâmetro único *k*. Cada grupo C_j engloba um conjunto de amostras tal que a soma das distâncias euclidianas seja minimizada.

$$C = \bigcup_{j=1}^k C_j \quad (1)$$

O algoritmo segue um processo iterativo, que começa localizando aleatoriamente o núcleo de cada grupo m_j . E atribui cada uma das amostras x_i , ao núcleo mais próximo calculando a soma das distâncias euclidianas $d_2(x_m, m_j)^2$, conforme (2).

Em cada iteração *t*, a re-atribuição das amostras ao núcleo é feito de acordo com (3) e uma atualização da localização

dos núcleos é feito através de (4) até atingir a convergência. Em cada iteração procura-se melhorar o agrupamento.

$$C^0 = \min_C \sum_{j=k} \sum_{m \in C_j} d_2(x_m, m_j)^2 \quad (2)$$

$$C_j^t = \{x_i | d_2(x_i, m_j^t) \leq d_2(x_i, m_{j^*}^t), \forall j^* = 1, 2, \dots, k\} \quad (3)$$

$$m_j^t = \frac{1}{|C_j^t|} \sum_{x_j \in C_j^t} x_j \quad (4)$$

O algoritmo de agrupamento *K-means* está baseado em dois fundamentos: 1) que o núcleo do agrupamento seja o seu centro de massa e, 2) que a distância de cada elemento ao seu núcleo é menor do que a qualquer outro (VanderPlas, 2016). Observe que, qualquer umas das métricas pode ser usado para calcular a distância entre dois vetores *n*-dimensionais: distância euclidiana, distância de Manhattan, distância de Chebyshev, distância de Minkowski, etc. (Hossain, 2017). Na Fig. 2, exemplifica-se a maneira como um conjunto de dados é organizado em 3 grupos de acordo a distância euclidiana.

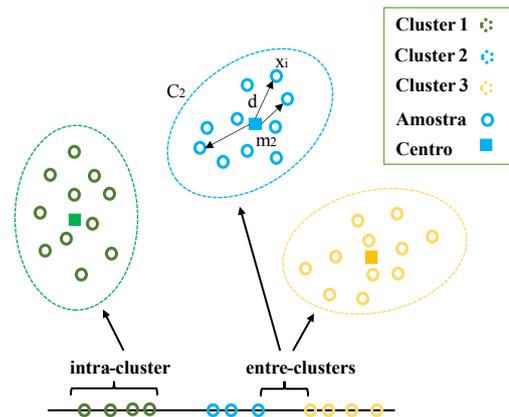


Figura 2. Identificação de três agrupamentos baseado no método *K-means*.

2.3 Pontuação da silhueta

Os métodos de agrupamento classificam objetos em grupos homogêneos. Dentro de cada agrupamento a variância tende a ser minimizada, porém maximizada entre os agrupamentos. A qualidade do agrupamento pode ser medida através da pontuação da silhueta (Rousseeuw, 1987). Uma pontuação que avalia quão bem os dados são agrupados, comparando a semelhança no seu próprio *cluster* e os outros *clusters*. A pontuação varia na faixa $[-1, 1]$, onde, 1 significa que existe uma boa semelhança entre os elementos do próprio *cluster* e fraca semelhança entre os *clusters* vizinhos, quanto maior a pontuação, melhores os desempenhos. A silhueta pode ser calculada utilizando as métricas de distância euclidiana e distância de Manhattan para avaliar a qualidade dos *clusters*. Em (5) apresenta-se a expressão para pontuação da silhueta $S(i)$.

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (5)$$

$$a(i) = \frac{1}{|C(I)|-1} \sum_{j \in C(I), i \neq j} d(i, j) \quad (6)$$

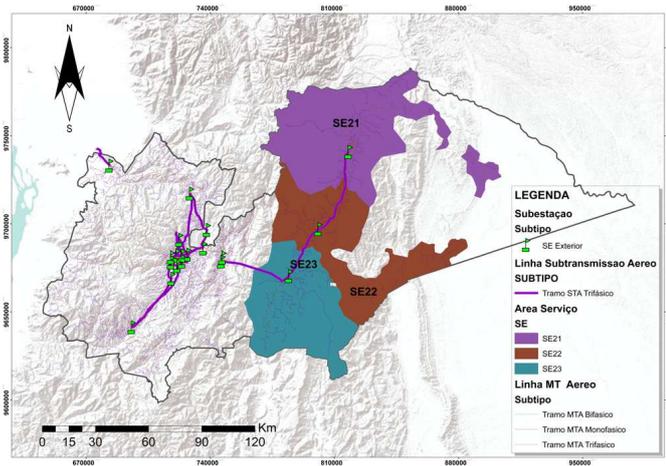
$$b(i) = \frac{\min_{J \neq I} 1}{|C(J)|} \sum_{j \in C(J)} d(i, j) \quad (7)$$

