Dimensionality Reduction of Environmental Data for Long-Term PV Performance Analysis Using Graph Based Methods

Srijani Mukherjee, Laurent Vuillon, Denys Dutykh, Ioannis Tsanakas

DIMENSIONALITY REDUCTION OF ENVIRONMENTAL DATA FOR LONG-TERM PV PERFORMANCE ANALYSIS USING GRAPH BASED METHODS

Srijani Mukherjee[1,2], Laurent Vuillon[2] , Denys Dutykh[3], and Ioannis Tsanakas[1,*]
[1]Univ. Grenoble Alpes, CEA, Liten, 73375 Le Bourget du Lac, France
[2]Univ. Savoie Mont Blanc, CNRS, LAMA, Chambéry, 73000, France
[3]Mathematics Department, Khalifa University, Abu Dhabi, PO Box 127788, United Arab Emirates
*Corresponding author: ioannis.tsanakas@cea.fr

ABSTRACT:

This study presents an innovative approach to solar photovoltaic (PV) performance analysis through the integration of temperature and solar irradiance data using a graph-based community detection method. While previous research often considered these factors independently, our method captures their complex interactions, offering a more comprehensive insight into the behaviour of solar PV systems. We introduce a novel data reduction technique that identifies 10 representative days from a year's worth of hourly data, significantly reducing computational requirements while preserving crucial microclimate patterns relevant to PV performance. Our methodology employs a graph-based community detection algorithm, representing days as nodes and similarities as edges, enabling the identification of non-spherical clusters that better represent complex weather patterns. We evaluated three normalization techniques—standard, min max, and robust—finding minmax normalization to yield the best-defined clusters for our dataset. The effectiveness of our approach is demonstrated using hourly average temperature and irradiance data. Results show that our method successfully captures annual weather patterns while substantially reducing data volume. This research will contribute to more efficient and accurate solar PV performance evaluations, balancing comprehensive data analysis with computational efficiency.

1 INTRODUCTION

The effectiveness of solar photovoltaic (PV) systems hinges on a thorough understanding of their interaction with the environment. Temperature and solar irradiance are two key factors that significantly affect power generation [1, 2]. A comprehensive knowledge of how these factors vary over time throughout the year is crucial for optimal system design. Solar irradiance, or sunlight intensity, directly dictates the electrical energy produced. Accurate modeling of irradiance variations caused by cloud cover, seasonality, and geographic location is essential for precise energy yield estimates [3]. Temperature plays an equally important role in PV system performance. As solar panels heat up, their efficiency declines, a phenomenon quantified by the temperature coefficient. Typically, for every degree Celsius increase in panel temperature, power output drops by more than 0.5% [4]. This inverse relationship between temperature and efficiency underscores the importance of considering local temperature patterns throughout the year. Accurate knowledge of these patterns is essential for generating realistic energy production estimates and optimizing system performance. However, analyzing yearlong environmental data is computationally expensive and time-consuming. Data reduction algorithms offer an efficient solution to this challenge. By selecting representative days, we can accelerate performance analysis and free up resources while providing meaningful insights into temporal trends. Various techniques have been proposed for data reduction, each having its own strengths and limitations. Typical Meteorological Year (TMY) analysis [5] and extreme value analysis (EVA) [6] are common approaches, but they may not capture the full spectrum of weather patterns. Clustering provides a more comprehensive data reduction approach for weather data, capturing both typical and extreme weather patterns [7].

In this paper, we address two key challenges in analyzing environmental data for solar PV performance analysis: information fusion and temporal representation. We present Graph-Oriented Information Fusion (GOIF), a novel approach to the information fusion of temperature and solar irradiance data. GOIF employs a community detection method and the PageRank technique to identify a subset of representative days that effectively capture diverse weather patterns throughout the year. This approach enables efficient analysis of large datasets while preserving crucial microclimate information. Importantly, GOIF represents an explainable AI approach, providing transparency in its decision-making process and allowing for interpretable results. By bridging the gap between comprehensive data analysis and practical applicability, our method offers a robust framework for improving solar photovoltaic performance predictions and system design optimizations.

2 METHODOLOGY

Our Graph-Oriented Information Fusion (GOIF) approach for analyzing solar PV performance data consists of several key steps. This section outlines the data source, preprocessing techniques, and the core algorithm used in our study.

