Research on Monitoring and Diagnosis Technology of Data Anomaly in Distribution Network

Wensi Huang*, Xin Lu, Jiandi Hu and Qiangbin Ye
State Grid Info-Telecom Great Power Science and Technology Co., LTD, Fuzhou Fujian 350003, China

*Corresponding author email: wensihuang@sgcc.com.cn

Abstract. Aiming at the problems of a large amount of abnormal data in the distribution network and the poor adaptability of traditional distribution network data, the paper proposes an intelligent distribution network status and abnormal data monitoring method, which has undergone data pre-processing and data fusion, data analysis and visualization, and state identification and processing, a total of 4 links, turning multiple electrical feature quantities into a single comprehensive feature quantity, monitoring the operation status of the distribution network, and according to the relationship between each node and the size of the local anomaly factor to achieve intelligence judgment and location of distribution network fault areas. The thesis realizes the detection and location of faults according to the size of each node's LOF value and the node's association relationship. After RTDS semi-physical closed-loop test, the accuracy and reliability of fault determination and location are high, which has certain reference value.

Keywords: Distribution network abnormal data, abnormal diagnosis, data fusion, online monitoring.

1. Introduction
In recent years, the country has vigorously developed smart grids, and the informatization of transmission and transformation systems has reached a very high level. But in the distribution network, due to the complexity of its use environment and structure, the level of intelligent informatization is low. With the rapid economic growth in China, rapid expansion and development of cities, insufficient funds for planning and construction of distribution networks, relatively weak distribution grid structure, insufficient distribution transformer capacity, and low reliability of power supply, the problems of distribution power supply are becoming increasingly prominent, and distribution automation the level needs to be further improved.

Power big data is a subset of big data in the power industry and one of the necessary conditions for the construction of the power Internet of Things. The massive data resources generated by the distribution network operation truly reflect the health of the distribution network operation, and effectively support real-time monitoring of the distribution network operation status, power fault location, power outage range location, and real-time line loss statistics, etc., continuously improving the quality of power supply services. To further meet the user's power needs, to achieve safe, reliable, and economical power supply. By carrying out distribution network data anomaly monitoring and diagnosis,
and constructing distribution network load characteristic portrait research, it can not only assist dispatching operations, accurately prepare orderly power consumption plans, but also optimize project projects and promote distribution network planning and construction. Realized the integrated development of equipment, marketing, and dispatching data, further improved the distribution network lean management level, improved power supply reliability management level, improved power grid quality service level, increased distribution network data asset value, and promoted the efficient operation of power grid data assets.

Therefore, based on big data analysis, this paper designs an intelligent distribution network state monitoring and fault processing method. This method transforms the fusion of multiple electrical feature quantities into a single comprehensive feature quantity and monitors the operation status of the distribution network. It can determine and locate the fault area according to the association status of each node and the size of the local anomaly factor. Its accuracy and reliability are high, and it can provide technical support for smart distribution network status monitoring and corresponding fault processing similar to big data [1].

2. Distribution network data abnormal monitoring and diagnosis

2.1. Abnormal file data
Abnormal analysis and inspection of station, bus, transmission line, distribution line, station area, high voltage user and low voltage user files. The following uses the file data of distribution lines and station areas as an example to illustrate the abnormal analysis of file data. The abnormal analysis and inspection items of distribution line file data include: starting station, starting switch, voltage level, status, and asset nature check, line wireless change relationship and line and user relationship (non-user assets, in operation, public special The change count is 0) Abnormal analysis, the line measurement point cannot find the corresponding switch or starting station, the line has multiple measurement points, the line has no evaluation measurement point, the line checklist is missing (non-user assets, the line is in operation), the checkpoint The comprehensive magnification of the table is null or 1 and other abnormal analysis and inspection.

2.2. Abnormal line loss
For the line loss and line loss rate data generated after line loss calculation, determine the bus balance rate, transmission line loss rate, and distribution line based on the line loss assessment index, line loss business knowledge, and the set line loss rate available threshold, etc. And abnormal data such as line loss rate in Taiwan area are divided into non-standard line loss rate, high loss, negative loss, super large loss, etc. Especially for super large loss, large negative loss, etc., it is directly judged as data abnormality, and then according to the diagnostic analysis function, further locate the problem data.

