A Method for Preprocessing State Data of Power Transmission and Transformation Equipment

Zhou Chao, Qin Jiafeng*, Lin Ying, Zheng Wenjie, Bai Demeng and Yang Yi

1State Grid Shandong Electric Power Research Institute, China

Corresponding author: *qinjiafeng0118@163.com

Abstract. In the era of big data, transportation and inspection data of power transmission and transformation equipment are characterized by diversity and richness. Massive data provide data support for state assessment of power transmission and transformation equipment, but at the same time, higher requirements are put forward for traditional data management and data quality model. How to clean the errors and invalid data in the transportation and inspection data, how to effectively repair the missing data, establish the data quality evaluation model, and improve the quality of the transportation and inspection data of the transmission and transformation equipment are of great significance to the equipment status evaluation. For the acquired data, this paper carries out preliminary cleaning work on the data, and carries out preliminary cleaning on the data through time series analysis and other technologies to ensure the validity, consistency and integrity of the data.

1. Introduction

In recent years, with the continuous expansion of grid and grid production information, equipment stand-books of the information and status information such as basic information of digital level is higher and higher, the enrichment of equipment testing means and produce the amount of data of power grid operation maintenance and equipment inspection exponentially, associated with the equipment state change of equipment condition monitoring, production management, operation scheduling, environmental meteorological data such as step by step to realize the integration of the on unified information platform sharing. Power transmission and transformation equipment status, related information is scattered in various business application system, different complex data structure, data interface, data communication difficulties, poor interactivity between platforms, lead to the spread of information and resources, serious heterogeneity, the vertical and horizontal sharing through difficulties, and the data quality, data extraction and fusion integration analysis is difficult, more difficult to effectively use the source data to realize comprehensive evaluation, the effect and efficiency of equipment condition assessment diagnosis. At present, the multi-level and multidimensional mass data of power transmission and transformation equipment status, power grid information and environment information generated by power grid enterprises provide a large amount of data basis for the effective evaluation of power transmission and transformation equipment status. However, these data have the following characteristics[1-4]:

(1) large data volume

Of power grid construction and operation for a long time, has formed a certain scale of power transmission and transformation equipment state information related data, the variety of devices
involved, including transformer, combined electrical equipment, circuit breaker, overhead lines, power cable and arrester, the equipment quantity is very large, and with the construction and development of smart grid, smart checking, online monitoring and testing system of electric equipment state information, and is closely related to the device status information such as the power grid, the environment of the great amount of data and rapid growth.

(2) diversified data types

The overall state information of power transmission and transformation equipment can be divided into four data types: structured data, unstructured data, historical/quasi-real-time data and GIS spatial data. The collection frequency and life cycle of each type of data vary from minute to hour, and even to monthly and annual levels.

(3) scattered data sources

Power transmission and transformation equipment status, related information is scattered in various business application system, including power transmission and transformation equipment condition monitoring, PMS, EMS, such as GIS, weather, lightning, decentralized deployment in different places, each system is managed by different units/departments, the data interface between each are not identical, platform, data communication difficulties, poor interactivity, lead to the spread of information and resources, serious heterogeneity, vertical and horizontal sharing through difficulties.

(4) inconsistent data model

The source data directly read from multiple business systems have problems such as inconsistent model and format, and the data of each system cannot be effectively integrated, which has become the bottleneck of subsequent data analysis.

(5) data quality needs to be improved

There are many kinds of data related to the state information of power transmission and transformation equipment, and their structures are complex and diverse. The source data has problems such as incomplete data, redundancy, conflict, error and omission, and anomaly, etc. The data quality needs to be improved.

