Modelo híbrido *fuzzy c-means* para clusterização de instalações de energia solar em terrenos contaminados norte-americanos

FRANCO, D. G. B. [a*], STEINER, M. T. A. [a]

a. Pontifícia Universidade Católica do Paraná, Curitiba

**Corresponding author, david.barros@pucpr.br*

Resumo

O presente artigo valeu-se de um modelo híbrido *fuzzy c-means* para clusterizar e definir locais aptos, em termos de área mapeada, distância até linhas de transmissão e incidência solar diária, para instalações de captação de energia solar no território continental norte-americano. Os dados utilizados são oriundos da *National Solar Radiation Database* (*NSRDB*), coleção de medidas horárias de radiação solar e dados meteorológicos, e do projeto *RE-Powering America's Land*, da *United States Environmental Protection Agency* (*EPA*), cujo propósito é identificar áreas abandonadas e contaminadas que são ideais para projetos de energia renovável. Inicialmente foi realizado o pré-processamento dos dados, para substituição de faltantes, normalização e análise dos componentes principais (*PCA*). Em seguida, foi aplicado o algoritmo híbrido proposto, de clusterização. Trata-se de um modelo *fuzzy c-means* inicializado por metaheurísticas, a saber, algoritmo genético, evolução diferencial e *particle swarm optimization* (*PSO*). O número de clusters foi validado por três métricas: *Calinski-Harabasz Index*, *Davies-Bouldin Index* e *Silhouette Coefficient*. Os três testes foram unanimes, indicando dois clusters como o número ideal, ou seja, um cluster para locais com potencial para alocação de instalações de captação de energia solar e outro para locais sem potencial. Como resultado da abordagem híbrida proposta, houve um incremento na velocidade de treinamento do algoritmo *fuzzy c-means*, que necessitou de um menor número de iterações para atingir o mesmo valor da função objetivo. Visualmente, pode-se perceber a predominância da alocação das instalações em estados de maior incidência média de radiação solar, sendo este, portanto, o fator predominante na convergência do algoritmo, o que está de acordo com o esperado. Por fim, são considerados os ganhos ambiental-econômico-social com a revitalização de terrenos improdutivos e contaminados para a implantação de usinas solares.

Palavras-chave: Clusterização; Fuzzy c-means; Metaheurísticas; Energia Solar; Reutilização do Solo.

1. Introdução

Áreas abandonadas, que possuem substâncias potencialmente perigosas ao meio ambiente ou à saúde humana, estão se tornando o centro de uma preocupação mundial (Bergius e Öberg, 2007; Greenberg e Lewis, 2000; Li et al., 2017; van Straalen, 2002). Entre estas áreas de risco, no contexto norte-americano, podemos citar: minas abandonadas, que incluem *spoil banks* (pilhas de rejeitos) e plantas de processamento de metais, geralmente contaminadas por metais pesados (Kovacs e Szemmelveisz, 2017); *brownfields* (campos marrons), que podem ser definidos como instalações industriais ou comerciais que apresentam dificuldades para reutilização, devido à presença de substâncias perigosas, poluentes ou contaminantes (U.S. Government Publishing Office, 2002); locais que se enquadram no *Superfund*, programa do governo federal norte-americano voltado para a localização e limpeza de

áreas contaminadas com substâncias perigosas ou poluentes (U.S. Government Publishing Office, 2015); aterros sanitários, que nos países desenvolvidos incluem, basicamente, restos de alimentos e embalagens (Rong et al., 2017); e áreas abrangidas pela legislação do *Resource Conservation and Recovery Act* (*RCRA*), de destinação de rejeitos sólidos (U.S. Government Publishing Office, 2011).

Ao mesmo tempo, a crescente ocupação do espaço urbano e rural têm-se tornado um problema no mundo moderno (Lambin e Meyfroidt, 2011; Morio et al., 2013), o que demanda maior eficiência na ocupação territorial, principalmente a reutilização de áreas abandonadas, que apresentam maior desafio (Morio et al., 2013; Nuissl e Schroeter-Schlaack, 2009; U.S. Government Publishing Office, 2002). Ainda mais se tais áreas possuem considerável tamanho e estão contaminadas, gerando, além dos riscos ambiental e de saúde, riscos econômicos (Apostolidis e Hutton, 2006; Cao e Guan, 2007; de Sousa, 2003; Kaufman et al., 2005; Morio et al., 2013). A partir do uso de informações públicas sobre as principais áreas em desuso com algum tipo de contaminação no território continental norte-americano, elaborou-se uma metodologia híbrida de *clusterização* (agrupamento) visando classificar tais áreas em apropriadas ou não para a implantação de usinas de captação de energia solar.

Uma vez que o aumento do consumo de energia entra em choque com as implicações do consumo de combustíveis fósseis e a consequente emissão de gases tóxicos e de efeito estufa, fazem-se necessários investimentos em pesquisa e desenvolvimento de novas fontes limpas e renováveis de energia (Almeida et al., 2017; Baños et al., 2011; Cadez e Czerny, 2016; Manzano-Agugliaro et al., 2012; Perea-Moreno et al., 2017). Desse modo, as energias renováveis, como a solar, por exemplo, têm-se colocado como fortes candidatas na nova corrida por produtividade e bem-estar ambiental e social (González et al., 2017; Lima et al., 2013), além de estarem em voga nos discursos políticos, empresariais e sociais em geral (Onat et al., 2014; Simas e Pacca, 2013). Neste cenário, a energia solar é considerada uma fonte abundante, gratuita e limpa (Fernández-García et al., 2015).