2.3. The bottom of the table and the power data are abnormal
Abnormal data analysis of the electricity generated by the meter reading data and the bottom data calculation, including the bottom of the meter collected by the electrical energy collection system and the calculated power, involving all the gateway models of stations, main transformers, transmission lines and distribution lines, including The bottom of the electricity information collection system and the amount of electricity generated by the calculation involve distribution lines (parts), stations, high-voltage users, and low-voltage users, including the power supply of the distribution lines and their bottoms, and the power supply of the districts and their bottoms, High-voltage users and low-voltage users sell electricity and its bottom, etc. Outlier detection is a big data mining technology. Generally, data mining is divided into four types, namely the discovery of correlation rules, category determination, category description, and outlier detection. The first three categories are all based on "the data are all the same and satisfy Based on the assumption of "the same model", however, the data often contains anomalies. Data mining for such anomalies is outlier detection. Outlier detection is divided into five detection methods based on statistical principles, density, distance, clustering, and artificial neural
network models. At present, it has been applied to credit card fraud, network intrusion detection, financial application and transaction fraud, etc. In recent years, it has become a research hotspot in the field of data mining research.

Outlier detection can also be used for the analysis of abnormal line loss data during the same period. It is more suitable for the abnormal detection of power and bottom data. The combination of file business attributes and power business sets the power and bottom threshold range to find the power and bottom that deviate from the normal range. Data. Currently, two detection methods based on statistical principles and based on distance solitary point detection are mainly used. Among them, detection methods based on statistical principles are divided into two types: distribution-based detection algorithms and depth-based detection algorithms. The former constructs a standard distribution to fit the data set, and then determines the isolation based on the probability distribution; the latter calculates geometry. As a basis, the outer layer objects are determined as isolated points by calculating the KD convex hulls of different layers; isolated point detection based on distance defines isolated points as the distance between the data set and most data objects is greater than a given The object point of the threshold is usually described as DB. If and only if the distance between at least pct. data points in the data set S and the P point is greater than dmin, the data object P point becomes an isolated point [2].

2.4. Abnormal data diagnosis
Through the abnormal analysis of the above files, line loss, meter bottom and power data, most of the abnormal data can be accurately located, and data correction and management can be performed. Some data, such as high line loss in stations, distribution lines, etc., cannot be clearly located. (Problem) Data, further diagnosis and analysis of abnormal data is needed, and the abnormal data is gradually located. For those that cannot be accurately located, list all data options with high probability problems to assist in the search for root cause abnormal data. The following uses the diagnosis of distribution line high-loss (including negative high-loss) as an example to illustrate the diagnosis and analysis process. The thesis is judged by the Lag bus criterion.

The Lag bus criterion is applicable to the case where the amount of electricity data at a single measurement point is not large, and the confidence probability is 99.50% . The method of discrimination is as follows. The electricity data \( x_i \) at the measurement point follows a normal distribution. The critical statistical coefficient \( G_0 \) at the time of the data quantity \( n \) is found in the following Table 1. The suspicious electricity data screened by the quartile method is located in the lower quartile and below, the upper four the quantile and above are recorded as \( k_{\text{min}}, k_{\text{max}} \). 1. First select the maximum and minimum values of suspicious data as \( x_{k_{\text{min}}}, x_{k_{\text{max}}} \) and 2. respectively; 2. Calculate the average \( \bar{x} \) and standard deviation \( \sigma \) of all \( n \) data at this time;

\[
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i
\]

\[
\sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}
\]

\[
G_{\text{min}} = \frac{(\bar{x} - x_{k_{\text{min}}})}{\sigma}
\]

\[
G_{\text{max}} = \frac{(x_{k_{\text{max}}} - \bar{x})}{\sigma}
\]

If \( G_{\text{min}} \geq G_{\text{max}} \) and \( G_{\text{min}} \geq G_0 \), then \( x_{k_{\text{min}}} \) is abnormal data; if \( G_{\text{max}} \geq G_{\text{min}} \) and \( G_{\text{max}} \geq G_0 \), then \( x_{k_{\text{max}}} \) is abnormal data; if \( G_{\text{min}} \in G_0 \) and \( G_{\text{max}} \in G_0 \), then \( x_{k_{\text{min}}}, x_{k_{\text{max}}} \) are normal data.

Repeat the above steps with the remaining data after removing the abnormal data, in order a) find the critical statistical coefficient at this time; b) select the maximum and minimum values of the remaining suspicious data; c) calculate the average and standard deviation of the current data; D) calculation and; e) comparative judgment. Until all suspicious data have been identified. Through the above method,
abnormal power can be screened out with high accuracy. As shown in Table 1, Lag bus critical value test table. As shown in Table 2 is the monthly electricity data of the power supply unit.

