Studies on the Screening Method of the Outliers of Wind Power in Wind Power Generation

Feng Gang
Electric Power Dispatching Center
State Grid ZhouShan Power Supply Company
Zhejiang, China

Liu Yao
Faculty of Science
Beijing Forestry University
Beijing, China
e-mail: 1284797299@qq.com

Li Hongjie
Faculty of Science
Beijing Forestry University
Beijing, China
e-mail: lhjlhj1991@163.com

Fei Jianping
Electric Power Dispatching Center
State Grid ZhouShan Power Supply Company
Zhejiang, China

Liu Lihua
Faculty of Science
Beijing Forestry University
Beijing, China
e-mail: 664923606@qq.com

Wang Dong
Electric Power Dispatching Center
State Grid ZhouShan Power Supply Company
Zhejiang, China

Yi Lei
Faculty of Science
Beijing Forestry University
Beijing, China
e-mail: 10487443@qq.com

Wang Hongqing *
Faculty of Science
Beijing Forestry University
Beijing, China
e-mail: hongqing991@126.com

Abstract—Through the analysis of wind power, reasonably adjusting the power is the key to stable operation of wind power generation. In this paper, first the wind power curve is studied on the basis of the least square method, so that the variation rules of wind power can be determined, and on the basis of a given reasonable threshold, the outliers and abnormal coefficient can be identified.secondly, we expand the data dimension and calculating the 15-day average of every time . Thus there is a new data set , and using the same method above can quickly locate the outliers existing range. Finally, time series iteration method is used to establish outliers identification model, and the points of wind power are found. Finally, the actual value on the abnormal point based on the outlier fitting model is estimated. Thus effective regulation of wind power can be achieved.

Keywords-The least square method; time series iteration method; outlier identification;The outlier fitting model; MATLAB

I. INTRODUCTION

The development and utilization of wind energy began in the 1970's , Some developed countries, such as America and Western Europe had to find new energy to substitute fossil energy under the pressure of oil crisis. They began to build wind farms and grid generation from 1980's, and then the wind energy became a new power energy. From the middle of 1980's, the wind power technology has made rapid development in the world. Wind power can not only reduce emissions of the main greenhouse gas, but also meet the growing global demand for energy[1].

The running data of wind farm is collected through the SCADA system. It may be interfered during the data collection, transmission or conversion, and because of the scheduling mechanism control, maintenance and other factors will also result in abnormal actual power of wind farm. Due to scheduling instruction, fan operation and other auxiliary information is difficult to obtain, the abnormal data cannot be effectively screened. Thus, choosing suitable methods and effectively eliminating the outliers in the data can make the wind power play to the best effect.

At present, research on the accuracy of wind power system data is roughly divided into two kinds:(1)a data validation method based on Neural Network[2],using parameter prediction model predict parameters; But the

neural network method always turns the feature of problems to digital, all reasoning to numerical calculation， and the result is bound to the loss of information. (2)Outlier detection based on wavelet analysis, in Literature [3] the high frequency components and low frequency components of the continuous data stream are separated, then the outliers in the data are found by combining with clustering method; however, once the wavelet function is selected, the property is fixed in the wavelet analysis, then it's difficult to accurately approximate the local features of the signal at different scales, there may be loss of original time domain features.

In this paper, the least square method to fit the data of wind power is used to determine the abnormal data and abnormal coefficient by formulating a reasonable threshold; in addition, a new method——the iterative algorithm in the time series analysis is used to distinguish the abnormal values. It does not only keep the time and domain characteristics and the hidden information of the original data, but also increases the fitting degree, improves the accuracy, and reduces the searching time of abnormal values.