Onde, $|C(I)|$ é o número de pontos pertencentes ao *cluster* $C(I)$, e $d(i, j)$ calcula a distância entre os pontos i, j no *cluster* $C(I)$. $a(i)$ é a distancia média entre i e todos os outros pontos no mesmo *cluster* e pode ser considerado como um quantificador de semelhança. Por outro lado, $b(i)$ calcula a dissimilaridade do ponto i para algum *cluster* $C(J)$, e, é expressada como a distância média entre i e todos os pontos em $C(J)$. O operador $\min\{\cdot\}$ em (7) garante ser tomada a distância média entre i e todos os outros pontos no próximo *cluster* mais próximo $C(J) \neq C(I)$.

Existem outros critérios para avaliação dos agrupamentos, entre eles estão o critério *Calinski&Harabasz* e o critério *Davies-Bouldin* (Liu et al., 2010). *Calinski&Harabasz* é calculada como a razão da soma da dispersão entre *clusters* e da dispersão inter-*cluster* para todos os *clusters*. Neste critério, procuram-se altas pontuações. O *Davies-Bouldin* calcula a semelhança média entre os *cluster*, valores baixos representam uma melhor separação entre os *clusters*.

3. ESTUDO DE CASO

O sistema de estudo considerado foi obtido da *Empresa Elétrica Regional Centro Sur C.A. (CENTROSUR)*, uma concessionária de energia elétrica no Equador. Os pontos de carga utilizados pertencem as subestações 21, 22, 23 localizadas nos municípios de Macas, Méndez e Limón respectivamente. A Fig. 3 apresenta o sistema de estudo da CENTROSUR.



trabalhar com pacotes e módulos, permitindo a troca de informações e funcionalidades de forma simples e eficiente. A conexão entre Python e Cymdist foi possível através do pacote cymy. Os algoritmos de agrupamento utilizados foram do pacote scikit-learn (Pedregosa et al., 2011).

Os dados de carga com suas características de localização geográfica e potência coletados são extraídos do Cymdist. Por questões de simplificação ilustrativa, escolhe-se o alimentador 0421 que pertence à subestação 21 para descrever suas características. A distribuição de 251 pontos de carga do alimentador mostra-se na Fig. 4, cujos eixos x e y representam a latitude e longitude, respectivamente.

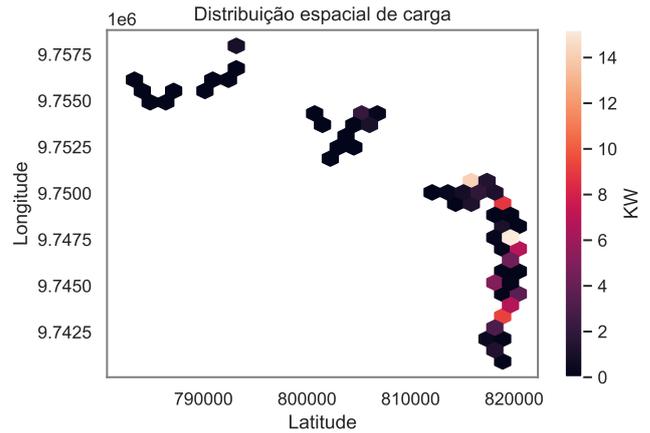


Figura 4. Mapa de distribuição de carga elétrica de um alimentador de teste.

O processo de agrupamento está dividido em três passos: 1) Pré-processamento de dados; 2) Ajuste dos algoritmos DBSCAN e *K-means* com base no conhecimento especializado; e 3) Avaliação da qualidade dos agrupamentos.

3.1 Pré-processamento de Dados

Esta etapa de processamento envolve fundamentalmente procedimentos de transformação de dados, como dimensionamento e normalização. Um mecanismo eficaz para mitigar a carga de armazenamento e processamento de dados.

