Research on distributed photovoltaic power prediction based on centralized photovoltaic output

Xiaoxiao Huang¹, Ruiqing Fan², Xiangji Guo³, Jian Chen³, Lingxu Guo³, Zhongchen Yuan¹, Kunyue Wang⁴ and Jiaan Zhang⁴,⁵

¹ State Grid Tianjin Electric Power Research Institute; ² State Grid Tianjin Wuqing Electric Power Supply Company; ³ State Grid Tianjin Electric Power Company; ⁴ Hebei University of Technology

Email: zhangjiaan@foxmail.com

Abstract. Distributed photovoltaic does not provide output data when it is connected to power grid due to cost constraints, which adversely affects power grid dispatching. Considering the solar irradiance correlation between output of centralized and distributed photovoltaic stations in a region, a method for predicting the output of distributed photovoltaic power generation is proposed. Sinusoid function is used to fit the output of unshielded photovoltaic system by using the least square method, and the unshielded coefficient is proposed to represent the variation of photovoltaic output. The output of distributed photovoltaic (DPV) can be predicted by combining the unshielded coefficient and the reference power curve. The method is verified by a provincial power grid in northern China, the results show that the method has high accuracy, especially for small fluctuations.

1. Introduction

With the approaching of energy crisis, photovoltaic power generation has developed rapidly with its unique advantages. In China, the total installed capacity of photovoltaic power is increasing. In 2017, the new installed capacity of photovoltaic power is as high as 53GW, of which distributed installed capacity accounts for nearly 40%. The influence of photovoltaic power generation, especially distributed photovoltaic power generation, in power grid cannot be ignored.

Although photovoltaic power generation has many advantages, its output fluctuation caused by seasonal, temperature, cloud cover and other random factors makes the large-scale integration of photovoltaic power plants into the grid still threaten the safe operation of the grid [1]. From the grid perspective, the PV output prediction is the basic guarantee of power grid to make reasonable scheduling. Photovoltaic output prediction is mainly divided into two categories, direct prediction and indirect prediction [2].

Direct prediction is to use historical data of photovoltaic output to find similar days at past to predict photovoltaic output, mainly by means of data processing, commonly used are support vector machine, Markov chain and BP neural network. Literature [3] use the second Renyi Entropy Criterion and Principal Component Analysis (PCA) to induce and reduce training data, and then use Least Square Support Vector Machine (LS-SVM) to classify weather to predict photovoltaic output. In literature [4], a multichain Markov Chain Monte Carlo method is proposed. Several Markov chains with complete conditional distribution are established to simulate the atmospheric state. The
photovoltaic output is predicted considering the correlation between multi-power stations. Literature [5] combines artificial neural network with analog set, and obtains prediction results through historical output and meteorological data training. Literature [6] proposes a deep learning model for photovoltaic output prediction, and uses supervised BP neural network as the conventional fitting layer for prediction. At the same time, a two-stage photovoltaic output prediction system with off-line training and online prediction is established. Literature [7] [8] uses principal component analysis to process historical data and other methods to predict the output by combining BP neural network. Literature [7] combines particle swarm optimization (PSO) and literature [8] combines genetic algorithm (GA).

Indirect prediction refers to the prediction of photovoltaic output by predicting the elements directly related to photovoltaic output, which is generally solar irradiance. Literature [9] presents a new model combining recursive neural network and wavelet theory. Diagonal recursive wavelet neural network (DRWNN) can predict the irradiance information to calculate the corresponding photovoltaic output. Literature [10] presents a photovoltaic ultra-short-term power prediction model which combines numerical weather prediction and ground cloud image. It can accurately predict the photovoltaic output in the next four hours. In literature [11], space clustering analysis based on neural network is used to estimate photovoltaic output in a region using satellite or numerical weather forecast data.