O presente trabalho está organizado da seguinte maneira: após esta seção introdutória, a Seção 2 apresenta a metodologia empregada, incluindo uma breve revisão de literatura, apresentação dos dados utilizados e descrição do modelo híbrido de clusteriação proposto. Na Seção 3 são apresentados e discutidos os resultados. Por fim, a Seção 4 sumariza as conclusões da pesquisa.

2. Metodologia

É apresentada a seguir a revisão de literatura referente ao processo de descoberta de conhecimento em base de dados (*KDD — Knowledge Discovery in Databases*), mineração de dados (*data mining*) e clusterização. Logo na sequência é apresentada a base de dados utilizada e o pré-processamento empregado. E, finalmente, o modelo proposto de clusterização híbrida.

2.1 Revisão de literatura

KDD é uma área do conhecimento dedicada à identificação e extração de padrões significativos de informação a partir de bases de dados (Fayyad et al., 1996). A aplicação do *KDD* se dá em múltiplos estágios, começando com a seleção, pré-processamento e transformação dos dados, que pode incluir remoção de *outliers*, substituição de dados faltantes, normalização, análise de componentes principais (*PCA — Principal Component Analysis*), entre outras técnicas, dependendo do algoritmo da etapa subsequente, de mineração (ou aprendizado), terminando com a interpretação dos resultados e a geração do conhecimento (Fayyad et al., 1996; Gamarra et al., 2016; Orriols-Puig et al., 2013).

A etapa de mineração de dados pode utilizar um ou vários algoritmos em busca de padrões, tendências e estruturas na base de dados, as quais podem assumir variadas formas, como equações, redes, grafos, conjuntos de regras, entre outros (Roiger, 2017; Witten et al., 2017). Nessa etapa de aprendizado pode-se adotar duas abordagens distintas: na primeira abordagem, de aprendizado supervisionado, consideram-se, *a priori*, estruturas e padrões pré-definidos; na segunda abordagem, de aprendizado não-supervisionado, não se consideram essas possíveis estruturas e padrões, deixando ao algoritmo a tarefa de identificar tais relações entre as variáveis (Orriols-Puig et al., 2013).

A clusterização se enquadra na segunda abordagem, de aprendizado não-supervisionado, na qual as instâncias são agrupadas com base em alguma regra inerente à sua estrutura, como a distância entre

elas (Bramer, 2016; Roiger, 2017). Simplificadamente, pode-se dizer que, após a definição do número de clusters e seus respectivos centros (chamados protótipos), há uma primeira etapa de designação das instâncias para o mais próximo protótipo, seguida pela otimização da localização do mesmo, minimizando a função objetivo (Aggarwal, 2015).

Há um variado número e tipo de algoritmos de clusterização (Halkidi et al., 2001; Xu e Wunsch II, 2005). Alguns deles são apresentados na Tabela 1.

Tabela 1. Algoritmos de clusterização

Tipo	Algoritmo	Referência
Particional	*K-Means*	(MacQueen, 1967)
	PAM (Partitioning Around Medoids), CLARA (Clustering Large Applications), CLARANS (Clustering Large Applications based on Randomized Search)	(Ng e Han, 1994)
	K-prototypes, K-mode	(Huang, 1998)
Hierárquico	*BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)*	(Zhang et al., 1996)
	CURE (Clustering Using Representatives)	(Guha et al., 1998)
	ROCK (Robust Clustering using linKs)	(Guha et al., 2000)
Density-based	*DBSCAN (Density Based Spatial Clustering of Applications with Noise)*	(Ester et al., 1996)
	DENCLUE (Density-based Clustering)	(Hinneburg e Keim, 2003, 1998)
Grid-based	*STING (Statistical Information Grid-based method)*	(Wang et al., 1997)
	WaveCluster	(Sheikholeslami et al., 1998)
Fuzzy (soft)	*Fuzzy c-means*	(Bezdek et al., 1984)
	EM (Expectation Maximization)	(Dempster et al., 1977)
Neural Network-based	*GLVQ (Generalized Learning Vector Quantization), SOFM (Self-Organizing Feature Maps)*	(Pal et al., 1993)
Kernel-based	*SVC (Support Vector Clustering)*	(Ben-Hur et al., 2001)

Os resultados da clusterização devem ser validados por alguma métrica, que pode ser externa, como rótulo de classes; interna, quando a avaliação se dá por característica inerentes aos dados, como variância e separação dos pontos em diferentes clusters; ou relativa, onde as métricas visam comparar o uso de diferentes parâmetros durante a execução do algoritmo. A Tabela 2 sumariza algumas dessas métricas (Amigó et al., 2009; Halkidi et al., 2001; Meilă, 2007; Zaki e Meira Jr., 2014).

