ISSN: 2414-1895 DOI: 10.6919/ICJE.202603_12(3).0040

Research Review on Photovoltaic Monitoring Data Processing Methods

Qiuping He*

Meteorological Information Center of Qinghai Province, Xining, Qinghai, China

*993032469@qq.com

Abstract

Photovoltaic power stations are subject to variable weather phenomena and equipment operational conditions during operation, resulting in a significant number of outliers and missing values in photovoltaic monitoring data. This poses substantial challenges for photovoltaic data analysis and evaluation, photovoltaic power forecasting, and photovoltaic equipment status analysis. This paper systematically outlines the fundamental principles and typical methods for anomaly detection and missing value imputation in PV monitoring data processing. It summarizes the current technical challenges and limitations in short-term PV monitoring data processing and provides insights into the key areas and research directions for future PV monitoring data processing methods.

Keywords

Photovoltaics; Data Processing; Missing Value Imputation; Feature Analysis.

1. Introduction

The Qinghai-Tibet Plateau covers approximately one-quarter of China's land area, spanning 25 degrees of longitude from east to west. Bordering the subtropics to the south and extending to mid-latitudes in the north, it boasts an average elevation exceeding 4,000 meters. As the world's most topographically complex plateau, it is known as the “Roof of the World” and the “Third Pole.” The plateau's intricate dynamic and thermal effects play a crucial role in shaping atmospheric circulation and weather patterns across China, East Asia, and the globe. Solar energy, valued for its cleanliness, sustainability, safety, and ubiquity, is recognized as one of the most promising future energy sources. Photovoltaic (PV) power generation, as a clean and renewable form of energy, has become an indispensable part of the global energy transition. However, solar power generation exhibits significant variability and randomness due to changing meteorological conditions. Since these conditions are influenced by topography, the resource potential of sites for new energy power plants must be assessed with the highest possible precision. Meteorological services can rapidly assess wind and solar resource conditions in designated areas, providing forecast data such as wind speed, wind direction, irradiance, and cloud cover required for renewable energy power forecasting. The higher the accuracy, precision, and timeliness of meteorological forecasts, the more preparation time is available for energy allocation. Consequently, the greater the value of meteorological information, the more conducive it is to the seamless integration and coordination of renewable energy with other power resources.

The Outline for High-Quality Development of Meteorological Services (2022–2035) explicitly calls for strengthening the rational development and utilization of climate resources. The 2023 Government Work Report emphasizes accelerating the construction of a new energy system. Providing services for renewable energy industries such as wind and solar power represents a key task for meteorological

ISSN: 2414-1895 DOI: 10.6919/ICJE.202603_12(3).0040

departments in supporting China's low-carbon energy transition, advancing national ecological civilization, and achieving carbon peaking and carbon neutrality. In 2022, the China Meteorological Administration issued the Guiding Opinions on Enhancing the Capacity for Climate Resource Protection and Utilization and the Action Plan for Improving Meteorological Service Capabilities for Wind and Solar Energy Resources (2021–2025). These documents set the goal of comprehensively enhancing the precision of monitoring, forecasting, and service delivery for wind and solar energy by 2025.

However, photovoltaic power plants are subject to variable weather conditions and equipment operational states during operation, resulting in a significant number of outliers in photovoltaic monitoring data. These outliers substantially impact subsequent data analysis and evaluation, photovoltaic power forecasting, and photovoltaic equipment condition analysis. Therefore, researching photovoltaic monitoring data processing methods is of paramount importance. PV monitoring data processing typically involves two stages: anomaly detection and missing value imputation. This paper introduces and analyzes the fundamental principles and typical methods for these two stages, while also outlining future research directions for PV monitoring data processing techniques.

2. Anomaly Detection in Photovoltaic Monitoring Data

Research on identifying outliers in photovoltaic monitoring data can be broadly categorized into three approaches: probability-statistical methods based on traditional analysis, machine learning methods based on data analysis, and multi-model fusion methods based on mechanism analysis. Based on data characteristics, PV data anomalies can be categorized into single-point anomalies, correlated anomalies, and clustered anomalies. Single-point anomalies represent outliers or outlying values where a single data point significantly deviates from the normal distribution. Correlated anomalies occur as abnormal data points under specific temporal, spatial, and environmental constraints. Clustered anomalies manifest as anomalous data values exhibiting a concentrated, homogenous distribution pattern influenced by anomalous events.