**Table 1. Lag bus critical test table**

| n   | 3   | 4   | 5   | 6   | 7   | 8   | 8   | 10  |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Go  | 1.155 | 1.496 | 1.764 | 1.973 | 2.139 | 2.274 | 2.387 | 2.482 |
| n   | 11  | 12  | 13  | 14  | 15  | 16  | 17  | 18  | 19  | 20  |
| Go  | 2.564 | 2.636 | 2.699 | 2.755 | 2.806 | 2.852 | 2.894 | 2.932 | 2.968 | 3.001 |

**Table 2. Monthly electricity data of power supply unit**

| Gate number | 06011504020317 | 06011504020323 |
|-------------|----------------|----------------|
| 06011504020317 | 5087500, 4675000, 4785000, 4111250, 4303750, 2213750, 5032500, 5802500, 7273750, 5843750 |
| 06011504020323 | 9308750, 9446250, 9721250, 7851250, 8992500, 8676250, 10106250, 10560000, 12856250, 12402500 |

When n=10. Analysis of the gate 06011504020317, power ranking: 2213750, 4111250, 4303750, 4675000, 4785000, 5032500, 5802500, 5843750, 7273750; analysis of the gate 06011504020323, power ranking: 7851250, 8676250, 8992500, 9308750, 9446250, 9721250, 10106250, 10560000, 12856250, 12402500;

**Figure 1.** Power sequencing when n=10

Through the quartile and Lag bus analysis results are shown in Table 3:

**Table 3. Interquartile and Lag bus analysis results**

| Gate number | 06011504020317 | 06011504020323 |
|-------------|----------------|----------------|
| Interquartile suspicious data | 2213750, 4111250, 4303750, 5802500, 5843750, 7273750 | 7851250, 8676250, 8992500, 12402500, 10560000, 12856250 |
| Abnormal data of the Lag bus criterion | no | no |

2.5. *Diagnosis method of linear change relationship*

The positive high loss and negative high loss of the line are often caused by the line change relationship, that is, the user's connection line relationship error. The following example illustrates the algorithm to find the user line change relationship error according to the impact of each user's power on the loss rate of the connection line. For example, the line loss rate, change situation of a certain line for 6 months and the power consumption of the users are given in Table 4 below.
Table 4. Line loss and user power

| Line Loss Rate | Increase or decrease from last month | User Power |
|---------------|-------------------------------------|------------|
|               | 1 | 2      | 3      | 4      | 5      | 6      |
| 11.24%        | - | -      | -      | -      | -      | -      |
| -74.1%        | -0.8534 | 323868 | 37409  | 13800  | 54732  | 163384 |
| -78.85%       | -0.0475 | 499996 | 57645  | 13469  | 86713  | 249070 |
| -78.16%       | 0.0069  | 496734 | 56796  | 14090  | -      | 247166 |
| 11.18%        | 0.8934  | 141840 | 0      | 9931   | -      | 0      |
| -49.87%       | -0.6105 | 305642 | 32715  | 11603  | -      | -      |

Substitute each user's power into the following formula, find the correlation coefficient $\rho$ with X and Y, and use $|\rho| > 0.3$ as the relevant judgment condition.

$$\rho = \frac{\sum_{i=1}^{5} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{5} (X_i - \bar{X})^2 \sum_{i=1}^{5} (Y_i - \bar{Y})^2}}$$

According to the condition of $|\rho| > 0.3$, it can be judged that the users with the wrong linear transformation relationship are user 1, user 2, user 3, and user 4.

3. Design of distribution network abnormal monitoring and diagnosis system

3.1. Overall framework

The overall architecture of the intelligent distribution network operation monitoring platform is shown in Figure 2. It consists of the provincial distribution network monitoring integrated platform and the provincial-level deployed enterprise bus through the distribution production management, marketing management, electricity collection system, and GIS. The system includes 12 distribution-related business systems, as well as the distribution automation system in the municipal area, the information interaction bus through the information security I/II area, and the distribution network production emergency repair command platform. The distribution automation system in the county area is composed of the information exchange bus and the enterprise. The bus runs through, integrates all distribution-related business systems, and realizes the optimal operation and self-healing control of distribution network, status monitoring and intelligent dispatch, risk early warning and evaluation decision, information sharing and open interaction, and production emergency repair command. The intelligent distribution network operation monitoring platform supports the smooth flow of power flow, information flow, and business flow. It integrates operation, monitoring, control, and management. The distribution automation system is the core, and the distribution network dispatching and distribution production management command are applications. The main body integrates related business systems and information systems based on the information interaction bus to achieve a high degree of sharing of distribution network data to meet the entire process and full business needs of the intelligent distribution network [3].
3.2. Data processing
With the implementation of distribution automation, electricity collection systems, energy efficiency management systems, GIS systems, and PMS systems in cities across the province, the data of each system has increased dramatically, especially the real-time monitoring data of the distribution automation system has increased in seconds. The amount reaches PB level. Using big data processing technology to meet the function and performance requirements of the intelligent distribution network operation monitoring platform.

