XVIII EDAO
Encontro para Debates de Assuntos de Operação
Florianópolis | SC

Método de Estimação de Strings Ausentes em Usinas Fotovoltaicas de Grande Escala usando Aprendizado de Máquina não Supervisionado

Bruno Castro Valle[2], Tiago Edmir Simão[1], Yago Castro Rosa[3], Arliones Hoeller Jr.[3], Fernando Santos Varela[1], Mário de Noronha Neto[3], Richard Demo Souza[2]
[1] ENGIE Brasil
[2] Departamento de Engenharia Elétrica e Eletrônica, Universidade Federal de Santa Catarina, Florianópolis/SC
[3] Área de Telecomunicações, Instituto Federal de Educação, Ciência e Tecnologia de Santa Catarina, São José/SC

SUMÁRIO

Este artigo propõe uma metodologia para detectar strings ausentes em sistemas fotovoltaicos de grande escala (VLSPV), utilizando apenas dados de corrente adquiridos no nível das stringboxes. Usando técnicas de análise de dados e aprendizado de máquina não supervisionado, o método proposto estima o número de strings ausentes por stringbox, comparando a corrente contínua de cada stringbox com stringboxes vizinhas dentro da mesma região. Esta abordagem preenche uma lacuna na literatura existente, oferecendo uma solução adaptada ao nível típico de instrumentação de plantas VLSPV. A avaliação com dados reais demonstrou que o método proposto, em todas as configurações, atinge uma precisão superior a 90% na detecção de strings ausentes, com as melhores configurações alcançando 99,3% de precisão. Essas descobertas oferecem informações valiosas para equipes de operações e manutenção, permitindo a identificação e mitigação de strings problemáticas em plantas VLSPV.

PALAVRAS CHAVE

1.Fotovoltaica. 2.Grande Escala. 3.Desempenho. 4.Falhas. 5.Manutenção.

Contato do autor principal:
EMAIL: bruno.castro.valle@gmail.com
TELEFONE: +55 (48) 99109-4275
LATTES: https://lattes.cnpq.br/9430815457041786

1. Introdução

Embora os sistemas fotovoltaicos exijam manutenção relativamente reduzida em comparação com outros sistemas (BADR; HAMAD et al., 2021), eles são suscetíveis a falhas ocasionais e perdas de energia, especialmente em sistemas fotovoltaicos em grande escala (LSPV). Tipicamente, esses sistemas LSPV possuem uma capacidade instalada superior a 1 MW (ELMALLAH et al., 2023), enquanto os sistemas fotovoltaicos em muito grande escala (VLSPV) têm capacidade superior a 100 MW (ITO et al., 2003). Essa distinção é crucial ao considerar desafios operacionais, custos e necessidades de manutenção. Falhas não detectadas podem causar grandes perdas de energia e monetárias (BADR; HAMAD et al., 2021).

Segundo (MAPS, 2023), o desempenho insuficiente devido a anomalias em sistemas fotovoltaicos quase dobrou de 1,61% em 2019 para 3,13% em 2022. Em sistemas VLSPV, o aumento foi ainda mais pronunciado, subindo de 1,10% em 2019 para 4,04% em 2022. Estimativas indicam uma perda anual de 2,5 bilhões de dólares para a indústria fotovoltaica, com uma perda média de 4.320 dólares por MW para sistemas VLSPV (MAPS, 2023). Se a perda de energia por anomalias continuar crescendo, espera-se que atinja quase 6% até 2025. O estudo (MAPS, 2023) identificou anomalias em strings, inversores, combinadores de strings (stringboxes) e rastreadores (trackers) como os quatro principais fatores de perda de receita devido a problemas em equipamentos.

