IOP Conference Series: Earth and Environmental Science

PAPER • OPEN ACCESS

Photovoltaic Generation Data Cleaning Method Based on Approximately Periodic Time Series

To cite this article: J Zhang *et al* 2017 *IOP Conf. Ser.: Earth Environ. Sci.* **63** 012008

View the article online for updates and enhancements.

Photovoltaic Generation Data Cleaning Method Based on Approximately Periodic Time Series

J Zhang[1,3], Sh Zhang[1,3], J Liang[1,3], B Tian[1,3], Z Hou[2,3] and B Zh Liu[2,3]

[1] Electric Power Research Institute, State Grid Ningxia Electric Power Company, Yinchuan, China
[2] School of Electrical and Electronic Engineering, North China Electric Power University, Beijing, China

[3] zhangjun@dky.nx.sgcc.com.cn, 0zhangsh0nx@163.com, liangjian@dky.nx.sgcc.com.cn, tianbei@dky.nx.sgcc.com.cn, zan_hou@163.com, bzliu@ncepu.edu.cn

Abstract. Data cleaning of photovoltaic (PV) power generation is an important step during data preprocessing for further utilization, such as PV power generation forecasting. The PV power generation data can be treated as a time series. An improved data cleaning method based on approximately periodic time series is proposed. First, the abnormal data in the PV data time series is classified with three types of the outliers. Then these three types of outliers are quantified based on the physical characters of PV power generation, and the effective corresponding cleaning implementations are described considering the rate capacity of PV station and period of PV data time series. Finally, the data cleaning method is tested on the PV generation data from a certain real power grid. The results show that this data cleaning method can effectively improve the PV data quality, and provide an effective support tool for the further application of PV data.

1. Introduction

Photovoltaic power generation has the advantages of no pollution, zero emission, and few limitation of geographical resources distribution. In recent years, the installed capacity of PV power generation continues to increase. But due to the randomness and volatility of PV power generation, large-scale grid-connected PV stations have a negative impact on power quality, systemic stability and reliability. In many photovoltaic (PV) power plants，problems of communication errors, equipment failures and PV power curtailment result in high proportion of outliers in measured PV power data, which is difficult for performance analysis of PV power plants and application of power data [1]. Studying the characteristics of PV power generation and then generating its simulation sequence is very important for assessing the planning and operation of the grid-connected PV stations.

In the application of PV generation data, a short-term power forecasting model was proposed based on adaptive fuzzy time series method to forecast the power of grid connected PV power generation system [2]. Based on the historical generation data of PV power generation system, the adaptive algorithm is adopted to match the data structure with the forecasting model.

But the PV data directly collected to the gird is often incomplete, noisy and inconsistent, so the original PV data quality often cannot meet the follow-up application requirements, such as PV power generation forecasting, so it is necessary to carry out data cleaning.

In the terms of big data cleaning in power system, paper [3] studied the types of the anomalies and then proposed an iterative data cleaning method based on time sequence analysis. Paper [1] proposed a new identification methodology of the PV data with outliers, which is based on Copula theory to describe the relationship between the measured global radiation and PV power. Machine identification models are proposed to identify three typical types of outliers, which are verified using measured data of artificial data and PV power plants. The effectiveness of applying the outliers identification methods is investigated through a day-ahead PV power forecasting application. Paper [4][5] proposed an effective data cleaning methods in power system.

In this paper, an improved data cleaning method is proposed based on approximately periodic time series. The abnormal data in the PV data time series is classified with three types of the outliers. Then these three types of outliers are quantified based on the physical characters of PV power generation, and the effective cleaning implementations are described considering the rated capacity of PV station and period of PV data time series. Finally, the data cleaning method is tested on the PV generation data from a certain power grid.

2. Principle of cleaning method

2.1. Basic definition

In order to achieve the PV power generation data cleaning effectively, we choose the time series theory as the basis of cleaning analysis.

Time series refer to a series of ordinal data recorded for a time-varying phenomenon according to the order of time intervals. Time series analysis is to explore all the information contained in the data, to study such a group of real data in the long-term fluctuations in the process of statistical regularity. Time series can be divided into standard periodic time series and approximately periodic time series.

Standard periodic time series, that is fixed cycle length, and for different periods of the same position on the value is exactly the same, $X(t)=X(t\text{-T})$， T is the period.