3.2.1. Distributed storage, computing and parallel processing. The intelligent power distribution network operation monitoring platform adopts distributed service design to realize data storage and data processing, and dispatches each distributed service through the total service deployed in the provincial area. The distribution network data of various cities and counties is still stored on the local server. The distributed service deployed in the prefecture and county parses the total service call command and calculates the data. After the data processing is completed, the processing results are returned to the total service. Summarize, calculate and store the processing results of each city and county. The interface server of each city and county calculates the data and summarizes the calculation results step by step to realize the distributed and balanced calculation to the server of each city and county node, reducing communication traffic and calculation burden. Through parallel computing, improve the reliability, availability and access efficiency of the intelligent distribution network operation monitoring platform [4].

The interface service of the total service of the intelligent distribution network operation monitoring platform adopts the load balancing method to realize that the same task runs on different interface
services at different stages. For multi-stage tasks, an independent algorithm and data model are used for each stage. During the task execution process, the unique sequence number of the task is used to ensure the execution order and execution results of each stage, thereby achieving parallel calculation and load balancing of the two interface services. Schematic diagram of data distribution storage and parallel computing is shown in Figure 3.

![Figure 3. Schematic diagram of distributed data storage and parallel computing](image)

3.2.2. Data display based on data caching technology. Through the use of distributed storage and computing, the amount of data reading and storage deployed on the intelligent power distribution network operation monitoring platform is reduced. In order to improve the access performance, the statistical summary data and high real-time data are divided into tables and partitions, and the cache tables and data statistics tables are established and stored in the provincial database to realize the efficient storage, analysis and display of big data.

(1) Use common buffer cache for all users’ common and common reports, graphics (with unique identification) and other data. Users can directly access the cached data through unique identification to improve query and calculation speed. (2) When users query for different parameters, provide special buffer cache data and generate unique serial numbers through request parameters. When other users access the same report or graphics, first generate serial numbers through parameters and then go to the special buffer. Find the data corresponding to the serial number. If the serial number matches, the data is returned directly; if it does not match, the data is obtained from the database through the data set and parameters defined in the report. Through the application of caching technology, the data reading speed is improved, and the rapid response of data display is realized [5].

3.2.3. Data processing algorithm. Data pre-processing is used for preliminary screening and pre-processing of the raw data uploaded by each sensing device, reducing the amount of irrelevant data and generating the required initial feature quantity matrix. This part mainly includes selecting feature quantity, constructing correlation matrix, and dealing with regional differences. In the process of selecting feature quantities, the electrical feature quantities selected in this paper are current and power. Among them, it covers the three-phase current, negative sequence current, and zero sequence current and the corresponding active power. The process of constructing the correlation matrix is as follows:
first, each terminal node $E_j$ in the distribution network is numbered, and the area $Z_i$ between the nodes is also numbered, and finally the matrix is constructed according to the rules shown in Table 4.

Table 5. The rules for constructing the network correlation matrix

| Associated value | connection relation                  |
|------------------|--------------------------------------|
| 0                | Node is not in area                  |
| 1                | The node is in the area, and the current/power of the node is in the area |
| -1               | The node is in the area, and the current/power of the node is outside the area |

The purpose of dealing with regional differences is mainly to amplify the difference between faults and normal nodes to facilitate fault identification. Data analysis and visualization present high-dimensional data in low-dimensional space and reduce the amount of data. This process will ensure that the relationship between the objects is basically unchanged, and then perform outlier detection to support online identification. Among them, focusing on the dimensionality reduction process, described as follows:

1) According to the Euclidean distance, the dissimilarity matrix $D$ between the objects in the high-dimensional matrix $W$ can be calculated:

$$D = d_{ij} = \left[ \sum_{k=1}^{n} (x_{ik} - x_{jk})^2 \right]^{0.5}$$ (4)

2) With the help of the above-mentioned dissimilarity matrix $D$, a centralized inner product matrix $B$ is further obtained:

$$B = b_{ij} = -\frac{1}{2} d_{ij}^2 - \frac{1}{n} \sum_{j=1}^{n} a_{ij} - \frac{1}{n} \sum_{i=1}^{n} a_{ji} + \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}$$ (5)

3) Solve the first 2 feature roots and the corresponding orthogonalized feature vectors in matrix $B$ to meet the following requirements:

$$\begin{align*}
\lambda_1 \geq \lambda_2 \geq 0 \\
\langle x'_i, x'_j \rangle = \lambda_i, 1 \leq i \leq 2
\end{align*}$$ (6)

4) Finally, the two-dimensional space representation $M = [x_1, x_2]$ after dimensionality reduction is obtained.

