Power Reconstructing Method of Distributed Photovoltaic Based on the Temporal and Spatial Correlation

Shangqiang Li[1,*], *Rongfu* Sun[2], *Ying* Qiao[1], and *Zongxiang* Lu[1]

[1]State Key Lab of Control and Simulation of Power Systems and Generation Equipment, Department of Electrical Engineering, Tsinghua University, 100084 Haidian District in Beijing, China
[2]State Grid Jibei Electric Power Company, 100031 Xicheng District in Beijing, China

Abstract. 55GW distributed photovoltaic have been installed in China, but nearly half are connected to the low voltage level of 380V, without real-time power data acquisition. The sequential power data is needed to be reconstructed based on some related monitoring data. Current researches focus on outliers recovery, but not reconstruction from none. This paper explores the temporal and spatial correlation of power between adjacent centralized photovoltaic stations and proposes a large-scale missing power data reconstructing method based on the time-delay power correlation, the spatial geometric characteristics of stations and the thought of ensemble learning. Finally, we verify the effectiveness of the proposed method by simulation based on the real photovoltaic power data. The proposed method can get the better effect of data reconstructing compared with the traditional method, which only use the power curves of the nearest CP station to reconstruct the power curves of the DP station according the capacity conversion.

1 Introduction

Until the end of 2018, 55GW distributed photovoltaic (DP) have been connected with distribution network in China, which challenge the operation of distribution network a lot [1-4]. To make clear characteristics from historical data is the very start of solving problems.

One of the biggest challenges in China is the worse data conditions [5-6]. There are half of DP are connected to the voltage level of 380V, which have no real-time power data acquisition. For example, 30% of DP installation in northern Hebei Province of China can only collect every electricity energy data, but not power data. In addition, the available meteorological data of DP is very few and low precision due to the high economic cost. Therefore, we need to reconstruct the large-scale missing data of DP.

Actually, some studies have been carried out in outliers recovery. Literatures [7-10] reconstructed the missing power data based on principal component analysis and neural network, using the data of CP stations around the target DP station and related weather factors. The current studies on the data reconstructing depend on the meteorological data and more suitable for repairing the small-scale missing data. In addition, some intelligent algorithms like neural network require data labels, which means that they cannot handle data reconstructing from none. However, we can learn from their analysis methods of the temporal and spatial characteristics between different photovoltaic stations.

As for the large-scale missing power data reconstructing, the traditional method in engineering is currently only using the power curves of the nearest CP station to reconstruct the power curves of the DP station according the capacity conversion. It is obvious that this method would bring great errors.

Therefore, this paper, based on the real power data of centralized/distributed photovoltaic power stations, proposes daily electric energy ratio, power cross-correlation coefficient, time-delay power cross-correlation coefficient and other indicators to mine the temporal and spatial characteristics between CP and DP from the perspective of historical power data. This paper also proposes a data reconstructing method based on the time-delay power correlation, the spatial geometric characteristics of stations and the thought of ensemble learning, which can get smaller errors than the traditional data reconstructing method.

2 Characteristic mining and data reconstruction

Without historical power curves in DP, it is impossible to reconstruct the missing data using relatively advanced learning algorithms of supervised learning. For the data reconstructing of DP, we can only use its daily electrical energy and the historical power curves of peripheral CP. Therefore, we adopt the idea of unsupervised learning, based on the time-delay power correlation, the spatial geometric characteristics of DP and CP, to reconstruct the missing data.

2.1 Definition of characteristic indicators

* Corresponding author: 363524255@qq.com

In order to explore the temporal and spatial characteristics of CP and DP historical power curves, we proposed station relative distance, daily electrical energy, ratio of daily electrical energy, power cross-correlation coefficient and time-delay power cross-correlation coefficient in this section.

Station relative distance (km) refers to the relative distance of each photovoltaic power station, which is calculated by latitude and longitude coordinates, reflecting the spatial characteristics of each station, as shown in (1). In the formula, the lg_1 and lg_2 represent longitude of two stations respectively. The la_1 and la_2 represent latitude of two stations respectively.