Traditional probabilistic statistical methods typically rely on principles of probability theory and data analysis. By observing data distribution characteristics and properties, they employ mean estimation and standard deviation calculation to identify outlier features. Reference [1] employs iterative quartile methods to screen outliers in discrete distributions, combining them with quadratic cluster analysis to enhance clustering effectiveness and detect collective outliers. Reference [2] employs the 3σ method for anomaly detection in photovoltaic monitoring data based on the central limit principle. However, its effectiveness is significantly constrained under varying environmental conditions due to factors like geographical settings and weather states. Reference [3] converts photovoltaic monitoring data into binary images through processing and utilizes basic mathematical morphological operations to identify and mark anomalies in the data. Reference [4] proposed a method integrating quartiles and optimal intra-group variance for identifying anomalies in PV monitoring data, tested and predicted using an LSTM (Long Short-Term Memory) network model.

Machine learning methods based on data analysis typically define relationships between data points by analyzing quantitative metrics such as distance, density, and distribution to locate and identify anomalous data. Reference [5] extracts PV anomaly features and classifies raw PV monitoring data using the KNN algorithm, with reclassification based on data environment and identification duration. Reference [6] proposes a pattern-based rapid anomaly detection method for WAMS, predefining standard features for abnormal and normal data via data characteristics, then converting judgment vectors for pattern matching. Reference [7] proposes a PV monitoring data processing method combining mathematical morphology denoising with the bisection interpolation correction technique, effectively enhancing the extraction and correction of anomalous data in raw PV datasets. Reference [8] establishes a tree-structured isolated forest anomaly detection model for efficient discrimination and perception of linearly complex and globally sparse data points.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202603_12(3).0040

Mechanism-based multi-model fusion approaches address complex multivariate relationships by establishing a multi-model fusion framework, enhancing the accuracy and efficiency of anomaly detection in PV monitoring data. Reference [9] analyzes seasonal fluctuation characteristics of array output power, identifies the correlation between irradiance and output power under seasonal variations, and establishes an Extreme Learning Machine (ELM) neural network to predict seasonal array power models. Reference [10] proposes a model for identifying temporal and coupling relationships in PV monitoring data based on a reconstruction-constrained generative adversarial network (GAN), combined with an optimization process using a cross-algorithm-catalyzed particle swarm optimization (PSO). Reference [11] constructs a GAN model to study the correlation, fluctuation patterns, and spatiotemporal characteristics of PV monitoring data, optimizing latent variables under constraints to obtain high-precision reconstructed data. Reference [12] establishes the probabilistic distribution relationship between irradiance and PV power using a Copula joint distribution function based on the dynamic illumination process, enabling dynamic PV data anomaly detection through power confidence intervals.

3. Filling Missing Values in Photovoltaic Monitoring Data

Photovoltaic data missingness can be categorized into three types based on the degree of missingness: completely random missingness, random missingness, and non-random missingness [13]. Completely random missingness refers to missing data that is entirely unrelated to the target variable, typically occurring only under ideal conditions. Random missingness usually indicates a unique dependency between missing values and the target variable, which is relatively common. Non-random missingness typically involves data associated with the target variable itself or other influencing factors, requiring special attention and consideration in research. Common missing value imputation methods include simple computational approaches such as direct deletion, mean imputation, hot-spot imputation, and regression-based imputation. However, these methods are constrained by the extent of data missingness and the magnitude of data variability. Common methods for filling missing values in photovoltaic data can be categorized into: time-series feature-based methods, correlation-based methods, spatial feature-based methods, and multi-dimensional combination-based methods [14].

Time-series feature-based methods leverage temporal dependencies within photovoltaic data to extract features for imputation. Interpolation methods are relatively straightforward, requiring only the original variable data to complete missing value filling, with multiple options available. These typically include mean imputation, linear regression-based filling, and data interpolation. However, this approach is constrained by the duration of missing photovoltaic data and the data collection frequency, resulting in poor filling performance for long-term missing data in time series. Reference [15] proposes a mean-filling interpolation method based on weather phenomenon classification. Reference [16] introduces a high-precision linear interpolation and Stineman interpolation method for minute- and hour-level data, further enhancing interpolation accuracy through structural modeling and smooth Kalman filtering. Reference [16] introduced an ensemble algorithm based on the CrossForest multi-time-series prediction model. It employs a sliding window sampling mechanism to distribute data filling tasks among multiple independent and identical weak learners, integrating their results through cross-validation.