Tabela 2. Métricas de validação da clusterização

Externa	Interna	Relativa
Purity	*BetaCV Measure*	*Calinski-Harabasz Index*
Maximum Matching	*Normalized Cut Measure*	*Gap Statistic*
F-Measure	*Modularity*	
Conditional Entropy	*Dunn Index*	
Normalized Mutual Information	*Davies-Bouldin Index*	
Variation of Information	*Silhouette Coefficient*	
Jaccard Coefficient	*Hubert Statistic*	
Rand Statistic		
Fowlkes-Mallows Measure		
Discretized Hubert Statistic		
Normalized Discretized Hubert Statistic		

2.2 Coleta e pré-processamento dos dados

Os dados utilizados são oriundos da *National Solar Radiation Database* (*NSRDB*), coleção de medidas horárias de radiação solar e dados meteorológicos, e do projeto *RE-Powering America's Land*, da *United States Environmental Protection Agency* (*EPA*), cujo propósito é identificar áreas abandonadas e contaminadas que são ideais para projetos de energia renovável. A Fig. 1 mostra a união entre os

dados de radiação solar direta (*DNI — Direct Normal Irradiance*) média para o período 1998-2014, medida em $kWh/m^2/dia$, e áreas contaminadas, maiores que 100 acres (~400 mil m^2), com potencial para instalação de usinas solares, num total de 5.063 pontos (minas abandonadas, a maioria).

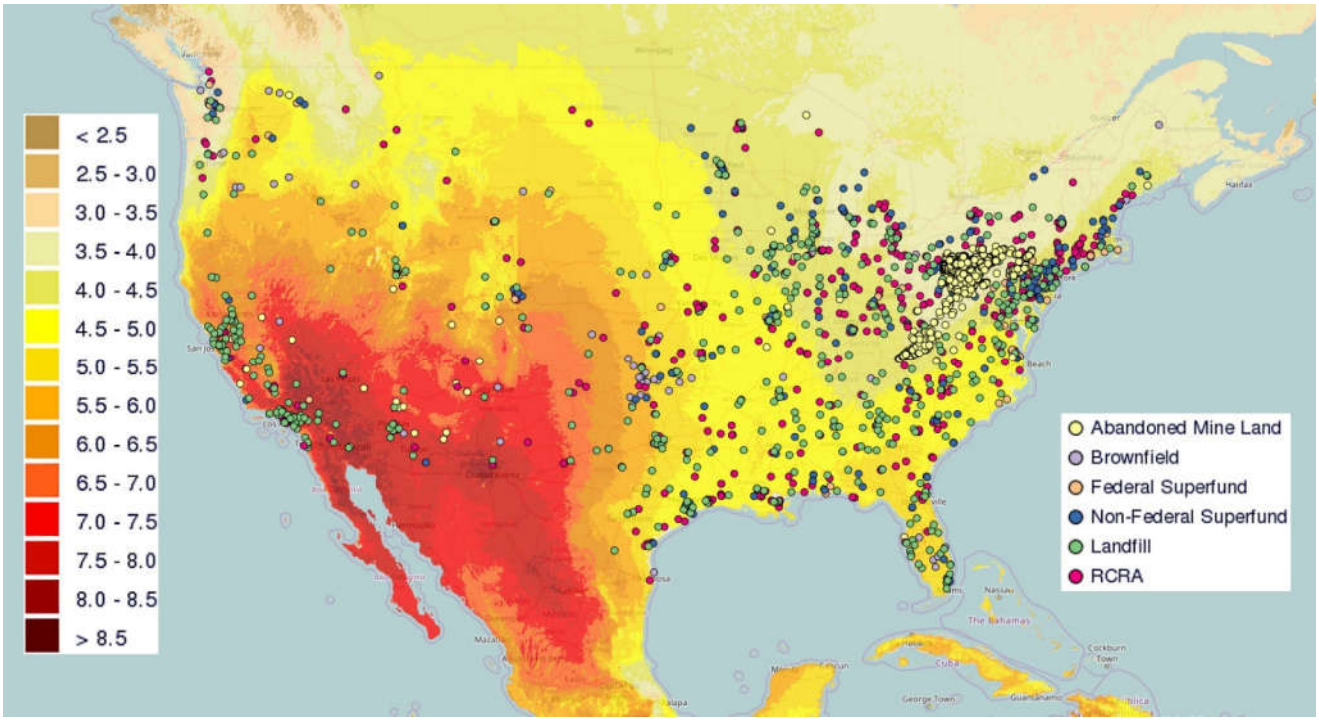

Fig. 1. Radiação solar direta e áreas contaminadas no território continental norte-americano.

As variáveis contidas na base de dados são: área mapeada, medida em acres; distância até linhas de transmissão, medida em milhas; e radiação solar direta, medida em $kWh/m^2/dia$, para escala de utilidade (geração na casa dos megawatts e possibilidade de exportação para a rede) e pequena escala (para atender as necessidades de uma única propriedade e inviabilidade de exportação para a rede).

Os valores faltantes, num total de 12 instâncias para a variável distância até linhas de transmissão, foram substituídos pela média do conjunto. Em seguida foi realizada a normalização dos dados, que passaram a ter média "0" e desvio-padrão "1", para dar sequência à análise dos componentes principais, que tem por objetivo diminuir o número de atributos, retirando os que são correlacionados entre si e que pouco contribuem para a variância do conjunto (Theodoridis e Koutroumbas, 2009).