(1)

Daily electrical energy refers to the sum of the daily power curve of a photovoltaic station, as shown in (2). In the formula, the n represents the name of a photovoltaic station. The d represents one day. The M represents total number of data points in a power curve. The $P_{n,d}(i)$ represents the active power of the station n at time of i on the day d.

$$Q_n(d) = \sum_{i=1}^{M} P_{n,d}(i) \tag{2}$$

Ratio of daily electrical energy refers to the ratio of days in which daily electric energy of CP is larger than that of DP, as shown in (3). In the formula, the D represents the total number of statistical days. The Q_{DP} represents daily electric energy of DP.

$$ratio_{n,DG} = \frac{\sum_{d=1}^{D} f(Q_n(d))}{D}$$
$$f(Q_n(d)) = \begin{cases} 1, Q_n(d) \geq Q_{DP}(d) \\ 0, others \end{cases} \tag{3}$$

Power cross-correlation coefficient refers to the correlation coefficient between a daily power curve of CP and that of DP, as shown in (4). In the formula, $P_n(d)$ and $P_{DG}(d)$ represent the power sequence of CP station n and DP station on the day d respectively. It should be noted that we only use the power sequence in 4 hours around 12 noon to calculate the correlation coefficient, which can reflect the characteristic of a photovoltaic power curve better.

$$\rho_{n,DG}(d) = \frac{\mathrm{cov}(P_n(d), P_{DG}(d))}{\sigma_{P_n(d)} \sigma_{P_{DG}(d)}} \tag{4}$$

Time-delay power cross-correlation coefficient refers to the correlation coefficient between a power curve of CP after a certain distance of time translation and that of DP, as shown in (5). In the formula, the dt represents the time translation distance. The $x_{\max}$ represents the maximum of the time translation distance, which should ensure that the power sequence used for calculation after time translation belongs to the same day.

$$\rho_{\mathrm{delay},n,DG}(d) = \frac{\mathrm{cov}(P_n^{'}(d,\mathrm{dt}), P_{DG}(d))}{\sigma_{P_n^{'}(d,\mathrm{dt})} \sigma_{P_l(d)}}, d = 1,2,...,D$$
$$P_n^{'}(d,\mathrm{dt}) = P_n(d, t+dt),$$
$$dt \in \{x \mid -x_{\max} \leq x \leq x_{\max}, x \in Z, x_{\max} \geq 0\} \tag{5}$$

2.2 Characteristics mining

In order to explore the temporal and spatial characteristics of CP and DP historical power curves, we select a DP station and 8 adjacent CP stations and take their historical power as the object of analysis. Fig. 1 shows their relative position and capacity data.

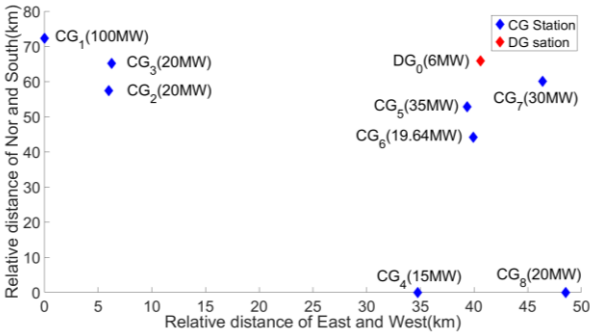


Fig. 1. Spatial Distribution Chart of Photovoltaic Stations.

(1) Characteristics about Electrical Energy

In order to compare the differences about *daily electrical energy* between DP and CP, we firstly convert the daily power curves of CP stations according to the capacity of DG_0 by (6). In the formula, S_n represents the capacity of CP station n.

$$P_{nc}(d) = P_n(d) / S_n \times S_{DG_0} = P_n(d) / S_n \times 6 \tag{6}$$

Furthermore, we calculate the *ratio of daily electrical energy* by (3) to get Fig. 2. We can know that the *ratio of daily electrical energy* of CP stations is close to 1 except for CG_6, which is caused by the better photovoltaic maintenance of CP stations than GP stations. However, the reason that the *ratio of daily electrical energy* of CG_6 is close to 0 may be because some inverters of CG_6 are not in operation.