The outlier detection process uses a local anomaly factor related detection method (based on density). The LOF value of the detection object is positively correlated with the outlier degree of the outlier. If the value is about 1, the outlier does not exist.

3.3. Fault judgment and treatment

Due to the operation of the power distribution network, protection misjudgement or even misoperation caused by sensor failure may occur. Therefore, in this paper, a single type of triggering fault conditions (sensors, power) is set and processed, and the required criteria are as follows: the fault start criterion is met, but the LOF at the generalized node does not meet the setting threshold. At this time, it is determined that the sensor is faulty, and the faulty node is the node where the maximum LOF value is located. The data processing centre sends an alarm message to each measurement and control terminal to ensure that the terminal is reliable and does not operate; if the criterion is met and the LOF value also reaches or exceeds the setting Threshold, at this time, it is determined that a power system failure has occurred, the public area where the physical node is located is located, and the data processing centre sends an action command to the node where the failure occurred, and further performs the isolation operation [6].
4. Case analysis
In order to verify the feasibility of the method proposed in this paper, a 10 kV intelligent distribution network with double DG is taken as the research object. Build relevant models in RTDS with actual parameters. Figure 4 shows a visual analysis diagram of the single-phase ground fault in the Z11 feeder section, the corresponding multidimensional scaling and the corresponding LOF value. Among them, 13 and 14 physical nodes and generalized nodes (17) all become outliers, and the corresponding LOF value reaches about 96.

At this time, it is determined that a fault related to the power system has occurred, and the fault is located in the Z11 area where the nodes 13 and 14 are located. The data processing centre sends tripping instructions to the faulty node terminal in time to isolate the Z11 area. In addition, this paper has also tested the failure of node 4 sensor failure and the failure of the bus node Z2 area (two-phase grounding). The monitoring and positioning effect is good, and the failure can be effectively and timely feedback and localization [7].

![Figure 4. Z11 feeder section fault monitoring matrix data chart](image)

5. Conclusion
Taking the emergence and application of big data in the smart grid as a background, this paper designs and proposes a set of smart distribution network state monitoring and fault processing methods based on big data technology. In condition monitoring, multiple electrical feature quantities are fused into a single comprehensive feature quantity to ensure identification accuracy. The LOF value is used instead of the determination of the electrical characteristic quantity to avoid the cumbersome setting calculation.

Acknowledgments
This paper is funded by the Technology project of State Grid Corporation of China, "Research and Application of Power grid planning Data Automatic Integration and Planning Execution Intelligent Control Technology Based on NEW ICT"

References
[1] Mumford, R. A. Macarthur, R., & Bonham, N. The role and challenges of new diagnostic technology in plant biosecurity. Food Security, 8 (1) (2016) 103-109.
[2] Brown, H. L. Does the use of diagnostic technology reduce fetal mortality Health Services Research, 54 (2) (2019) 321-323.
[3] Andryukov, B. G., Karpenko, A. A., Matosova, E. V., & Lyapun, I. N. Raman spectroscopy as a modern diagnostic technology for study and indication of infectious agents (review). Sovremennye Tehnologii v Medicine, 11 (4) (2019) 161.

[4] Bryony, H. Caroline, M., Aoife, C., & O’Kennedy Richard. Developments in point-of-care diagnostic technology for cancer detection. Diagnostics, 8 (2) (2018) 39-48.

[5] None. Sw icers more common for diagnostic technology than drugs. PharmacoEconomics & Outcomes News, 793 (1) (2017) 30-30.

[6] None. Nice set to recommend heart disease diagnostic technology. Pharmacoconomics & Outcomes News, 761 (1) (2016) 4.

[7] Liu, J., Wang, J., Ruan, W., Lin, C., & Chen, D. Diagnostic and gradation model of osteoporosis based on improved deep u-net network. Journal of Medical Systems, 44 (1) (2020) 1-7.