Correction of Predictive Power for Photovoltaic Plant based on Meteorological and Geographical Correlations

Correction of Predictive Power for Photovoltaic Plant based on Meteorological and Geographical Correlations

GUO Hui[1, a], YANG Guoqing[1], YAO Lixiao[1] and ZHANG Shujie[2]

[1]Institute of Water Resource and Hydro-electric Engineering, Xi'an University of Technology, Xi'an 710048, Shaanxi, China
[2] State Grid Qinghai Electric Power Company Electric Power Research Institute, Xining 810000, Qinghai, China

[a] Email address:593717210@qq.com (GUO Hui)

Abstract. A abnormal data repairing method based on adjacent power plant and integrated similar days of BP neural network is presented. Some factors influencing power generation such as geographical position, temperature and day type are considered. By means of selecting adjacent power plants with high power correlation to plant to be repaired by Pearson product-moment correlation coefficient and the combination of Grey Relational Analysis and curve similarity is used to select similar days. Finding out the integrated similar days' data of the adjacent power plant that is in conformity with the day to be repaired, corresponding BP neural network model is built, then the diverse learning speed algorithm are employed to repair abnormal data. The actual abnormal data in PV prediction power repairing results for Qinghai district show that the proposed method possesses better repairing accuracy.

1. Introduction

With the rapid development of modern electric power technology, many types of large-scale data with its rapid growth, as the information and the value of mining from large data, while the production and life will get more support from the data, data mining technology[1]~[2] has been rapid development in the trend. Photovoltaic power generation data because of its long distance transmission, transmission variables, the system, the environment and other objective causes and human factors will lead to data missing or abnormal fluctuations[3]~[4]. Therefore, in order to ensure the normal operation of the power system, it is necessary to detect, identify and correct the abnormal data, so as to provide the basis for the optimization of power generation and power grid planning.

The correlation method of pv generation power forecast can provide reference for data repair. In the field of power prediction research,[5] divides the historical data into four different weather types. It makes use of each type of historical data and builds 4 prediction sub-models respectively based on SVM. According to the weather forecast information, the corresponding model is selected for forecasting. Due to the distribution characteristics of power system, the traditional data anomaly detection method is divided into the detection, identification, detection and identification before the calculation. The innovation graph approach[6] is also a method before estimation. This method analyzes the spatial correlation by introducing the innovation vector associated with graph theory. Because of the difficulty of telemetry and remote data synchronization method to traditional anomaly recognition, then some scholars have studied some new data recognition method based on Data Mining. In the area of neural networks, [7] is based on the structure of back propagation neural

network, which can eliminate the interference estimate before, so it is easier to measure different patterns of error identification. This method is convenient for real-time monitoring of the network, and then more accurate identification of abnormal data. However, this method is difficult to be widely used in practical applications because of its subjectivity. Reference[8] establishes a neural network large data detection method based on time series analysis and unsupervised learning. This method detects the dynamic data by integrating historical data and current data. However, when the external environment changes greatly, it is possible to generate false positives of abnormal data. In the area of fuzzy theory and cluster analysis, [9] uses the dynamic clustering method in fuzzy mathematics to identify abnormal data. Its advantage is to eliminate the generation of residual pollution. This is because it uses the data difference between measuring time points and standard residuals to cluster. In the field of data correction, Professor Ding Ming of HeFei University of Technology proposed the selection of "similar days" based on BP neural network, and predicted the output power[10] of PV plant from historical data analysis .

2. Basic idea

2.1. The principle of selecting similar days

Similar days are based on days' output data for the PV station, and these days must have similar meteorological conditions. In the end, these similar days were used to correct the output data of the repairing day. According to the similarity degree between different dates' historical weather[11] to determine the 'similar days[12]~[13]'.There is a set $W=\{w_1, w_2, w_3, w_4\}$ to represent Meteorological factors that impact PV output, and w_1 indicates the type of weather, w_2 indicates the highest temperature, w_3 indicates the lowest temperature, w_4 indicates the average temperature.

