Journal of Information Organization

DLINE JOURNALS

Print ISSN : 2278 – 6503
Online ISSN: 2278 – 6511

JIO 2026: 16 (3)
https://doi.org/10.6025/jio/2026/16/3/99-121

Spatial Clustering of the NREL PVDAQ Photovoltaic Systems: Geographic Structure of a National Performance Dataset

Olaniyi Olawale Omoyajowo, Ezendu Ariwa and Yue Yong
Department of Computer Science and Technology
University of Bedfordshire, United Kingdom

ABSTRACT

The increasing deployment of photovoltaic (PV) systems has created a growing need for reliable, geographically representative datasets for performance assessment, degradation analysis, and power forecasting. However, the spatial organization of large PV monitoring archives is often not well understood, which can introduce sampling bias and limit the generalizability of predictive models. This study presents a spatial clustering analysis of the National Renewable Energy Laboratory (NREL) Photovoltaic Data Acquisition (PVDAQ) public dataset, examining the geographic structure of 1,862 PV systems distributed across the United States. Using latitude and longitude as clustering variables, three complementary unsupervised methods are applied: K-Means, DBSCAN, and hierarchical Ward clustering. Internal validation using silhouette scores indicates strong and consistent spatial structure. K-Means identifies 16 macro regions with a silhouette score of 0.701, hierarchical Ward clustering identifies 17 similar macro regions with a silhouette score of 0.697, and DBSCAN identifies 34 denser local clusters while classifying 5.16% of systems as spatial outliers. The resulting clusters correspond to recognizable deployment and climatic regions, with pronounced concentration in California and the Northeast corridor. The findings demonstrate that the PVDAQ network is not randomly distributed but exhibits a clear geographic architecture. These spatial strata provide a foundation for stratified sampling, regional performance analysis, outlier detection, and future spatially aware PV forecasting models, including deep learning and physics informed machine learning approaches.

Keywords: Photovoltaic Systems, PVDAQ Dataset, Spatial Clustering, K-Means, DBSCAN, Hierarchical Ward Clustering, Geographic Structure, PV Forecasting, Physics-Informed Machine Learning

Received: 7 January 2026, Revised: 29 April 2026, Accepted: 5 June 2026

1. Introduction

High quality spatiotemporal characterization of energy systems, including photovoltaic (PV) and concentrating solar power (CSP) technologies, has historically been limited by sparse, incomplete, or inaccessible datasets. Privacy restrictions and paywalls have further constrained access to system level information, thereby hindering regional and global scale investigations of an increasingly important component of the evolving energy landscape. These limitations have affected research and applications related to energy system design, distribution, monitoring, and performance assessment. In response, numerous studies and initiatives have sought to improve the spatiotemporal characterization of solar energy through remote sensing, manual digitization, crowdsourcing, and machine learning techniques [1-17]. In parallel, organizations have developed databases containing value added attributes to support a wide range of energy related applications. [18-21].

The increasing availability of photovoltaic field performance data provides an important opportunity to improve understanding of how PV systems perform under diverse operating and environmental conditions. Improved analysis and reporting of PV field performance can increase confidence that deployed systems will operate as expected. This requirement has become particularly important with the emergence of advanced module technologies, including passivated emitter and rear cell (PERC), heterojunction (HJT), and bifacial modules, which introduce new degradation mechanisms and performance characteristics. In this context, the FY19–21 Reducing Uncertainty project leveraged the rapidly expanding PV fleet to develop improved models and knowledge concerning the field performance of both established and emerging PV technologies [22].

Beyond performance characterization, photovoltaic power prediction has become an important component of renewable energy management. Accurate forecasting supports energy management and contributes to the reliable integration of solar generation into power systems. Artificial neural networks (ANNs) have consequently been explored for modeling PV ageing and forecasting power generation. Such approaches consider multiple factors affecting PV output and employ a range of neural network architectures, including recurrent neural networks, autoencoders, and convolutional neural networks. Real world PV power data can be collected and preprocessed to facilitate the training and evaluation of these models [23].

The importance of accurate PV forecasting has increased alongside the rapid transformation of the global energy structure. Photovoltaic generation has emerged as a central component of green and low carbon power systems [24]. Nevertheless, PV output is characterized by considerable stochasticity and volatility, creating substantial challenges for reliable grid scheduling [25]. Consequently, effective PV forecasting requires more than increasing algorithmic complexity; it also requires a detailed understanding of the physical evolution of PV components under specific climatic and environmental conditions. Meteorological forecasting is particularly important because accurate prediction of weather related factors can enhance solar energy output forecasting [26] [M. Zhang et]. Furthermore, environmental and operational conditions exert substantial physical influences on PV-module performance [27].

1.1 Deep Learning for PV Forecasting and Spatiotemporal Modeling

Recent developments in deep learning have demonstrated considerable potential for PV forecasting. Among these approaches, convolutional neural networks (CNNs) have shown effectiveness in extracting features and supporting multi time scale power prediction through their end to end mapping capabilities [27]. These developments indicate that deep learning can capture complex relationships within PV datasets while reducing the need for manually engineered representations of system behavior.

Despite their predictive capabilities, deep learning models have frequently been criticized for their limited interpretability. Recent research has therefore explored architectures that incorporate static information and self attention mechanisms to develop more interpretable forecasting models, thereby addressing some of the “black box” limitations associated with deep learning in power system applications [28]. Such approaches are particularly relevant to PV forecasting because reliable prediction should ideally be accompanied by an understanding of the factors governing the predicted behavior.

The incorporation of temporal and spatial dependencies has further expanded the capabilities of deep learning for energy system analysis. For example, a transient stability assessment method based on GCN LSTM has been proposed to improve the perception of dynamic system evolution by jointly modeling the spatiotemporal characteristics of disturbance responses [29]. Similarly, Ghimire S et al. [30] employed a CNN LSTM architecture to achieve joint spatiotemporal feature modeling. These studies demonstrate the growing recognition that complex energy systems cannot always be adequately represented through purely temporal or purely spatial models. Instead, their behavior is influenced by interactions among spatial configuration, temporal evolution, environmental conditions, and system states.

1.2 Integration of Physical Mechanisms into PV Forecasting

As research in PV forecasting has progressed, increasing attention has been directed toward integrating physical mechanisms with data driven learning. A recent review by J. H. Yousif [31] systematically categorized PV forecasting approaches into three major classes: physics informed machine learning, optimized physical models, and physics guided models. The review emphasized that incorporating scientific theories and physical knowledge can substantially reduce the dependence of predictive models on large volumes of training data. Empirical investigations have likewise demonstrated the advantages of hybrid modeling strategies over single-model approaches for predicting electrical performance [32].

The motivation for such hybrid approaches arises partly from the limitations of conventional data-driven models. Although these models can achieve high statistical accuracy, their inherent “black box” nature can result in a lack of physical consistency. According to the Physics Informed Machine Learning (PIML) theory proposed by Karniadakis et al, purely data driven models may produce predictions that violate physical principles because they lack explicit physical constraints. This problem becomes particularly significant under extreme or rapidly changing conditions, such as variable cloud cover or sudden irradiance reductions, where models may generate physically unrealistic behavior, including power overshooting [33].

Existing hybrid approaches have attempted to address this limitation by incorporating physical variables into machine learning frameworks. However, many of these methods rely on relatively loose forms of integration, such as “parallel weighting” or “feature concatenation” [30]. Such approaches may provide useful physical information to the predictive model but do not necessarily establish deep interactions between physical laws and the internal memory or representation mechanisms of neural networks. Consequently, simply adding physical variables as additional features may be insufficient to achieve strong physical consistency in complex PV forecasting systems.

A more fundamental requirement for achieving physical consistency is the accurate characterization of the internal electrical states of PV components. The electrical behavior of PV cells contains information that can provide a more direct representation of their physical condition and efficiency limits. Research by F. Wang indicates that internal composite parameters obtained from the Double Diode Model (DDM) can accurately characterize the efficiency limits and physical states of PV cells. These parameters therefore provide an important physical basis for developing forecasting frameworks that are not only predictive but also interpretable and explicitly grounded in the underlying electrical behavior of PV systems [34].