O aumento no desempenho insuficiente devido a anomalias levou a uma busca por métodos de detecção e classificação de falhas em sistemas fotovoltaicos, área que não recebia tanta atenção em instalações de grande escala (ZHAO et al., 2019). Um dos desafios em usinas de grande escala está na escassez de instrumentação detalhada para cada módulo fotovoltaico, geralmente adquirida por meio de um sistema de Supervisão, Controle e Aquisição de Dados (SCADA). O sistema SCADA frequentemente carece de granularidade para monitorar módulos ou strings individuais (ANDERSON et al., 2022), pois a instrumentação necessária muitas vezes implica em investimentos consideráveis. Nessa escala, os dados são tipicamente coletados no nível de uma stringbox (ZHAO et al., 2019).

Este trabalho foca na estimativa do número de strings ausentes por stringbox em usinas VLSPV utilizando técnicas de análise de dados e aprendizado de máquina não supervisionado. Nossa metodologia se baseia exclusivamente nos dados de corrente contínua coletados de cada stringbox, eliminando a necessidade de instrumentação por string, módulo ou dados mais granulares. Esta abordagem preenche uma lacuna na literatura existente, pois metodologias similares adequadas para usinas VLSPV ainda não foram identificadas.

Aproveitando os dados de corrente da saída das stringboxes, o método proposto compara-os com as stringboxes vizinhas para estabelecer uma média estimada de corrente saudável por string. A partir dessa comparação, estima-se o número de strings operacionais e,

consequentemente, a quantidade de strings ausentes, auxiliando o pessoal da usina a identificar strings problemáticas. A avaliação do método comparando-o com dados reais mostrou boa precisão, sendo capaz de detectar cerca de 90% das strings ausentes, um grande benefício para as equipes de operação e manutenção de usinas VLSPV.

2. Metodologia

Esta seção descreve o método proposto para estimar strings faltantes em stringboxes em usinas VLSPV, baseado em análise de dados e aprendizado de máquina, e estruturado em quatro fases. A abordagem incorpora parâmetros ajustáveis para acomodar a variabilidade entre plantas e a disponibilidade de dados fornecidos pelo sistema SCADA, garantindo uma estratégia flexível e eficiente aplicável a várias instalações fotovoltaicas.

Especificamente, o método segmenta a planta em agrupamentos geográficos para isolar efeitos locais, como microclima e envelhecimento. Em seguida, estima a corrente média de uma string saudável dentro de cada agrupamento, servindo como referência para comparação com a corrente observada em cada stringbox. Isso permite estimar a quantidade de strings operacionais e faltantes. O método utiliza exclusivamente dados da saída dos stringboxes via sistema SCADA e é executado diariamente.

2.1. Pré-processamento

O pré-processamento de dados é essencial para preparar dados brutos para análises e modelagem. Isso inclui limpeza, transformação e organização dos dados para garantir um formato adequado à extração de informações significativas, aumentando a precisão e confiabilidade dos resultados. Neste projeto, o foco inicial é processar os valores de corrente da saída de cada stringbox na planta, juntamente com os *timestamp* (registro de data e hora da medida) e a identificação da stringbox.

A primeira etapa envolve a limpeza de valores ausentes e inválidos. Um desafio significativo é o mau funcionamento dos trackers, que deve ser evitado ao analisar os dados. Assumindo que trackers com falhas são configurados para a posição de repouso (*stow* - horizontal), utilizamos uma abordagem baseada na posição dos trackers, identificando uma janela de tempo em que trackers operacionais e em repouso estão alinhados horizontalmente com o sol. Isso permite isolar o problema dos trackers com falhas das strings ausentes.

Essa janela de tempo é determinada através de uma simulação de rastreamento utilizando a biblioteca PVlib (ANDERSON et al., 2023) em Python, considerando a localização da usina, o tipo de instalação de trackers e a época do ano. A simulação fornece a posição dos trackers solares filtrada para a janela de tempo em que os trackers estão entre -5 e 5 graus em relação ao

horizonte, geralmente ao redor do meio-dia. Durante essa janela, calcula-se a mediana dos valores de corrente de cada stringbox em um determinado dia, resultando em um valor diário para cada stringbox (I_n). A corrente estimada em cada string conectada ao n-ésimo stringbox é então obtida pela fórmula

$$I_{n,m} = \frac{I_n}{M_n}, \quad (1)$$

onde M_n é o número de strings conectadas ao n-ésimo stringbox.