The grey relation degree method was used to analyze the meteorological correlation degree, $\zeta_i(k)$ represents the correlation coefficient between the date to be repaired and Index k on Day i.

$$\zeta_i(k)=\frac{\min\limits_{i=1}^{n}\min\limits_{k=1}^{m}|w_0(k)-w_i(k)|+\rho\cdot\max\limits_{i=1}^{n}\max\limits_{k=1}^{m}|w_0(k)-w_i(k)|}{|w_0(k)-w_i(k)|+\rho\cdot\max\limits_{i=1}^{n}\max\limits_{k=1}^{m}|w_0(k)-w_i(k)|} \tag{1}$$

In the formula, $w_o(k)$ represents the reference object, $w_i(k)$ represents the comparison object i(n objects in total) .k represents the KTH index (m indicators in total) of the object, and the resolution coefficient ρ is 0.5.

Calculating the correlation coefficient of elements corresponding to each index and reference sequence respectively. Then we calculate their mean, which is correlation degree. In order to facilitate comparison with shape similarity, equation (2) is used to indicate the degree of meteorological correlation. The smaller the value, the higher the degree of correlation.

$$O_i=1-\frac{1}{m}\sum_{k=1}^{m}\zeta_i(k) \tag{2}$$

Because the degree of meteorological correlation only indicates the similarity of weather condition, temperature and other parameters. It does not fully show that the power curve is similar. The power curve can be compared by introducing a judgment function which is curve similarity.

S_{ij} represents the curve similarity of power curves on Day i and Day j. $P_i=\{p_{i1}, p_{i2}, \ldots, p_{in}\}$ represents the power generation of Day i.

$$S_{ij}=\frac{1}{n}\sum_{k=1}^{n}|(p_{ik}-p_{jk})-\frac{1}{n}\sum_{k=1}^{n}(p_{ik}-p_{jk})| \tag{3}$$

In the formula, p_{ik} represents the power data of the sampling point k and has been normalized. The closer S goes to 0, the smaller the difference.

Combined with the meteorological correlation degree and the curve similarity degree, the comprehensive similarity function is defined:

$$T_{ij} = \alpha O_i + \beta S_{ij} \tag{4}$$

When T_{ij} is less than threshold η, it satisfies the condition of comprehensive similarity day. In the repair process, the historical power and meteorological factors have different effects on the correction accuracy. Therefore,α and β should be chosen carefully.

2.2. The principle of selecting similar stations

Several adjacent PV stations in the same geographical area are in extremely similar meteorological conditions, and their weather patterns, temperature and sunshine radiation intensity are extremely similar. Therefore, there is power correlation between different power stations. It is possible to greatly improve the correction accuracy by combining the data of the similar days and the historical output data of similar power stations. It can increase the credibility of data correction.

Suppose the output power data of station A is $P_A = \{p_{A1}, p_{A2}, ..., p_{An}\}$.In the same geographical area, adjacent power stations are B, C and D. The output power respectively is

$$P_B = \{p_{B1}, p_{B2}, ..., p_{Bn}\}, P_C = \{p_{C1}, p_{C2}, ..., p_{Cn}\}, P_D = \{p_{D1}, p_{D2}, ..., p_{Dn}\} \tag{5}$$

In statistics, the Pearson product-moment correlation coefficient is used to measure the correlation (linearly related) between two variables, X and Y, and its value is between -1 and 1.Suppose there are two variables X, Y, after averaging they are $\overline{X}$ 、 $\overline{Y}$ .The Pearson correlation coefficient between two variables can be calculated by the following formula:

$$\rho_{X,Y} = \sum_{i=1}^{n}(x-\bar{x})(y-\bar{y}) \Bigg/ \left(\sum_{i=1}^{n}(x-\bar{x})^2 \sum_{i=1}^{n}(y-\bar{y})^2 \right)^{1/2} \tag{6}$$

The $\rho = 0$ represents X and Y completely irrelevant. The $\rho < 0$ represents X, Y negative correlation. The $\rho > 0$ represents X and Y positive correlation. In this paper, $\rho > 0$ is defined as highly correlated and satisfies the condition of similar days. The threshold σ is selected according to the sample, which will affect the number of samples and ultimately the precision.