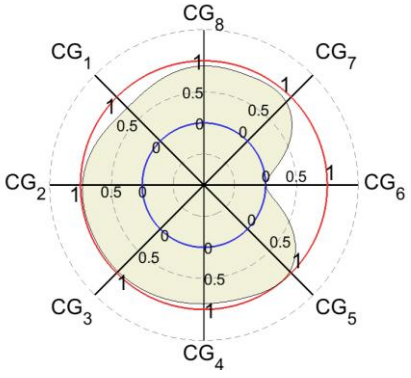


Fig. 2. Analysis Chart of Ratio of Daily Electrical Energy.

In general, under the same installed capacity, there are great differences between the *daily electrical energy* of CP and DP. Thus, when we use the power data of CP to reconstruct the missing power data of DP, we should consider the differences about electrical energy rather than the direct capacity conversion.

(2) Characteristics about Correlation

Time-delay power cross-correlation coefficient reflects temporal characteristics of power between different stations. Fig. 3 shows the *time-delay power cross-correlation coefficients* for four typical days of CG_7 and the corresponding power curves of CG_7 and

DG_0. The power curves of CG_7 are converted according to the capacity of DG_0. We can know that the correlation coefficient increases first and then decreases with the increase of time delay, and reaches its maximum on one point called the best delay time. However, this trend would weaken with the increase of power curve fluctuation. When the power curves is smooth, maybe in a sunny day, the station relative determines the best delay time. When the power curve is smooth, maybe in a sunny day, the best delay time is determined by the station relative. When the power curve fluctuates greatly, maybe in an overcast day, the best delay time may also be influenced by the speed of wind or other weather conditions.

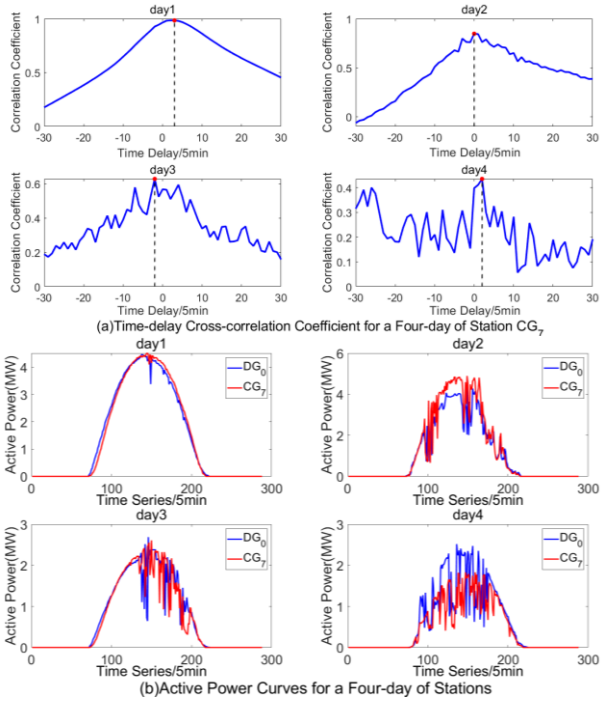


Fig. 3. Analysis Chart of Time-delay Power Cross-correlation Coefficient.

Furthermore, we calculate the average *time-delay power correlation coefficients* of each CP station, as shown in Fig. 4. In general, the time-delay correlation between CP power curves and DP power curves is obvious, which can be used for data reconstructing.

(3)Other Characteristics

When we do data reconstructing, we don't know the power data of DP stations, so that we can't directly use its power curves to calculate the *time-delay power cross-correlation coefficient*. In order to use time-delay characteristic to reconstruct data, we need to find other power curves similar to the missing power curves. Thus, we calculate the correlation coefficients between the daily power curves of other CP stations and those of CG_7. Compared this correlation coefficient and *power cross-correlation coefficient* calculated before, we can know that for the different CP stations, The ratio of days where their difference is less than 0.1 is shown in the Table 1.

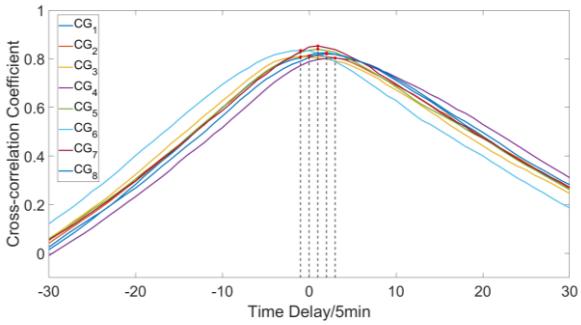


Fig. 4. Analysis Chart of the Average Time-delay Power Cross-correlation Coefficient.