Correlation-based missing value imputation establishes relationship models between target variables and associated variables by analyzing coupling relationships among photovoltaic data, power generation data, system resources, and equipment operation data, enabling imputation based on associated variable data. Reference [17] leverages generative adversarial networks to learn historical data patterns, investigates feature correlations, and optimizes generators using function constraint techniques to enhance missing data imputation accuracy. Reference [10] constructs a Wasserstein divergence GAN learning model based on convolutional neural networks. It extracts temporal and correlation feature information along with distribution patterns from PV data. An identity mapping residual block is introduced to prevent gradient vanishing during training, and the reconstruction loss function is iteratively optimized. This maximizes the approximation of reconstructed sequences to

ISSN: 2414-1895 DOI: 10.6919/ICJE.202603_12(3).0040

actual sequences. Furthermore, it integrates CSPSO to optimize the latent variable input of the generator, thereby enhancing reconstruction accuracy. Reference [18] proposes an improved cloud segmentation model for PV missing data imputation. It comprehensively analyzes solar irradiance's impact on PV power output. Based on the volatility of missing PV power data and the refined cloud model results, it constructs a PV power missing data completion model to fill missing values in PV power data.

The spatial feature-based missing value filling method leverages the spatial correlation among PV data. It identifies spatially correlated sites through correlation analysis, constructs a relationship model, and achieves spatial feature-based missing data filling for PV data. Reference [19] proposes a BPNN-based model for filling missing irradiation and power values in PV power plants. Reference [20] constructs remote correlation relationships among distributed PV power data by studying spatio-temporal correlations between PV stations, then employs the XGBoost method to optimize model latency effects and enhance data filling accuracy. Reference [21] establishes a spatiotemporal correlation and dynamic graph attention network fusion model for PV power data imputation based on dynamic coupling relationships among power plants within a region. It employs a variational modal decomposition model optimized by Bayesian algorithms to decompose raw PV data, thereby improving the accuracy of spatiotemporal correlation mining in PV output.

A multidimensional combination-based missing value filling method reconstructs data by constructing a multidimensional combination model, achieving spatial feature filling for missing photovoltaic data. Reference [22] employs Pearson's product-moment correlation coefficient to measure data similarity, proposing a BPNN photovoltaic monitoring data filling model based on historical similar days and adjacent geospatial locations. Reference [23] enhances the time-series feature extraction capability of PV data using the “Prophet” model and establishes an iterative update random forest model to improve the capture of spatial residual information in PV data.

4. Challenges and Recommendations in Photovoltaic Monitoring Data Processing

As climate change grows increasingly complex, extreme weather events have become more frequent in recent years, further exacerbating the randomness and volatility of photovoltaic data. This trend has also raised higher demands for photovoltaic power forecasting and fault analysis. Although various photovoltaic data processing methods have been widely applied in practice, numerous technical challenges and limitations persist in photovoltaic monitoring data processing due to the diversity of monitoring equipment and the complexity of meteorological influences. Through a systematic review of outlier detection and missing value imputation methods in PV monitoring data processing, this paper identifies the following technical challenges:

1) Constrained by the diversity of anomalous data distribution characteristics, traditional probabilistic statistical outlier detection methods struggle to efficiently process large-scale PV monitoring data and complex interrelated features. These methods exhibit relatively low accuracy, are computationally intensive, and are heavily influenced by manual expertise and computational logic. Further research is needed to develop multi-dimensional analysis of PV data's multi-parameter correlations, enabling personalized analysis and mining of feature data to construct more stable and accurate outlier detection models for PV monitoring data.

2) Machine learning methods based on data analysis circumvent the complexity of target objects by identifying anomalies through clustering and classification of PV monitoring data. However, such approaches suffer from limited portability, high demands on model data and training costs, and weak capabilities in handling high-dimensional data, extracting key features, and modeling nonlinear relationships. A fully rational, highly generalizable machine learning feature extraction method remains elusive, constituting a key focus for future research.