The approximately periodic time series is a time series whose seasonal trend is an approximate periodic function. The concept of the N-order skeleton with approximate periodic function, and then the concept of approximate periodic time series is given in paper [6-7]. These two papers gave the definition as follows:

Let $\{\mathrm{f}\,(t),\, t \geq 0\}$ be an approximately periodic function. Denoting $t_0 \geq 0$ and $t_\mathrm{k}= t_0 +k\Delta t$, where $\Delta t \geq 0$ and $k =1,2,\cdots$, if there exists a natural number N such that $\min\{\, T_k\text{-}T_{k\text{-}1}\}/\Delta t, \geq N$, where $k \geq 1$ and $T_k\text{-}T_{k\text{-}1}$ is the length of k^{th} approximate period, $k =1,2,\cdots$, then $\{\, f(t_\mathrm{k}),\, k \geq 0\,\}$ is called an N-order skeleton of $f(t)$. Generally, N needs to be greater than 5.

The experimental data in this paper are obtained from the PV power data collected at a real PV power plant, the sample interval is 1min, and the measured PV power data curve is shown in Figure 1. It can be seen that the original PV power fluctuates greatly and cannot be directly used for the hourly average PV power prediction modelling. Moreover, there are abnormal data in the PV power generation data, such abnormal data will affect the follow-up application, such as PV power prediction. Data cleaning of these time series is an effective way to solve the problem.

According to the problem mentioned above, from the view of equipment operation, a higher failure rate may occur with the outdoor PV power generation equipment, inverters and a variety of communication equipments which are facing the various surrounding environment, weather and seasonal changes. Especially those equipments with poor installation process and production quality the failure rate will be higher. They cannot guarantee the generation of electricity and other monitoring information in real time and correct transmission.

2.2. Approximately periodic time series

State data of the PV generation under normal operation state generally show the character of periodic time series. Similarity can be observed with the observation points after s time interval. It can be fitted by a time series as ARIMA (p, ds, q).

2.2.1. Compression of PV generation data. A real PV generation curve is shown in Figure 1. This is a typical periodic time series. From the point of view of solving the problem briefly, the period is constant as 24 hours. But the PV equipment only generates active power in good sunny conditions during the day. So the PV equipment doesn't work in almost half the time of the day. In order to save storage space, simplify the workload of data analysis, a large number of zero-value data is not necessary to store and handle. The PV generation curve which the zero value has been removed is shown in Figure 2.

From Figure 1, PV power generation data has a strong randomness and volatility, also the periodicity is very apparent.

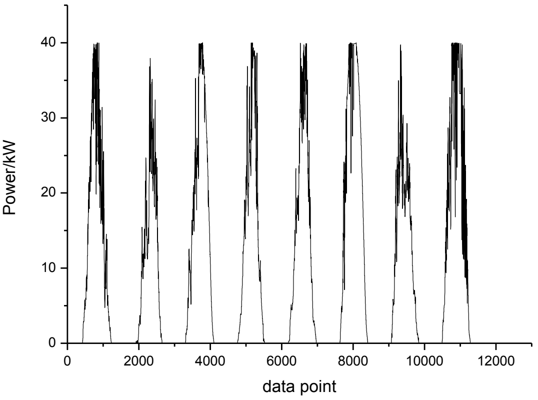

Figure 1. PV generation data under normal operation with zero-value data.

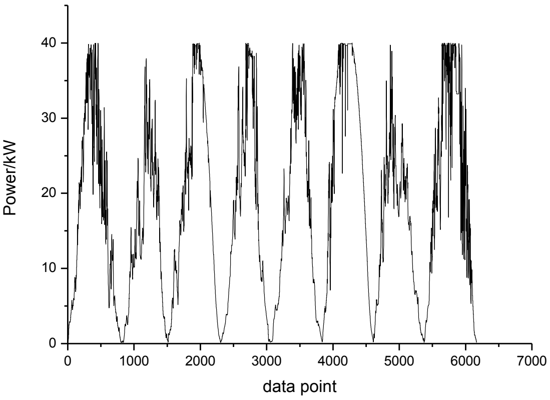

Figure 2. PV generation data under normal operation without zero-value data.

Compare the curve in Figure 1 with that of Figure 2, data storage capacity decreases by about 50%. The useful information in the data is completely retained and not missing. The compressed PV data curve is more convenient to adapt the time series analysis method.