II. DATA FITTING BASED ON LEAST SQUARE METHOD

Supposing that the error $\delta_i = S^*(x_i) - y_i (i = 0,1,...,m)$, $\delta = (\delta_1, \delta_2, ..., \delta_m)^T$,and $\varphi_0(x), \varphi_1(x), ..., \varphi_n(x)$ is linearly independent functions in $C[a,b]$,then we can find a function $S^*(x)$ in $\varphi = span\{\varphi_0(x), \varphi_1(x), ..., \varphi_n(x)\}$, and the $S^*(x)$ meets the condition of

$$\|\delta\|_2^2 = \sum_{i=0}^{m} \delta_i^2 = \sum_{i=0}^{m} [S^*(x_i) - y_i]^2 = \min_{S(x)\in\varphi} \sum_{i=0}^{m} [S(x_i) - y_i]^2 \quad (1)$$

Where $S(x) = a_0\varphi_0(x) + a_1\varphi_1(x) + ... + a_n\varphi_n(x) \quad (n < m)$. This is the method of least squares in curve fitting[4].

A. Establish Model

We take a data set $(x_i, y_i)(i = 10,1,2,...,m)$, containing m+1 elements, in which x_i is time and y_i is power. Then we use the polynomial of $p_n(x) = a_0 + a_1x + a_2x^2 + ... + a_nx^n$, the approximating function $y_i = f(x_i)$. The approximating method is to make the distance minimum based on the discrete norm of $d(a_0, a_1, ..., a_n) = \sum_{i=0}^{m} (y_i - p_n(x_i))^2$. Therefore we get the linear equation about $a_0, a_1, ..., a_n$:

$$s_0a_0 + s_1a_1 + ... + s_na_n = t_0$$
$$s_1a_0 + s_2a_1 + ... + s_{n+1}a_n = t_1$$
$$.....$$
$$s_na_0 + s_{n+1}a_1 + ... + s_{2n}a_n = t_n$$

Where $s_k = \sum_{i=0}^{m} x_i^k, 0 \le i \le 2n; t_k = \sum_{i=0}^{m} y_i x_i^k, 0 \le i \le n$. Supposing that the threshold we set before is σ ， then we decide y_i is outlier if y_i satisfies $|p(x_j) - Y_i| > \sigma$.

B. Prove and Analysis

The data originates from the Jiangsu coastal wind power plants, a total of 34848 groups of wind monitoring data from December 1, 2008 to March 31, 2009, provided by the Electric Power Research Institute. Due to the large amount of data, we use the grouping method to simplifying the process, and also improve the accuracy of the model. The specific steps are as follows:

(1)Take 150 data to fit in turn, that is, the first group is the data from 1st to the 150th, the second group is from the 151th to the 200th, and the third one is from the 101th to the 250th, and so on .There are 233 groups.

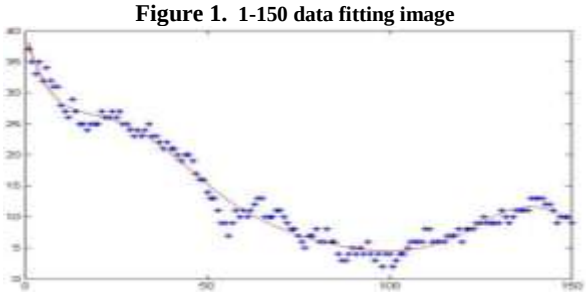
(2)Use least square method to construct 9 times polynomial approximation on the 150 data we take, where m=9, n=9.

(3)Because the middle part of the curve is fitted the best, we take the middle 50 data of the first group (that is 51-100) and the middle 50 data of the second group (that is 101-150) end to end，and so on；

(4)As that, until the end of the data.

