Research on the Quality Evaluation Method of Transmission and Transformation Inspection Data

Qin Jiafeng\textsuperscript{1,*}, Gu Chao\textsuperscript{1}, Lv Juntao\textsuperscript{2}, Bai Demeng\textsuperscript{1}, Zheng Wenjie\textsuperscript{1} and Liu Meng\textsuperscript{1}

\textsuperscript{1}State Grid Shandong Electric Power Research Institute, China; \textsuperscript{2}State Grid Shandong Electric Power Company, China; *qinjiafeng0118@163.com

Abstract. In the era of big data, transportation and inspection data of power transmission and transformation equipment have the characteristics of diversity and richness. Massive data provide data support for state assessment of power transmission and transformation equipment, but at the same time, it also puts forward higher requirements for traditional data management and data quality mode. In this paper, the application of density distribution function estimation and information entropy statistical method in data quality evaluation of transmission and transformation equipment is studied. From the perspective of data integrity, validity and consistency, a five-dimensional data quality evaluation model is established to realize the information checking and quality evaluation of transmission, transformation and transportation inspection based on data characteristics and business rules. On the one hand, it can be compared and evaluated before and after data cleaning, on the other hand, it can also be convenient to find problems in the process of data collection.

1. Introduction

With the rapid development of power grid size, grid capacity continues to expand, power transmission and transformation equipment shipment inspection data exponentially, large number of devices and the contradiction between the limited manpower will be more and more big, the traditional way is given priority to with artificial information work already cannot adapt to the development of intelligent power grid equipment shipment inspection requirements, needs to use perceptive technology to improve to ensure data accuracy, integrity, and work efficiency. How to establish the data quality evaluation model and improve the data quality of transmission and transformation equipment is of great significance to the equipment status evaluation\textsuperscript{[1-3]}.

To evaluate data quality, this paper considers the data quality in data field meaning including the rules and data value range of the data itself, as well as the contact between different fields, including field, the decision of the relationship between dependencies, etc., on the one hand, can be cleaned for data comparison before and after evaluation, on the other hand also can easily find the problems existing in the data collection process\textsuperscript{[4-5]}. After data pre-processing, a five-dimensional data quality evaluation model can be established from the perspectives of data validity, redundancy, information content, integrity and accuracy, so as to realize the information checking and quality evaluation of transmission, transformation and transportation inspection based on data characteristics and business rules.
In this paper, a method of data quality inspection for transmission and transformation is proposed, mainly from the following five aspects: validity, redundancy, information quantity, completeness and accuracy, and a data quality evaluation device is developed. Firstly, an objective index is given for each aspect to measure the performance of data in this aspect, and then a scoring function is used to map this index to the space of \([0,5]\) to obtain a score of \(0~5\). The following is a detailed description of how to evaluate the performance of the data in these five aspects[6-10].

2. Five dimensions of data quality assessment
The five dimensions of data quality assessment are validity, redundancy, information quantity, completeness and accuracy.

2.1. Validity
There is often some invalid data in the data, for the on-line monitoring data of the substation equipment, if all the gas data in a record is zero, then the record is invalid. Or for a certain type of gas, when its value is less than 0 or appears as -9999, it can also be considered as invalid data. The existence of these invalid data will interfere with data analysis, so the evaluation of data validity is an important quality evaluation index.

Firstly, the invalid ratio is defined as:

\[
\alpha_{\text{inv}} = \frac{\text{Number of invalid points}}{\text{Number of effective points}}
\]  

(1)

The invalid ratio should be inversely proportional to the score of validity. Since the total score of validity is 5, the validity score should meet:

\[
\text{Score(effectiveness)} = \frac{1}{\theta \alpha_{\text{inv}} + 0.2}
\]

(2)

Where \(\theta\) is a parameter related to the slope of the curve. The larger the \(\theta\), the faster the score declined as the void ratio increased. Therefore, the selection of \(\theta\) is subjective to some extent. For a certain type of data, several sets of data points \((\alpha_{\text{inv1}}, \text{score}_1), (\alpha_{\text{inv2}}, \text{score}_2), \ldots, (\alpha_{\text{invk}}, \text{score}_k)\) can be given by experts' scoring, and the most suitable \(\theta\) can be found by the method of least squares fitting. Generally, for transformer online monitoring data, take \(\theta=1\).