2.1 Feature Extraction

Our approach begins with extracting informative features from daily temperature (T) and irradiance (Q) data. For each day i, we compute a feature vector fi, comprising several statistical measures. Mean (μ) represents the average value for each variable $\mu_{ix} = (\Sigma_{k=1}^{M} x_k) / M$. Standard deviation (σ) measures data dispersion around the mean $\sigma_{ix} = \sqrt{[(\Sigma_{j=1}^{M} (x_j - \mu_i)^2) / M]}$. Minimum (min) and Maximum (max), provide data range and identify potential outliers. Quartiles (Q^1, Q^2, Q^3) offer insights into data distribution and central tendency.

10.4229/EUPVSEC2024/4CV.1.27
020411-001

The comprehensive feature vector for day i is:
$f_i = [\mu^T_i, \sigma^T_i, min^T_i, max^T_i, Q^{1T}_i, Q^{2T}_i, Q^{3T}_i, \mu Q_i, \sigma Q_i, minQ_i, maxQ_i, Q^1Q_i, Q^2Q_i, Q^3Q_i]$

To balance model performance and computational efficiency, we focus on mean temperature (μ^T_i) and mean irradiance (μ^Q_i), creating a simplified two-dimensional feature vector $f_i = [\mu^T_i, \mu^Q_i] \in \mathbb{R}^2$. This approach results in a characteristic matrix $F \in \mathbb{R}^{Nx2}$, where N is the total number of days, effectively representing average daily weather conditions throughout the year.

2.2 Data Source and Preprocessing

For this study, we utilized hourly average temperature and irradiance data from the solar panel installation at INES (Institut National de l'Énergie Solaire), located in Le Bourget-du-Lac, France (Latitude: 45.64395818844°N, Longitude: 5.875884919217°E, Elevation: 233m). The dataset spans a full year. Prior to applying our algorithm, we applied three different normalization techniques to align all features and improve clustering performance. **Standard normalization,** transforms data to have zero mean and unit variance, defined as $z = (x - \mu) / \sigma$, where μ is the mean and σ is the standard deviation of the feature. **Min max normalization**, scales data to a fixed range [0, 1], given by $z = (x - min(x)) / (max(x) - min(x))$, where min(x) and max(x) are the minimum and maximum values of the feature, respectively. **Robust normalization,** centers data around the median and scales it based on the interquartile range, expressed *as* $z = (x - median(x)) / IQR(x)$, where IQR(x) is the interquartile range of the feature.

2.3 Algorithm Implementation

Graph-Oriented Information Fusion (*GOIF*) algorithm consists of four main steps.

Graph Construction: We represent our dataset as an undirected graph G = (V, E), where V is the set of nodes, each representing a day in the dataset and E is the set of edges, encoding similarities between weather profiles. We employ a k-nearest neighbors (k-NN) approach to determine edge connections. For each node v_i (i = 0 to N-1, where N is the total number of days), k nearest neighbors Ni(k) are determined based on euclidean distances to all other nodes. Edges are then created between vi and each vj ∈ Ni(k). A node may have ≥ k edges due to reciprocal connections. Here k is a user-defined parameter determining the number of nearest neighbors.

Community Detection: We apply the Louvain Modularity Maximization algorithm [8] to identify distinct communities within the graph, optimizing the modularity of graph partitions to effectively group similar days together. By adjusting the resolution parameter, we can control the number of communities identified; in this study, we set the resolution to 0.9, resulting in 10 distinct communities. This parameter can be modified to achieve different numbers of desired communities based on specific analysis needs.

PageRank Application: Within each identified community, we apply the PageRank algorithm [9] to determine the most central node. The PageRank algorithm iteratively calculates the importance of each node based on the importance of its incoming connections, assigning higher scores to nodes that are linked to by many high-scoring nodes. PageRank scores are initialized uniformly and updated iteratively based on the equation:

$$PR(i) = (1-d)/N + d * \Sigma(PR(j) / deg(j))$$

Where PR(i) is the PageRank score of node i, d is the damping factor (typically set to 0.85) modeling random navigation, N is the total number of nodes, and the sum is over all nodes j that have an edge to node i. This approach allows us to identify the most influential or representative day within each weather pattern cluster.