Pearson product moment correlation coefficient is used to calculate the correlation coefficient of A and B each day, and then calculate the correlation coefficient of A and C, D station. On the basis of $\rho > \sigma$, the most relevant power station is selected as the similar power station, and its output data is selected as the training sample of neural network.

2.3. The principle of selecting Neural network training samples

The change of the output power for PV station is extremely complex due to the change of the weather. It is a quick and effective method to repair the output power by using the neural network. Neural network input samples can be divided into three types:

(1) $T_{ij} > \eta$, $\rho < \sigma$, In this case, the output power data of the similar days for station are selected as the neural network training samples. And then repair the data of be repairing day.

(2) $T_{ij} < \eta$, $\rho > \sigma$, Selecting the output data of the day to be repaired for similar power stations as the training sample. And then repair the data of be repairing day.

(3) $T_{ij} > \eta$, $\rho > \sigma$, Selecting the output data of the similar days for similar power stations as a training sample. And then repair the data of be repairing day.

In conclusion, the process of training sample selection is shown in Figure 1.

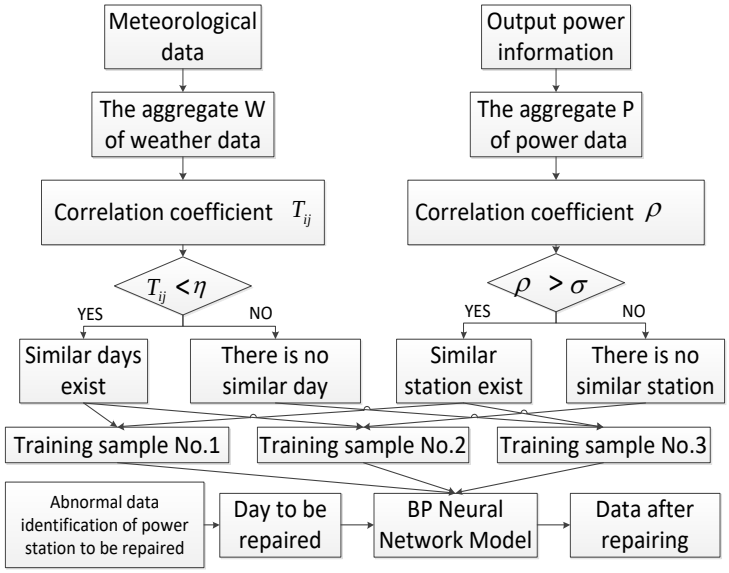

Figure 1. Flow chart of BP neural network training samples selecting

3. Detection of Abnormal Data

3.1. Vacancy data detection

Generally, vacancy data appear in the data transmission link. It may be caused by communication abnormality and information record error.

Suppose the output power of the station is

$$P = \{p_1, p_2, ..., p_n\} \tag{7}$$

If there is $p_i = \varnothing (i = 1,2,..., n)$, it will be judged as vacant data.

3.2. Ultra range data detection

Data transmission errors and storage exceptions are all will cause real data error record. There will be some data beyond the maximum power or less than zero.

Such data can be identified by setting upper and lower bounds.

If there is

$$p_i > p_{\text{Maximum output power}} \parallel p_i < 0 (i = 1,2,..., n) \tag{8}$$

It will be judged as ultra range data.

3.3. Abnormal fluctuation data

Due to similar stations in the same area have a great reference value for abnormal data detection. Using its similarity, the abnormal fluctuation can be detected.

After identifying and removing the missing data and ultra range data, then using the Pearson product distance correlation coefficient distinguish similar stations of the be repairing station. Selecting one of the highest degree of similar stations. Then put two stations' data normalized and subtracted.