Table 1. The ratio of days where the difference between the two correlation coefficients is less than 0.1.

CG1	CG2	CG3	CG4	CG5	CG6	CG8
0.808	0.838	0.798	0.869	0.788	0.869	0.818

We can know that in the majority of cases, the difference between the two correlation coefficients is less than 0.1. Therefore, we can use the power curves of the CP stations near DG_0 to calculate the *time-delay power cross-correlation coefficient*.

2.3 Data reconstructing method

According to the analysis of characteristics about the power curves of DP and CP, we propose a DP power data reconstructing method based on the electrical energy data of DP and CP, the power curves of CP, the spatial location of the stations and the correlation between the power curves of DP and those of CP.

In addition, traditional DP power data reconstructing method is using the power curves of nearest CP station, which would bring the great accuracy sometimes, such as in the sunny days. However, overall speaking, this model would bring the large variance. Therefore, we lead the idea of bagging in ensemble learning into the data reconstructing method and use several near-centralized photovoltaic power curves to reconstruct the missing distributed photovoltaic power data.

The specific data reconstruct process is as shown in Fig. 5. There are 9 steps in this method, of which the first two steps screen out target DP and CP stations: DG_t, CG_{tt1}, CG_{tt2} and CG_{tt3}. And other steps involve some calculation formulas, which are described in detail below.

Step4, step5 and step6 are in the cycle. We would introduce the situation that i is equal to one, to which other situations are similar.

In step4, we calculate the best delay time, dt_1 and dt_2, at the biggest time-delay power cross-correlation coefficients between the power curves of CG_{tt1} and those of CG_{tt11} and CG_{tt12} by (7). In the formula, CG_{tt11} is CG_{tt2} and CG_{tt12} is CG_{tt3}. $P_{CG_{tt1}}(t)$ and $P_{CG_{tt1j}}(t)$ respectively represent the power curves of the CP station CG_{tt1} and CG_{tt1j} on the day day_t. The explanations of the other variables in (7) are mentioned in (5).

$$\arg\max_{dt_j} \left| corr(P_{CG_{tt1}}(t), P'_{CG_{tt1}j}(t)) \right|$$
$$s.t. \quad P'_{CG_{tt1}j}(t) = P_{CG_{tt1}j}(t + dt_j), j = 1 \ or \ 2 \qquad (7)$$
$$dt_j \in \{x \mid -x_{\max} \le x \le x_{\max}, x \in Z, x_{\max} \ge 0\}$$