Representative Day Selection: We select the node with the highest PageRank score in each community as the representative day for that cluster. This process results in a set of 10 representative days that capture the essential characteristics of the annual temperature and irradiance patterns.

2.4 Evaluation Metric

To assess cluster quality, we use the Average Intra-Cluster Standard Deviation ($\bar{\sigma_I}$) as our primary metric. This is calculated in three steps.

1. Standard Deviation within Each Cluster (σ): For each cluster k, we compute the standard deviation of temperature (T) and irradiance (Q).
$\sigma_{k,T} = \sqrt{[\Sigma(T_i - \bar{T}_k)^2 / (N_k - 1)]}$
$\sigma_{k,Q} = \sqrt{[\Sigma(Q_i - \bar{Q}_k)^2 / (N_k - 1)]}$
Where N_k is the number of days in cluster k, T_i and Q_i are daily values, and $\bar{T}_k$ and $\bar{Q}_k$ are cluster means.

2. Intra-Cluster Standard Deviation (σ_I): We average the standard deviations of temperature and irradiance.
$\sigma_{I,k} = (\sigma_{k,T} + \sigma_{k,Q}) / 2$

3. Average Intra-Cluster Standard Deviation ($\bar{\sigma_I}$): We calculate the mean of $\sigma_{I,k}$ across all K clusters.
$\bar{\sigma_I} = \Sigma\, \sigma_{I,k} / K$

A lower $\bar{\sigma_I}$ indicates tighter clusters with smaller variations, reflecting a more effective normalization technique for the community detection process. Thus we aim to provide a comprehensive and efficient approach to analyze environmental data, balancing data reduction with the preservation of crucial microclimate information.

3 RESULTS AND DISCUSSION

Our Graph-Oriented Information Fusion (GOIF) approach yielded several significant findings in the analysis of environmental data. This section presents our key results and discusses their implications.

We evaluated three normalization techniques, standard, minmax, and robust normalization. The effectiveness of each technique was assessed using the Average Intra-Cluster Standard Deviation ($\bar{\sigma_I}$) metric.

Normalization Technique	$\bar{\sigma_I}$
Standard	1.55
Minmax	1.25
Robust	1.60

10.4229/EUPVSEC2024/4CV.1.27
020411-002

Table 1: Comparison of Normalization Techniques

As seen in Table 1, Minmax normalization produced the lowest $\bar{\sigma_i}$ value, indicating that it resulted in the most coherent clustering of days. The Louvain Modularity Maximization algorithm identified 10 distinct communities within our dataset which is illustrated in Figure 1, with nodes colored by community.

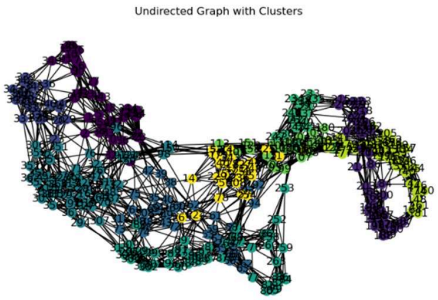
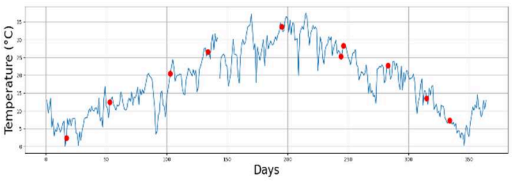
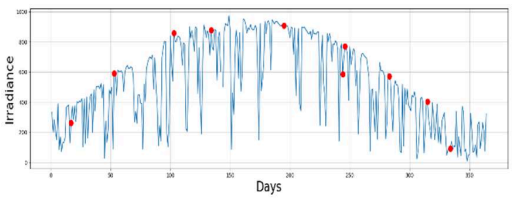


Figure 1: Community Graph with Minmax Normalization

The PageRank algorithm successfully identified the most central node within each community, representing the most typical day for that weather pattern. Figure 2 shows the distribution of daily average temperature and irradiance over the year, with the 10 representative days marked as red dots.