In view of the above characteristics, such as multiple sources, heterogeneous information, large quantity and various attributes, it is difficult to guarantee data integrity, validity and consistency. In view of the acquired data, this paper carries out preliminary cleaning of the data, and realizes the cleaning of massive data based on the time series model.[4-6]

2. Data cleaning based on time series model

The detection of the state quantity of power transmission and transformation equipment is completed by each sensor, but the original data uploaded to the database for state evaluation after the underlying pre-processing can be considered as the characteristic quantity data arranged according to time series[7].The uniform format of these data is "time. Feature quantity = value", so it can be considered that all state quantities collected form a continuous and complete time series of one unit or multiple variables, as shown in matrix X.

$$X = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{m1} & X_{m2} & \cdots & X_{mn} \end{bmatrix}$$  \hspace{1cm} (1)

Where, $X_{mn}$ represents the value of the state quantity m at time $T_n$. The state data of power transmission and transformation equipment under normal operation generally present the following three rules, which can be applied to the time series method:

1) the amplitude of state variables changes little, such as wire tension, grounding current, and gas C2H2 in oil. These state variables are stationary sequences that can be directly fitted with ARMA(m, n).

2) the state quantity shows a slow rising trend. For example, the gas CO and CO2 in the oil can be converted into a stationary sequence by difference method and fitted with ARIMA(p, d, q).
3) the state quantity changes periodically, and the observation points after s time interval are similar in time series, such as oil temperature, wire temperature, etc., which can be fitted by ARIMA(p, ds, q).

According to the operation characteristics of power transmission and transformation equipment, anomalies in the state data are usually manifested in two forms: 1) anomalies that can be used for data cleaning, namely noise points and missing values; 2) abnormal data caused by interference in the running state of the equipment.

3. The implementation process
Since the existence of abnormal data will cause deviation in the estimation of time series parameters, the iterative test method is used to clean the observed time series in the case of unknown time and number of noise points and no model parameters in advance[8].

1) assuming no outliers exist, a time series model is established for the observation series M, and the estimated model is used to calculate the initial residual, namely:

\[ \hat{e}_t = \pi(B)Z_t = \frac{\varphi(B)\nabla^d}{\theta(B)}Z_t \] (2)

Where, \( Z_t \) is the observation sequence, \( \hat{e}_t \) is the residual sequence of the initial fitting, \( \theta(B) \) and \( \varphi(B) \) are the stationary and reversible operators of the initial fitting, and \( \nabla = 1 - B \) is the delay operator. The initial estimate of the residual variance is

\[ \sigma_e^2 = \frac{1}{n} \sum_{j=1}^{n} e_j^2 \] (3)

2) observation and fitting residual sequence. If the residual sequence shows horizontal migration from a certain time point and is much larger than the previous residual value, the original time series needs to be fitted with an intervention model and skip to step 7. Otherwise jump to the outer loop.

3) In the outer loop, using the estimated model, t=1,2...,n, calculate the test statistics of each observation point \( T_{io} \) and \( T_{ao} \). By definition \( \lambda_T = \max \left\{ \left| T_{io}^T \right|, \left| T_{ao}^T \right| \right\} \), here T is the time when the maximum value occurs. At that time, when \( \lambda_T > C \) it indicates the existence of abnormal data, into the inner loop to correct the data.

4) In the inner loop, correct the data.

When \( \lambda_T = \left| T_{io}^T \right| > C \), where C is A pre-determined normal number, usually between 3 and 4, it can be determined that there is abnormal data(AO),and its influence on model fitting at time T is \( \omega_{ao} \):

\[ AO: \quad \omega_{ao} = \frac{e_t - \sum_{j=1}^{n-1} \pi_j e_{t+j}}{\sum_{j=0}^{n-1} \pi_j^2} \] (4)

By adding outlier model \( Z_t = X_t + \omega_{i o}^{(T)} = \theta(B) + \omega_{i o}^{(T)} \) to modify the original time series data, the new time series \( Z_t \) is obtained:

\[ Z_t = Z_t - \omega_{ao} I_{ao}^{(T)} \] (5)

And the new residual \( \hat{e}_t \) is obtained by \( AO: \quad e_t = \omega_{ao} \pi(B)I_{ao}^{(T)} + a_t \), correction:

\[ \hat{e}_t = e_t - \omega_{ao} \pi(B)I_{ao}^{(T)} \] (6)

When \( \lambda_T = \left| T_{io}^T \right| > C \), it is determined that there is abnormal data IO at time T, and its influence on model fitting is :

\[ IO: \quad \omega_{io} = e_t \]. The following equation is used to correct the data.
\[ Z_t = X_t + \omega(\frac{\partial(B)}{\varphi(B)N^d})I_t^{(B)} = \theta(B) + \omega I_t^{(B)}(a_t + \omega I_t^{(B)}) \]  