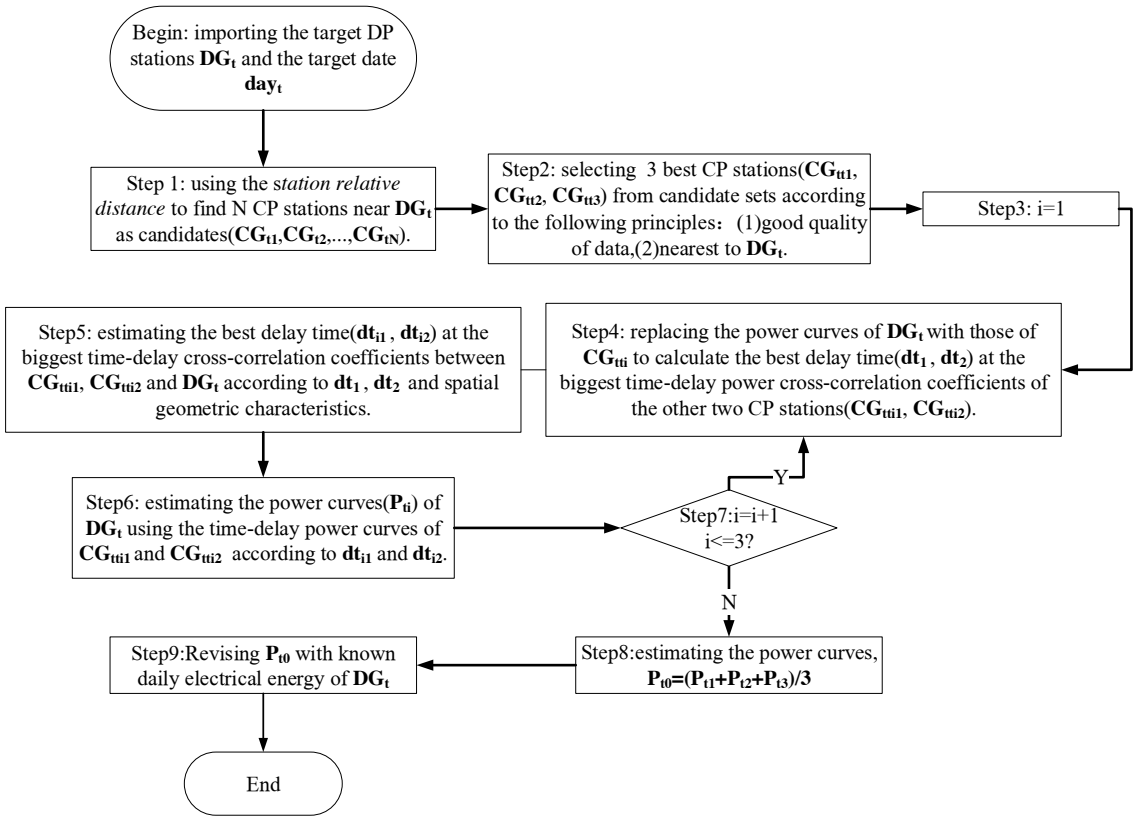


Fig. 5. Flow Chart of Data Reconstructing Method.

In step5, we use dt_1, dt_2 and spatial geometric characteristics to calculate the best delay time, dt_{11} and dt_{12}, at the biggest time-delay cross-correlation coefficients between CG_{tt11}, CG_{tt12} and DG_t. The spatial geometric characteristics of stations CG_{tt1}, CG_{tt11} and CG_{tt12} are shown in Fig. 6. We can calculate dt_{11} and dt_{12} by (8). The meanings of variables in the formula are shown in Fig. 6.

$$dt_{1j} = dt_j \times \cos(\theta_j) \times (1 - \frac{l_0}{l_j \times \cos(\theta_j)}), j = 1 \ or \ 2 \qquad (8)$$

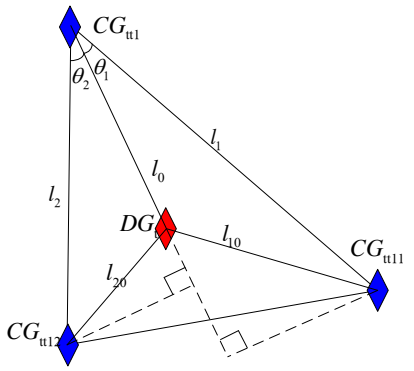


Fig. 6. The Spatial Geometric Characteristics.

In step6, we estimate the power curves P_{t1} of station DG_t by (9). In the formula, $P_{t11}(t+dt_{11})$ and $P_{t12}(t+dt_{12})$ respectively represent the power curves after the time translation of stations CG_{tt11} and CG_{tt12}. Q_t, Q_{t11} and Q_{t12} respectively represent the electrical energy of stations DG_t, CG_{tt11} and CG_{tt12} on the day day_t. The meanings of other variables in the formula are shown in Fig. 6.

$$P_{t1}(t) = \frac{l_{20} \times Q_t}{(l_{10} + l_{20}) \times Q_{t11}} P_{t11}(t + dt_{11}) + \frac{l_{10} \times Q_t}{(l_{10} + l_{20}) \times Q_{t12}} P_{t12}(t + dt_{12}) \qquad (9)$$

We can get P_{t1}, P_{t2} and P_{t3} at the end of the cycle. In step8, we can calculate the power curve of station DG_t by (10).

$$P_{t0} = (P_{t1} + P_{t2} + P_{t3}) / 3 \qquad (10)$$

In step9, we revise P_{t0} with known daily electrical energy of DG_t by (11). In the formula, Q_{t0} represent the daily electrical energy of DG_t calculated by P_{t0}.

$$P_t = P_{t0} - P_{t0} \times (Q_{t0} - Q_t) / Q_{t0} \qquad (11)$$

2.4 Evaluation indicators of data reconstructing effect

We can't get the real power curves of DP stations whose power data is large-scale missing, so we can't evaluate the data reconstructing effect in the actual situations. However, in order to evaluate the data reparing method proposed by us, we assume that we don't know the

power curves of the DP station and use the data reparing method mentioned before to get the power curves. Then we use the actual power curves and the reparing power curves to calculate some indicators to evaluate the data reconstructing method, including *daily mean absolute error*, *daily mean relative error* and *daily mean capacity relative error*.