Os pontos de carga estão definidos por mais de um componente de natureza distinta, os quais tem diferentes unidades de medida. Por exemplo, coordenadas geográficas e unidades de potência elétrica. Uma vez que os dados são multidimensionais, cada um dos componentes devem ser levados para uma escala comum. A padronização das dimensões é feita removendo a média e escalonando para uma variância unitária de acordo com (Sebastian Raschka, 2014).

3.2 Ajuste dos Algoritmos de Agrupamento

Algoritmo DBSCAN: o fato de o DBSCAN ser um algoritmo de busca de vizinhança é necessário definir o número mínimo de amostras que criem os pontos centrais do agrupamento. Além disso, ao depender dessas vizinhanças, a implementação precisa definir o raio da vizinhança, uma distância máxima entre duas amostras para que uma seja considerada próxima da outra.

Figura 3. Área de serviço da CENTROSUR.

A base de dados do sistema está hospedada em uma plataforma Cymdist (CYME, 2022), onde encontra-se a rede de média tensão composta por 11 alimentadores, 546 transformadores de distribuição, e 2831 pontos de carga concentrada, além dos dispositivos de proteção e manobra. As análises foram desenvolvidas nas linguagem de programação Python devido a sua capacidade para

Para um determinado conjunto de dados, o número mínimo de amostras que formam um agrupamento pode ser especificado usando o conhecimento do domínio e familiaridade com os dados. Mas, geralmente, o valor definitivo esta entre quatro e duas vezes a dimensão das amostras, isto é $\{4, 2 \times dim\}$ (Che et al., 2017).

O raio da vizinhança depende de uma função de distância, chamada de gráfico K_{dist} conforme mostrado na Fig. 5, em que a linha em azul apresenta a relação entre o número de amostras n e o raio de vizinhança classificados de acordo com seu comprimento. Seja ε o raio de vizinhança definido como a distância de uma amostra x ao seu n -ésimo vizinho mais próximo. Se classificar as distâncias em ordem crescente, todas as amostras com um valor $K_{dist}(x) < \varepsilon$ são pontos centrais do agrupamento. Então, um agrupamento contém exatamente $n + 1$ amostras para quase todas as amostras. Frequentemente, o ponto escolhido $K_{dist}(x)$ está entre a curvatura máxima e o último ponto da curva do gráfico K_{dist} . Na Fig. 5, a faixa vermelha define o intervalo de valores para ε onde ocorre a curvatura máxima no gráfico K_{dist} . Todas as amostras com um valor $K_{dist}(x) > \varepsilon$ (à esquerda do limite) são considerados *outliers*, enquanto os outros pontos (à direita do limite) são atribuídos a algum *cluster* (Ester et al., 1996).

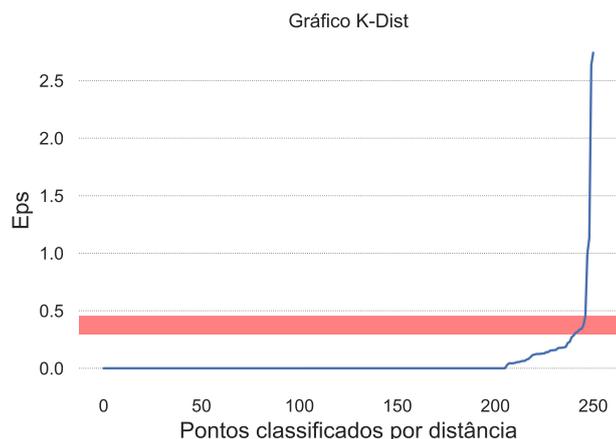


Figura 5. Gráfico K_{dist} para determinar o raio de vizinhança (ε).

Algoritmo K -means: as métricas de distância utilizadas pelo algoritmo identificam de que maneira os dados estão inter-relacionados. Estas métricas aproveitam a capacidade do k -means de fazer quantização vetorial para calcular a similaridade entre os dados multidimensionais. Identificar quais medidas são apropriadas para obter uma função de distância é o principal objetivo.