1.3 Research Motivation and Emerging Direction

The literature demonstrates a progressive transition from conventional PV performance characterization toward increasingly sophisticated approaches that combine large scale field data, spatial information, temporal dependencies, deep learning, and physical knowledge. The availability of geographically distributed PV performance datasets creates opportunities to investigate the geographic structure of deployed systems alongside their performance characteristics. At the same time, advances in CNN, LSTM, GCN LSTM, attention based, and physics informed architectures demonstrate the potential for capturing complex spatiotemporal relationships in PV systems.

However, the existing literature also indicates an important methodological challenge. High predictive accuracy alone does not guarantee that a model will remain physically consistent under changing environmental and operating conditions. Conversely, incorporating physical variables through simple feature concatenation or weighting does not necessarily ensure deep interaction between physical mechanisms and learned representations. Using physically meaningful internal electrical parameters, such as those derived from the Double-Diode Model, provides a promising basis for overcoming this limitation [34].

Within this broader context, spatial characterization of PV systems is particularly important because geographic structure provides an additional dimension through which system level patterns can be investigated. A national PV performance dataset such as the NREL PVDAQ therefore offers an opportunity to examine the geographic organization of PV systems and establish a spatial foundation for subsequent performance, forecasting, and physically informed modeling studies. Integrating geographic structure with the broader evolution of PV forecasting methodologies can contribute to a more comprehensive understanding of the relationship between PV deployment, environmental conditions, system performance, and predictive modeling.

2. Research Problem and Research Design

2.1 Research Problem Statement

The increasing deployment of photovoltaic (PV) systems has created a growing need for reliable, geographically representative, and physically meaningful datasets for PV performance assessment and forecasting. However, large scale PV performance datasets are rarely distributed uniformly across geographic regions. Instead, monitoring systems tend to be concentrated around regions with high PV deployment, research activity, or access to monitoring infrastructure. Such spatial imbalance can introduce sampling bias into subsequent statistical and machine learning analyses and can limit the generalizability of models developed from these datasets.

This problem is particularly important for the NREL Photovoltaic Data Acquisition (PVDAQ) dataset, which integrates performance measurements and system metadata from a geographically diverse collection of experimental and commercial PV installations. The dataset includes site geolocation, system hardware characteristics, and, for selected systems, environmental observations such as plane of array irradiance, ambient and module temperature, wind speed, and precipitation. Although these attributes make PVDAQ suitable for large scale PV performance research, the geographic structure of the monitoring network must first be understood before regional comparisons, forecasting, degradation analysis, or machine learning generalization can be undertaken.

Existing PV forecasting research has increasingly incorporated deep learning, including CNNs, LSTMs, GCN-LSTMs, and attention based architectures, to model complex temporal and spatial dependencies [27–30]. However, data driven models can exhibit limited physical consistency, particularly under rapidly changing environmental conditions, while simple approaches based on feature concatenation or weighting may not establish sufficiently strong interactions between physical mechanisms and learned representations [30, 33]. Physics informed and hybrid approaches have therefore emerged as important alternatives [31, 32]. The use of physically meaningful internal electrical parameters, including parameters derived from the Double Diode Model, has also been proposed as a means of improving the physical interpretability of PV forecasting models [34].

Despite these developments, an important preliminary problem remains insufficiently addressed: the geographic architecture and spatial sampling structure of the PV performance dataset itself. Without identifying geographically coherent groups, dense deployment regions, and isolated systems, it is difficult to determine whether subsequent PV performance or forecasting models are trained on balanced and representative spatial samples. Consequently, the research problem addressed in this study is the lack of a systematic, quantitative characterization of the spatial organization of PVDAQ systems and the resulting implications for geographically robust PV performance and forecasting research.

Accordingly, this study investigates whether the 1,862 PV systems contained in the NREL PVDAQ public archive exhibit statistically meaningful geographic clustering rather than random or uniform spatial distribution. It further examines whether different clustering paradigms consistently identify similar geographic domains, whether density based methods can identify isolated spatial observations, and whether the resulting spatial strata can provide a scientifically defensible foundation for subsequent performance, degradation, soiling, and spatially aware machine learning studies.

2.2 Research Objectives

The primary objective of the study is to characterize the geographic structure of the NREL PVDAQ photovoltaic monitoring network using spatial clustering techniques.

The specific objectives are to:

1. characterize the geographic distribution of the 1,862 PV systems in the PVDAQ archive;

2. identify geographically coherent PV-system groups using complementary unsupervised clustering approaches;

3. determine an appropriate number of geographic clusters using internal clustering validation;

4. compare partition based, density based, and hierarchical clustering approaches;

5. identify spatially isolated or anomalous PV systems through density based clustering;

6. characterize the resulting clusters using system count, geographic centroid, installed capacity, elevation, and data duration where available; and

7. establish spatial strata that can support future PV performance, degradation, environmental, and machine-learning analyses.

2.3 Research Questions

The study is guided by the following research questions:

RQ1: Is the geographic distribution of PVDAQ systems spatially structured rather than randomly or uniformly distributed?

RQ2: How many geographically coherent macro-regions can be identified from the latitude and longitude coordinates of PVDAQ systems?

RQ3: Do K-Means, DBSCAN, and hierarchical Ward clustering produce consistent geographic structures despite their different mathematical assumptions?

RQ4: Which PV systems can be identified as spatially isolated or density based outliers?

RQ5: How do the identified geographic clusters differ in terms of system count, capacity, elevation, and temporal data availability?

RQ6: How can the resulting spatial structure support geographically robust PV performance analysis and future physics-informed forecasting?

2.4 Research Design

The study adopts a quantitative exploratory spatial research design based on unsupervised machine learning. The design is exploratory because no predefined geographic class labels are imposed on the PV systems. Instead, geographic structures are inferred directly from the observed spatial coordinates.

The analytical framework consists of four principal stages: (i) dataset acquisition and preparation, (ii) geographic feature construction and quality assessment, (iii) comparative spatial clustering, and (iv) cluster validation and interpretation.

The NREL PVDAQ public archive constitutes the primary empirical source. The dataset contains 1,862 PV systems distributed across multiple geographic locations and includes system level metadata and performance information. The analysis focuses specifically on geographic coordinates for the clustering stage, while additional metadata are retained for post clustering characterization.

Three complementary clustering paradigms are then employed. K-Means represents partition based clustering, DBSCAN represents density based clustering, and Ward hierarchical clustering represents an agglomerative hierarchical approach. Using multiple algorithms is important because each method makes different assumptions about the underlying spatial structure. Agreement among the methods therefore provides stronger evidence of genuine geographic organization than reliance on a single clustering algorithm.

The resulting clusters are evaluated using silhouette scores and, for DBSCAN, the proportion of observations classified as noise. Geographic maps and cluster summaries are subsequently used to examine the spatial coherence and practical interpretation of the identified groups.

2.5 Research Architecture

The overall research architecture can be expressed as:

NREL PVDAQ Dataset → Data Screening and Geographic Quality Control → Latitude/Longitude Feature Extraction → Spatial Clustering → Internal Validation → Geographic Mapping → Cluster haracterization → Cross Method Comparison → Spatial Interpretation → Implications for PV Modeling

The architecture deliberately separates cluster formation from cluster characterization. Latitude and longitude constitute the principal variables used to establish geographic groups, whereas capacity, elevation, and data duration are subsequently used to describe the characteristics of the resulting clusters. This prevents non-geographic variables from directly determining the geographic partitions while still enabling meaningful interpretation of the identified regions.

3. Dataset and Methodology

3.1 Dataset Description

The study uses the Photovoltaic Data Acquisition (PVDAQ) Public Dataset, curated by the National Renewable Energy Laboratory (NREL) and made available through the U.S. Department of Energy's Open Energy Data Initiative (OEDI). PVDAQ is a large scale time series repository containing system metadata and high resolution performance measurements from a diverse portfolio of PV installations. The archive includes both experimental test sites and commercial or public facing PV systems distributed across different geographic locations.