3) Mechanism-based multi-model fusion anomaly detection methods, influenced by environmental and equipment factors, often rely solely on single models for superficial identification across all

ISSN: 2414-1895 DOI: 10.6919/ICJE.202603_12(3).0040

feature data sequences. They fail to account for characteristic differences between sequences, and highly volatile elements significantly impact PV anomaly detection outcomes. When weighting multiple models, assigning combination weights requires multifaceted consideration.

4) Time-series feature-based missing value imputation methods are constrained by the duration of missing data and data collection frequency in PV data, yielding poor results for filling long-term missing data. Furthermore, existing PV monitoring data contain numerous variables.

5) Correlation-based missing value imputation requires prior correlation analysis to mine relationships among photovoltaic data. The accuracy of the relationship model depends on the results of this correlation analysis. Due to the complexity of interrelated features, traditional neural network relationship models demand substantial amounts of normal data for training, resulting in high training costs. Furthermore, reconstructing data based on correlations necessitates preserving the integrity and accuracy of relevant variables, imposing high demands on the original data.

6) Spatial feature-based missing value imputation requires comprehensive consideration of temporal relationships, spatial relationships, and correlations among raw data. Identifying neighboring objects to establish high-precision neighborhood models is challenging, and the impact of time delays on model accuracy must be accounted for. This approach demands strict integrity and accuracy of neighborhood data.

7) Multi-dimensional combination-based missing value imputation must fully account for the multi-parameter correlations in PV monitoring data. This involves collecting multi-dimensional, multi-temporal, and multi-type data information to extract feature vectors influencing PV power variations. These feature vectors require personalized analysis, mining, and information fusion. Construct a multidimensional weighted model using multi-objective optimization algorithms to calculate optimal weight coefficients. Implement dynamic weighting to optimize model performance, conduct in-depth analysis of the relationships between feature vectors, monitoring elements, and objective functions, thereby enhancing the precision and accuracy of missing value imputation in photovoltaic monitoring data.

5. Conclusion

Photovoltaic power plants are subject to variable weather conditions and equipment operational states during operation, resulting in a significant number of outliers and missing values in monitoring data. This poses substantial challenges for data analysis and evaluation, photovoltaic power forecasting, and equipment condition assessment.

This paper systematically elaborates on the fundamental principles and typical methods for anomaly detection and missing value imputation in PV monitoring data processing. It summarizes the current technical challenges and limitations in PV data processing and outlines future research directions.

References

[1] Zhao Yongning, Ye Lin, Zhu Qianwen. Characteristics and Processing Methods of Anomalous Data Clusters in Wind Farm Curtailment [J]. Automation of Electric Power Systems, 2014, 38 (21): 39-46.

[2] ZHAO Y, LEHMAN B, BALL R, et al. Outlier detection rules for fault detection in solar photovoltaic arrays[C]//2013 Twenty-Eighth Annual IEEE Applied Power Electronics Conference and Exposition (APEC). Long Beach, CA, USA, 2013: 2913-2920.

[3] Hao Ying, Dong Lei, Wang Lijie et al. An Abnormal Data Identification Algorithm for Photovoltaic Power Generation Curtailment Based on Mathematical Morphology Denoising [J]. Transactions of the Chinese Society for Electrical Engineering, 2022, 42 (21): 7843-7855.

[4] Li Huanlong. Short-Term Photovoltaic Power Generation Data Processing and Power Forecasting for Anomalous Data [D]. Qinghai University, 2024. DOI:10.27740/d.cnki.gqhdx.2024.000988..

[5] Lin Lixin, Yu Yanhua, Tu Jianfeng. Anomaly Detection Method for Network Data Flows Based on an Improved KNN Algorithm [J]. Information and Computers(Theoretical Edition), 2023, 35 (08): 108-110.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202603_12(3).0040

[6] Wan Chulin, Chen Haoyong, Guo Manlan. Pattern Recognition-Based Active Power Error Data Processing for WAMS [J]. Power System Technology, 2017, 41(3): 922-930.