2.2. Agrupamento geográfico

O objetivo desta fase é estabelecer microrregiões dentro da usina fotovoltaica que exibem perfis de irradiação solar ou desempenho esperado semelhantes, permitindo resultados mais precisos, uma vez que as strings nessas regiões devem ter desempenho similar. Esta abordagem é relevante devido à vasta área física ocupada por usinas VLSPV, que cobrem áreas substanciais. Consequentemente, essas usinas podem enfrentar condições de irradiação solar variadas em diferentes seções, como observado em (ZHAO et al., 2019), que abrange mais de 400.000 painéis fotovoltaicos espalhados por uma ampla área. O agrupamento geográfico visa mitigar essas variações, proporcionando uma análise mais precisa.

Para alcançar este objetivo, o método requer coordenadas de latitude e longitude para cada stringbox, permitindo o agrupamento automatizado. Utilizando o bem conhecido algoritmo k-means (MACQUEEN, 1967), empregamos os dados geográficos para fins de agrupamento. Especificando o número desejado de clusters (agrupamentos) geográficos, denotado aqui por K, esta abordagem assegura um método direto e adaptável. A partir de agora, cada cluster é tratado de forma independente nas fases subsequentes. Podemos observar um exemplo de agrupamento geográfico na Figura 1, onde uma área é separada em $K = 4$ clusters. Em nosso contexto, os pontos representam a posição de cada Grupo de Inversores (ITS).

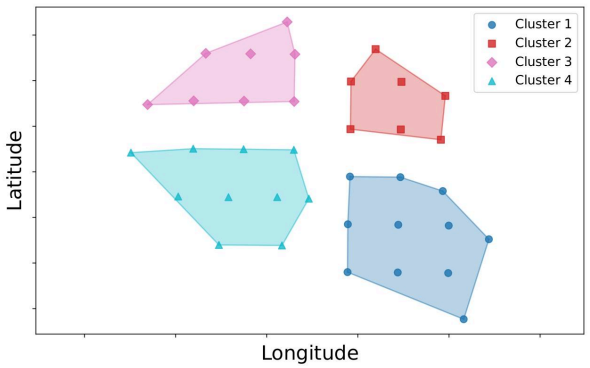


Figura 1 - Ilustração mostrando o agrupamento geográfico com $K = 4$ realizado usando k-means, com cada ponto representando um ITS.

2.3. Agrupamento de correntes

Esta fase visa estabelecer um valor estimado para a corrente de uma string saudável, tratando cada cluster geográfico separadamente. Denotamos a corrente estimada de uma string saudável no k-ésimo cluster geográfico, $k \in 1, ..., K$, como H_k. Para obter essa estimativa, realizamos um agrupamento das correntes de strings do k-ésimo cluster geográfico usando o algoritmo k-means, separando-as em L clusters. O cluster com a maior corrente mediana é selecionado como o cluster saudável, resultando em H_k. A Figura 2 ilustra um exemplo de agrupamento de correntes com $L = 4$ clusters. Cada ponto representa a corrente média da string em uma stringbox, e o eixo das abscissas é ordenado pelo número do ITS, onde cada ITS contém várias stringboxes. Os pontos no formato quadrado representam o cluster com a maior corrente média por string, sendo a corrente mediana desse cluster a referência de string saudável estimada para o cluster geográfico.