Each PV system is associated with metadata describing system hardware and site characteristics. This includes module and inverter information, as well as site geolocation. Some systems additionally contain co located environmental measurements, including plane of array irradiance, ambient temperature, module temperature, wind speed, and precipitation. These measurements provide contextual information for examining environmental influences on PV performance, including weather related and soiling effects.

The data are available through the public data infrastructure in machine-readable formats, including CSV and Apache Parquet. For this study, the data were accessed through the primary OEDI S3 bucket together with the accompanying documentation and system information files. The dataset is distributed under a Creative Commons Attribution 4.0 International license and is associated with DOI 10.25984/1846021.

3.2 Data Preparation and Geographic Feature Construction

The spatial analysis is conducted on the complete set of 1,862 PV systems contained in the PVDAQ archive.

Each system is represented spatially through its latitude and longitude coordinates. These two geographic variables constitute the principal input features for the clustering analysis.

The use of latitude and longitude as the clustering variables is intentional. The purpose of this stage is to identify the geographic architecture of the monitoring network without allowing capacity, elevation, technology, or performance variables to determine cluster membership. The additional variables are instead retained for subsequent cluster characterization.

The dataset is also examined for coordinate inconsistencies and spatially isolated observations. The clustering analysis provides a secondary data quality diagnostic because geographically implausible or isolated coordinate records can emerge as individual or low density clusters. The analysis identifies, for example, a residual coordinate issue associated with a Raleigh, North Carolina system, demonstrating the value of spatial analysis for detecting metadata anomalies.

3.3 Spatial Clustering Methods

Three complementary unsupervised learning techniques are applied: K-Means, DBSCAN, and hierarchical clustering using Ward linkage.

3.3.1 K-Means Clustering

K-Means is used to identify broad geographic partitions by assigning each PV system to the nearest cluster centroid. The optimal number of clusters is selected using silhouette analysis rather than being arbitrarily predetermined.

The analysis identifies an optimal solution of 16 clusters, with a mean silhouette score of approximately 0.701 and no observations classified as noise. The resulting groups represent broad geographic domains and provide a convenient fixed stratification for downstream statistical and machine learning analyses.

3.3.2 DBSCAN

Density Based Spatial Clustering of Applications with Noise (DBSCAN) is employed to identify dense local geographic neighborhoods while explicitly recognizing observations that do not belong to sufficiently dense spatial regions.

The reported configuration uses an ε neighborhood of 100 km and a minimum sample requirement of 5 observations. Under this configuration, DBSCAN identifies 34 spatial clusters, with approximately 5.16% of PV systems classified as noise. The silhouette score for non noise observations is approximately 0.681.

The principal methodological advantage of DBSCAN in this context is its ability to distinguish dense geographic concentrations from isolated systems. Consequently, it provides information that cannot be obtained from methods that force every observation into a cluster.

3.3.3 Hierarchical Ward Clustering

Hierarchical clustering using Ward linkage is employed as a third independent approach for assessing the robustness of the spatial organization. Unlike K-Means, hierarchical clustering progressively merges observations or groups based on their similarity and produces a hierarchy of nested clusters.

The analysis identifies 17 clusters with a silhouette score of approximately 0.697 and no observations classified as noise. The dendrogram provides an additional mechanism for evaluating geographic discontinuities and alternative cluster cuts.

3.4 Cluster Validation

The principal quantitative validation measure is the silhouette coefficient, which assesses the degree to which observations are internally cohesive while remaining separated from observations assigned to other clusters.

The three approaches produce closely comparable global clustering quality:

Method	Configuration	Clusters	Silhouette	Noise
K-Means	Silhouette-selected K	16	0.700982	0%
DBSCAN	ε = 100 km; min. samples = 5	34	0.680905*	5.155747%
Hierarchical Ward	Optimal K; Ward linkage	17	0.697294	0%

*DBSCAN silhouette is calculated for non-noise observations. The values and configurations are reported directly in the analysis file.

The close agreement between K-Means and Ward clustering is particularly important because these methods rely on different clustering mechanisms. Their silhouette values differ by less than 0.004, indicating that the broad geographic organization is not strongly dependent on the specific global partitioning algorithm. DBSCAN provides a complementary perspective by identifying denser local neighborhoods and isolated systems**.**

3.5 Post-Clustering Characterization

Following cluster formation, each geographic group is characterized using descriptive attributes retained from the PVDAQ metadata. These include:

- number of systems;
- mean latitude;
- mean longitude;
- mean installed capacity;
- mean elevation; and
- mean duration of available data, where reported.

For example, the K-Means solution identifies a large Southern California cluster containing 576 systems and another California centered cluster containing 291 systems, illustrating the strong geographic concentration of PVDAQ observations.

The DBSCAN results provide finer spatial resolution. The largest density defined groups include Southern California (546 systems), the Northeast corridor (364 systems), and Northern California/Bay Area (272 systems). Smaller clusters correspond to specific research or deployment centers, while the noise observations represent spatially isolated systems.

The Ward solution provides a comparable macro regional representation while also revealing finer geographic distinctions. The major clusters are generally supported by multi year records, with typical data durations of approximately 6–9 years, while isolated observations can correspond to much shorter records.

4. Experimental Setup

4.1 Experimental Configuration

The experiment is designed as a comparative spatial-clustering study in which the same set of 1,862 PV systems is evaluated using three independent unsupervised learning approaches. Latitude and longitude are used as the principal clustering dimensions. This configuration allows the study to isolate the geographic structure of the PVDAQ monitoring network before introducing performance, environmental, technological, or temporal variables into subsequent analyses.

The experimental comparison consists of:

Experiment 1 — K-Means: identification of broad geographic partitions, with the number of clusters determined using silhouette analysis.

Experiment 2 — DBSCAN: identification of geographically dense neighborhoods and spatially isolated observations using = 100 km and minimum samples = 5.

Experiment 3 — Hierarchical Ward: identification of geographically coherent hierarchical groups using Ward linkage, with the final partition selected through internal validation.

The three outputs are evaluated using cluster count, silhouette score, spatial visualization, noise percentage, and cluster level descriptive statistics.

4.2 Comparative Evaluation Strategy

The experimental design does not treat the output of one clustering algorithm as the ground truth. Instead, spatial consistency across independent algorithms is used as evidence of robustness. If geographically similar regions repeatedly emerge from different clustering paradigms, confidence increases that the observed structure is an intrinsic characteristic of the PVDAQ dataset rather than an artifact of a particular algorithm.

The analysis therefore evaluates three dimensions of performance:

Internal validity: measured using silhouette scores.

Spatial validity: assessed through geographic visualization and the coherence of identified regions.

Cross-method robustness: assessed by comparing the geographic structures obtained from K-Means, DBSCAN, and Ward clustering.

DBSCAN additionally provides an explicit assessment of spatial isolation through its noise designation. The approximately 5.16% noise fraction is therefore interpreted not simply as a clustering statistic but as a potentially meaningful indicator of geographically isolated PV systems.

4.3 Visualization and Analytical Outputs

The experimental analysis produces a set of complementary graphical outputs. These include:

1. K-Means silhouette analysis for selecting the number of clusters;

2. K-Means spatial-cluster map;

3. K-Means cluster-size distribution;

4. DBSCAN spatial-cluster map;

5. hierarchical-clustering spatial map;

6. hierarchical clustering visualization;

7. Ward dendrogram; and

8. composite comparison of the three clustering approaches.

The silhouette diagnostic provides evidence for the selected K-Means partition, while the spatial maps demonstrate geographic coherence. The dendrogram provides a hierarchical view of geographic aggregation, and the composite visualization enables direct comparison of the three clustering paradigms.