(7)

Then the influence of IO can be eliminated, that is

\[ Z_t = Z_t - \frac{\theta(B)}{\varphi(B)N^d} \omega I_t^{(B)} \]

(8)

The new residuals are obtained by IO: \( e_t = \omega I_t^{(B)} + a_t \):

\[ \tilde{e}_t = e_t - \omega I_t^{(B)} \]

(9)

The iterative method is used to identify and correct all noise points in the time series. Test statistics \( T_{io} \) and \( T_{ao} \) of each observation point were calculated again based on the modified residuals \( \tilde{e}_t \) and residuals standard deviation \( \sigma_t^2 \), and step 4 was repeated until all abnormal data were identified. When \( \lambda_t < \lambda \), it indicates that the out-of-step loop has fixed the abnormal data and the inner loop ends.

5) Suppose \( k \) abnormal data are identified at time \( t_1, t_2, ..., t_k \) after the inner loop is finished, and their effects are \( \omega_{t_1}, \omega_{t_2}, ..., \omega_{t_k} \) respectively. Meanwhile, the abnormal data are modified to obtain a new time series \( Z_t^{(i)} \) (1 in the upper right corner represents the sequence obtained by the first iteration of the outer loop). At this point, return to 3), enter the outer loop, re-estimate the time series parameter \( \hat{\theta}^{(i)}(B), \hat{\varphi}^{(i)}(B), \pi^{(i)}(B) \), and obtain the residual \( \hat{e}_t^{(i)} \) of the time series model:

\[ \hat{e}_t^{(i)} = \pi^{(i)}(B)[Z_t^{(i)} - \sum_{j=1}^{k} \omega_{j}^{(i)} \nu_{j}(B)I_t^{(B)}] \]

(10)

\[ \nu_{j}(B) = \begin{cases} \frac{\theta^{(i)}(B)}{\varphi^{(i)}(B)N^d} & \text{IO} \\ 1 & \text{AO} \end{cases} \]

(11)

Test statistics were calculated according to the re-estimated time series parameters. When \( \lambda_t < \lambda \), the outer loop ended, and when \( \lambda_t < \lambda \), it re-entered the outer loop, until all abnormal data were repaired.

6) After the end of the last outer cycle, a joint estimation was conducted for the time series \( Z_t \) with modified noise points, and a model of fitting outliers was obtained:

\[ Z_t = \sum_{j=1}^{k} \omega_{j} \nu_{j}(B)I_t^{(B)} + \frac{\theta(B)}{\varphi(B)N^d} a_t \]

(12)

Where, each parameter \( \theta(B), \varphi(B), \omega_{j}, \nu_{j} \) was obtained in the last iteration. The purpose of this joint estimation is to verify whether the mathematical model of data cleaning is similar to the real data, that is, the fitting residual is within the acceptable range. At this point, the data at the abnormal time point in the formula is regarded as the "modified" data to replace the original data, while the data at other time points still retains the original value.

7) Use the time series intervention model of the following formula to fit the original data and figure out the occurrence time of the intervention point.

\[ \frac{\omega B^d}{1 - \delta B} S_t^{(B)} \]

(13)

4. Data optimization

Data optimization is the processing of abnormal data such as filling, removing singular value, smoothing and envelope fitting[9-10].
1) data supplement

The lost data, constant data, negative value and overrange data identified in the data quality discrimination are replaced by null values. The null data and the null values existing in the original data all need to be filled by the method of data filling. Filling a gap is a "make-believe" job, so a more conservative approach is adopted. Due to the poor horizontal comparability between different measurement points of oil chromatographic data, it is difficult for this kind of data to estimate the lost value with similar data, and it is impossible to calculate the lost value with its historical data. Therefore, a simple linear estimation method is adopted, that is, the data changes linearly during the loss. This method is relatively conservative, but has few changes to the statistical information of the data.