However, both direct and indirect forecasting cannot be separated from historical data. For the current distributed photovoltaic system in China, due to cost and other factors, most distributed photovoltaic systems lack corresponding historical data. The main factors affecting photovoltaic output are solar height, irradiance, temperature, humidity and so on. These factors are highly consistent with regional centralized and distributed photovoltaic power plants. Thus, a method of predicting distributed photovoltaic output based on centralized photovoltaic output prediction in the same region is proposed.

2. Theory
In order to know the influencing factors of photovoltaic output, the mathematical model of photovoltaic power generation is given.

\[
I_E = n_p I_I - n_p I_R \left[ e^{\frac{g(U + I_p R_p)}{AKT_n}} - 1 \right]
\]

(1)

In the formula, \( I_E \) is the load current, \( I_I \) is the photogenerated constant current affected by irradiance, \( I_R \) is the reverse saturated current, \( A \) is the ideal factor, \( q \) is the electronic charge, \( K \) is the Poppetmann constant, \( T \) is the temperature of photovoltaic cells, \( R_s \) is the series equivalent resistance, \( n_s \) is the series number, \( n_p \) is the parallel number. It can be seen that irradiance and temperature are the main factors affected by external factors, of which irradiance has a greater impact. Irradiation is also affected by season and weather.

The season, weather, temperature and humidity of centralized and distributed photovoltaic power stations in the same region are similar or even the same. The correlation coefficients are used to quantify and analyze the correlation of centralized and distributed photovoltaic power output in typical weather.

Figure 1 shows the output of centralized and distributed photovoltaic in typical weather. The curves of centralized photovoltaic generation are solid and the distributed are dotted. It can be seen that although there are differences, the general trend is close to the same.

Pearson, Kendall and Spearman are the commonly used correlation analysis methods in statistics. They are used to calculate the output correlation coefficients of centralized and distributed photovoltaic plants. The results are shown in Table 1. It can be seen that the centralized and distributed photovoltaic output in the same region has a very high consistency, and it can be considered that the variation of distributed photovoltaic output is consistent with that of centralized photovoltaic output with the allowable accuracy.
Figure 1. Contrast of photovoltaic output curves in typical weather.

Table 1. Photovoltaic output correlation coefficient.

| Type     | Kendall Pearson | Spearman |
|----------|-----------------|----------|
| Sunny    | 0.9559          | 0.9981   | 0.9965   |
| Cloudy   | 0.8780          | 0.9732   | 0.9725   |
| Rainy    | 0.8386          | 0.9711   | 0.9516   |
| Yearly   | 0.9236          | 0.9810   | 0.9818   |

3. Method

3.1. Calculation of the unshielded coefficient

3.1.1. The output curve of unshielded PV system. Although photovoltaic output has a high fluctuation, its output curve is basically similar to the sinusoidal function under the weather without shade. The unshielded curve is the maximum value of historical output data in a period of time which the solar elevation is similar. For example, there are N time points per day according to the data interval in a total of M days. It can be represented as a matrix of M rows and N columns. \( P_{ij} \) is the historical data of output of photovoltaic power station at day I and point J. \( W \) is the output data of the unshielded curve.

\[
W = \left\{ w_1, w_2, \ldots, w_j, \ldots, w_n \right\},\quad \begin{cases} 
  \begin{align*}
  w_1 &= \max \left\{ P_{11}, P_{12}, \ldots, P_{1M} \right\} \\
  w_2 &= \max \left\{ P_{21}, P_{22}, \ldots, P_{2M} \right\} \\
  &\vdots \\
  w_n &= \max \left\{ P_{n1}, P_{n2}, \ldots, P_{nM} \right\}
  \end{align*}
\end{cases} \tag{2}
\]

According to the shape of photovoltaic output curve in sunny days, the output power array of unshielded days is fitted with sinusoidal function to solve the influence of extreme value. The function is as follows.

\[
\sum_{s=1}^{n} a_s \sin(b_s x + c_s), \quad x \in (1, 2, \ldots, n), s \in N^+
\tag{3}
\]

Different S will have different fitting effect. Taking the autumn output data of a photovoltaic power plant as an example, S is taken from 1 to 8, and the differences between the curves obtained by different fitting functions are compared. The results are calculated with SSE, RMSE and R-square, respectively, as shown in Table 2. It can be seen that when S increases gradually from 1 to 4, SSE and RMSE decrease significantly, and R-square approaches to 1, which means that the fitted function is
closer to the actual data. As shown in Figure 2, when $S$ is greater than 4, the curve is wavy, which is not in accordance with the actual situation. So the best value of $S$ is 4.