(a)



(b)

Figure 2: Distribution of (a) Temperature and (b) Irradiance with Representative Days

As evident from Figure 2, the representative days are well distributed throughout the year, capturing both seasonal variations and extreme weather events. This distribution validates the effectiveness of our GOIF approach in selecting days that comprehensively represent the annual weather patterns. Our method successfully reduced a year's worth of hourly data to 10 representative days, achieving a data reduction ratio of 36.5:1. Despite this significant reduction, the selected days maintain the essential characteristics of the annual temperature and irradiance patterns, as demonstrated by the low $\bar{\sigma_i}$ value and the distribution of representative days. The GOIF approach offers several advantages for solar PV performance analysis. By reducing the dataset to 10 representative days, our method significantly decreases computational requirements for subsequent analyses. The selected days capture both typical and extreme weather patterns, providing a balanced dataset for performance evaluations. The distribution of representative days across the year ensures that seasonal variations in temperature and irradiance are adequately represented. The graph-based approach and PageRank algorithm provide transparency in the selection process, allowing for interpretable results.

4 CONCLUSIONS AND FUTURE WORKS

Despite promising results, our approach has limitations. The current study is based on data from a single location, and the generalizability of the method to diverse geographical areas needs further investigation. Additionally, the optimal number of representative days may vary depending on the specific application and desired level of detail. Future work is focused on validating the GOIF method across diverse geographical locations and climates, exploring the impact of different numbers of representative days on analysis accuracy. We will incorporate additional environmental variables, such as wind speed or humidity, into the GOIF framework and other PV system specific parameters. Our aim is to develop a user-friendly tool for implementing the GOIF approach in solar PV system design and optimization.

In conclusion, our Graph-Oriented Information Fusion approach demonstrates significant potential for enhancing solar PV performance analysis. By effectively reducing data while preserving crucial information, GOIF enables more efficient and comprehensive evaluations of solar PV systems, potentially leading to improved system designs and more accurate energy yield predictions.

5 REFERENCES

[1] S. Dubey, J. N. Sarvaiya, and B. Seshadri, "Temperature dependent photovoltaic (pv) efficiency and its effect on pv production in the world – a review," Energy Procedia, vol. 33, pp. 311–321, 2013, pV Asia Pacific Conference 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1876610213000829

[2] C. Cornaro and A. Andreotti, "Influence of average photon energy index on solar irradiance characteristics and outdoor performance of pv modules," Progress in Photovoltaics: Research and Applications, vol. 21, 04 2012.

[3] K. Hasan, S. B. Yousuf, M. S. H. K. Tushar, B. K. Das, P. Das, and M. S. Islam,
"Effects of different environmental and operational factors on the pv performance: A
comprehensive review," Energy Science & Engineering, vol. 10, no. 2, pp. 656–675,
2022. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/ese3.1043

[4] B. R. Paudyal and A. G. Imenes, "Investigation of temperature coefficients of pv modules through field measured data," Solar Energy, vol. 224, pp. 425–439,

2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0038092X21004837

[5] O. Kilanko, S. O. Oyedepo, J. O. Dirisu, R. O. Leramo, P. Babalola, A. K. Aworinde, M. Udo, A. M. Okonkwo, and M. I. Akomolafe, "Typical meteorological year data analysis for optimal usage of energy systems at six selected locations in Nigeria," International Journal of Low-Carbon Technologies, vol. 18, pp. 637–658, 05 2023. [Online]. Available: https://doi.org/10.1093/ijlct/ctad014.

[6] D. Clarkson, E. Eastoe, and A. Leeson, "The importance of context in extreme value analysis with application to extreme temperatures in the U.S. and Greenland," Journal of the Royal Statistical Society Series C: Applied Statistics, vol. 72, no. 4, pp. 829–843, 02 2023. [Online]. Available: https://doi.org/10.1093/jrsssc/qlad020

[7] A. Arroyo, V. Tricio, E. Corchado, and A. Herrero, A Comparison of Clustering Techniques for Meteorological Analysis, 05 2015, pp. 117–130.

[8] S. Fortunato, "Community detection in graphs," Physics Reports, vol. 486, no. 3, pp. 75–174, 2010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0370157309002841

[9] Solanki, S., Verma, S., Chahar, K. (2022). A Comprehensive Study of Page-Rank Algorithm. In: Bhateja, V., Tang, J., Satapathy, S.C., Peer, P., Das, R. (eds) Evolution in Computational Intelligence. Smart Innovation, Systems and Technologies, vol 267. Springer, Singapore. https://doi.org/10.1007/978-981-16-6616-2_1