After that we use matlabR2010a，operate all data by adopting the method above，the following is partly fitting results (**Fig .1**)：

Figure 1. 1-150 data fitting image

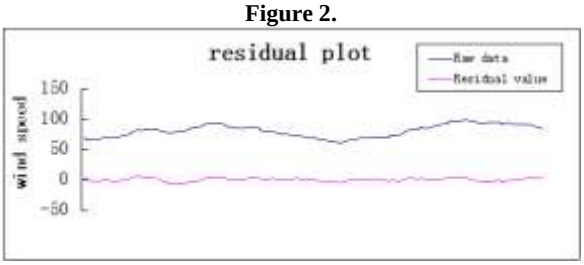


At the same time, we get the coefficient matrix of the step function, thus the final 9 order approximation polynomial is:

$$y = 64.9332 + 1.2756x - 0.1665x^2 + 0.0209x^3 - 0.0011x^4 + (2.56E-05)x^5 - (3.34E-07)x^6 + (2.40E-09)x^7 + -(8.95E-12)x^8 + (1.36E-14)x^9$$

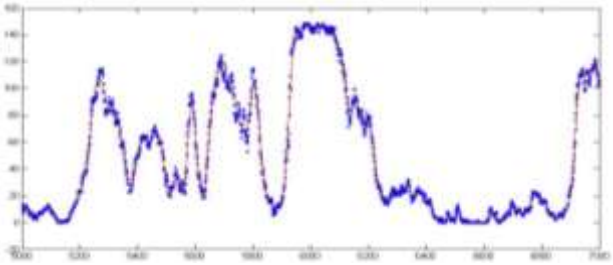
The residuals chart of the original data and fitting data:

Figure 2.



Due to the large amount of data (all data is 34848)，we intercept the fitting image of the 5000th to the 7000th ， as following：

Figure 3.



Analysis：Using the 9 order polynomial fitting achieves the best effect, which is more than 90%.By the residuals chart of the data, we can find the residual curve fluctuates around 0, and the relative data fluctuation is very small. This can also illustrate that the fitting degree is very good.

III. INSPECTION OF OUTLIERS BASED ON LEAST SQUARE METHOD

Using all of the original data to search the outliers is a method based on one-dimensional space. We expand the dimension of the data. We use the average of 5 days (15 days) data instead of the original data at each time to form a new dataset. Then using the above method can be more quickly to determine the interval of the outliers.

In this case, we suppose that for data of each time, the arithmetic average of five data: the two data prior to this data, the two data next to this data and this data take place of the original data of each time. For example, we render the data of 100th time y_{100} a new value Y_{100}, $Y_{100}=(y_{98}+y_{99}+y_{100}+y_{101}+y_{102})/5$ to replace the original data. Then, apply previous Least Square Method to analyze new data, where in the same way, denote σ as threshold value and assume that if Y_i satisfies $\left|p(x_j)-Y_i\right|>\sigma$, there are outliers between $y_{i-2}, y_{i-1}, y_i, y_{i+1}, y_{i+2}$.

In the same way, we can use the arithmetic average of fifteen data to replace the original data. And in this case, if Y_i satisfies $\left|p(x_j)-Y_i\right|>\sigma$, we assume that there are outliers between these fifteen data which are near Y_i , and this is the range of outliers that we find.

Applying this method, inspect data 101st -250th. Figure 4 shows the figure of new data and fitting curve. Table 1 shows the outliers where we assume that σ =0.1 and use arithmetic average of fifteen data to replace the original data.

Figure 4.

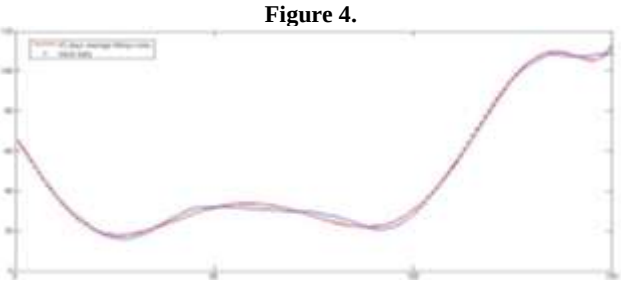


Table 1.

Time	Abnormal coefficient	Time	Abnormal coefficient
181	12.65%	179	11.13%
183	12.62%	129	10.89%
182	12.61%	128	10.42%
180	12.50%	178	10.23%
184	11.17%	127	10.08%
184	11.17%	127	10.08%

The figure and table show that the abnormal coefficients of time 178th to 184th are much big, which means that it is very possible that outliers exist between their overlap parts. On the other, the abnormal coefficients of 127th to 129th are also big, which means it is also possible that outliers exist between their overlap parts.