[7] Ren Yulu, Cheng Yushu, Wang Shushu, et al. Ultra-Short-Term Prediction Method for Photovoltaic Output Based on MMD-OHCP Data Preprocessing [J]. Power Grids and Clean Energy, 2025, 41(02): 93-99.

[8] LONG H, SANG L W, WU Z J, et al. Image-based abnormal data detection and cleaning algorithm via wind power curve［J］. IEEE transactions on sustainable energy, 2020, 11(2): 938-946

[9] Yutong Han, Ningbo Wang, Ming Ma, et al. A PV power interval forecasting basedon seasonal model and nonparametric estimation algorithm[J]. Solar Energy, 2019,184:515-526.

[10] Yin Hao, Ding Weifeng, Chen Shun, et al. A Method for Reconstructing Missing Photovoltaic Data Based on Generative Adversarial Networks and Crossed Particle Swarm Optimization [J]. Power System Technology, 2022, 46 (04): 1372-1381.

[11] Wang Shouxang, Chen Haiwen, Pan Zhixin, et al. A Method for Reconstructing Missing Measurement Data in Power Systems Using an Improved Generative Adversarial Network [J]. Chinese Journal of Electrical Engineering, 2019, 39(01): 56–64+320. DOI: 10.13334/j.0258-8013.pcsee.181282.

[12] YANG M, HUANG X. Abnormal data identification algorithm for photovoltaic power based on characteristics analysis of illumination process[J]. Automation of electric power systems, 2019, 43(6): 64-69.

[13] GONG Shanghong, PAN Tinglong, WU Dinghui, et al. Research on MCMC-Based Method for Filling Missing Photovoltaic Data in Microgrids [J]. Renewable Energy, 2018, 36(03): 346-350. DOI:10.13941/j.cnki.21-1469/tk.2018.03.005.

[14] Cao Min. Research on Data Cleaning for New Energy Power Generation Based on Artificial Intelligence [D]. Southeast University, 2023. DOI:10.27014/d.cnki.gdnau.2023.005025.

[15] Layanun V, Suksamosorn S, Songsiri J. Missing-data imputation for solarirradiance forecasting in Thailand.2017 56th Annual Conference of theSociety of Instrument and Control Engineers of Japan (SICE): IEEE; 2017. p.1234-9.

[16] Demirhan H, Renwick Z. Missing value imputation for short to mid-termhorizontal solar irradiance data. Applied Energy. 2018;225:998-1012.

[17] He Jianping. Method for Filling Missing Data in Photovoltaic Systems Based on Multi-Time-Series Prediction Models [D]. North China Electric Power University, 2025. DOI:10.27139/d.cnki.ghbdu.2025.000082.

[18] Zhang, H. P., Liu, J. Q., Duan, Z. W., et al. Research on PV Power Missing Data Imputation Based on an Improved Cloud Segmented Model [J]. Renewable Energy, 2020, 38(12): 1590-1596. DOI:10.13941/j.cnki.21-1469/tk.2020.12.005.

[19] Lin S, Li P, Xue W, et al. Recognition and reconstruction of photovoltaic output abnormal databased on geographic correlation[C]/ 2021 3rd Asia Energy and Electrical EngineeringSymposium (AEEES). Chengdu, China: IEEE, 2021: 942-948.

[20] Qiao Ying, Sun Rongfu, Ding Ran, et al. Short-Term Power Forecasting for Distributed PV Plant Clusters Based on Data Augmentation (Part I): Methodological Framework and Data Augmentation [J]. Power System Technology, 2021, 45(5): 1799-1808.

[21] Ren Hui, Yu Guangfa, Qiang Hanyue, et al. Regional PV Output Data Quality Enhancement and Ultra-Short-Term Power Forecasting Based on DGAT-Trans Combined Algorithm [J/OL]. Journal of Solar Energy, 1-11 [2026-03-04]. https://doi.org/10.19912/j.0254-0096.tynxb.2024-2352.

[22] Guo Hui. Repairing Faulty Data in Photovoltaic Power Stations Based on Clustering of Power Stations and Similar Meteorological Days [D]. Xi'an: Xi'an University of Technology, 2019.

[23] Li H, Li M, Lin X, He F, Wang Y. A spatiotemporal approach for traffic dataimputation with complicated missing patterns. Transportation research part C:emerging technologies. 2020;119:102730.