2.2. Redundancy
Redundancy refers to some duplicate records contained in the data. As shown in finger 1, the gas content of the on-line monitoring data of a certain device is exactly the same on several consecutive records, and the time of each record is exactly the same. These duplicate records waste a lot of storage space and interfere with subsequent data analysis, so redundancy of data is also an important indicator of quality assessment.

| H2 | CH4 | C2H4 | C3H2 | CO | CO2 | O2 | N2 | TOTALHYDROCARBON | LSHANDLED | C2H6 | RESAVE_TIME      |
|----|-----|------|------|----|-----|----|----|------------------|------------|------|------------------|
| 17.990 | 0.000 | 0.010 | 0.000 | 2.100 | 29.030 | 0.000 | 0.000 | 0.020             | 0.010      | 2019/1/1 1:00:09 |
| 17.990 | 0.000 | 0.010 | 0.000 | 2.100 | 29.030 | 0.000 | 0.000 | 0.020             | 0.010      | 2019/1/1 1:00:09 |
| 17.990 | 0.000 | 0.010 | 0.000 | 2.100 | 29.030 | 0.000 | 0.000 | 0.020             | 0.010      | 2019/1/1 1:00:09 |
| 17.990 | 0.000 | 0.010 | 0.000 | 2.100 | 29.030 | 0.000 | 0.000 | 0.020             | 0.010      | 2019/1/1 1:00:09 |
| 17.990 | 0.000 | 0.010 | 0.000 | 2.100 | 29.030 | 0.000 | 0.000 | 0.020             | 0.010      | 2019/1/1 1:00:09 |

**Finger 1. Duplicate records**

The repetition rate is defined as:
\[ \alpha_{ry} = \frac{\text{Number of duplicate records}}{\text{Number of valid records}} \quad (3) \]

Similarly, the score for redundancy should be inversely proportional to the repetition rate. The higher the repetition rate, the lower the data quality and the lower the score of redundancy. Conversely, the lower the repetition rate, the higher the data quality and the higher the score of redundancy. Therefore, when the total score is 5, the relationship between repetition rate and redundancy score is also applicable to the model:

\[ \text{Score(Redundancy)} = \frac{1}{\theta \alpha_{ry} + 0.2} \quad (4) \]

\( \theta \) can also be obtained by using the least squares fitting method of expert rating.

2.3. Information quantity
Data is a form of information, and each data point contains certain information. Taking C2H6 data as an example, the content of C2H6 is 0 most of the time, and only in a few days, the content of C2H6 is greater than 0. Intuitively, although those points with a value of 0 also express the state of C2H6, they contain less information, while those points with a value of non-0 contain more information. More extreme, if the content of C2H6 is all 0, then the information contained in the data set is more limited.

Firstly, based on Shannon’s theory, a quantization formula of information quantity can be given:

\[ I(x) = -\log_2 p(x) \quad (5) \]

Where \( p(x) \) is the probability of point \( x \). For invalid data points, we simply set \( I(x) \) to 0, so we only need to estimate \( I(x) \) for each valid data point \( x \).

Normalize the data first, that is

\[ x' = \frac{x}{\mu_{valid}} \quad (6) \]

Where \( \mu_{valid} \) is the mean of all valid data, so we can still call \( x' \) as \( x \). Suppose the number of days corresponding to point \( x_i \) is \( t_i \), and \( x_j \) satisfies the normal distribution \( \mathcal{N}(\mu, \sigma_i^2) \), where

\[ \mu_i = \frac{(t_{i+1} - t_i) * x_{i-1} + (t_i - t_{i-1}) * x_{i+1}}{t_{i+1} - t_{i-1}} \quad (7) \]

\[ \sigma_i^2 = \begin{cases} \sum_{x_{j,\text{neighbor}(x_i)}} (x_j - \mu_{\text{nei}})^2 + m_i \times \max_{x_{j,\text{nei}}}(x_j - \mu_{\text{nei}})^2, & \text{if } \text{neighbor}(x_i) \geq 4 \\ \sum_{x_{j,\text{validdata}}}(x_j - \mu_{\text{nei}})^2 + m_i \times \max_{x_{j,\text{validdata}}}(x_j - \mu_{\text{nei}})^2, & \text{if } \text{neighbor}(x_i) < 4 \end{cases} \quad (8) \]