Para as variáveis testadas, apenas a radiação solar direta em pequena escala foi removida, uma vez que não contribuía para a variância do conjunto de dados (apenas 0,5829%, contra 50,5270% para área mapeada, 25,9960% para distância até linhas de transmissão e 22,8941% para radiação solar direta em escala de utilidade), ficando o conjunto final, utilizado nos testes, com três variáveis.

2.3 Modelo híbrido evolucinário-fuzzy c-means

Na clusterização *fuzzy*, ou *soft*, cada instância possui graus de participação, no intervalo $[0,1]$, em cada um dos clusters, de modo que a soma total das participações, para cada instância, é igual a "1". Após a execução do algoritmo, adotou-se o arredondamento simples da matriz de partição *fuzzy*, que armazena os graus de participação para cada instância em cada cluster, de modo que os valores maiores ou iguais a "0.5" foram arredondados para "1" e os valores menores que "0.5" foram arredondados para "0" (uma vez que são apenas dois clusters). Poder-se-ia adotar que os valores intermediários de graus de participação (compreendidos entre 0.45 e 0.55, por exemplo) seriam representantes de um terceiro cluster, intermediário. Como, porém, o objetivo do trabalho (respaldado por três métricas distintas) era se terem apenas dois clusters, tal procedimento não foi realizado.

A função objetivo empregada foi a *fuzzy c-means functional*, base para uma ampla família de algoritmos de clusterização *fuzzy* (Babuška, 1998; Bezdek, 1981; Dunn, 1974). O valor desta função pode ser entendido como uma medida da variância total entre cada instância e o centro de seu respectivo cluster — o protótipo (Babuška, 1998). A minimização desta função representa um problema de otimização não-linear, que pode ser solucionado de inúmeras maneiras, entre elas: algoritmos genéticos (Babu e Murty, 1994), *simulated annealing* (Desarbo, 1982) e *grouped coordinate minimization* (Bezdek et al., 1987; Hathaway e Bezdek, 1991). A mais popular, entretanto, é a *fuzzy c-means* (Babuška, 1998).

Uma vez que o algoritmo *fuzzy c-means* padrão pressupõe a inicialização aleatória da matriz de partição *fuzzy*, existe a necessidade de um maior número de iterações para alcançar a solução final. Pensando nessa limitação, nós propomos a inicialização da matriz de partição *fuzzy* através de três metaheurísticas distintas: algoritmo genético (Goldberg, 1989) e evolução diferencial (Storn, 1996; Storn e Price, 1997), estratégias evolutivas, e *particle swarm optimization* (*PSO*) (Kennedy et al., 2001; Kennedy e Eberhart, 1995; Poli, 2008), estratégia de enxames.

3. Resultados

Como esperado, a proposta sugerida conseguiu reduzir o número de iterações. A versão clássica do algoritmo *fuzzy c-means* demandou 28 iterações, enquanto que a inicializada por algoritmo genético e *PSO* demandou 23 iterações, e apresentaram a mesma convergência. A versão inicializada por evolução diferencial precisou de 22 iterações. Todas alcançaram o mesmo valor final para a função objetivo (8.737,8). Confira a Fig. 2 (a).

Percebe-se que, quando o algoritmo é inicializado pelas metaheurísticas, o valor da função objetivo é cerca de 18,7% mais baixo (10.076,0) do que quando inicializado aleatoriamente (12.394,0), decorrência do treinamento inicial realizado pelas metaheurísticas. Nota-se também, pelo comportamento da curva de treinamento, que o algoritmo *fuzzy c-means* realiza um refinamento da solução a partir de certo nível da função objetivo (aproximadamente 10.000).

Quanto ao resultado final da clusterização, houve divergência em apenas 1 instância entre as 5.063. Para o algoritmo clássico, 835 instâncias foram classificadas como pertencentes ao *cluster 1* e 4.228 como pertencentes ao *cluster 2*. Para o método proposto, 836 instâncias foram alocadas para o *cluster 1* e 4.227 para o *cluster 2*, conforme ilustrado pela Fig. 2 (b).