2) elimination of singular value

Before the singular value is eliminated, it is necessary to identify which values are singular. According to 3σ rule in signal processing, the singular value is evaluated and processed. According to Chebyshev's inequality, if the expectation of random variable X is \( \mu \) and the variance is \( \sigma^2 \), then for any given positive number \( \varepsilon \), there is:

\[
P\left(\left|X - \mu\right| \geq \varepsilon \right) \leq \frac{\sigma^2}{\varepsilon^2}
\]

In the case that the variance is known, if I take \( \varepsilon = 3\sigma \), then:

\[
P\left(\left|X - \mu\right| \geq 3\sigma \right) \leq \frac{\sigma^2}{9\sigma^2} \approx 0.111
\]

Therefore, there is the rule of 3σ: for any given distribution, as long as the expectation and variance exist, the probability that the value of random variable X deviates from \( \mu \) by more than 3 times the mean variance is less than 0.111. From the probability point of view, the data of 0.3% farthest from the data distribution center is considered as singular value, and only the original data of 99.7% is retained. In practice, we can replace \( \sigma \) with sample standard deviation S. After using the aa rule to identify the singular value in the data column, the mode is used to replace it, and the original characteristics of the data are retained.

3) smooth data

For the random noise and white noise in the data, the method of sliding average is used to remove the noise, and a low pass filter whose filter coefficient is equal to the reciprocal of the span is used. The treatment method is as follows:

\[
\begin{align*}
y_1 &= x_1 \\
y_2 &= (x_1 + x_2 + x_3) / 3 \\
&\vdots \\
y_i &= (x_{i-2} + x_{i-1} + x_i + x_{i+1} + x_{i+2}) / 5 \\
&\vdots \\
y_{n-1} &= (x_{n-2} + x_{n-1} + x_n) / 4 \\
y_n &= (x_{n-2} + x_{n-1} + x_n) / 3
\end{align*}
\]

For the data of the edge position of the step point, the method of reducing the smooth area is adopted.

5. The experimental results

Select the annual monitoring data of a transformer to draw its distribution diagram before and after quality improvement, as shown in figure 1. Figure in the distribution of original data is shown in blue and figure illustrates the monitoring data in 180 days or so singular value, red curve said after the data quality improvement of data distribution, it can be seen after the data quality improvement, singular value in 180 days was eliminated, and the data after the data quality improvement compared to the original data better characterization of oil chromatographic data trends.
6. Conclusion
In this chapter, the preprocessing method of transportation inspection data of transmission and transformation equipment is presented, including data cleaning technology, feature extraction, filling and related steps. Based on the data of the early stage of the observation, found to be existing in the data cleaning is invalid, repetition, error and missing problem, for this kind of situation, the overall data pretreatment process adopted for data cleaning first remove the invalid data or duplicate data, then the data is wrong data restoration and missing data filling process. In the process of data repair, the data to be repaired is the timing sequence data that has been cleared of outliers and wrong data points after pre-processing. For the reliable timing sequence operation data after pre-processing, the missing part of the data is repaired.

References
[1] Ramsay J O, Silverman B W. Applied functional data analysis: methods and case studies[M]. Springer, 2007: 77.
[2] Wold S, Esbensen K, Geladi P. Principal component analysis[J]. Chemometrics and intelligent laboratory systems, 1987, 2(1-3): 37-52.
[3] Leng X, Müller H G. Classification using functional data analysis for temporal gene expression data[J]. Bioinformatics, 2005, 22(1): 68-76.
[4] Lian H. Nonlinear functional models for functional responses in reproducing kernel Hilbert spaces[J]. Canadian Journal of Statistics, 2007, 35(4): 597-606.
[5] Hall P, Hosseini - Nasab M. On properties of functional principal components analysis[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2006, 68(1): 109-126.
[6] Li Y, Hsing T. Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data[J]. The Annals of Statistics, 2010, 38(6): 3321-3351.
[7] Ghanem R G, Spanos P D. Stochastic finite elements: a spectral approach[M]. Courier Corporation, 2003: 17-20.
[8] Creven P, Wahba G. Smoothing noisy data with spline function number[J]. Math, 1979, 31: 377-403.

[9] Golub G H, Heath M, Wahba G. Generalized cross-validation as a method for choosing a good ridge parameter[J]. Technometrics, 1979, 21(2): 215-223.

[10] Allen D M. The relationship between variable selection and data augmentation and a method for prediction[J]. Technometrics, 1974, 16(1): 125-127.