Getting the power Different-value every 15 minutes. Calculating the mean and standard deviation of power Different-value. When the difference between the power difference and the average value is greater than 1.5 times the standard deviation, the data is interpreted as abnormal fluctuation data. That is to say, the power station should be modified after normalization $P'_A = \{p'_{A1}, p'_{A2}, ..., p'_{An}\}$.The B power station has the highest similarity degree. After the normalization of the data is $P'_B = \{p'_{B1}, p'_{B2}, ..., p'_{Bn}\}$ .Power difference is $E'_{A-B} = \{e'_{A1-B1}, e'_{A2-B2}, ..., e'_{An-Bn}\}$ .Its average value is $\overline{E'_{A-B}}$ and standard deviation is σ .

If exist

$$e'_{Ai-Bi} - \overline{E'_{A-B}} > 3\sigma \tag{9}$$

It will be identified as abnormal volatility data

4. Example Analysis

The use of sections to divide the text of the paper is optional and left as a decision for the author. Where the author wishes to divide the paper into sections the formatting shown in table 2 should be used.

On the basis of the previous work, the paper has repaired the power data of A power station in Qinghai area. Firstly, selecting the A power station as the repairing power station and screening all of the power stations within the radius of 10 km from the A power station. The power stations with different terrain orientation and large degree of slope fluctuation are filtered out from the A power station. The three stations with the highest correlation coefficient are B, C and D respectively, and their installed capacity is 20MW, 20MW, 10MW and 20MW respectively. Taking A station as the revised power station, and using the power data of a natural month in May 2016 to correct the abnormal data.

4.1. Abnormal data detection

After the detection of abnormal data, it is found that the A power station has data defects on 27th and 22nd.

4.1.1. May 27th:

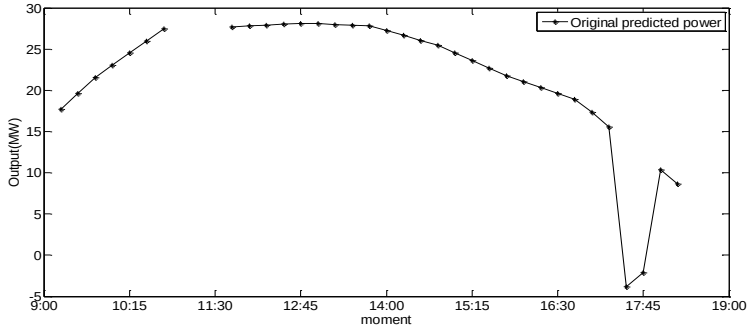

Figure 2. Original predicted power of a station on May 27th

There is vacancy data at 11:00-11:30 on May 27, and power negative value during 17:30-17:45, which is data beyond the normal range (As shown in Figure 2).

4.1.2. May 22th: After calculation, the answer to difference between data of two moments from 16:30 to 16:45 on May 22 and the output power of similar power station minus mean value are respectively

0.4067 and 0.4049.The power difference is 0.0735.They are all 3 times larger than the standard deviation of 0.0959, so classify them as abnormal fluctuation data (As shown in Figure 3).

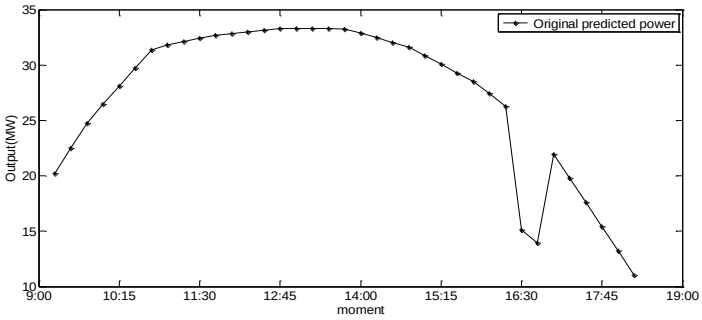

Figure.3.Original predicted power of a station on May 22nd

4.2. Calculation of similarity degree

4.2.1. Calculation of integrated similarity. Through simulation, the integrated similarity days of May 27 are respectively 10, 12 and 28 days. And the integrated similarity values are shown in Table 1.