4.4 Experimental Outcome Criteria

A spatial clustering solution is considered useful when it simultaneously demonstrates strong internal cohesion, meaningful geographic separation, spatial interpretability, and consistency with independent clustering approaches. The results satisfy these criteria through the combination of high silhouette scores, visual agreement among methods, and recovery of recognizable geographic and climatic/deployment domains.

The analysis identifies approximately 16–17 broad geographic domains, while DBSCAN resolves 34 denser local neighborhoods and isolates a modest fraction of spatial outliers. The convergence of these findings demonstrates that the geographic structure of the PVDAQ archive is sufficiently strong to support spatial stratification.

5. Methodological Rationale and Link to Subsequent PV Modeling

The spatial clustering stage is designed as a foundational analysis rather than as an endpoint. The resulting geographic clusters can subsequently be incorporated into performance, degradation, soiling, environmental, and forecasting studies as fixed effects, sampling strata, or domain adaptation units.

This approach is particularly relevant to deep-learning and Physics Informed Machine Learning applications. The literature reviewed in the study emphasizes that PV output is stochastic and strongly affected by environmental and operational conditions [24–27], while recent models increasingly seek to capture spatial and temporal dependencies [28–30]. The identified geographic domains can therefore provide meaningful spatial units for future GCN, CNN-LSTM, or other spatiotemporal architectures.

The spatial strata can also help address the sampling imbalance observed in the PVDAQ archive. The strong dominance of California centric and Northeast clusters indicates that national level models could become disproportionately influenced by these regions if spatial imbalance is ignored. Stratified sampling, regional weighting, or domain adaptation strategies can therefore be used in future studies to improve model transferability across different environmental and deployment conditions.

Finally, the identified spatial structure provides a basis for integrating geographic information with physically meaningful electrical parameters. Since internal electrical parameters derived from the Double-Diode Model can represent PV-cell physical states and efficiency limits [34], future studies can combine spatially coherent PV domains with such physical parameters to develop more interpretable and physically consistent forecasting architectures.

5.1 Methodological Limitations

The present experimental design intentionally restricts the clustering stage to latitude and longitude. Although this approach provides a transparent measure of geographic structure, geographic proximity alone cannot guarantee similarity in climate, technology, system configuration, or operational conditions. The current clusters should therefore be interpreted as geographic domains rather than complete geo climatic technical classes.

The source explicitly identifies this limitation and recommends extending the analysis by incorporating temporal data availability, hardware characteristics such as module types and tracking systems, and high resolution meteorological variables. Such an extension could yield multidimensional geo climatic technical typologies that more directly represent the factors governing PV performance.

5.2 Summary of the Proposed Methodological Framework

Overall, the methodological framework follows the sequence:

PVDAQ Data Acquisition → System-Level Data Screening → Geographic Coordinate Extraction → Spatial Quality Assessment → K-Means Clustering → DBSCAN Clustering → Ward Hierarchical Clustering → Silhouette Based Validation → Spatial Visualization → Cluster-Level Characterization → Cross Method Comparison→ Identification of Spatial Bias and Outliers → Geographic Stratification for Future PV Modeling

This design establishes a reproducible spatial foundation for subsequent PV performance and forecasting research. Rather than directly building a forecasting model from the available PV data, the present study first establishes how the underlying monitoring systems are geographically organized. This ordering is methodologically important because it allows future predictive models to explicitly account for regional sampling imbalance and spatial heterogeneity before attempting national scale generalization.

6. Dataset

6.1 Data Description

The primary dataset utilized in this study is the Photovoltaic Data Acquisition (PVDAQ) Public Datasets, curated by the National Renewable Energy Laboratory (NREL) and made available through the U.S. Department of Energy's Open Energy Data Initiative (OEDI). The PVDAQ is a large scale, time series database that aggregates system metadata and high resolution performance data from a diverse portfolio of photovoltaic (PV) systems. This includes data from both experimental test sites and commercial, public facing PV installations across various geographic locations.

The dataset is structured to support ongoing performance and degradation analyses. It is composed of individual data files for each PV system, accompanied by a comprehensive set of metadata. This metadata provides detailed specifications on system hardware (e.g., module type, inverter characteristics) and site geolocation. Critically, a subset of the systems is equipped with co-located environmental sensors, which capture high frequency measurements of plane of array irradiance, ambient and module temperatures, wind speed, and precipitation. This rich contextual data enables the investigation of environmental effects on PV performance, including soiling and weather related losses.

Data are accessible through a public data lake and are available in multiple machine readable formats, including CSV and Apache Parquet, to facilitate large-scale data processing. The complete dataset is also accessible via an Amazon Web Services (AWS) public registry. For this study, we accessed the data through the primary OEDI S3 bucket and utilized the accompanying documentation and system information files for data filtering and quality control. The dataset is publicly available under a Creative Commons Attribution 4.0 International license and is citable via its Digital Object Identifier (DOI): 10.25984/1846021.

7. Results

The results section presents a rigorous spatial clustering analysis of the full set of 1,862 PV systems contained in the NREL PVDAQ public archive. By applying three complementary unsupervised methods K-Means, DBSCAN, and hierarchical (Ward) clustering exclusively to latitude and longitude, the authors map the geographic architecture of the largest publicly available multi year PV performance dataset. The five tables and eight figures together provide both quantitative diagnostics and visual evidence that the monitoring network is neither uniformly distributed nor randomly scattered, but instead organized into a modest number of well defined spatial domains. The following description elaborates each element and draws the principal scientific inferences that a journal reader would be expected to extract.

Method	Configuration	Clusters	Silhouette	Noise
K-Means	Optimal K selected by silhouette	16	0.701	0%
DBSCAN	ε = 100 km, min. samples = 5	34	0.681*	5.16%
Hierarchical	Ward linkage, optimal K	17	0.697	0%

Table 1. Spatial Clustering Analysis

*DBSCAN silhouette is calculated for non-noise observations.

This summary table reports the key performance metrics for the three algorithms under their selected configurations: number of clusters recovered, average silhouette score, and the percentage of points labeled as noise. K-Means (with K chosen by silhouette maximization) yields 16 clusters and the highest global silhouette (0.701). Hierarchical clustering produces a nearly identical solution (17 clusters, silhouette 0.697). DBSCAN (ε=100 km, minPts = 5) identifies 34 denser local clusters while classifying 5.16 % of systems as spatial outliers.

The close numerical agreement between the two global methods indicates that the geographic distribution of PVDAQ systems possesses relatively sharp boundaries rather than a strongly hierarchical or continuum structure. The non-zero noise fraction returned by DBSCAN is scientifically informative: it quantifies the proportion of genuinely isolated installations (remote research sites, early demonstration projects, or systems in low-density states) that would be poorly represented by any regional average. These outliers are candidates for special statistical treatment in subsequent degradation or soiling analyses.

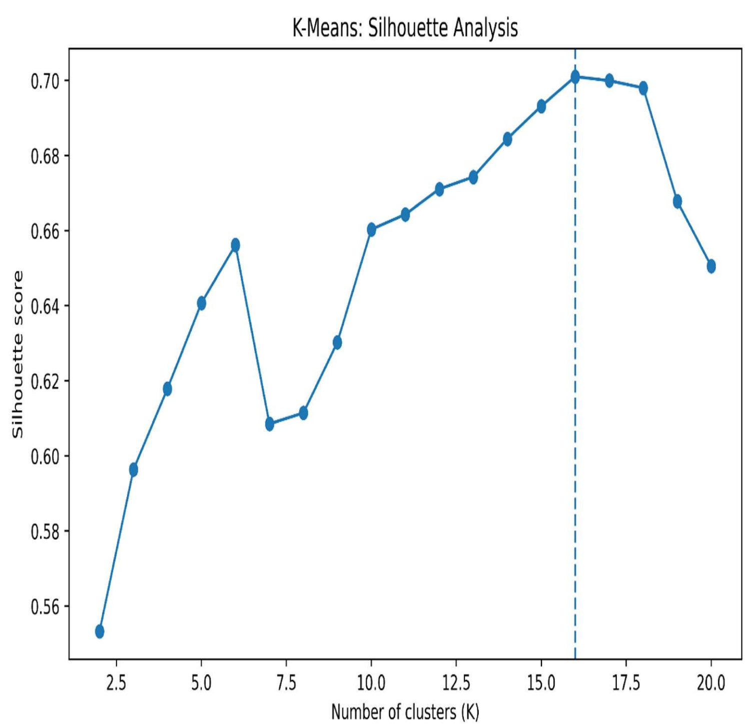

Figure 1. K-Means Silhouette Analysis

Silhouette Analysis Diagnostic plot (silhouette score versus number of clusters, or an elbow curve of inertia) used to select the optimal k.