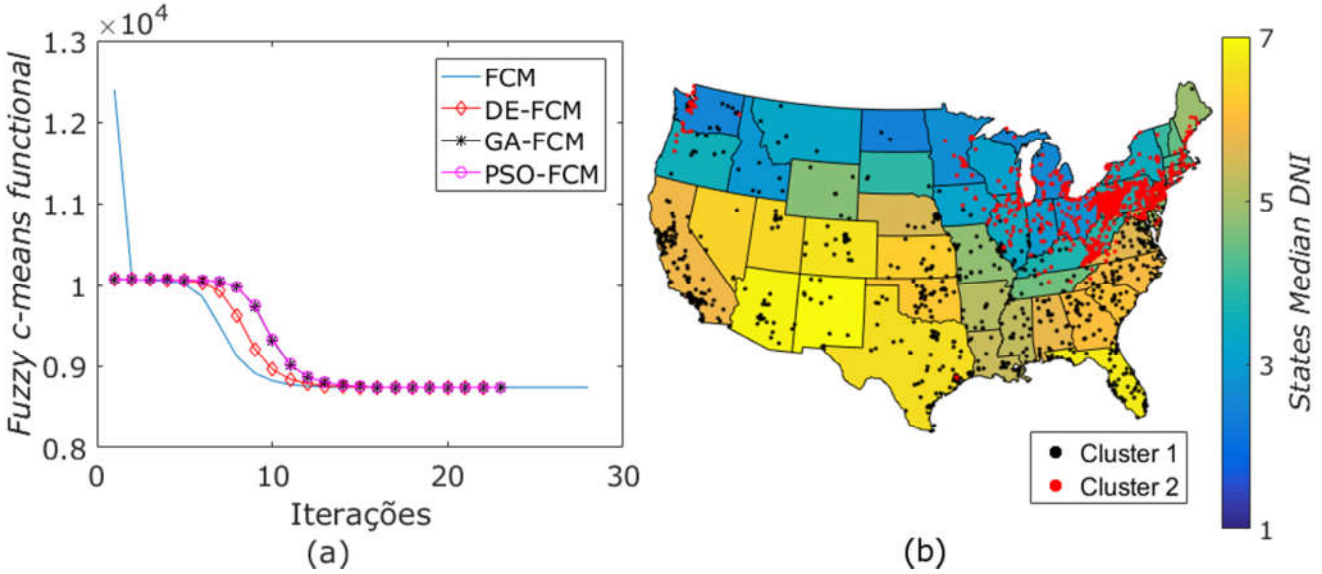

Fig. 2. (a) Resultados da estratégia proposta (*DE-FCM*, inicializado por evolução diferencial; *GA-FCM*, inicializado por algoritmo genético; e *PSO-FCM*, inicializado por *PSO*) e da versão clássica (*FCM*). (b) Resultado final da clusterização e incidência solar mediana para o território continental americano.

Visualmente, pode-se perceber a predominância da alocação das instalações em estados de maior incidência média de radiação solar, sendo este, portanto, o fator predominante na convergência do algoritmo, o que está de acordo com o esperado (embora este atributo contribua menos com a variância total do conjunto).

O número de clusters foi validado por três métricas: *Calinski-Harabasz Index*, *Davies-Bouldin Index* e *Silhouette Coefficient*. Os três testes foram unanimes, indicando dois clusters como o número ideal, ou seja, um cluster para locais com potencial para alocação de instalações de captação de energia solar (principalmente nos estados ensolarados) e outro para locais sem potencial (estados com menor incidência solar).

Para análise da diferença entre os clusters, a Tabela 3 apresenta a estatística para cada um deles. Nota-se que as instâncias alocadas para o *cluster 1* possuem maior média para as variáveis área mapeada (Var. 1) e incidência solar, em escala de utilidade (Var. 3) e em pequena escala (Var. 4). Para a distância até linhas de transmissão (Var. 2), o *cluster 2* obteve a menor média, uma diferença de 11,6% em relação ao *cluster 1*. É interessante notar que o *cluster 2*, mesmo possuindo mais de cinco vezes o número de instâncias do *cluster 1*, possui menores desvio padrão e variância, para todas as variáveis, um dos indicativos de qualidade (compacidade) de uma clusterização (Babuška, 1998).

Tabela 3. Estatística para os clusters encontrados

	Cluster 1 (836 instâncias)				**Cluster 2 (4227 instâncias)**			
	Var. 1	Var. 2	Var. 3	Var. 4	Var. 1	Var. 2	Var. 3	Var. 4
Mínimo	100,80	0,00	3,47	4,17	100,04	0,00	2,17	2,86
Média	**347,74**	2,50	**4,76**	**5,20**	255,50	**2,21**	3,30	4,21
Máximo	996,00	63,84	7,75	6,68	999,56	49,88	4,36	4,79
Desvio padrão	236,30	5,54	0,86	0,48	**147,32**	**4,07**	**0,17**	**0,12**
Variância	55.839,95	30,66	0,73	0,23	**21.702,83**	**16,54**	**0,03**	**0,01**

4. Conclusão

Como resultado da abordagem híbrida proposta, houve um incremento na velocidade de treinamento do algoritmo *fuzzy c-means*, que necessitou de um menor número de iterações para atingir o mesmo valor da função objetivo. Os dois clusters resultantes apresentaram características estatísticas que validam a inicialização pelas metaheurística testadas, uma vez que o *cluster 1* apresentou maiores médias para três das quatro variáveis analisadas, sendo o cluster das instâncias com maior potencial para instalação de usinas solares. Enquanto que o *cluster 2* apresentou melhor compacidade, tanto em termos das variáveis analisadas quanto em relação à localização geográfica de suas instâncias.

