Method of Analyzing Transformer DC Magnetic Bias Based on Big Data Cleaning and Dimensionality Reduction

Pei Yuan*, Xiaowen Wu, Tao Wan, Huisheng Ye and Yun Liu
State Grid Hunan Electric Power Company Limited Research Institute, Changsha, Hunan, 410007, China
*Corresponding author’s e-mail: ypsnow87@sina.com

Abstract: Transformer DC magnetic bias is difficult to accurately analyze and predict because it is caused by complex factors and has a wide influence. This paper combines big data with DC bias magnetic analysis method, making a breakthrough in the DC bias prediction model that only considers a single factor. From the perspective of multi-system coupling, this paper analyzes the modeling of DC magnetic bias based on multi-dimensional big data, and extracts the key influencing factors of transformer DC bias through multi-system big data cleaning and dimensionality reduction technology, thus providing an effective basis for “one-button sequence control” bias magnetic treatment.

1. Introduction
When UHV DC transmission and rail transit are in operation, DC current will penetrate into the earth and into the transformer cores in the adjacent substations, resulting in greater loss, severe harmonics, local overheating, mechanical vibration and noise increase, or the phenomenon of DC bias, which hinders the safe and stable operation of the transformer [1-2]. Meantime, the current will be transmitted to larger substations through the transmission network of the intricate power system, thus leading to the spread of hidden troubles. Due to the differences in earth barrier and soil resistivity in different areas, this part of the current is “hidden” and “dispersive”, thus it is difficult to measure and locate using existing detection methods.

The method adopted in this paper breaks the limits of the traditional method using static power parameters by dynamically accessing the multi-sourced data [3-4]. The density anomaly detection data cleaning is used to remove the abnormal data, while the principal component analysis method is used to reduce the space dimension of the data feature and reduce calculation amount. In this way, this paper realizes the evaluation of different influencing factors of multiple systems of DC bias.

2. Data cleaning

2.1. Data source
This paper combines the data of the power industry and other industries to carry out data cleaning. The data of the power industry is taken from the 110 kV-and-above substations of Hunan Power Grid in 2018 (geographic information, design drawings of the grounding grid and distribution map of electrical equipment, grounding resistance value, model and parameters of the transformer, busbar parameters, station current, historical data of the operation mode of the station wiring, etc.), high-voltage transmission lines (geographic information, tower grounding resistance, parameters of wires and lightning conductor, wiring operation mode, etc.), DC transmission system (line current, converter
station, DC line and the geographic information of DC grounding electrode, parameters of the wiring model, grounding resistance, etc.), the data of the load current and power of the traction station of Changsha Rail Transit 1#~4# lines, the data of neutral point current, noise and vibration of grounding power transformer in Changsha Power Grid, and the pipeline data (geographic information, piping parameters, anti-corrosion coatings, cathodic protection for piping systems and configuration parameters of auxiliary facilities).

The data of other industries taken from the location and acceleration of the rail transit train, Changsha Rail Transit 1#~4# lines (geographic information, configuration and parameters of the track system, transition resistance, train parameters and annual operation time, electrical information of the traction station, etc.); meteorological data of the area for prediction; 14 geomagnetic first-level fixed observatories in the National Geomagnetic Network Center and the published data (magnetic declination, magnetic dip, total intensity, horizontal intensity, north component, east component and vertical intensity); relative change in magnetic declination, horizontal intensity, and vertical intensity). This part of data has storage of approximately 108 GB, as shown in Table 1.

| Data information | Scope | Source | Data magnitude | Data content |
|------------------|-------|--------|----------------|--------------|
| **Power industry** | Data of substation, tower, etc | 5.0 GB | Basic parameters of transformer, winding and iron core, excitation current, wiring operation information, substation geographical location information, soil resistivity, line-tower coordinate information, geographical information and electrical parameters of HVDC transmission system, etc. |
| Metro traction station data | 1.0 GB | Geographic information and electrical parameters of metro tracks and traction stations |
| Transformer on-line monitoring data | 8.62 TB | Substation transformer neutral dc current, excitation current, oil chromatography, transformer vibration displacement, acceleration and noise online monitoring data |
| Meteorological data | 20 GB | Meteorological data for hunan province |
| Geomagnetic data | 8.0 GB | From 2016 to 2018, the magnetic declination Angle, magnetic inclination Angle, total strength, horizontal strength, north component, east component and vertical strength, as well as the relative changes of horizontal strength and vertical strength |
| Metro operation data | 80.0 GB | Record data of subway running current, train speed/acceleration, running time and position, and annual running time of trains |

2.2. Data cleaning based on density anomaly detection

The principle of the density-based anomaly detection method is: any data point with dense points in its local surroundings is considered to be normal, while the outlier data point is one that is distant from the nearest neighboring points of the normal data point (a certain threshold is usually used to define the distance) [6]. This paper uses the local outlier factor method for data cleaning.

$$\text{reach-dist}(q, p) = \max \{ d(q, p), k - \text{distance}(p) \}$$

The value of reach-dist(q, p) is the reachable distance between data points p and q; d(q, p) is the Euclidean distance between points p and q; and k-distance(p) is the distance between points p and q.