Table 1. Integrated similarity values on May 27th

Integrated similarity values	10	12	28
27	0.4897	0.4770	0.4227

The integrated similarity values of these three days were all less than 0.5. The minimum on day 28 indicates the highest degree of correlation. And the 27th value is 0, indicating complete correlation.

Similarly, the similar days of May 22 are respectively 21, 24and 25.Their Integrated similarity values are shown in the Table 2.

Table 2. Integrated similarity values on May 22nd

Integrated similarity values	21	24	25
22	0.3948	0.3749	0.3818

The lowest value of the 24th day indicates the highest degree of correlation, and the overall value is less than 0.4. The May 22 composite similarity was higher than the May 27 composite similarity day.

4.2.2. Coefficient calculation of similar power station. The simulation results show that the Correlation coefficient values between a power station and the other three power stations is shown in Table 3.

Table 3. Correlation coefficient of power station

Correlation Coefficient	B	C	D
A	0.8275	0.9455	0.9857

According to the number of samples, the σ was 0.9.After discriminating the B, C and D of the power stations in the same area, the similar power stations are C and D. Selecting the highest degree of power station, that is D power station. The original predictive power of similar power stations in similar days is used as training samples of neural networks.

4.3. Data correction

After the correction of neural network, the repaired prediction curve, original prediction curve and actual output curve of May 27 are shown in Figure 4.

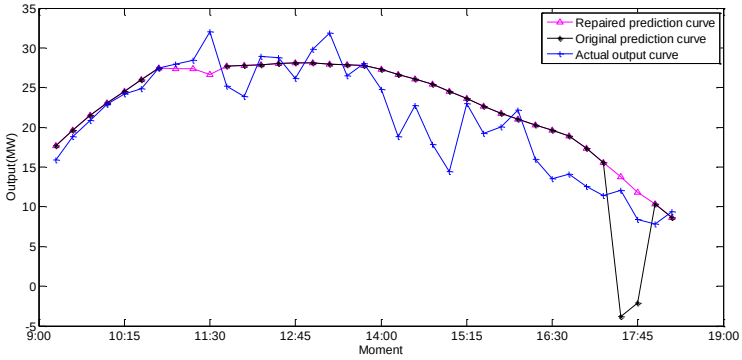

Figure 4. Actual output power and predictive power of a station before and after correction on May 27th

The rms error are calculated for the prediction curve before and after correction.

In addition to the vacancy period, the whole period error of the prediction curve before repair is 4.8417, and error after repair is 3.6023. Data is missing from 11:00 to 11:30. After the completion of the vacancy data, the error is 3.1851.During the abnormal period from 17:30 to 17:45, the error of the prediction curve before repair is 13.4537, and the after repair is 2.7084.It can be seen that the error value between the prediction curve and the actual output is effectively reduced.

Figure 5 shows the repaired predicted curve, the original predicted curve and the actual output curve on May 22nd.

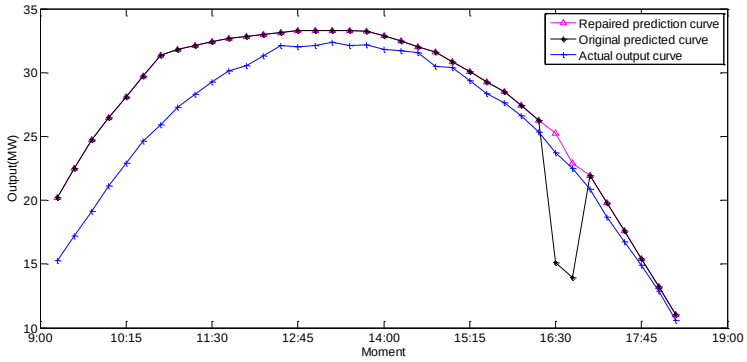

Figure 5. Actual output power and predictive power of a station before and after correction on May 22nd

After the error calculation, the whole period error of prediction curve before repair is 3.3672, and the error after repair is 2.7257. During the abnormal period from 16:30 to 16:45, the error of the prediction curve before repair is 8.5751, and the after repair is 1.1103.It can be seen that the repaired prediction curve has a high reduction accuracy for the abnormal period. The correction of abnormal data in the original prediction curve effectively reduces the error with the actual output.