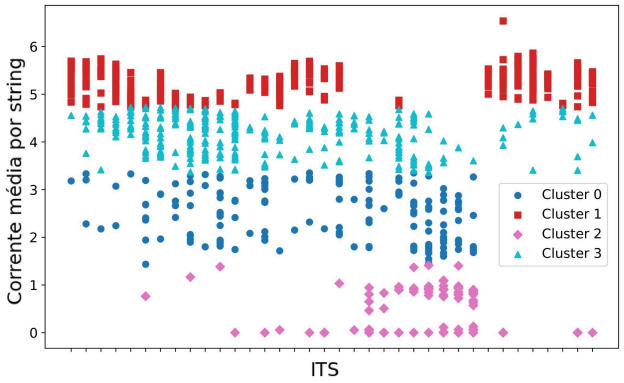


Figura 2 - Um exemplo de agrupamento de correntes com $L = 4$ clusters, usando k-means, onde cada ponto representa a corrente média da string em uma determinada stringbox.

2.4 Strings ausentes

Com a corrente estimada de uma string saudável em cada cluster geográfico, a estimativa do número de strings faltantes pode prosseguir. Primeiro, para cada stringbox n pertencente ao k-ésimo cluster geográfico, o número de strings conectadas dentro daquele stringbox pode ser determinado como

$$M_n^{con} = \frac{I_n}{H_k}. \quad (2)$$

Em seguida, o número de strings faltantes em cada stringbox pode ser estimado como

$$M_n^{mis} = M_n - M_n^{con}. \quad (3)$$

Embora o resultado seja um número real, na prática, ele deve ser um inteiro, obtido por arredondamento. A Figura 3 apresenta um diagrama resumindo a metodologia proposta, que inclui parâmetros ajustáveis, como o tamanho do cluster geográfico (K) e o tamanho do cluster de corrente (L), para melhor acomodar variações de desempenho e sensibilidade do método.

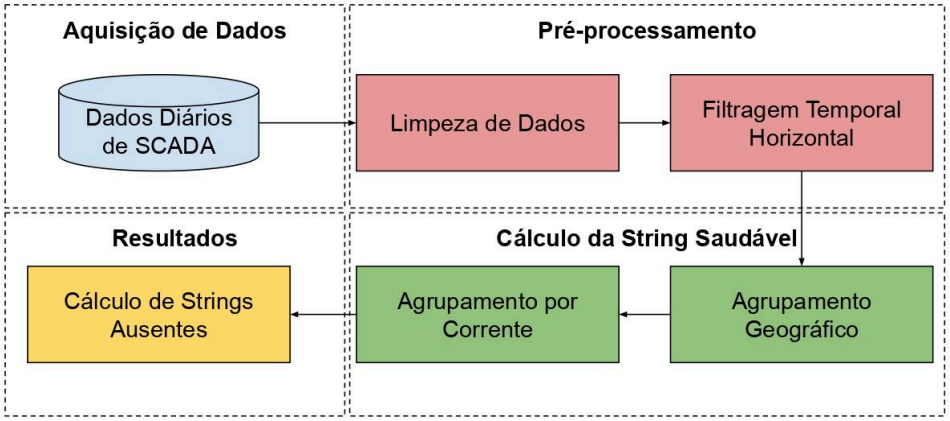


Figura 3 - Diagrama ilustrando a metodologia proposta, delineando a aquisição de dados, pré-processamento, estimativa da string saudável e, finalmente, o cálculo das strings faltantes.

3. Resultados

Nesta seção são apresentados os resultados da aplicação do método proposto a dados reais de uma usina VLSPV, incluindo o efeito dos parâmetros ajustáveis. Os resultados são discutidos em termos de eficácia na estimativa de strings faltantes, análise de perdas de corrente e sua aplicabilidade em diferentes condições operacionais da usina. Além disso, são exploradas as variações de desempenho observadas em função dos parâmetros ajustáveis do método, como o tamanho dos clusters geográficos e de corrente.

O método proposto possui dois parâmetros ajustáveis: o número de clusters geográficos K e o número de clusters de corrente L. Esses parâmetros podem ser utilizados para, sob a disponibilidade de dados de referência, ajustar o método às condições de uma usina específica. A seguir, ilustramos o comportamento do método proposto sob diferentes configurações.