Daily mean absolute error(DMAE) can be calculated by (12). In the formula, the M represents total number of data points in a power curve. The P_{act} and P_{rep} respectively represent the actual power curve and the reparing power curve.

$$DMAE = \frac{1}{M}\sum_{i=1}^{M}|P_{act}(i) - P_{rep}(i)| \tag{12}$$

Daily mean relative error(DMRE) can be calculated by (13).

$$DMRE = \frac{1}{M}\sum_{i=1}^{M}\frac{|P_{act}(i) - P_{rep}(i)|}{P_{act}(i)} \tag{13}$$

Daily mean capacity relative error(DMCRE) can be calculated by (14). In the formula, S_{DG} represents the capacity of the DP station.

$$DMCRE = \frac{1}{M}\sum_{i=1}^{M}\frac{|P_{act}(i) - P_{rep}(i)|}{S_{DG}} \tag{14}$$

3 Case study

3.1 Simulation data

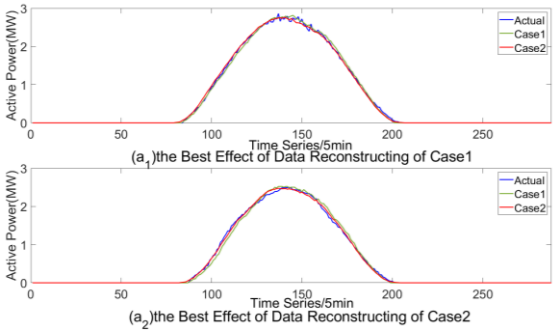
The simulation data originates from 8 CP stations and a DP station in Hebei Province, China, including their capacity, daily power curves, daily electrical energy and geographical position. It should be noted that the power curves of the DP station are only used for the evaluation of the data reconstructing but not for the reconstructing. Their basic information is shown in Fig. 1.

3.2 Simulation analysis

In order to evaluate the effect of data reconstructing, we propose some evaluation indicators and design two cases, whose explanations are following.

Case1: only using the power curves of the nearest CP station to reconstruct the power curves of the DP station according the capacity conversion.

Case2: using the data reconstructing method proposed in this paper based on the temporal and spatial characteristics between the power curves of the CP stations and those of the DP station.



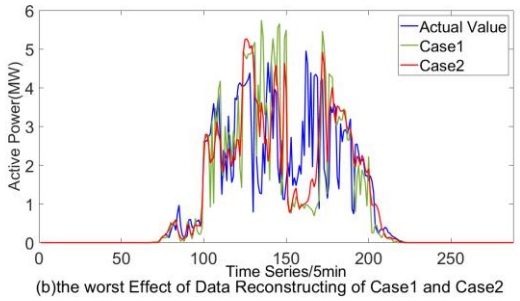


Fig. 7. The Best and Worst Results of Case1 and Case2.

Fig. 7 shows the simulation results of case1 and case2. Fig. 7(a_1) shows the best effect of data reconstructing of case1 and Fig. 7(a_2) shows the best effect of data reconstructing of case2.

We can know that case 1 and case2 can get the great effect of data reconstructing no matter in the best situation of case1 or in the best situation of case2. But actually, case2 can get the better effect of data reconstructing according to the evaluation indicators, which are shown in Table 2 and Table 3.

Table 2. The evaluation indicators of case1 and case2 when case1 gets the best effect of data reconstructing.

Indicators	DMAE	DMRE	DMCRE
Case1	0.0488MW	9.62%	0.81%
Case2	0.0426MW	6.46%	0.71%

Table 3. The evaluation indicators of case1 and case2 when case2 gets the best effect of data reconstructing.