Table 2. Photovoltaic output correlation coefficient.

| $S$ | SSE      | RMSE  | R-square |
|-----|----------|-------|----------|
| 1   | 2.972e+06 | 143.2 | 0.9923   |
| 2   | 2.277e+06 | 126.6 | 0.9941   |
| 3   | 1.6e+06   | 107.3 | 0.9958   |
| 4   | 1.362e+06 | 100.1 | 0.9965   |
| 5   | 1.318e+06 | 99.54 | 0.9966   |
| 6   | 1.248e+06 | 97.98 | 0.9967   |
| 7   | 1.25e+06  | 99.23 | 0.9967   |
| 8   | 1.145e+06 | 96.11 | 0.997    |

Figure 2. Comparison of fitting curves of different fitting equations.

3.1.2. Formula of the unshielded coefficient. Because the capacity of centralized and distributed photovoltaic is different, the first step is normalization.

The traditional normalization method mostly divides the output of photovoltaic power directly by the installed capacity of photovoltaic power. This normalization method includes the natural change rule of photovoltaic, which makes it impossible to reflect its shading degree. A new method of photovoltaic output normalization is proposed. The photovoltaic output is divided by the unshielded output in this period, and the unshielded coefficient of photovoltaic output is obtained to express the shading degree of photovoltaic power station. As shown in formula (4), $u_{ij}$ is the unshielded coefficient of photovoltaic power station at i-day j-hour, the closer it approaches 1, the less occlusion. $P_{ij}$ is the output of photovoltaic power station at I-day j-hour, and $w_j$ is the output power of the j-hour without occlusion during that period.

$$
\mu_{ij} = \begin{cases} 
\frac{P_{ij}}{w_j} & P_{ij} \leq w_j \\ 1 & P_{ij} > w_j 
\end{cases}
$$  \hspace{1cm} (4)

3.2. Distributed photovoltaic output prediction

Due to cost constraints, most of the distributed photovoltaic power plants connected to the grid currently do not have power prediction and recording devices. For the grid, the historical data of distributed photovoltaic is only the daily cumulative generation capacity.

The cumulative generating capacity is calculated by integrating the output power, which can be shown in Formula (5). $t$ is the time interval of historical data. $E_i$ is the cumulative daily generating capacity of day i.
\[\sum_{j=1}^{n} f_{tj} / 60 = E_t \]  

(5)

Because the centralized and distributed photovoltaic output have very high consistency in the trend of change, it can be considered that the uncovered coefficients of the two are equal when estimating. As shown in Formula (6), where \( K_{Dij} \) is the unshielded coefficient of distributed photovoltaic output and it is equal to the centralized unshielded coefficient \( K_{Cij} \).

\[ K_{Dij} = K_{Cij} = P_{Cij} / W_{ij}, \quad i \in (1, 2, \ldots, m), \quad j \in (1, 2, \ldots, n) \]  

(6)

Combining formula (3) and formula (6), formula (7) can be obtained.

\[ K_{Dij} = K_{Cij} = P_{Cij} / \sum_{i=1}^{n} a_i \sin(b_i j + c_i), \quad i \in (1, 2, \ldots, m), \quad j \in (1, 2, \ldots, n) \]  

(7)

According to formula (5), the multivariate primary equations are obtained as follows.