A função de distância definida em (8) foi utilizada para agrupar os pontos de carga de acordo com sua localização geográfica, distância elétrica, e potência. Para que os resultados sejam comparáveis, o número de agrupamentos é estabelecido de acordo ao número de agrupamentos obtidos pelo DBSCAN.

$$d_2(x_m, m_j)^2 = \sqrt{\frac{(X_{xm} - X_{mj})^2 + (Y_{xm} - Y_{mj})^2}{(Z_{xm} - Z_{mj})^2 + (P_{xm} - P_{mj})^2}} \quad (8)$$

Onde, a localização geográfica de um ponto de carga está definida pelas coordenadas (X_{xm}, Y_{xm}) respeito ao núcleo (X_{mj}, Y_{mj}) . A distância elétrica define-se como o espaço percorrido através dos trechos da rede elétrica entre o ponto de carga e o núcleo $(Z_{xm} - Z_{mj})$. Por fim, também é considerado mais o componente de potência das cargas P_{xm} .

Continuando com os dados de carga apresentados na Fig. 4 o DBSCAN trabalha com uma dimensão de três. Note-se como a diferença do DBSCAN o k -means facilmente pode ser escalado de três a quatro dimensões adicionando um componente $(Z_{xm} - Z_{mj})$ na função de distância.

3.3 Discussão de Resultados

Nesta seção, é discutida a sensibilidade dos algoritmos DBSCAN e k -means para os pontos de carga apresentados na Fig. 4. Os resultados são verificados nos 11 alimentadores.

O DBSCAN dispõe de dois parâmetros de ajuste, ε e $min_samples$. Considerando o gráfico K_{dist} da Fig. 5 o valor do ε está em torno de 0,40. Portanto, o ε é variado no intervalo $[0,30; 0,45]$, para três cenários com $min_samples$ em 4, 5 e 6. A variação de $min_samples$ se deve ao fato de estar dentro de 4 e duas vezes a dimensão das amostras $2 \times dim = 6$, isto é $[4; 6]$.

Os três cenários são organizados na Tabela 1, de tal forma que na leitura da esquerda à direita apresentase a variação de ε mantendo fixo $min_samples$, e na leitura de cima para baixo apresentase a variação de $min_samples$ mantendo fixo ε . O intervalo de variação de ε é de 0,30 a 0,45 com incrementos de 0,5.

Dos resultados, o DBSCAN diminui o número de *clusters* e aumenta o número de *outliers* a medida que $min_samples$ aumenta. O ε também influencia diretamente no número de agrupamentos e *outliers*. Se ε aumenta, o número de agrupamentos diminui junto com o número de *outliers*. Na Tabela 1 apresenta-se a influência dos parâmetros de ajuste nos resultados e o tempo decorrido. Para cada cenário, a pontuação da silhueta apresenta uma melhoria (quanto maior pontuação, melhores *clusters* são obtidos) quando o número de *clusters* diminui. No cenário 2, quando o número estimado de *clusters* é 4 a pontuação da silhueta alcança as melhores pontuações. Da perspectiva das pontuações da silhueta, quanto maior a pontuação menor o número de *outliers*. O tempo de corrido não tem maior variação e está relacionado ao número de *clusters*.

Além do conjunto de dados, o algoritmo k -means tem como parâmetro de entrada o número de *clusters* k . O número k é definido de acordo com a quantidade de *clusters* estimados pelo DBSCAN. A correspondência de k é utilizada para comparar o k -means com o DBSCAN. Na Tabela 2, apresentam-se as métricas de *Davies-Bouldin*, *Calinski&Harabasz* e *Pontuação da silhueta*. Na pontuação de silhueta, para $k = 4$, é obtido o melhor particionamento. O mesmo acontece para o critério de *Davies-Bouldin*, $k = 4$ produz o melhor particionamento. Porém, para o critério *Calinski&Harabasz*, $k = 3$ produz o melhor particionamento. O tempo decorrido está diretamente relacionado ao número de *clusters*.

Tabela 1. Resultados na aplicação de DBSCAN com diferentes eps e $min_samples$.

Cenário 1	$eps = 0,30$ $min_samples = 4$	$eps = 0,35$ $min_samples = 4$	$eps = 0,40$ $min_samples = 4$	$eps = 0,45$ $min_samples = 4$
Número de clusters	10	9	6	5
Outliers	21	12	9	9
Pontuação da silhueta	0,217	0,288	0,485	0,538
Tempo decorrido (<i>seg</i>)	0,19	0,18	0,16	0,15
Cenário 2	$eps = 0,30$ $min_samples = 5$	$eps = 0,35$ $min_samples = 5$	$eps = 0,40$ $min_samples = 5$	$eps = 0,45$ $min_samples = 5$
Número de clusters	7	6	4	4
Outliers	33	25	19	14
Pontuação da silhueta	0,328	0,277	0,490	0,524
Tempo decorrido (<i>seg</i>)	0,25	0,24	0,22	0,22
Cenário 3	$eps = 0,30$ $min_samples = 6$	$eps = 0,35$ $min_samples = 6$	$eps = 0,40$ $min_samples = 6$	$eps = 0,45$ $min_samples = 6$
Número de clusters	7	6	4	4
Outliers	38	29	23	14
Pontuação da silhueta	0,359	0,272	0,482	0,524
Tempo decorrido (<i>seg</i>)	0,16	0,16	0,14	0,14

Tabela 2. Resultados na aplicação de k-means com diferentes valores de k agrupações.