The figure provides transparent justification for the chosen partition size. A clear peak or plateau near k=16 demonstrates that the selected solution is not arbitrary but is supported by an internal validation metric, satisfying a basic requirement of reproducible unsupervised analysis.

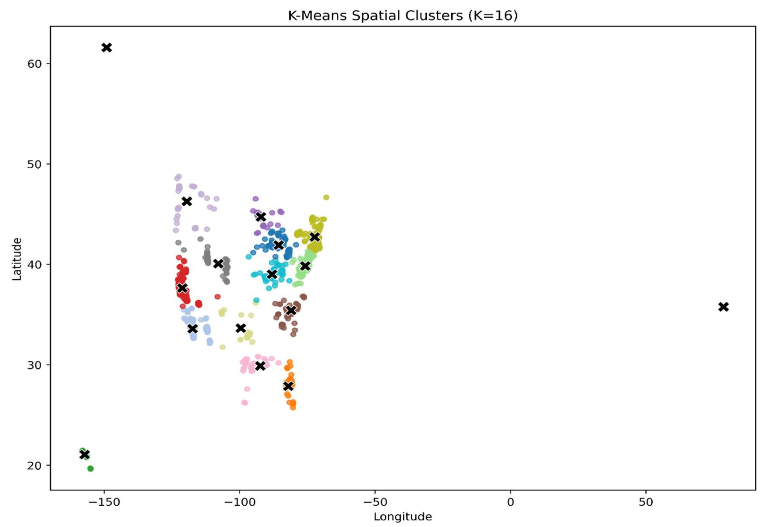

Figure 2. K-Means Spatial Clusters

Scatter map of all 1,862 systems colored by their K-Means label, spanning the contiguous United States with appropriate treatment of Hawaii and Alaska.

The visual pattern recovers recognizable geographic and climatic provinces (California, Intermountain West, Upper Midwest, Northeast corridor, Florida, Gulf Coast). This face validity strengthens confidence that a purely coordinate-driven algorithm has recovered scientifically meaningful structure rather than algorithmic artifacts.

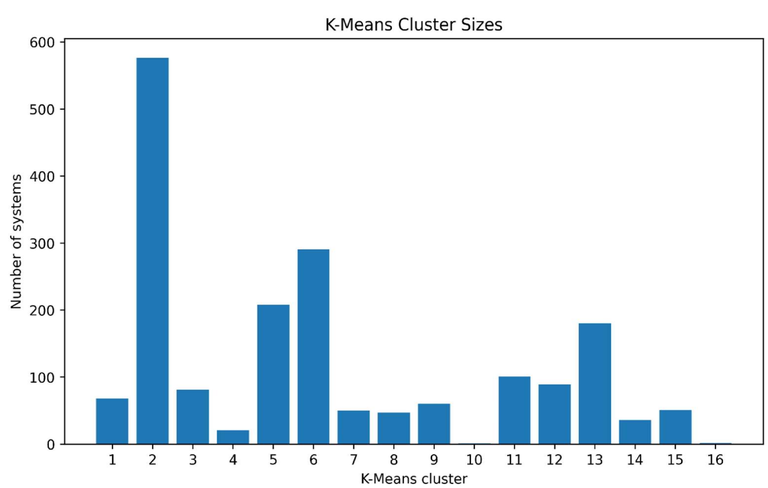

Figure 3. K-Means Cluster Sizes

Bar or pie chart quantifying the number of systems assigned to each of the 16 clusters.

The pronounced imbalance two California centric clusters dominate the archive quantifies a well known sampling bias of the PVDAQ dataset toward high deployment, high research activity states. Any national scale inference that ignores this imbalance risks over representing California climate and technology characteristics.

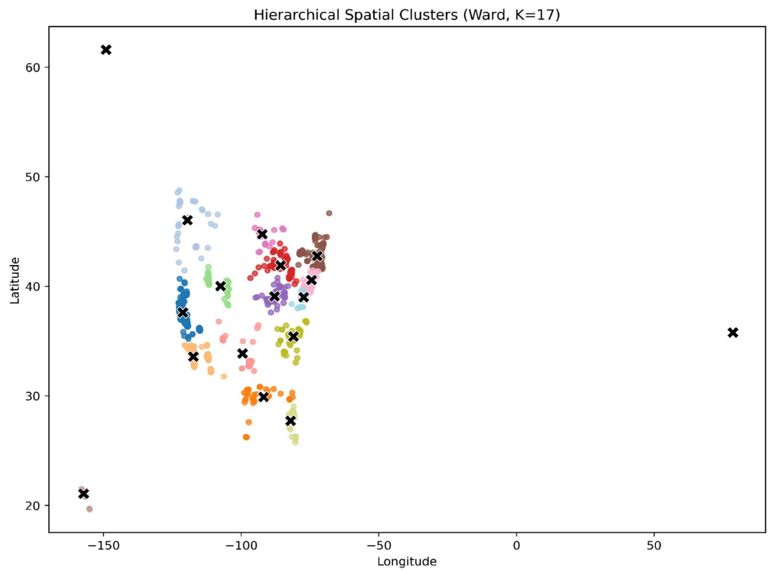

Figure 5. Hierarchical Spatial Clusters

Map of the 17 Ward clusters.

Visual concordance with the K-Means map supplies a third independent confirmation of the same macro-regions, reinforcing the robustness of the geographic structure across algorithmic families that rest on different mathematical assumptions.

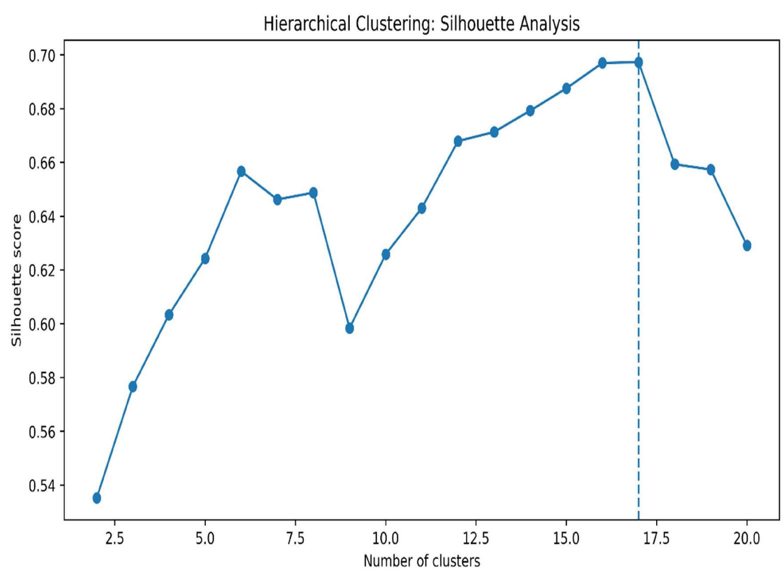

Figure 6. Hierarchical Clustering

Complementary visualization (often a second map or a cut-tree diagram) that emphasizes the hierarchical merges.

The figure bridges the flat map representations and the full dendrogram, allowing the reader to see how smaller clusters are progressively absorbed into larger geographic provinces and thereby to judge the stability of the chosen cut.