3.1. Agrupamento geográfico

O número de microrregiões (K) no agrupamento geográfico influencia a quantidade de stringboxes dentro de cada cluster de corrente, usadas para estimar valores de corrente saudável. Menos clusters geográficos resultam em conjuntos de dados maiores para calcular a corrente saudável, podendo incluir mais stringboxes de alto desempenho e aumentando a corrente "saudável" estimada. Isso pode diminuir as estimativas de strings operacionais e aumentar as de

strings faltantes. Por outro lado, mais clusters geográficos permitem uma análise mais granular, acomodando microclimas ou efeitos de baixo desempenho, resultando em um desempenho mais uniforme entre as stringboxes dentro de cada cluster. Isso levaria a uma corrente saudável menor e a menos strings faltantes estimadas. No entanto, poucos clusters geográficos podem gerar resultados não confiáveis devido à falta de dados suficientes para um agrupamento de corrente preciso. Portanto, escolher um K adequado é crucial na abordagem proposta.

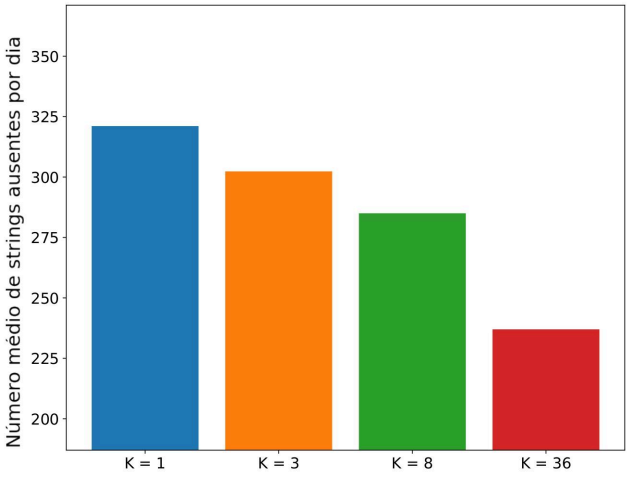


Figura 4 – Número médio estimado de strings faltantes por dia, para diferentes números de clusters geográficos K para um número fixo de clusters de corrente $L = 2$, em um período de 15 dias.

O impacto da escolha do parâmetro K está apresentado na Figura 4. Ao considerar um único cluster geográfico ($K = 1$), todas as stringboxes são agrupadas em uma única região, aumentando a corrente saudável de referência e, consequentemente, aumentando a estimativa de strings faltantes, possivelmente gerando mais detecções falsas. À medida que o número de clusters geográficos aumenta, a corrente saudável estimada naturalmente diminui, pois as stringboxes de alto desempenho podem agora ser divididas em diferentes clusters geográficos, resultando em uma estimativa menor de strings faltantes.

Utilizar um número muito grande de clusters geográficos pode subestimar a corrente saudável de um cluster, reduzindo a estimativa de strings faltantes e, potencialmente, deixando strings com falhas não detectadas.

3.2. Agrupamento de correntes

Ao contrário dos clusters geográficos, um maior número de clusters de corrente eleva o critério de filtragem, levando a uma corrente saudável estimada mais alta. Isso ocorre porque apenas o cluster com a maior mediana de corrente é usado para estimar a corrente saudável por string. Portanto, à medida que o número de clusters de corrente aumenta, a mediana do cluster

"saudável" tende a subir, resultando em um menor número estimado de strings operacionais e, consequentemente, uma maior estimativa de strings faltantes.