Número de clusters	7	6	5	4	3
Critério Davies-Bouldin	0,7099	0,5792	0,5427	0,5129	0,5707
Criterio Calinski & Harabasz	327,3986	388,1579	453,623	562,927	596,445
Pontuação da silhueta	0,417	0,486	0,477	0,531	0,506
Tempo decorrido (<i>seg</i>)	92,508	86,501	63,368	74,054	32,989

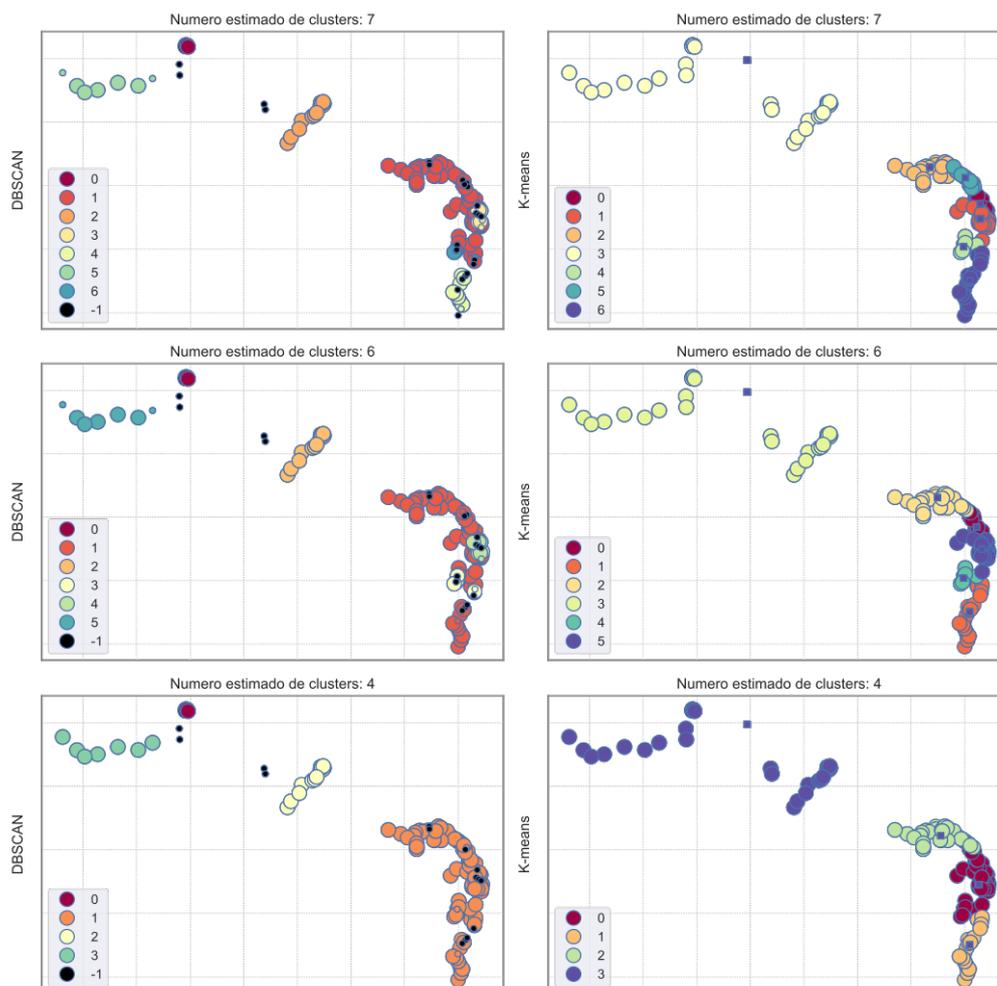


Figura 6. Resultados do DBSCAN e o K-means correspondentes com os valores da Tabela 1 e 2.