\[
\begin{align*}
& w_{d1} K_{11} + w_{d2} K_{12} + \cdots + w_{dn} K_{1n} = E_1 / T \\
& w_{d1} K_{21} + w_{d2} K_{22} + \cdots + w_{dn} K_{2n} = E_2 / T \\
& \vdots \\
& w_{d1} K_{m1} + w_{d2} K_{m2} + \cdots + w_{dn} K_{mn} = E_m / T
\end{align*}
\]  

(8)

Array \( w \), the output of distributed photovoltaic in unshielded day, can be fitted by the known unshielded coefficient matrix \( K \) and the daily cumulative generating capacity array \( E \) by the least square method. But this method has too many unknowns, which makes it difficult to calculate. In combination with (7), the following equations is given as (9). The reduction of unknowns greatly reduces the computational difficulty and the demand for data, and also improves the prediction accuracy.

\[
\begin{align*}
& \sum_{i=1}^{n} \left( \sum_{j=1}^{n} a_i \sin(b_i j + c_i) \right) K_{ij} = E_1 \\
& \sum_{i=1}^{n} \left( \sum_{j=1}^{n} a_i \sin(b_i j + c_i) \right) K_{2j} = E_2 \\
& \vdots \\
& \sum_{i=1}^{n} \left( \sum_{j=1}^{n} a_i \sin(b_i j + c_i) \right) K_{nj} = E_m
\end{align*}
\]  

(9)

Then, according to the centralized photovoltaic output prediction results, the distributed photovoltaic output prediction results can be obtained, as shown in (10).

\[ P_{Dji} = W_{Dji} K_{Cij} \]  

(10)

4. Case study

In order to verify the effectiveness of this method, the output data of centralized and distributed photovoltaic in a certain area of Hebei Province in the same year are tested and analyzed. The tool used for curve fitting is MATLAB. The historical data range from September 1, 2015 to August 31, 2016, and the time interval of photovoltaic output is 5 minutes. The centralized historical data of typical weather such as sunny, cloudy, rainy days are used to estimate the distributed output data and the accuracy of prediction is evaluated with MSE and MAPE as comparison with the distributed actual historical output data. The formula is shown in (11) (12), where \( f_t \) is the predicted value, \( p_t \) is the actual value, and \( N \) is the total number of data.

\[ \text{MSE} = \frac{1}{N} \sum_{t=1}^{N} (f_t - p_t)^2 \]  

(11)

\[ \text{MAPE} = \frac{1}{N} \sum_{t=1}^{N} \left| \frac{f_t - p_t}{p_t} \right| \times 100\% \]  

(12)
The results of distributed photovoltaic output prediction and measured data in typical weather are compared as shown in Figure 3, and the results of typical weather prediction accuracy evaluation indicators are shown in Table 3.

![Comparison of Forecast Results in Typical Weather](image)

**Figure 3.** Comparison of Forecast Results in Typical Weather.
Table 3. Accuracy of prediction in typical weather.

| Type   | MSE       | MAPE         |
|--------|-----------|--------------|
| Sunny  | 122.5744911 | 0.063290407 |
| Overcast | 46.94705138  | 0.171896028 |
| Cloudy | 4783.646009 | 0.211157686 |
| Rainy  | 3298.106635 | 0.227774791 |

The forecast results of distributed output in sunny days are basically the same as that of actual output, with an average deviation rate of only 6.3%. This is because the radiance, temperature and humidity of centralized and distributed photovoltaic power stations in the same area are almost the same in sunny days, the change of output of centralized power stations is basically the same as that of distributed power stations.

The output of cloudy day is lower, so the MSE calculated is the lowest. And for the same reason, even if the deviation value is small, the MAPE can reach 17.1%. In fact, from Figure 3 (b), we can see that the accuracy of forecast result in cloudy day is acceptable. This is the same as in sunny days, mainly because there are not too many sudden changes in output, so the prediction results using this method are well behaved.

However, there are still differences between the predicted and measured data on cloudy and rainy days, and the MAPE is more than 20%. The reason is that although centralized and distributed generation have a very high correlation in the trend of output change, there are some differences after all. When the interval between points is small, the effect of volatility is magnified.