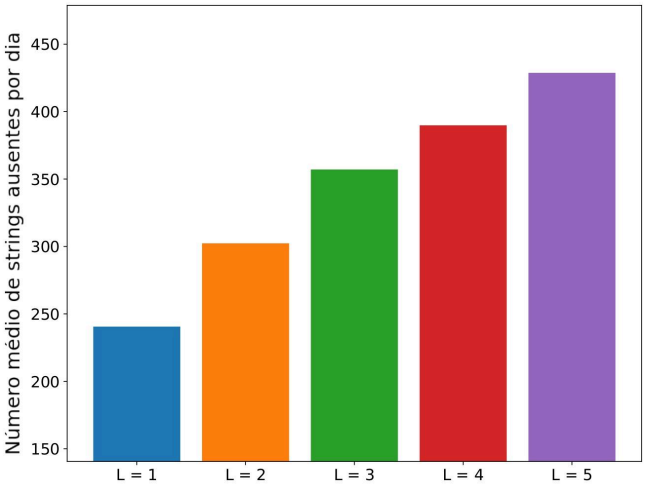


Figura 5 – Número estimado de strings faltantes por dia, para diferentes números de clusters de corrente L para um número fixo de $K = 4$ clusters geográficos, em um período de 15 dias.

A Figura 5 ilustra o impacto de K na estimativa de strings faltantes. Consideramos variar o número de clusters de corrente entre $L = 1$ e $L = 5$, enquanto mantemos o número de clusters geográficos fixo em $K = 4$. Novamente, os resultados representam a média do número estimado de strings faltantes por dia ao longo de um período de 15 dias. Os resultados reforçam a intuição de como o método se comporta para diferentes números de clusters de corrente, de modo que o número estimado de strings faltantes aumenta com L. Definir o par L, K de melhor desempenho só é possível com base em dados de referência.

3.3. Validação com dados em campo

O método proposto foi validado usando dados de campo rotulados coletados ao longo de um único dia, onde foram registrados o estado de um subconjunto de strings da usina VLSPV. A coleta desses dados é cara e demorada, justificando a realização do teste em apenas um dia. Durante esse período, foram observadas várias falhas significativas, incluindo um inversor com defeito afetando oito stringboxes, um fusível aberto no barramento de um inversor e danos menores em várias strings, como fusíveis queimados e strings abertas, impactando um total de 265 strings.

As métricas usadas para avaliar o desempenho do método são:

- MAE (Erro Absoluto Médio): o erro absoluto médio da estimativa do número de strings faltantes;
- FP (Falsos Positivos): o número de strings operacionais marcadas como faltantes;
- FN (Falsos Negativos): o número de strings faltantes marcadas como operacionais;
- TP (Verdadeiros Positivos): o número de strings faltantes marcadas como faltantes;
- TN (Verdadeiros Negativos): o número de strings operacionais marcadas como operacionais;
- P (Precisão), $P = TP/(TP + FP)$, a proporção de strings marcadas como faltantes e que estão em falha;
- R (Recall), $R = TP/(TP + FN)$, a proporção de strings faltantes que foram corretamente marcadas;
- F1-Score: a média harmônica entre P e R, calculado como $F_1 = 2PR/(P + R)$.

A Tabela 1 mostra os resultados da aplicação da abordagem proposta para estimar strings faltantes, considerando dados reais rotulados. Durante os testes, variamos o número de clusters geográficos e de corrente, K e L, respectivamente. Note que o número de strings faltantes em cada stringbox é aproximado para o inteiro mais próximo. Primeiramente, observe que o desempenho do método, sendo um algoritmo não supervisionado, é bastante robusto, já que todas as configurações analisadas alcançaram um F-Score acima de 90%. Além disso, a variação dos parâmetros ajustáveis oferece a possibilidade de ajustar o modelo com base na métrica mais relevante em cada cenário. O menor MAE foi alcançado, por exemplo, ao separar a planta em $K = 4$ clusters geográficos, com $L = 3$ clusters de corrente, também alcançando o maior F-Score. Curiosamente, o mesmo desempenho foi obtido com $K = 1$ e $L = 3$. Um desempenho notável também foi obtido com $K = 4$ e $L = 2$, alcançando o mesmo MAE, mas com tanto R quanto P acima de 90%, um bom equilíbrio entre FP e FN.

Tabela 1 - Validação com dados reais para vários números de clusters geográficos (K) e de corrente (L).