Uma comparação entre a Tabela 1 e a Tabela 2 mostra que o algoritmo *k-means* requer maior tempo computacional. Considerando que o tempo exigido pelo *k-means* é $O(l \times k \times m, n)$, onde l é o número de iterações, k é o número de agrupamentos, m é o número de pontos e, n o número de componentes.

Uma análise gráfica dos resultados é apresentada na Fig. 6, onde, os gráficos do DBSCAN correspondem ao segundo cenário. Observa-se que o particionamento DBSCAN segue a distribuição espacial das cargas e o *k-means* segue partições com forma esférica. Em relação as potencialidades, o DBSCAN é robusto no descobrimento de padrões baseados na densidade, reconhecendo conglomerados tanto de potência quanto de localização mas é muito sensível aos parâmetros de ajuste. O *k-means* têm a robustez da escalabilidade para adicionar dimensões ao custo de tempo de execução. Outra característica é que com a troca da distância elétrica pela distância linear no algoritmo *k-means* são definidas as seções de rede que pertencem a vizinhança de cada *cluster*, permitindo assim clusterizar a carga e particionar a rede simultaneamente, o que não acontece com o DBSCAN.

4. CONCLUSÕES

No escopo de definir agrupamentos de carga elétrica baseado em componentes elétricos e espaciais, este trabalho compara a aplicação do método DBSCAN e *K-means*, assim como o ajuste dos parâmetros de controle do desempenho desses métodos. Uma rede de média tensão real foi tomada em consideração na análise desses métodos.

O gráfico K_{dist} pode ser um bom recurso de suporte à seleção do valor inicial do *eps*, que é o parâmetro principal e carrega a maior sensibilidade do algoritmo DBSCAN. O algoritmo *K-means*, mesmo sendo facilmente escalável, não pode ser utilizado diretamente do pacote *scikit learn* como no caso do DBSCAN, porém, é possível aproveitar grande parte de suas funções.

Os resultados mostram que o algoritmo DBSCAN é robusto a *outliers* mas o ajuste dos parâmetros *min_samples* e *eps* são sensíveis e dependentes do conhecimento dos dados. Tal sensibilidade condiciona a automação do algoritmo. O algoritmo *K-means* permite facilmente adicionar componentes a serem avaliados na função de proximidade e afinar o particionamento, porém o custo computacional é afetado.

O trabalho atual abordou a análise de dois métodos de agrupamento utilizando informação espacial. Em trabalhos futuros, a integração dos algoritmos de mineração de dados, com os métodos de agrupamento, deve ser realizada com os modelos de redes elétricas baseados no SIG para definir as coberturas das MGs baseado na distribuição de carga, assim, obtendo um modelo de sistematização de dados das concessionárias de energia elétrica.

RECONHECIMENTO

Este trabalho foi apoiado pelas instituições brasileiras: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, CNPQ) - Código Financeiro 001, e pela Fundação de Pesquisa de São Paulo (FAPESP), processos 2015/21972-6 e 2017/02831-8

Os autores agradecem à CENTROSUR, concessionária de distribuição elétrica do Equador, pelas informações fornecidas e pelo apoio colaborativo.