Indicators	DMAE	DMRE	DMCRE
Case1	0.0771MW	11.94%	1.28%
Case2	0.0401MW	6.09%	0.67%

Fig. 7(b) shows the worst effect of data reconstructing of case1 and case2. Case1 and case2 get the worst effect of data reconstructing in the same day. Specifically speaking, case2 can get the better effect of data reconstructing according to the evaluation indicators, which are shown in Table 4.

Table 4. The evaluation indicators of case1 and case2 in the worst simulation results.

Indicators	DMAE	DMRE	DMCRE
Case1	0.9616MW	61.37%	16.03%
Case2	0.8017MW	53.64%	13.36%

We compares the DMCRE of case2 with those of case1 in different days in Fig. 8. We can know that case2 can get smaller DMCRE and show the better effect of data reconstructing in most of days.

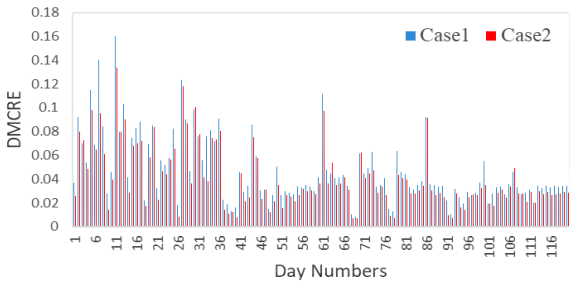


Fig. 8. The DMCRE of Case1 and Case2 in different days.

4 Conclusion

This paper defines some characteristic indicators to describe the temporal and spatial characteristics between the power curves of CP stations and those of DP stations, and proposes DP data reconstructing method based on the temporal and spatial characteristics. Finally, we verify the effectiveness of the proposed method by simulation based on the real photovoltaic power data. We can get the following conclusions:

(1) Converted to the same capacity, the daily electrical energy of CP is slightly larger than that of DP in most of cases.

(2) There is the correlation and the time-delay correlation between the power curves of DP and those of CP, which are related to the station relative distance.

(3) Compared with the traditional method, the proposed method can reduce the errors of data reconstructing effectively.

This work was supported by the project, named Research on Power Prediction Technology of Distributed Generation Cluster in the Power Grid of Jibei (SGJB0000TKJS1900140).

References

1. Y Wang, Q Chen, T Hong, et al. Review of Smart Meter Data Analytics: Applications, Methodologies, and Challenges. IEEE Transactions on Smart Grid(2018).
2. Y Lei, Z Li, Z Lu, etc. Review on the Research of Distributed Generation Technology and Its Impacts on Electric Power System. Southern Power System Technology, **5**, 46. (in Chinese, 2011)
3. J H Angelim ,C M Affonso. Impact of distributed generation technology and location on power system voltage stability. IEEE Latin America Transactions, **14**, 1758.(2016)
4. X Shen, M Cao. Research on the Influence of Distributed Power Grid for Distribution Network. Transactions of China Electrotechnical Society, **30**, 346. (in Chinese, 2015)
5. H Wang, L Ge, H Li, F Chi. A Review on Characteristic Analysis and Prediction Method of Distributed PV. Electric Power Construction, **38**, 1. (in Chinese, 2017)
6. Y Gong, Z Lu, Y Qiao, Q Wang. An Overview of Photovoltaic Energy System Output Forecasting Technology. Automation of Electric Power Systems, **40**, 140. (in Chinese, 2016)
7. S Gong, T Pan, D Wu, Z Ji. Research on missing data imputation of Mcro-Grid PV system based on MCMC. Renewable Energy Resources, **36**, 346. (in Chinese, 2018)
8. Zarzo M, Martí P. Modeling the variability of solar radiation data among weather stations by means of principal components analysis. Applied Energy, **88**, 2775.(2018)
9. R Yu, N Chen, N Miao, D Dang. A Repair Method for PV Power Station Output Data Considering Weather and Spatial Correlations. Power System Technology, **41**, 2229. (in Chinese, 2017)
10. B Lu, Q Fang, G Wang. Methods for identification and restoration of abnormal power data in distributed household rooftop photovoltaic device. Electric Drive Automation, **40,** 1. (in Chinese, 2018)