In summary, in sunny and cloudy days, which the fluctuation of output is relatively small, the method of predicting distributed photovoltaic output by using centralized photovoltaic output variation is more accurate. While in cloudy and rainy weather with large fluctuation, the prediction accuracy will decrease. However, in the absence of distributed output history data, this method can still use centralized output prediction to predict distributed photovoltaic output cheaply and quickly, and can ensure acceptable accuracy.

5. Conclusions
In the distribution network, the proportion of installed photovoltaic devices is rising, so that the impact of output fluctuation on the grid is increasing. Especially, the lack of data in distributed photovoltaic plants makes it difficult to make rational scheduling planning for the grid. Considering the correlation between distributed photovoltaic and centralized photovoltaic output changes in the same region, a method of predicting distributed photovoltaic output by centralized photovoltaic output prediction is proposed. Based on the historical data of centralized photovoltaic output and the accumulated daily generation capacity of distributed photovoltaic, an improved least squares method is used to fit the output curve of distributed photovoltaic in unshielded day. Distributed photovoltaic output is calculated and predicted based on the unshielded coefficient of centralized photovoltaic output prediction. The validity of this prediction method is proved by the measured data of typical weather, which shows that this method is cheap, fast and has good accuracy. Most important of all is that the method can effectively complete the distributed photovoltaic output prediction with few data.

Acknowledgment
This work was supported by the project “(KJ18-1-05)Study of the influence of the increasing distributed new energy power stations on urban power grid dispatching and operation” of State Grid Tianjin Electric Power Company in China.
References

[1] Xie E Z 2016 A Study on the Model of Forecasting for Photovoltaic Power Generation Considering Meterologic Elements [D] Harbin University of Science and Technology

[2] WANG H K, GE L J, LI H W, et al 2017 A Review on Characteristic Analysis and Prediction Method of Distributed PV [J] Electric Power Construction 38(7) 1-9

[3] Malvoni M, De Giorgi M G, Congedo P M 2016 Photovoltaic forecast based on hybrid PCA–LSSVM using dimensionality reducted data [J] Neurocomputing 211 72–83

[4] FAN A J, GAO S, FANG J C 2018 A PV Power Time Series Generating Method Considering Correlation Characteristics Based on Multi Markov Chain Monte Carlo Method [J] Jiangsu Electrical Engineering 37(6) 55-61

[5] Cervone G, Clemente-Harding L, Alessandrini S, et al 2017 Short-term photovoltaic power forecasting using Artificial Neural Networks and an Analog Ensemble [J] Renewable Energy 108 274-286

[6] SHI J Q, ZHANG J H 2017 Ultra Short-Term Photovoltaic Refined Forecasting Model Based on Deep Learning [J] Electric Power Construction 38(6) 28-35

[7] Hou S B, Wang K H, Shi K B 2018 Short-term output power forecast of Photovoltaic power generation system based on similar day and principal component analysis [J] Renewable Energy Resources 36(1) 15-21

[8] Tongyu X U, Yiming M A, Yingli C, et al 2016 Short term forecasting of photovoltaic output power based on principal component analysis and genetic optimization of BP neural network [J] Power System Protection and Control

[9] Cao J, Lin X 2008 Study of hourly and daily solar irradiation forecast using diagonal recurrent wavelet neural networks [J] Energy Conversion and Management 49(6) 1396-1406

[10] ZHU X, JU R R, CHENG X, et al 2015 A Very Short-term Prediction Model for Photovoltaic Power Based on Numerical Weather Prediction and Ground-based Cloud Images [J] Automation of Electric Power Systems 39(6) 4-10

[11] Pierro M, De Felice M, Maggioni E, et al 2017 Data-driven upscaling methods for regional photovoltaic power estimation and forecast using satellite and numerical weather prediction data [J] Solar Energy 158 1026-1038