5. Conclusion

The power generation of photovoltaic power station is stochastic with the fluctuation of meteorological factors. It is very difficult to distinguish the abnormal data from the normal data. The abnormal data mainly come from acquisition, measurement, transmission phase interference and fault generation.

The paper mainly revised the original forecast power of May 27th from 11:00 to 11:30, 17:30 to 17:45 and 16:30-16:45 of May 22th.After calculating the Pearson product distance correlation coefficient of the B, C and D power output data that the adjacent power station A within 10 kilometers

of the same area radius of the revised power station, that the D power station has the highest degree of correlation.

Considering the three similar days of 27th and the three similar days of 22th, the power output of D power station in similar days is used as the neural network sample. Finally, the original predictive power data of A power station is repaired. Through the error analysis of the calculation results, the error interval of 27th and 22th of abnormal fluctuations were reduced by 10.7453 and 7.4648 respectively. Effectively reduce the error value with the actual output.

Acknowledgments
Project Supported by the Science and Technology Project of SGCC: Research on data government of PV power stations and optimization of power generation plan.（NYB11201704444）

References
[1] YANG Mao, XIONG Hao, YAN Gan-gui, MU Gang. Real-time prediction of wind power based on data mining and fuzzy clustering[J].Power System Protection and Control,2013,41(1): 1-6.
[2] Schlechtingen M, Santos I F, Achiche S.Using data-mining approaches for wind turbine power curve monitoring: a comparative study[J]. IEEE Transactions on Sustainable Energy, 2013, 4(3):671-679(in Chinese).
[3] ZHOU Yongning, YE Lin, ZHU Qianwen. Characteristics and Processing Method of Abnormal Data Clusters Caused by Wind Curtailments in Wind Farms[J]. Automation of Electric Power Systems,2014,38(21):39-46.
[4] ZHU Qianwen,YE Lin, ZHAO Yongning, LANG Yansheng, SONG Xuri. Methods for elimination and reconstruction of abnormal power data in wind farms[J].Power System Protection and Control,2015,43(3):38-44.
[5] Shi Jie, Lee Weijen, Liu Yongqian, et al. Forecasting power output of photovoltaic systems based on weather classification and support vector machines[J]. IEEE Transactions on Industry Applications, 2012, 48(3):1064-1069.
[6] ZHANG Yongchao. Research on bad data detection and identification methods in power system[D].Southwest Jiaotong University,2009.
[7] TEEUWSEN S P.Neural network based multi-dimensional feature forecasting for bad data detection and feature restoration in power system[C].IEEE Power Engineering Society General Meeting, 2006:18-22.
[8] YAN Yingjie, SHENG Gehao, CHEN Yufeng. An Method for Anomaly Detection of State Information of Power Equipment Based on Big Data Analysis[J].Proceedings of the CSEE,2015,35(1):52-59.
[9] LIU Hao, HOU Boyuan. The Application Of Fuzzy Mathematics At Bad Data Detection And Iden Tification Of State Estimation[J].Joural of Shandong Architectural And Civil Engineering In Stitute,1996,8(3):50-57.
[10] Ding Ming, Wang Lei, Bi Rui. A short-term prediction model to forecast output power of photovoltaic system based on improved BP neural network[J]. Power System Protection and Control,2012,40(11):93-99(in Chinese).
[11] Dai Qian,Duan Shanxu, Cai Tao,et al. Short-term PV generation system forecasting model without irradiation based on weather type clustering[J]. Proceedings of the CSEE, 2011, 31(34):28-35(in Chinese).
[12] Wang Fei, Mi Zengqiang, Yang Qixun, Zhao Hongshan.Power Forecasting Approach Of PV Plant Based On ANN And Relevant Data[J].ACTA ENERGIAE SOLARIS SINICA,2012,33(7):1171-1177.
[13] İzgi E, Öztopal A, Yerlib B, et al. Short-mid-term solar power prediction by using artificial neural networks[J]. Solar Energy,2012,86(2):725-733.