É importante ressaltar a importância do estudo em relação a uma melhor alocação dos parcos recursos disponíveis para energias renováveis (em relação às outras energias convencionais, como os combustíveis fósseis). Com a clusterização pode-se determinar *a priori* os locais com melhor potencial e, respectivamente, retorno sobre o investimento feito. Entre os principais benefícios decorrentes da instalação de usinas solares, podemos citar a própria economia com custos de energia, seguido dos ganhos ambientais, com redução de emissões diretas e indiretas de gases do efeito estufa, além da criação de empregos. Outro importante benefício é a limpeza de áreas antes contaminadas e improdutivas, que geravam riscos para a saúde e o meio ambiente. Especificamente para o caso norte-americano podemos citar, ainda, os subsídios estatais, como o *Solar Renewable Energy Certificate* (*SREC*), uma *commodity* energética (específica para geração solar) não tangível que é emitida ao se gerar $1\,MWh$ de energia de fonte solar e que pode ser vendida no mercado (Bird et al., 2011).

A título de ilustração, com a implantação de 190 unidades de geração de energia renovável (em seu potencial máximo), pelo programa *RE-Powering America's Land*, seria possível reduzir as emissões de gases do efeito estufa em 1,7 milhão de toneladas de CO_2 por ano, o equivalente (em termos de dióxido de carbono equivalente) a 500 mil toneladas de lixo ou 193 milhões de galões de gasolina. Com relação a empregos, o *2015 Solar Jobs Census* estimou em 209 mil o número de empregados na industrial solar norte-americana. Tudo isso ressalta o potencial de ganhos ambiental-econômico-social das energias renováveis e, em especial, a solar (EPA, 2016).

Referências

Aggarwal, C.C., 2015. Data Mining: The Textbook. Springer International Publishing. doi:10.1007/978-3-319-14142-8

Almeida, C.M.V.B., Agostinho, F., Huisingh, D., Giannetti, B.F., 2017. Cleaner Production towards a sustainable transition. J. Clean. Prod. 142, 1–7. doi:10.1016/j.jclepro.2016.10.094

Amigó, E., Gonzalo, J., Artiles, J., Verdejo, F., 2009. A comparison of extrinsic clustering evaluation metrics based on formal constraints. Inf. Retr. Boston. 12, 461–486. doi:10.1007/s10791-008-9066-8

Apostolidis, N., Hutton, N., 2006. Integrated water management in brownfield sites: More opportunities than you think. Desalination 188, 169–175. doi:10.1016/j.desal.2005.04.114

Babu, G.P., Murty, M.N., 1994. Clustering with evolution strategies. Pattern Recognit. 27, 321–329. doi:10.1016/0031-3203(94)90063-9

Babuška, R., 1998. Fuzzy Modeling for Control. Springer International Publishing, New York. doi:10.1007/978-94-011-4868-9

Baños, R., Manzano-Agugliaro, F., Montoya, F.G., Gil, C., Alcayde, A., Gómez, J., 2011. Optimization methods applied to renewable and sustainable energy: A review. Renew. Sust. Energ. Rev. 15, 1753–1766. doi:10.1016/j.rser.2010.12.008

Ben-Hur, A., Horn, D., Siegelmann, H.T., Vapnik, V., 2001. Support Vector Clustering. J. Mach. Learn. Res. 2, 125–137.

Bergius, K., Öberg, T., 2007. Initial screening of contaminated land: A comparison of US and Swedish methods. Environ. Manage. 39, 226–234. doi:10.1007/s00267-006-0005-4

Bezdek, J.C., 1981. Pattern Recognition with Fuzzy Objective Function Algorithms, SIAM Review. Springer US. doi:10.1007/978-1-4757-0450-1

Bezdek, J.C., Ehrlich, R., Full, W., 1984. FCM: The fuzzy c-means clustering algorithm. Comput. Geosci. 10, 191–203. doi:10.1016/0098-3004(84)90020-7

Bezdek, J.C., Hathaway, R.J., Howard, R.E., Wilson, C.A., Windham, M.P., 1987. Local convergence analysis of a grouped variable version of coordinate descent. J. Optim. Theory Appl. 54, 471–477. doi:10.1007/BF00940196

Bird, L., Heeter, J., Kreycik, C., 2011. Solar Renewable Energy Certificate (SREC) Markets: Status and Trends.

Bramer, M., 2016. Principles of Data Mining, 3º ed, Undergraduate Topics in Computer Science. Springer-Verlag London, London. doi:10.1007/978-1-4471-7307-6

Cadez, S., Czerny, A., 2016. Climate change mitigation strategies in carbon-intensive firms. J. Clean. Prod. 112, 4132–4143. doi:10.1016/j.jclepro.2015.07.099

Cao, K., Guan, H., 2007. Brownfield redevelopment toward sustainable urban land use in China. Chinese Geogr. Sci. 17, 127–134. doi:10.1007/s11769-007-0127-5

de Sousa, C.A., 2003. Turning brownfields into green space in the City of Toronto. Landsc. Urban Plan. 62, 181–198. doi:10.1016/S0169-2046(02)00149-4

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. J. Roy. Stat. Soc. B Met. 39, 1–38.

Desarbo, W.S., 1982. Gennclus: New models for general nonhierarchical clustering analysis. Psychometrika 47, 449–475. doi:10.1007/BF02293709

Dunn, J.C., 1974. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. J. Cybern. 3, 32–57. doi:10.1080/01969727308546046

EPA, 2016. RE-Powering America's Land Initiative: Benefits Matrix [WWW Document]. URL https://goo.gl/XMov1T (acessado 3.3.17).