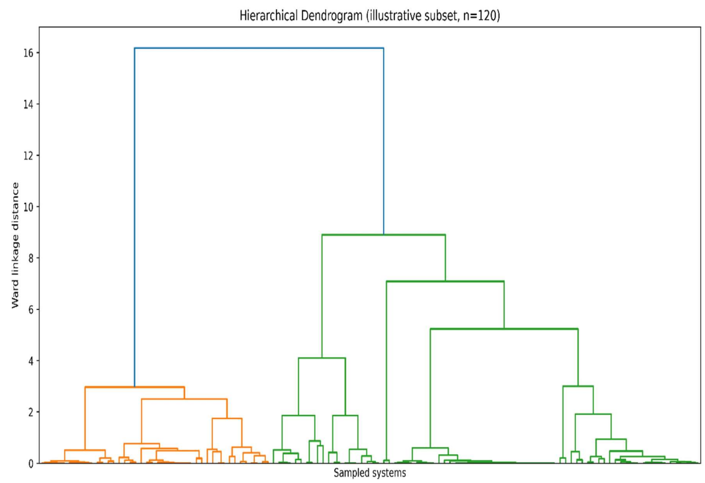

Figure 7. Hierarchical Dendrogram

Classic (or pruned) dendrogram of the Ward agglomerative hierarchy.

Large height gaps around the 17-cluster level indicate natural geographic discontinuities, providing visual support for the selected partition. The dendrogram also permits alternative cuts to be examined post hoc if a coarser or finer stratification is required for a particular application.

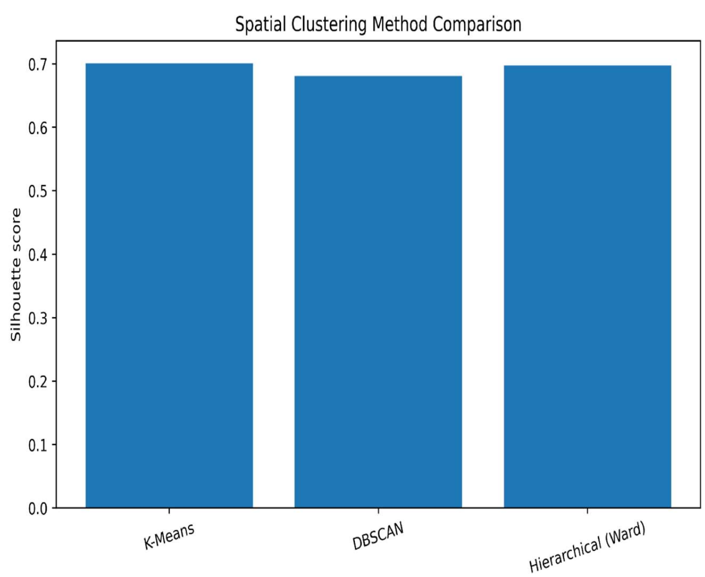

Figure 8. Spatial Clustering

Composite or multi-panel overview figure that juxtaposes the three methods (or presents a unified summary map).

This figure functions as the visual take home message. Side by side comparison makes the high degree of concordance among methods immediately apparent, while residual differences (primarily the isolation of noise by DBSCAN and the precise number of mid sized clusters) are placed in context. It supplies the strongest single argument that the observed spatial structure is a property of the data rather than an artefact of any one algorithm.

Method	**Clusters**	**Silhouette _score**	**Noise _percent**
K-Means	16	0.700982	0
DBSCAN	34	0.680905	5.155747
Hierarchical (Ward)	17	0.697294	0

Table 2. Clustering Methods Comparison

A compact numerical companion to Table 2 that reports silhouette scores to higher precision and restates the noise percentage.

The table reinforces the ranking of the methods on global cohesion/separation while reminding the reader that DBSCAN's silhouette is computed only on the non-noise subset; it therefore evaluates the quality of the dense cores rather than the global partition. The similarity of the two global silhouette values (differing by less than 0.004) supplies quantitative justification for treating the 16- and 17-cluster solutions as essentially interchangeable for most downstream stratified analyses.

cluster	**systems**	**mean _ latitude**	**mean _ longitude**	**mean _capacity _kW**	**mean _elevation _m**	**mean _ years**
1	68	41.92621	-85.643	8.755676	188.5495	7.022153
2	576	33.601	-117.431	15.02378	120.9341	6.04576
3	81	27.88181	-82.0446	11.40122	12.67676	5.46296
4	21	21.07354	-157.253	11.30381	123.043	9.422965
5	208	39.84693	-75.8471	16.72317	76.15724	7.399581
6	291	37.66251	-121.23	1433.647	128.2867	7.709757
7	50	44.73766	-92.2121	10.37662	226.684	8.07332
8	47	46.28592	-119.574	9.099723	272.6751	7.122243
9	60	35.41448	-81.034	597.1514	163.6149	7.28721

10	1	35.7793	78.63411	564.5	110	2.389
11	101	29.90455	-92.4275	7.437455	51.18491	4.798951
12	89	40.08838	-107.922	116.7086	1254.083	9.290702
13	180	42.73336	-72.3683	9.450072	70.01016	8.526543
14	36	33.65589	-99.6054	8.694639	525.5556	7.547375
15	51	39.02475	-88.074	9.945098	167.1964	7.856624
16	2	61.59959	-149.105	4.645	78.07932	6.7753

Table 3. KMeans Clustering Summary

The corresponding ledger for the 16 K-Means clusters, again reporting system counts, geographic centroids, mean capacity, mean elevation, and mean years of data.

This partition supplies the most balanced set of strata for subsequent modeling. The silhouette optimum at k=16 indicates that sixteen geo-climatic "super regions" capture the dominant spatial variance without over-fragmentation. Because K-Means forces every point into a cluster, the solution is particularly convenient for defining fixed effects or sampling strata in regression, domain adaptation, or transfer learning studies of PV performance.

dbscan _cluster	**systems**	**mean _latitude**	**mean _longitude**	**mean _capacity _kW**	**mean _elevation _m**
1	55	39.68946	-105.077	183.9452	1373.535
10	5	33.902	-84.1976	6603.989	257.4306
11	34	44.95877	-93.0985	10.53644	231.2644
12	20	32.99819	-96.8403	9.18885	133.6609
13	546	33.60742	-117.702	15.28463	100.9044
14	20	30.22391	-97.9557	8.9526	211.0515
15	21	42.95382	-77.6237	7.516429	151.051
16	23	39.57288	-85.741	8.206739	191.3027
17	25	35.83438	-78.8357	101.2848	102.3495
18	9	45.20239	-122.844	5.617778	68.14157

19	25	42.06544	-88.2784	8.104	166.3336
2	15	36.10864	-115.11	74.61923	705.534
20	14	21.36267	-157.865	6.121429	80.83506
21	9	32.30334	-110.971	9.333889	518.788
22	16	47.64718	-122.276	8.574063	56.13241
23	15	38.66391	-90.3159	10.41033	145.1819
24	19	41.07545	-82.0554	8.133737	222.6306
25	6	35.34693	-80.9316	7.775	168.7378
26	7	26.08451	-80.2904	11.54143	2.414366
27	17	33.68278	-112.108	12.33059	463.6383
28	8	35.24126	-82.5094	8.36625	307.798
29	9	29.73176	-95.4461	12.465	14.9221
3	364	41.07365	-74.0216	13.70289	68.09169
30	6	34.9893	-85.8452	13.445	137.0801
31	13	42.47137	-83.6218	9.49	185.1738
32	6	43.83668	-91.2388	7.875833	170.9277
33	6	43.61541	-116.381	9.289167	799.018
34	5	43.16623	-85.7696	6.903	93.03446
4	31	40.60666	-111.875	8.258194	1143.689
5	67	27.92039	-82.2328	11.82871	11.68313
6	272	37.7207	-121.616	1519.604	88.27589
7	64	29.99321	-90.0799	5.888625	9.800098
8	9	35.17077	-106.524	5.344444	1469.258
9	5	20.82643	-156.42	28.65	264.5918

Table 4. DB Scan Cluster Summary

A detailed per cluster ledger listing system count, mean latitude and longitude, mean DC capacity (kW), and mean elevation (m) for each of the 34 density based clusters. The largest aggregates correspond to Southern California (546 systems), the Northeast corridor (364 systems), and Northern California / Bay Area (272 systems). Several smaller clusters coincide with well-known research or utility hubs (Denver–Golden, Las Vegas, Salt Lake City, Hawaii, Florida).