The time series shown in Figure 2 is a typical approximately periodic time series. It exhibits certain periodic characteristics, but its period length is not a fixed value. So it is necessary to extract the approximate period.

2.2.2. Period extraction of approximately periodic time series. For an approximately periodic time series (t, X_t), $X_t=f(t)$, there is a time transformation function $t'=g(t)$, let $(t', X_{t'})$, $X_{t'}=f(g(t))=f(t)$, this time series has the standard periodicity.

According to Figure 2, the time series $\{T_k, k = 0, 1, 2,\cdots\}$ can be calculated and shown in Table 1. The period is approximate constant.

Table 1. Time series T_k and the approximate period.

k	T_k	T_k -T_{k-1}
1	805	805
2	1512	707
3	2305	793
4	3059	754
5	3835	776
6	4610	775
7	5373	763

From Table 1, the approximate periodic character is obvious. If we want to know the exact value at any time in the time interval (T_k, T_{k+1}), the time transform function $g(t)$ is necessary to deduced. The function $g(t)$ can achieve the task of PV data cleaning. The specific formula of $g(t)$ can be a fitting estimation method based on paper [6-7].

3. Process of cleaning method

The existence of abnormal data will make the parameter estimation deviation of PV data time series. Therefore, the time series and the number of unknown outliers are not known. Combined with the physical properties of the PV data, different cleaning strategies are used to clean the observed PV data. The abnormal data in the time series can be divided into innovational outlier, additive outlier and level shift outlier and a combination of three types of outliers.

Suppose X_t is a time series with no outliers which obey the distribution of ARIMA (*p, d, q*) [2][8]. It can be expressed as

$$X_t = \frac{\theta(B)}{\varphi(B)\nabla^d}\alpha_t \tag{1}$$

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q \tag{2}$$

$$\varphi(B) = 1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_q B^q \tag{3}$$

Where B = delay operator; $\theta(B)$ = smoothing operator; $\nabla = 1 - B; \varphi(B)$ = reversible operator; α_t is independent white noise sequence with the same distribution.

When Z_t is used to represent the observed time series, the formula for the noise point at time T (pulse generation time) can be expressed as the following three outlier models [2][8].

3.1. Innovational outlier (IO)

IO affects all PV data after time T, and their effects are related to the model form of Z_t. Through the system described by the dynamic characteristics of the impact of the latter series. The IO model is expressed as

$$Y_t = \frac{\theta(B)}{\varphi(B)\nabla^d}\left(\alpha_t + \omega I_t^{(T)}\right) = Z_t + \omega\left(\frac{\theta(B)}{\varphi(B)\nabla^d}\right) I_t^{(T)} \tag{4}$$

Where $I_t^{(T)} = \begin{cases} 1, & t = T \\ 0, & t \neq T \end{cases}$.

For a certain PV data time series, this type of IO usually appears as a zero value during the daytime or a value greater than the rated power. So according to this physical character, it is easily to complete the data cleaning work.

3.2. Additive outlier (AO)

$$Y_t = Z_t + \omega I_t^{(T)} = \frac{\theta(B)}{\varphi(B)\nabla^d}\alpha_t + \omega I_t^{(T)} \tag{5}$$

AO only affects the PV data at time T where the disturbance occurs, and does not affect the sequence value after that time. The missing value in the time series can be thought of as an AO through an unknown ω.

The same cleaning strategy can be used to AO as well.

3.3. Level shift outlier (LO)

$$Y_t = Z_t + \frac{\omega}{1-B} I_t^{(T)} = \frac{\theta(B)}{\varphi(B)\nabla^d}\alpha_t + \frac{\omega}{1-B} I_t^{(T)} \tag{6}$$

Level shift outliers affect all observations after *T*, and the effect is the same.

A special strategy may be needed to handle this type of outliers. Depending on the length of time of the missing data, a suitable cleaning procedure should be used. An example shown in Section 4 is adapted to illustrate the special cleaning process.

4. Simulation case

In order to verify the practicability and effectiveness of the data cleaning method proposed in this paper, the original PV data at the moment of t =1312min is set 347.69kW as the point of failure, which is shown in Figure 3. While the rated capacity of PV station is 40kW. From t=1860min to t=6096min, the real PV data is replaced by a constant value as 34.14kW. Another type of outlier is set as IO, which begins at the time of 7103min with the value of 10.32kW. The modified curve is shown in Figure 4.