IV. INSPESTION OF OUTLIERS BASED ON ITERATION METHOD[5]

Before you begin to format your paper, first write and save the content as a separate text file. Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

A. Mathematic Model of Outliers

Among the on-line monitoring data of wind power, according to their properties and generating mechanisms, outliers can be classified into 2 kinds[6].

1) Additional outliers

This kind of outliers is isolated. Normally, these outliers generate because equipments have been interfered externally or disturbed by themselves. Their appearance won't affect adjacent observed data. As for time series which is based on on-line monitoring value, this kind of outliers is nonessential and they are not related to internal structure of time series.

2) New Informational Outliers

This kind of outliers will appear aggregated because they will affect a series of observed data through correlation of time series. And the generating mechanism of this kind of outliers is that the structure of dynamical system has changed so that the internal structure of time series will change abnormally and generate these outliers.

Z_t is observed series, and X_t is a series without outliers。 Assume that { X_t } fits an ARMA(p,q) model: $\varphi(B)X_t=\theta(B)a_t$,where $\varphi(B)=1\text{-}\varphi_1 B\text{-}...\text{-}\varphi_p B^p$, $\theta(B)=1\text{-}\theta_1 B\text{-}...\text{-}\theta_q B^q$, are smooth and invertible operator without common factor, { a_t } is white noise series which are mutually independent and follow same

distribution $N(0,\sigma_a^2)$. With these preconditions, denote Additional Outliers (AO) model:

$$Z_t=\begin{cases} X_t, & t\neq T \\ X_t+\omega, & t=T \end{cases}=X_t+\omega I_t^{(T)}=\frac{\theta(B)}{\phi(B)}a_t+\omega I_t^{(T)} \qquad (2)$$

where $I_t^{(T)}=\begin{cases}1, t=T\\ 0, t\neq T\end{cases}$ is an indicative function to describe whether outlier exists at time T. And denote New Informational Outliers (IO) model:

$$Z_t=X_t+\frac{\theta(B)}{\phi(B)}\omega I_t^{(T)}=\frac{\theta(B)}{\phi(B)}(a_t+\omega I_t^{(T)}) \qquad (3)$$

Therefore, AO just affect Tth observed value while IO affect all of after time t=T value $Z_t, Z_{t+1},\ldots$ which are through the system described by $\frac{\theta(B)}{\varphi(B)}$.

B. Steps of iterative method

Step 1[7]: We assume that outlier does exist. By modeling the sequence of Z_t, we can estimate the residuals. That is $\hat{e}_t=\hat{\pi}(B)Z_t=\frac{\hat{\phi}(B)}{\hat{\theta}(B)}Z_t$ with $\hat{\phi}(B)=(1-\hat{\phi}_1B-\ldots-\hat{\phi}_pB^p)$, $\hat{\theta}(B)=(1-\hat{\theta}_1B-\ldots-\hat{\theta}_qB^q)$. Let $\hat{\sigma}_\alpha^2=\frac{1}{n}\sum_{t-1}^{n}\hat{e}_t^2$,there is the initial estimate of σ_a^2.