The local reachability density is recorded as lrd, as shown in Equation (2):
In this formula, $N_k(q)$ represents the set of $k$ points closest to the data point $q$; $lrd_k(q)$, the local reachable density defined in equation (2) measures the sparseness of the data point $q$ within its first $k$ nearest points: large $lrd_k(q)$ value indicates the dense distribution of point $q$ in $k$ points, thus a normal point, while small $lrd_k(q)$ value indicates sparse distribution of point $q$ in $k$ points, thus an outlier point.

LOF (local outlier factor): LOF identifies the degree of outliers of data points and indicates of the probability of an outlier point.

$$\text{LOF}_k(q) = \sum_{p \in N_k(q)} \frac{lrd_k(p)}{lrd_k(q)}$$

LOF represents a density comparison showing the density difference between the data point $q$ and the whole. Studies have shown that if a LOF value much larger than 1 indicates a significantly difference between the density of the point $q$ and the overall density, thus considering the point $q$ as an outlier; a LOF value close to 1 means that the difference between the point $q$ and the whole is small, thus considering the point $q$ as a normal one; a LOF value smaller than 1 indicates that the density of point $p$ is higher than that of its surrounding points and $p$ is a dense point. Anomalies are divided into global anomaly and local anomaly: the values of the former on is the value comparing with all values in the dataset, and are often easy to spot when cleaning data; the values of the latter one is what seem normal in the whole data set, but abnormal when viewing the surrounding points. The common abnormal data of influencing factors of DC bias is shown in Figure 1.
This paper conducts data cleaning of data of the neutral point DC current monitoring of a transformer based on the density anomaly detection standard and obtains the data correction curve and local abnormal points using the local outliers method with LOF > 2, as shown in Figure 2.

![Data for local anomalies](image)

**Figure 2.** Local abnormal distribution of transformer neutral current and data correction curve

### 3. Data dimensionality reduction

There are two factors that this paper focuses on while simplifying and reducing the dimension of big data using Principal Component Analysis (PCA): first, there are a lot of redundant information in the 35-dimensional feature space that affects DC bias (see 2) [5], but a correlation exists between the features; second, the present storage and computing capacity of general computers is not enough for calculating complicated high-dimensional data of DC bias.

| number | 1 | 2 | 3 | 4 | 5 |
|--------|---|---|---|---|---|
| data   | transformer neutral point current | transformer acceleration | transformer displacement | transformer noise | transformer neutral point isolation voltage |
| number | 6 | 7 | 8 | 9 | 10 |
| data   | transformer operating power | excitation current component | substation geographical location | earth resistivity | line-tower coordinate information |
| number | 11 | 12 | 13 | 14 | 15 |
| data   | Dc grounding pole geographic information | substation grounding resistance | train load current | train acceleration | train speed |
| number | 16 | 17 | 18 | 19 | 20 |
| data   | Train real time position | transformer basic parameters | temperature | relative humidity | rainfall |
| number | 21 | 22 | 23 | 24 | 25 |
| data   | Wind grade | magnetic declination | magnetic dip | total geomagnetic intensity | horizontal intensity |
For 35-dimensional vectors, the purpose of principal component analysis is to linearly combine raw data vectors:

\[
\begin{align*}
I_1 &= v_{11}x_1 + v_{12}x_2 + \ldots + v_{1H}x_H \\
I_2 &= v_{21}x_1 + v_{22}x_2 + \ldots + v_{2H}x_H \\
&\vdots \\
I_H &= v_{H1}x_1 + v_{H2}x_2 + \ldots + v_{HH}x_H
\end{align*}
\]

(4)

Satisfy the constraint that the quadratic sum of coefficient \(v\) is equal to 1:

\[
\sum_{i=1}^{H} v_i^2 = 1
\]

(5)

There is no correlation between the principal components \(I\):

\[\text{cov}(I_i, I_j) = 0, \quad i \neq j, \quad i, j = 1, 2, \ldots, H\]

(6)