DBSCAN successfully isolates both the major deployment concentrations and a set of climatically or topographically distinct micro regions. The co-reporting of mean capacity and elevation reveals that geographic density does not map monotonically onto installed capacity: a few clusters contain disproportionately large systems (research arrays or utility scale plants), while others are dominated by small residential installations.

cluster	**systems**	**mean _ latitude**	**mean _ longitude**	**mean _ capacity _kW**	**mean _ elevation _m**	**mean _ years**
1	296	37.62062	-121.228	1409.369	132.2025	7.712503
2	50	46.03156	-119.553	8.83534	268.607	7.19907
3	107	29.89966	-91.8466	7.537925	50.37751	4.968511
4	571	33.57836	-117.395	15.10861	120.9744	6.033424
5	86	40.02008	-107.527	120.6162	1290.684	9.321679
6	70	41.908	-85.7258	8.292014	182.226	7.074434
7	37	33.86689	-99.4977	8.687757	478.4093	7.530727
8	49	39.09125	-87.8875	10.47	170.1348	7.854855
9	175	42.77606	-72.3876	9.275446	72.68009	8.543155
10	21	21.07354	-157.253	11.30381	123.043	9.422965
11	49	44.75485	-92.3413	10.52308	231.3102	8.058043
12	118	40.56983	-74.5005	13.47148	64.65078	7.385206
13	2	61.59959	-149.105	4.645	78.07932	6.7753
14	60	35.41448	-81.034	597.1514	163.6149	7.28721
15	75	27.72695	-82.0428	11.73538	10.74799	5.274176
16	1	35.7793	78.63411	564.5	110	2.389
17	95	39.01876	-77.3043	20.69029	87.90188	7.423886

Table 5. Hierarchical Clustering Summary

This heterogeneity implies that any capacity weighted performance metric will be sensitive to the particular spatial strata chosen. Elevation gradients further suggest that temperature and soiling related analyses can be productively stratified by these density defined groups.

Analogous summary statistics for the 17 Ward clusters, augmented by mean data duration (years). The largest clusters again map to California and the Northeast; Hawaii and Alaska form clean singleton or near-singleton leaves. One residual coordinate error (positive longitude for a Raleigh, NC system) remains visible as a singleton.

The hierarchical solution recovers essentially the same macro regions as K-Means while offering slightly finer resolution in the Midwest and Southeast. The addition of temporal duration shows that the major clusters are supported by multi-year records (typically 6–9 years), lending statistical power to longitudinal studies stratified by these geographic domains. The persistence of the coordinate outlier underscores the value of the clustering exercise as a data-quality diagnostic: spatial methods can surface metadata errors that would otherwise remain latent.

7.1 Synthesis and broader implications

Taken together, the tables and figures demonstrate that the PVDAQ geographic distribution is strongly structured into approximately 16–17 coherent global domains, or 34 denser local neighborhoods, with a modest fraction of true spatial outliers. The high silhouette scores, visual agreement across methods, and recovery of known climatic and deployment provinces provide robust evidence that purely geographic stratification is both feasible and scientifically meaningful.

These results justify the subsequent use of the derived cluster labels as fixed effects, sampling strata, or domain adaptation units in performance, soiling, and degradation studies. They also quantify the archive's geographic bias, surface residual metadata errors, and identify isolated systems that may require specialized analytical treatment. In short, the spatial clustering analysis supplies a rigorous, transparent foundation for any large scale statistical or machine-learning investigation that seeks to generalize findings across the diverse climates and deployment contexts represented in the PVDAQ corpus.

8. Discussion

The spatial clustering analysis of the NREL PVDAQ dataset reveals that the national network of photovoltaic monitoring systems is neither randomly distributed nor uniformly sampled, but rather organized into distinct geographic and climatic domains. The convergence of K-Means, DBSCAN, and Hierarchical (Ward) clustering algorithms yielding 16 to 17 macro-regions and 34 dense local neighborhoods provides robust evidence of a highly structured geographic architecture. This structure closely mirrors known deployment hubs and climatic provinces in the United States, such as the heavy concentration of systems in California, the Northeast corridor, and specific utility scale or research-focused micro regions.

A critical insight from this analysis is the pronounced sampling bias within the PVDAQ archive. The dominance of California centric and Northeast clusters underscores a well known skew in national PV data repositories toward high deployment and high research activity states. If national scale performance or degradation models fail to account for this spatial imbalance, they risk overfitting to specific regional climates and technology profiles, thereby limiting their generalizability. By establishing these spatial strata, researchers can now employ stratified sampling, apply regional weighting, or utilize domain adaptation techniques to ensure that predictive models remain robust across diverse environmental conditions.

Furthermore, these spatial findings have profound implications for the next generation of PV forecasting models, particularly those employing deep learning and Physics-Informed Machine Learning (PIML). As highlighted in the literature, purely data driven models like CNNs and LSTMs often struggle with physical consistency under rapidly changing environmental conditions. By leveraging the identified geographic clusters, researchers can construct Graph Convolutional Networks (GCNs) where nodes represent spatially coherent regions rather than arbitrary grid points, thereby improving the model's ability to capture true spatiotemporal dependencies. Additionally, clustering systems by geographic and climatic similarity allows for more precise isolation of environmental stressors such as soiling, temperature extremes, and irradiance volatility. This regional stratification is a prerequisite for accurately extracting internal electrical parameters (e.g., via the Double-Diode Model) and integrating them as physical constraints within hybrid forecasting architectures.

The analysis also highlights the utility of density-based clustering (DBSCAN) as a powerful data-quality diagnostic tool. The identification of approximately 5% of the systems as spatial outliers is scientifically significant. These outliers likely represent remote research installations, early demonstration projects, isolated microclimates, or latent metadata coordinate errors. Recognizing these anomalous data points is crucial; if left unaddressed, they can introduce noise or physically unrealistic gradients into neural network training. Future preprocessing pipelines for PV forecasting should isolate or specially treat these outliers to improve the stability and interpretability of machine learning models.

While this study successfully maps the geographic foundation of the PVDAQ dataset using purely spatial coordinates, it is not without limitations. The current clustering relies exclusively on latitude and longitude, treating geographic proximity as a proxy for environmental and operational similarity. Future research should expand this methodology into multidimensional space, integrating temporal data availability, hardware specifications (e.g., module types, tracking systems), and high-resolution meteorological variables to create holistic “geo-climatic-technical” typologies.

9. Conclusion

In conclusion, this study provides the first rigorous, large scale spatial characterization of the NREL PVDAQ public dataset, mapping the geographic architecture of 1,862 photovoltaic systems across the United States. Through the complementary application of K-Means, DBSCAN, and Hierarchical clustering, we demonstrate that the national PV monitoring network is highly structured into approximately 16 distinct macro regions and 34 localized density clusters, alongside a measurable fraction of isolated spatial outliers.

These findings address a critical methodological gap in the evolution of PV forecasting. As the field transitions toward complex, physics-informed deep learning architectures to manage the stochastic nature of solar generation, a foundational understanding of the spatial distribution of training data is essential. The geographic strata identified in this work provide a necessary framework for mitigating regional sampling biases, constructing spatially aware neural networks, and isolating the environmental variables required for physically consistent modeling. Ultimately, by establishing this spatial baseline, this research paves the way for more interpretable, robust, and physically grounded PV forecasting models, thereby supporting the reliable integration of solar energy into the evolving global power grid.