Use the process of cleaning method mentioned in Section 3, a time series can modeled and the outlier type can be recognized. The cleaning result is shown in Table 2.

Table 2. The cleaning result.

Outlier Type	Outliers Interval	
	Begin time/min	End time/min
AO	1312	1312
LO	1860	6096
IO	7103	7241

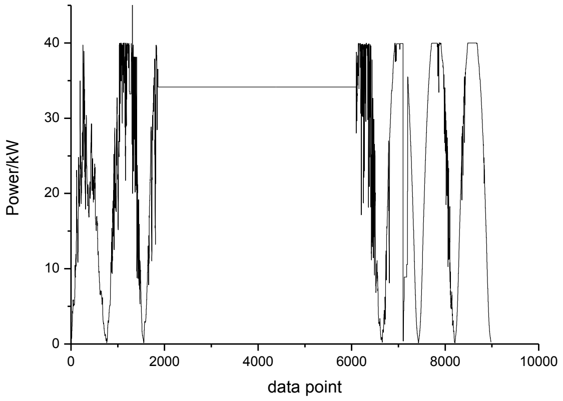

Figure 3. PV generation data with outliers.

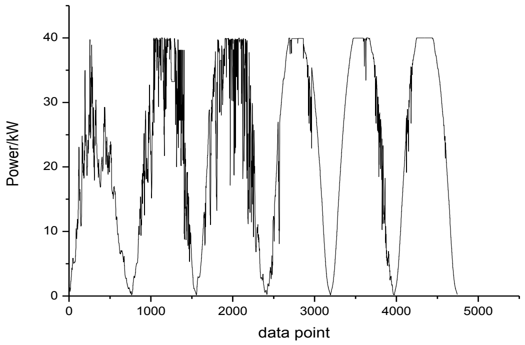

Figure 4. PV generation data without outliers after cleaning.

For the AO, it is greater than the rated capacity of the PV equipment. So the cleaning method is adapted to set this data as 40kW.

For the LO, the time interval continues for a long time, more than a few cycles. So the idea of revision or repair is likely unreasonable. The best way is to remove these large number of wrong values. During the cleaning process, the full cycle must be maintained.

5. Conclusions

Data cleaning of PV power generation is an important step during data preprocessing for further utilization. An improved data cleaning method based on approximately periodic time series is proposed. The main ideas are as follows:

The abnormal data in the PV data time series is classified with three types of the anomalies. Different types of outliers have its own cleaning steps. The approximately periodic time series use the time transform function to achieve the outliers correction considering the rated capacity of PV station. The data cleaning method is tested on the PV generation data from a certain real power grid, which show that this data cleaning method can effectively improve the PV data quality, and provide an effective support tool for the further application of PV data.

6. References

[1] Gong Y F, Lu Z X, Qiao Y, Wang Q and Cao X 2016 Copula Theory Based Machine Identification Algorithm of High Proportion of Outliers in Photovoltaic Power Data. *Automation of Electric Power Systems.* 40(9), 16-23

[2] Yang Z C, Zhu F, Zhang Ch L, Ge L and Yuan X D 2014. Photovoltaic Power Generation Short-term Power Forecasting based on Adaptive Fuzzy Sequence Method. *Journal of Nanjing Institute of Technology*, 12(1), 6-13

[3] Yan Y J, Sheng G H, Chen Y F, Jiang X Ch, Guo Zh H and Qin Sh P 2015 Cleaning Method for Big Data of Power Transmission and Transformation Equipment State Based on Time Sequence Analysis. *Automation of Electric Power Systems.* 39(7), 138-144

[4] Wang Zh L and Hu Y H 2012 *Application of time series analysis*. Beijing:Science Press

[5] Chen J Y, LI W Y and LAU A 2010 Automated Load Curve Data Cleansing in Power Systems. *IEEE Trans on Smart Grid*. 1(2), 213-221

[6] Diao Y L, Sheng W X, Liu K Y, He K Y and Meng X L 2015 Research on Online Cleaning and Repair Methods of Large-Scale Distribution Network Load Data. *Power System Technology.* 39(11), 3134-3140

[7] Wu Sh J, Zhu X Y and Yang X 2015 Analysis of Approximately Periodic Time Series. *Chinese Journal of Applied Probability and Statistics* 31(2), 199-212

[8] Hong Sh H 2015 *Periodicity Recognition and Extraction of approximately periodic time series* Shanghai East China Normal University