Determine the importance of principal components using variance decreasing. During dimension reduction, the principal component dimension \(d'\) under the new output coordinates is determined according to user judgment. Usually, set a corresponding threshold, meaning that the selected dimension can maintain the variance contribution of the information of the original sample, such as confidence = 99%, then select \(d\) based on the following formula:

\[
\frac{\sum_{i=1}^{d'} v_i}{\sum_{i=1}^{d} v_i} \geq \varepsilon
\]

(7)

The goal of PCA is to maintain the information of the influencing factors affecting the original dataset of DC bias as much as possible when loss is inevitable, simplify the high-dimensional data [7], and to reduce the variance of the cost function by selecting the few variables that contribute the most to the information of data sample. The relationship between the maintained principal component and the original data sample is:

1) To achieve dimensionality reduction, the retained principal component needs to be smaller than the dimension of the original data set.
2) Each principal component is a linear combination of the original variables without correlation.
3) Keep as much information of the original data samples as possible while minimizing losses.

4. Conclusion
Table 3 shows the cumulative variance percentage of each component after conducting dimensionality reduction of the multidimensional big data set in Table 2. The confidence of the first 12 principal components using the PCA algorithm exceeds 99%, meaning it includes more than 99% of the original data information, so 12 principal components can be used to conduct dimensionality reduction. In addition, the confidence of the first 8 principal components exceeds 95%. When the data set is large, dimensionality reduction can greatly reduce storage requirements and calculation amount. Table 3 shows the time consumed and time for clustering of the two dimensionality reductions (12 principal...
component dimensionality reduction and 8 principal component dimensionality reduction) in this example.

![Figure 3. Comparison of dimensionality reduction effect](image)

Table 3. Performance comparison

| method                  | variance (%) | dimension reduction time (s) | clustering time (min) | correlation analysis time (min) | total time (min) |
|-------------------------|--------------|------------------------------|-----------------------|--------------------------------|-----------------|
| no dimension reduction  | 0.0          | 0.0                          | 1136.5                | >10,000                        | >11136.5        |
| PCA-12                  | 99.2         | 6237.5                       | 201.5                 | 87.6                           | 6526.6          |
| PCA-8                   | 95.9         | 4510.3                       | 143.6                 | 180.9                          | 4834.8          |

Table 4. Results of principal component analysis (in order of confidence importance)

| number | 1          | 2                          | 3                        | 4                        |
|--------|------------|-----------------------------|--------------------------|--------------------------|
| data   | transformer neutral point current | transformer acceleration | train load current | train real-time position |
| number | 5          | 6                           | 7                        | 8                        |
| data   | transformer neutral point isolated dc voltage | DC grounding pole geographic information | substation geographical location | earth resistivity |
| number | 9          | 10                          | 11                       | 12                       |
| data   | substation grounding resistance | transformer vibration displacement | transformer noise | excitation current component |

In conclusion, 12 principal components (see Table 3) is sufficient to represent more than 99% of the original big data, while 8 principal components (see Table 4) can represent more than 95%. PCA performs well in dimensionality reduction in clustering and correlation analysis in terms of time,
which helps to carry out the analysis of system correlation of the original big data. In this paper, the big data method is applied to provide the “one-button sequence control” operation and solution for the prevention and control of transformer DC magnetic bias.

**Acknowledgments**

This paper thanks for the financial support of the 2018 science and technology guide project of State Grid Corporation of China《Research and application of the key technology of one-button sequence control》(Project Number: 5216A018002).

**References**

[1] Jinliang He. Grounding Technology of Power System [M], Beijing: Science Press, 2006: 124-210.
[2] Guangrun Xie. Grounding Technology of Power System [M], Beijing: Hydraulic and Electric Power Press, 1991:77-105.
[3] Xinxue Liu, Wanxing Sheng, Dongxia Zhang Et Al. Research on Big Data Application Demand And Scenario Analysis of Smart Distribution Network [J]. Chinese Journal of Electrical Engineering, 2015,35 (02) : 287-293.
[4] Teng Zhao, Yan Zhang, Dongxia Zhang. Big Data Application Technology and Prospect Analysis Of Smart Distribution Network [J]. Power Grid Technology, 2014,38 (12) : 3305-3312.
[5] Ma Jinxin, Dawalibi F. P. Analysis Of Grounding Systems in Soils With Cylindrical Soil Volumes [J], IEEE Trans on Power Delivery, 2000, 15 (3) : 913-918.
[6] Fang Liu, Jianpeng Qi, Yanwei Yu, Lei Cao Et Al. Density-Based Top-N Local Anomaly Detection Algorithm [J]. Acta Automatica Sinica, 2019,2:1-16.
[7] Di Xiao. Joint Application of Wavelet Transform and PCA Analysis in Vibration Signal Processing [M]. Kunming University of Science and Technology, 2016.