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, in: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining - KDD '96. AAAI Press, Portland, p. 226–231.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., 1996. From data mining to knowledge discovery in databases. AI Mag. 37–54. doi:10.1145/240455.240463

Fernández-García, A., Rojas, E., Pérez, M., Silva, R., Hernández-Escobedo, Q., Manzano-Agugliaro, F., 2015. A parabolic-trough collector for cleaner industrial process heat. J. Clean. Prod. 89, 272–285. doi:10.1016/j.jclepro.2014.11.018

Gamarra, C., Guerrero, J.M., Montero, E., 2016. A knowledge discovery in databases approach for industrial microgrid planning. Renew. Sust. Energ. Rev. 60, 615–630. doi:10.1016/j.rser.2016.01.091

Goldberg, D.E., 1989. Genetic Algorithms in Search, Optimization & Machine Learning. Addison-Wesley Professional, Boston.

González, M.O.A., Gonçalves, J.S., Vasconcelos, R.M., 2017. Sustainable development: Case study in the implementation of renewable energy in Brazil. J. Clean. Prod. 142, 461–475. doi:10.1016/j.jclepro.2016.10.052

Greenberg, M., Lewis, M.J., 2000. Brownfields Redevelopment, Preferences and Public Involvement: A Case Study of an Ethically Mixed Neighbourhood. Urban Stud. 37, 2501–2514. doi:10.1080/00420980020005442

Guha, S., Rastogi, R., Shim, K., 2000. Rock: a robust clustering algorithm for categorical attributes. Inf. Syst. 25, 345–366. doi:10.1016/S0306-4379(00)00022-3

Guha, S., Rastogi, R., Shim, K., 1998. CURE: An Efficient Clustering Algorithm for Large Databases, in: Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data - SIGMOD '98. ACM, New York, p. 73–84. doi:10.1145/276304.276312

Halkidi, M., Batistakis, Y., Vazirgiannis, M., 2001. On clustering validation techniques. J. Intell. Inf. Syst. 17, 107–145. doi:10.1023/A:1012801612483

Hathaway, R.J., Bezdek, J.C., 1991. Grouped coordinate minimization using Newton's method for inexact minimization in one vector coordinate. J. Optim. Theory Appl. 71, 503–516. doi:10.1007/BF00941400

Hinneburg, A., Keim, D.A., 2003. A general approach to clustering in large databases with noise. Knowl. Inf. Syst. 5, 387–415. doi:10.1007/s10115-003-0086-9

Hinneburg, A., Keim, D.A., 1998. An efficient approach to clustering in large multimedia databases with noise, in: Proceedings of 4th International Conference on Knowledge Discovery and Data Mining - KDD '98. AAAI Press, New York, p. 58–65.

Huang, Z., 1998. Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Min. Knowl. Discov. 2, 283–304. doi:10.1023/A:1009769707641

Kaufman, M.M., Rogers, D.T., Murray, K.S., 2005. An empirical model for estimating remediation costs at contaminated sites. Water. Air. Soil Poll. 167, 365–386. doi:10.1007/s11270-005-0214-0

Kennedy, J., Eberhart, R., 1995. Particle swarm optimization, in: Proceedings of International Conference on Neural Networks ICNN '95. IEEE, p. 1942–1948. doi:10.1109/ICNN.1995.488968

Kennedy, J., Eberhart, R.C., Shi, Y., 2001. Swarm intelligence, Science. Morgan Kaufmann.

Kovacs, H., Szemmelveisz, K., 2017. Disposal options for polluted plants grown on heavy metal contaminated brownfield lands: A review. Chemosphere 166, 8–20. doi:10.1016/j.chemosphere.2016.09.076

Lambin, E.F., Meyfroidt, P., 2011. Global land use change, economic globalization, and the looming land scarcity. Proc. Natl. Acad. Sci. U. S. A. 108, 3465–3472. doi:10.1073/pnas.1100480108

Li, X., Jiao, W., Xiao, R., Chen, W., Liu, W., 2017. Contaminated sites in China: Countermeasures of provincial governments. J. Clean. Prod. 147, 485–496. doi:10.1016/j.jclepro.2017.01.107

Lima, F., Ferreira, P., Vieira, F., 2013. Strategic impact management of wind power projects. Renew. Sust. Energ. Rev. 25, 277–290. doi:10.1016/j.rser.2013.04.010

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations, in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, Berkeley, p. 281–297.

Manzano-Agugliaro, F., Sanchez-Muros, M.J., Barroso, F.G., Martínez-Sánchez, A., Rojo, S., Pérez-Bañón, C., 2012. Insects for biodiesel production. Renew. Sust. Energ. Rev. 16, 3744–3753. doi:10.1016/j.rser.2012.03.017

Meil□, M., 2007. Comparing clusterings: an information based distance. J. Multivar. Anal. 98, 873–895. doi:10.1016/j.jmva.2006.11.013

Morio, M., Schädler, S., Finkel, M., 2013. Applying a multi-criteria genetic algorithm framework for brownfield reuse optimization: Improving redevelopment options based on stakeholder preferences. J. Environ. Manage. 130, 331–346. doi:10.1016/j.jenvman.2013.09.002

Ng, R.T., Han, J., 1994. Efficient and effective clustering methods for spatial data mining, in: Proceedings of the 20th International Conference on Very Large Data Bases - VLDB '94. Morgan Kaufmann, San Francisco, p. 144–155.