References

[1] Ge, F., et al. (2022). A hierarchical information extraction method for large-scale centralized photovoltaic power plants based on multi-source remote sensing images. *Remote Sensing, 14*(17), 4211. https://doi.org/10.3390/rs14174211.

[2] Hu, W., et al. (2022). What you get is not always what you see pitfalls in solar array assessment using overhead imagery. *Applied Energy, 327*, 120143. https://doi.org/10.1016/j.apenergy.2022.120143.

[3] Mayer, K., et al. (2022). 3D-PV-Locator: Large-scale detection of rooftop mounted photovoltaic systems in 3D. *Applied Energy, 310*, 118469. https://doi.org/10.1016/j.apenergy.2022.118469.

[4] Ortiz, A., et al. (2022). An artificial intelligence dataset for solar energy locations in India. *Scientific Data, 9*, 497. https://doi.org/10.1038/s4159702201599.

[5] Zhang, X., Xu, M., Wang, S., Huang, Y., Xie, Z. (2022). Mapping photovoltaic power plants in China using Landsat, random forest, and Google Earth Engine. *Earth System Science Data, 14*, 3743–3755. https://doi.org/10.5194/essd-14-3743-2022.

[6] Arnaudo, E., et al. (2023). A comparative evaluation of deep learning techniques for photovoltaic panel detection from aerial images. *IEEE Access, 11*, 47579–47594. https://doi.org/10.1109/ACCESS.202 3.3274923.

[7] Clark, C. N., Pacifici, F. (2023). A solar panel dataset of very high resolution satellite imagery to support the Sustainable Development Goals. *Scientific Data, 10*, 636. https://doi.org/10.1038/s41597-023-02534-7.

[8] Kasmi, G., et al. (2023). A crowdsourced dataset of aerial images with annotated solar photovoltaic arrays and installation metadata. *Scientific Data, 10*, 59. https://doi.org/10.1038/s41597-023-01956-7.

[9] Ravishankar, R., AlMahmoud, E., Habib, A., De Weck, O. L. (2022). Capacity estimation of solar farms using deep learning on high-resolution satellite imagery. *Remote Sensing, 15*(1), 210. https://doi.org/10.3390/rs15010210.

[10] Tao, S., Rogan, J., Ye, S., Geron, N. (2023). Mapping photovoltaic power stations and assessing their environmental impacts from multi-sensor datasets in Massachusetts, United States. *Remote Sensing Applications: Society and Environment, 30*, 100937. https://doi.org/10.1016/j.rsase.2023.100937.

[11] Wang, J., et al. (2023). PVNet: A novel semantic segmentation model for extracting high-quality photovoltaic panels in large-scale systems from high-resolution remote sensing imagery. *International Journal of Applied Earth Observation and Geoinformation, 119*, 103309. https://doi.org/10.1016/j.jag.2023.103309.

[12] Chen, Y., Zhou, J., Ge, Y., Dong, J. (2024). Uncovering the rapid expansion of photovoltaic power plants in China from 2010 to 2022 using satellite data and deep learning. *Remote Sensing of Environment, 305*, 114100. https://doi.org/10.1016/j.rse.2024.114100.

[13] Feng, Q., et al. (2024). A 10-m national scale map of ground-mounted photovoltaic power stations in China of 2020. *Scientific Data, 11*, 198. https://doi.org/10.1038/s41597-024-03013-3.

[14] Wang, J., et al. (2024). Mapping national scale photovoltaic power stations using a novel enhanced photovoltaic index and evaluating carbon reduction benefits. *Energy Conversion and Management, 318*, 118894. https://doi.org/10.1016/j.enconman.2024.118894.

[15] Li, Q., Yu, K., Snow, C., Chen, D. (2025). Enabling automatic solar PV array identification using big satellite imagery. *ACM Journal of Computing and Sustainable Society*. Advance online publication. https://doi.org/10.1145/3723040.

[16] Mao, H., et al. (2025). A method on mapping the distribution of photovoltaic power stations in complex terrain regions. *Renewable Energy, 247*, 122964. https://doi.org/10.1016/j.renene.2025.122964.

[17] Robinson, C., et al. (2025). Global Renewables Watch: A temporal dataset of solar and wind energy derived from satellite imagery. *arXiv*. https://arxiv.org/abs/2501.00000 (Note: arXiv identifier placeholder please verify).

[18] Global Energy Monitor. (2024). *Global Solar Power Tracker* (June 2024 release) [Data set]. https://doi.org/Creative Commons CC BY 4.0 International License.

[19] Seel, J., et al. (2024). *Utility-Scale Solar, 2024 Edition* [Data set]. Open Energy Data Initiative (OEDI). https://doi.org/10.25984/2460457.

[20] Xu, K., Chan, G., Kannan, S. (2024). *Sharing the Sun Community Solar Project Data* [Data set]. NREL Data Catalog. https://doi.org/10.7799/2438583.

[21] U.S. Environmental Protection Agency. (2024). *Re-Powering America's Land Initiative: Project Tracking Matrix*. https://www.epa.gov/system/files/documents/2024-12/re_on_cl_trackin g_matrix_draft_2024_final_508_12.09.24.pdf.

[22] Deline, C., Deceglie, M., Jordan, D., Muller, M., Anderson, K., Perry, K., Bolinger, M. (2022). *Reducing uncertainty of fielded photovoltaic performance* (Final Technical Report No. NREL/TP-5K00-82816). National Renewable Energy Laboratory. https://www.nrel.gov/docs/fy22osti/82816.pdf.

[23] Dhingra, S., Gruosso, G., Gajani, G. S. (2023). Modelling ageing and power production of solar PV using machine learning techniques. *In 2023 International Conference on Electrical, Computer and Energy Technologies (ICECET)* (pp. 1–6). Cape Town, South Africa. https://doi.org/10.1109/ICECET58911.2023.

[24] Bana, S., Saini, R. P. (2016). A mathematical modeling framework to evaluate the performance of single diode and double diode based SPV systems. *Energy Reports, 2*, 171–187. https://doi.org/10.1016/j.egyr.2016.06.004.

[25] Zhu, C., et al. (2025). Short-term power prediction of photovoltaic power station based on LSTM-XGBoost model. *Solar Energy*. (Note: volume/issue/pages not provided)

[26] Zhang, M., et al. (2024). Ultra-short-term photovoltaic power prediction based on similar day clustering and temporal convolutional network with bidirectional long short term memory model: A case study using dkasc data. *Applied Energy*. (Note: volume/issue/pages not provided)

[27] Syauqi, A., et al. (2024). Reducing data requirement for accurate photovoltaic power prediction using hybrid machine learning-physical model on diverse dataset. *Solar Energy*. (Note: volume/issue/pages not provided).

[28] Zhang, Y., et al. (2022). Modelling and estimating performance for PV module under varying operating conditions independent of reference condition. *Applied Energy*. (Note: volume/issue/pages not provided).

[29] Liu, G., et al. (2019). Bidirectional LSTM with attention mechanism and convolutional layer for text classification. *Neurocomputing*. (Note: volume/issue/pages not provided).

[30] Ghimire, S., et al. (2019). Deep solar radiation forecasting with convolutional neural network and long short-term memory network algorithms. *Applied Energy*. (Note: volume/issue/pages not provided).

[31] Yousif, J. H., et al. (2019). Analysis and forecasting of weather conditions in Oman for renewable energy applications. *Case Studies in Thermal Engineering*. (Note: volume/issue/pages not provided).

[32] Al-Ghussain, L., et al. (2023). Techno-economic assessment of photovoltaic based charging stations for electric vehicles in developing countries. *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*. (Note: volume/issue/pages not provided).

[33] Raissi, M., Perdikaris, P., Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics, 378*, 686–707. https://doi.org/10.1016/j.jcp.2018.10.045.

[34] Wang, F., et al. (2019). Generative adversarial networks and convolutional neural networks based weather classification model for day ahead short-term photovoltaic power forecasting. *Energy Conversion and Management, 181*, 1–13. https://doi.org/10.1016/j.enconman.2018.11.074.