REFERÊNCIAS

- Adil, A.M. and Ko, Y. (2016). Socio-technical evolution of decentralized energy systems: A critical review and implications for urban planning and policy. *Renewable and Sustainable Energy Reviews*, 57, 1025–1037.
- Benmouiza, K. and Chekane, A. (2013). Forecasting hourly global solar radiation using hybrid k-means and nonlinear autoregressive neural network models. *Energy Conversion and Management*, 75, 561–569.
- Blanco, M.I. (2009). The economics of wind energy. *Renewable and sustainable energy reviews*, 13(6-7), 1372–1382.
- Borges, C.L.T. (2012). An overview of reliability models and methods for distribution systems with renewable energy distributed generation. *Renewable and sustainable energy reviews*, 16(6), 4008–4015.
- Che, L., Zhang, X., Shahidepour, M., Alabdulwahab, A., and Abusorrah, A. (2017). Optimal interconnection planning of community microgrids with renewable energy sources. *IEEE Transactions on Smart Grid*, 8(3), 1054–1063. doi:10.1109/TSG.2015.2456834.
- CYME (2022). Cyme power engineering software. URL <https://www.cyme.com/software/cymdist/>.
- de Castro, N. and Dantas, G. (2017). Distributed generation: International experiences and comparative analyses. *PUBLIT, Rio de Janeiro*.
- Ester, M., Kriegel, H.P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, 226–231.
- Flach, P. (2012). *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*. Cambridge University Press. URL <https://books.google.com.ec/books?id=VJdQLwEACAAJ>.
- Gadelha Filho, T., Silvia, C., Aleksandar, D., Massimo, B., and Marco, M. (2021). Rural electrification planning based on graph theory and geospatial data: A realistic topology oriented approach. *Sustainable Energy, Grids and Networks*, 28, 100525.
- Hable, M. (2009). Distribution network planning for high load density areas. In *CIRED 2009 - 20th International Conference and Exhibition on Electricity Distribution - Part 1*, 1–4.
- Hossain, A.S. (2017). Customer segmentation using centroid based and density based clustering algorithms. In *2017 3rd International Conference on Electrical Information and Communication Technology (EICT)*, 1–6. doi:10.1109/EICT.2017.8275249.
- Kroposki, B., Sen, P.K., and Malmedal, K. (2013). Optimum sizing and placement of distributed and renewable energy sources in electric power distribution systems. *IEEE Transactions on Industry Applications*, 49(6), 2741–2752. doi:10.1109/TIA.2013.2262661.
- Lawanson, T., Karandeh, R., Cecchi, V., Wartell, Z., and Cho, I. (2018). Improving power distribution system situational awareness using visual analytics. In *SoutheastCon 2018*, 1–6. doi:10.1109/SECON.2018.8479204.
- Liu, Y., Li, Z., Xiong, H., Gao, X., and Wu, J. (2010). Understanding of internal clustering validation measu-

- res. In *2010 IEEE International Conference on Data Mining*, 911–916. doi:10.1109/ICDM.2010.35.
- Meehan, B. (2013). *GIS for Enhanced Electric Utility Performance*. Power Engineering. Artech House. URL <https://books.google.com.ec/books?id=-wEgAgAAQBAJ>.
- Mulenga, E., Bollen, M.H., and Etherden, N. (2020). A review of hosting capacity quantification methods for photovoltaics in low-voltage distribution grids. *International Journal of Electrical Power Energy Systems*, 115, 105445. doi:<https://doi.org/10.1016/j.ijepes.2019.105445>.
- Panapakidis, I.P. (2016). Clustering based day-ahead and hour-ahead bus load forecasting models. *International Journal of Electrical Power & Energy Systems*, 80, 171–178.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Rousseeuw, P.J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65. doi:[https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7). URL <https://www.sciencedirect.com/science/article/pii/0377042787901257>.
- Sebastian Raschka (2014). About feature scaling and normalization. URL https://sebastianraschka.com/Articles/2014_about_feature_scaling.html#standardization-and-min-max-scaling.
- Si, C., Xu, S., Wan, C., Chen, D., Cui, W., and Zhao, J. (2021). Electric load clustering in smart grid: Methodologies, applications, and future trends. *Journal of Modern Power Systems and Clean Energy*, 9(2), 237–252.
- Song, P., Xu, Z., Dong, H., Cai, H., and Xie, Z. (2017). Security-constrained line loss minimization in distribution systems with high penetration of renewable energy using upfc. *Journal of Modern Power Systems and Clean Energy*, 5(6), 876–886.
- VanderPlas, J. (2016). *Python Data Science Handbook: Essential Tools for Working with Data*. O'Reilly Media. URL <https://books.google.com.ec/books?id=6omNDQAAQBAJ>.
- Yu, L., Shi, D., Guo, X., Jiang, Z., Xu, G., Jian, G., Lei, J., and Jing, C. (2018). An efficient substation placement and sizing strategy based on gis using semi-supervised learning. *CSEE journal of power and energy systems*, 4(3), 371–379.
- Zain ul Abideen, M., Ellabban, O., and Al-Fagih, L. (2020). A review of the tools and methods for distribution networks' hosting capacity calculation. *Energies*, 13(11), 2758.
- Zambrano-Asanza, S., Cando, D.J., Chuqui, F.H., Sango, J., and Franco, J.F. (2021). Graph partitioning-based clustering for the planning of distribution network topology using spatial- temporal load forecasting. In *2021 IEEE PES Innovative Smart Grid Technologies Conference - Latin America (ISGT Latin America)*, 1–5. doi:10.1109/ISGTLatinAmerica52371.2021.9543010.