Nuissl, H., Schroeter-Schlaack, C., 2009. On the economic approach to the containment of land consumption. Environ. Sci. Policy 12, 270–280. doi:10.1016/j.envsci.2009.01.008

Onat, N.C., Kucukvar, M., Tatari, O., 2014. Integrating triple bottom line input-output analysis into life cycle sustainability assessment framework: The case for US buildings. Int. J. Life Cycle Assess. 19, 1488–1505. doi:10.1007/s11367-014-0753-y

Orriols-Puig, A., Martínez-López, F.J., Casillas, J., Lee, N., 2013. Unsupervised KDD to creatively support managers' decision making with fuzzy association rules: A distribution channel application. Ind. Mark. Manag. 42, 532–543. doi:10.1016/j.indmarman.2013.03.005

Pal, N.R., Bezdek, J.C., Tsao, E.C.-K., 1993. Generalized Clustering Networks and Kohonen's Self-Organizing Scheme. IEEE Trans. Neural Networks 4, 549–557. doi:10.1109/72.238310

Perea-Moreno, A.-J., García-Cruz, A., Novas, N., Manzano-Agugliaro, F., 2017. Rooftop analysis for solar flat plate collector assessment to achieving sustainability energy. J. Clean. Prod. 148, 545–554. doi:10.1016/j.jclepro.2017.02.019

Poli, R., 2008. Analysis of the Publications on the Applications of Particle Swarm Optimisation. J. Artif. Evol. Appl. 2008, 1–10. doi:10.1155/2008/685175

Roiger, R.J., 2017. Data Mining: A Tutorial-Based Primer, 2º ed. CRC Press.

Rong, L., Zhang, C., Jin, D., Dai, Z., 2017. Assessment of the potential utilization of municipal solid waste from a closed irregular landfill. J. Clean. Prod. 142, 413–419. doi:10.1016/j.jclepro.2015.10.050

Sheikholeslami, G., Chatterjee, S., Zhang, A., 1998. Wavecluster: A multi-Resolution Clustering Approach for Very Large Spatial Databases, in: Proceedings of 24rd International Conference on Very Large Data Bases - VLDB '98. Morgan Kaufmann, San Francisco, p. 428–439.

Simas, M., Pacca, S., 2013. Energia eólica, geração de empregos e desenvolvimento sustentável. Estud. Avançados 27, 99–116. doi:10.1590/S0103-40142013000100008

Storn, R., 1996. On the usage of differential evolution for function optimization, in: Proceedings of North American Fuzzy Information Processing - NAFIPS '96. IEEE, p. 519–523. doi:10.1109/NAFIPS.1996.534789

Storn, R., Price, K., 1997. Differential Evolution – A Simple and Efficient Heuristic for global Optimization over Continuous Spaces. J. Glob. Optim. 11, 341–359. doi:10.1023/A:1008202821328

Theodoridis, S., Koutroumbas, K., 2009. Pattern Recognition, 4th ed. Academic Press.

U.S. Government Publishing Office, 2015. 42 U.S.C. 9601-9628 - Hazardous Substances Releases, Liability, Compensation [WWW Document]. United States Code, 2012 Ed. Suppl. 3, Title 42 - Public Heal. Welfare, Subchapter I. URL https://goo.gl/y0ki6N (acessado 3.3.17).

U.S. Government Publishing Office, 2011. 40 C.F.R. 239-282 - Solid Wastes [WWW Document]. Code Fed. Regul. (annual ed). URL https://goo.gl/UBCLDF (acessado 3.3.17).

U.S. Government Publishing Office, 2002. Public Law 107-118 - Small Business Liability Relief and Brownfields Revitalization Act [WWW Document]. H.R. 2869. URL https://goo.gl/UK19n2 (acessado 3.3.17).

van Straalen, N.M., 2002. Assessment of soil contamination: a functional perspective. Biodegradation 13, 41–52. doi:10.1023/a:1016398018140

Wang, W., Yang, J., Muntz, R.R., 1997. STING: A Statistical Information Grid Approach to Spatial Data Mining, in: Proceedings of 23rd International Conference on Very Large Data Bases - VLDB '97. Morgan Kaufmann, San Francisco, p. 186–195.

Witten, I.H., Frank, E., Hall, M.A., Pal, C.J., 2017. Data Mining: Practical Machine Learning Tools and Techniques, 4º ed. Morgan Kaufmann.

Xu, R., Wunsch II, D., 2005. Survey of Clustering Algorithms. IEEE Trans. Neural Networks 16, 645–678. doi:10.1109/TNN.2005.845141

Zaki, M.J., Meira Jr., W., 2014. Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press.

Zhang, T., Ramakrishnan, R., Livny, M., 1996. BIRCH: An Efficient Data Clustering Databases Method for Very Large, in: Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data - SIGMOD '96. ACM, New York, p. 103–114. doi:10.1145/233269.233324