K	L	MAE	FP	FN	R %	P %	F1 %
1	2	0,487	23	16	92,0	94,3	93,2
1	3	0,438	31	4	89,5	98,5	93,8
1	4	0,700	54	2	83,1	99,3	90,4
1	5	0,700	54	2	83,1	99,3	90,4
4	2	0,438	22	13	92,3	95,3	93,8
4	3	0,438	31	4	89,5	98,5	93,8

4	4	0,625	48	2	84,7	99,3	91,4
4	5	0,650	50	2	84,1	99,3	91,1
36	2	0,450	25	11	91,4	96,0	93,6
36	3	0,525	33	9	88,9	96,7	92,7
36	4	0,713	48	9	84,7	96,7	90,3
36	5	0,688	51	4	83,9	98,5	90,6

4. Conclusão

A abordagem proposta enfrenta o desafio de detectar strings ausentes em plantas VLSPV utilizando técnicas de análise de dados e aprendizado de máquina. O método estima a quantidade de strings ausentes por stringbox, auxiliando o pessoal no local a identificar strings problemáticas sem a necessidade de instrumentação granular cara e impraticável em instalações de grande escala. Baseando-se exclusivamente em dados de corrente coletados de cada stringbox, a técnica se mostra eficiente e economicamente viável.

A validação do método contra dados reais mostrou uma precisão acima de 90% na detecção de strings ausentes em diversas configurações. Observamos que o método funciona bem em dias com alta irradiância, mas requer ajustes no número de clusters geográficos (K) em dias com baixa irradiância para obter estimativas mais precisas. A abordagem demonstrou-se robusta e ajustável às necessidades específicas das operações e equipes de manutenção, proporcionando um equilíbrio entre a detecção de falsos positivos e falsos negativos. Em suma, a metodologia proposta não apenas facilita a detecção de strings ausentes, mas também oferece informações valiosas para a operação e manutenção de plantas VLSPV, melhorando a eficiência e reduzindo as perdas energéticas associadas.

BIBLIOGRAFIA

[1] BADR, Mohamed M.; HAMAD, Mostafa S. et al. Fault Identification of Photovoltaic Array Based on Machine Learning Classifiers. IEEE Access, Institute of Electrical and Electronics Engineers Inc., v. 9, p. 159113–159132, 2021. DOI: 10.1109/ACCESS.2021.3130889.

[2] ELMALLAH, Salma et al. Shedding light on large-scale solar impacts: An analysis of property values and proximity to photovoltaics across six U.S. states. Energy Policy, Elsevier Ltd, v. 175, abr. 2023. ISSN 03014215. DOI: 10.1016/j.enpol.2023.113425.

[3] ITO, Masakazu et al. A preliminary study on potential for very large-scale photovoltaic power generation (VLS-PV) system in the Gobi desert from economic and environmental viewpoints. v. 75. [S.l.: s.n.], 2003. P. 507–517.

[4] MAPS, Raptor. 2023 Global Solar Report. 2023. Disponível em: https://raptormaps.com/resources/2023-global-solar-report.

[5] ZHAO, Yingying et al. Hierarchical anomaly detection and multimodal classification in large-scale photovoltaic systems. IEEE Transactions on Sustainable Energy, Institute of Electrical and Electronics Engineers Inc., v. 10, p. 1351–1361, 3 jul. 2019. ISSN 19493037. DOI: 10.1109/TSTE.2018.2867009.

[6] ANDERSON, Kevin et al. A Method for Estimating Time-Series PV Production Loss from Solar Tracking Failures. IEEE Journal of Photovoltaics, IEEE Electron Devices Society, v. 12, p. 119–126, 1 jan. 2022. ISSN 21563403. DOI: 10.1109/JPHOTOV.2021.3123872.

[7] ANDERSON, Kevin, Hansen, C., Holmgren, W., Jensen, A., Mikofski, M., and Driesse, A. “pvlib python: 2023 project update.” Journal of Open Source Software, 8(92), 5994, (2023). DOI: 10.21105/joss.05994.

[8] MACQUEEN, J. B. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, v. 1, p. 281–297, 1967.