Step 2 : Use estimated model to calculate $\hat{\lambda}_{1,t}$ and $\hat{\lambda}_{2,t}$ .Let $\hat{\lambda}_T=\max_t\max_i\{|\hat{\lambda}_{i,t}|\}$, Where T is the time of occurrence of maximum. If $\hat{\lambda}_T=|\hat{\lambda}_{1,T}|>C$ (C is a positive constant predetermined), then there is an outlier of AO type at T, with $\hat{\omega}_{AT}$ expressed its influence. We can use $X_t+\omega I_t^{(T)}$ to revise the data: $\tilde{Z}=Z_t-\hat{\omega}_{AT}I_t^{(T)}$ .Let AO： $e_t=\omega\pi(B)I_t^{(T)}+a_t$, define a new residual value $\tilde{e}=\hat{e}_t-\hat{\omega}_{AT}\hat{\pi}(B)I_t^{(T)}$. If $\hat{\lambda}_T=|\hat{\lambda}_{2,T}|>C$, then there exists an outlier of IO type at T with the influence of $\hat{\omega}_{IT}$. We can use $Z_t=X_t+\frac{\theta(B)}{\phi(B)}\omega I_t^{(T)}$ to revise the data and eliminate the effect of IO, which is $\tilde{Z}=Z_t-\frac{\hat{\theta}(B)}{\hat{\phi}(B)}\hat{\omega}_{IT}I_t^{(T)}$. Let IO: $e_t=\omega I_t^{(T)}+a_t$ define a new residual value： $\tilde{e}=\hat{e}_t-\hat{\omega}_{IT}I_t^{(T)}$. Thus, we can calculate the new estimates $\tilde{\sigma}_a^2$ from the corrected residuals.

Step 3: On the basis of the revised residuals and $\tilde{\sigma}_a^2$, we calculate $\hat{\lambda}_{1,t}$ and $\hat{\lambda}_{2,t}$. Then repeat step 2, until all the outliers are identified.

Step 4 : We assume that there are k outliers at time T_1, $T_2,\ldots,T_k$ is tempted to identify after step 3. We deal with these moments as known values to estimate the parameters of outliers ω_1, $\omega_2,\ldots\omega_k$, . At the same time we estimate the parameters of time series, and use the model L $Z_t=\sum_{j=1}^{k}\omega_j\nu_j(B)I_j^{(T_j)}+\frac{\theta(B)}{\phi(B)}a_t$. Among them, at $t=T_j$, if outlier is AO type, then $\nu_j(B)$ =1; if outlier is IO type, then $\nu_j(B)=\frac{\theta(B)}{\phi(B)}$. Thus it leads to a new residual: $\hat{e}_t^{(1)}=\hat{\pi}^{(1)}(B)[Z_t-\sum_{j=1}^{k}\omega_j\hat{\nu}_j(B)I_t^{(T_j)}]$. So the revised estimation of σ_a^2 can be calculated.

Repeat steps 2 to 4.Until all the outliers are identified, estimate their impact at the same time. So you get the following fitting model of outliers : $Z_t=\sum_{j=1}^{k}\hat{\omega}_j\hat{\nu}_j(B)I_t^{(T_j)}+\frac{\hat{\theta}(B)}{\hat{\phi}(B)}a_t$. Among them , $\hat{\omega}_j$, $\hat{\phi}(B)=(1-\hat{\phi}_1B-\ldots-\hat{\phi}_pB^p)$ and $\hat{\theta}(B)=(1-\hat{\theta}_1B-\ldots-\hat{\theta}_qB^q)$ are obtained in the last iteration.

C. Results and analysis

We set the threshold parameter C=2.5.We use MATLAB software to write the iterative algorithm package of outliers recognition model. The above method is applied to test the sequence with outliers. We examine all the outliers in the data with 34848. Results are as follows:

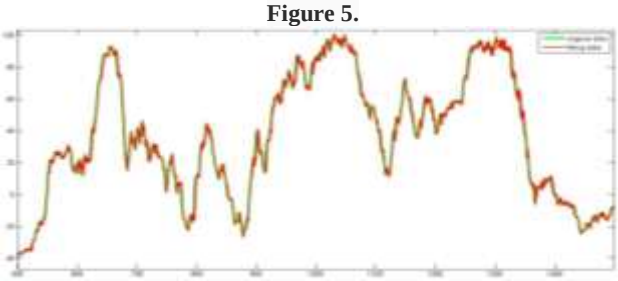
Table 2. outliers detected

Num of iterations	num	outliers types 1:AO 2:IO	inpact factor	num of iterations	num	outliers types 1:AO 2:IO	inpact factor
1	5932	2	27.33539	13	15181	2	19.69036
2	15182	2	25.06777	14	31855	2	-18.7368
3	3607	2	23.70482	15	25188	2	-18.4197
4	24653	2	23.65092	16	25104	2	-18.3697
5	24657	2	-23.59477	17	22594	2	18.2194
6	22224	2	23.08763	18	2878	2	17.9904
7	31854	2	-21.93443	19	11335	2	-17.9791
8	24732	2	20.98653	20	31846	2	17.62073
9	31862	2	-20.86744	21	33743	2	17.48789
10	25202	2	20.26142	22	24658	2	-16.9551
11	7873	2	-20.2145	23	29332	2	16.9051
12	11333	2	-20.13833				

Through the check of the outliers we found 0 IO outlier and 23 AO outliers. Therefore these outliers have no effect on the following data. Among them the two most influential outliers which have been found are No. 5932 and No.15182. Compared with the original data, the two outliers are more abnormal than the adjacent data.

Finally, we use the outlier fitted model to estimate the true value of the outliers that have been identified. The final fitting results by running MATLAB software is shown below:

Figure 5.



Among them, the green part is the original data image; the red part is the fitting data images.

V. OUTLOOK AND SUMMARY

In this paper, large data of wind power generation is piecewise fitting by using the method of least squares. Using 9 order polynomial fitting improves the fitting degree, and maintains the time domain characteristics and basic information hidden of original data; Using 15-day-averages of curve fitting can be more quickly to determine the existence range of outliers, and reduces the searching time of outliers; at the same time, the use of time series analysis to check of the outliers improves the accuracy. In the future, according to the need of engineering, by accurate definition of outliers we will present an effective screening algorithm of outliers. Then we can realize the online recognition of the outliers of power on wind farm, and develop the online recognition software package of the data with abnormal time series.

ACKNOWLEDGMENT

The other authors wish to express their thanks to the corresponding author--Associate Professor Hongqing Wang for his useful discussion with them and his encouragement.

REFERENCES

[1] Chen Zhenghong, Xu Yang, Xu Peihua. Wind power prediction technology andbusiness system [M]. Beijing: China Meteorological Press, 2013.1:14

[2] Si Fengqi, Xu Zhigao. Measurement data based on auto associative neural network self-tuning test methods [J]. Proceedings of the CSEE of China, 2002, 22(6): 152-155

[3] Xu Xuesong, Shen Honghong, Tao Fan et al. Uncertain data stream outlier data detection based on wavelet analysis[J]. Software Guide，2011，10(11):40-42.

[4] Li QingyangWang Nengchao, Yi. Numerical analysis (Fifth Edition) [M]. Beijing: Tsinghua University press，2008.12:73

[5] George E P Box, Gwilym M Jenkins and Gregory C Reinsel. Time Series Analysis: Forecasting and Control(Third Edition)[M]. Prentice-Hall, Inc 1994.

[6] Fox A J. outliers in time series, J. Roy. Statist. Soc. 1972, B34, 350-363.

[7] William W.S.Wei.TIME SERIES ANALYSIS —— Univariate and Multivariate Methods (Second Edition) [M]. 2009.4:214-226.

[8] Li Fudong, Wu Min, Feng Gaoyi. Wind power slope event classification and prediction based on statistical analysis and multiple support vector machines[J]. Journal of Shanghai Jiao Tong University，2012，,46(12):1971-1976.

[9] Guo Jiwei, Xie Jingdong, Tang Guoqing.Transformer on-line monitoring method based on the analysis of the abnormal value[J]. Electric power system and automation，2003,15(6)：64-66

[10] Wu Xueguang, Zhang Xuecheng, Yin Yonghua et al. The mathematical model for analyzing the dynamic stability of wind power systemand its application [J]. Power system technology，1998,22(6):68-72.

[11] Ni Jingfeng, Liu Lihua, Gu YujiongAnomaly detection method of measurement data sequence based on the method of least squares support vector calculation[J]. Journal of North China Electric Power University，2008,35(3):62-66.