\( (x_{i-1}, t_i) \) and \( (x_{i+1}, t_{i+1}) \) are the two effective observation points before and after the closest to \( (x_i, t_i) \), respectively. Here, in addition to the duplicate records and all 0 records are regarded as invalid observations, the points whose distance between the observed value and the mean value is more than 3 times the sample variance are considered as invalid observations.
represents the set of effective observation points in $[t_i - 3, t_i + 3]$ period, $\mu_{nei}$ represents the mean of all points in $\text{neighbor}(x_i)$, $m_i$ represents the number of days without effective observation points in $[t_i - 3, t_i + 3]$ period, namely $m_i = 7 - |\text{neighbor}(x_i)|$, and $\text{validdata}$ represents the set of all effective observation points. Then $p(x_i)$ can be estimated as follows:

$$p(x_i) = \begin{cases} \int_{x_i-0.05}^{x_i+0.05} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(x_i - \mu_i)^2}{2\sigma_i^2}\right)dx, & \sigma_i \neq 0 \\ 1, & \sigma_i = 0 \end{cases}$$ (9)

So you get $I(x_i) = -\log_2 p(x_i)$ as the estimate of the amount of information at point $x_i$. The amount of information $I(x)$ obtained in this way has the following properties:

(i) $I(x) \geq 0$, the amount of information will not be negative.

(ii) $I(x_i) = 0$ if and only if the number of effective observation points in $[t_i - 3, t_i + 3]$ period is greater than or equal to 4, and the values of these effective observation points are all equal. In other words, if there are records on more than 4 days in the last week of $x_i$, and these records are equal to $x_i$, the information of $x_i$ is considered to be 0.

(iii) Generally, if the degree of $x_i$ is similar to that of $x_j$, and there is a missing point in the last week of $x_i$, and there is no missing point in the last week of $x_j$, then $I(x_i) > I(x_j)$. Intuitively, since the value of missing points can be estimated by $x_i$, the information of partial missing points is contained in $x_i$. Therefore, in the case of similar fluctuation degree, the information at $x_i$ should be larger than that at $x_j$.

(iv) $I(x)$ has a great relationship with the fluctuation degree of the data set. Generally speaking, the more volatile the data set is, the larger $I(x)$ is.

The average information on data set $V$ is defined as:

$$\text{avg}_V I(V) = \frac{\sum_{x \in V} I(x)}{|V|}$$ (10)

The score of the data on the quantity of information can be given based on the average quantity of information. Since the information growth effect is similar to the population growth effect, the logistic growth model can be used to describe the relationship between average information amount and score:

$$\text{Score}(\text{Information quantity}) = \frac{5c \cdot e^{r(\text{avg} - 1)}}{5 - c + c \cdot e^{r(\text{avg} - 1)}}$$ (11)

Where $c$ and $r$ are parameters, the parameter selection method is the same as the previous two indicators.

2.4. Completeness

Integrity refers to the absence of data, which is an important indicator for data quality assessment. A simple indicator of integrity is:
\[ \text{Loss rate} = \frac{\text{Number of missing records}}{\text{Number of total records}} \tag{12} \]

Based on this consideration, the above definition of information can be used to estimate the amount of information that the missing point should have. The larger the value, the greater the impact of the missing on data quality. For the missing point at time \( t_i \), suppose it satisfies the normal distribution \( N(\mu_i, \sigma_i^2) \), where

\[
\mu_i = \frac{(t_{i+1} - t_i) x_{i+1} + (t_i - t_{i-1}) x_{i-1}}{t_{i+1} - t_{i-1}} \tag{13}
\]

\[
\sigma_i^2 = \begin{cases} 
\sum_{x_j \in \text{neighbor}(t_i)} (x_j - \mu_{\text{nei}})^2 + m_i \times \max_{x_j \in \text{neighbor}(x_i)} (x_j - \mu_{\text{nei}})^2 & \quad \text{if } |\text{neighbor}(t_i)| \geq 4 \\
\sum_{x_j \in \text{neighbor}(t_i)} (x_j - \mu_{\text{nei}})^2 + m_i \times \max_{x_j \in \text{validdata}} (x_j - \mu_{\text{nei}})^2 & \quad \text{if } |\text{neighbor}(t_i)| < 4 
\end{cases} \tag{14}
\]

\((x_{i-1}, t_{i-1})\) and \((x_{i+1}, t_{i+1})\) are the two effective observation points before and after the time when they are closest to \( t_i \). \( \text{neighbor}(t_i) \) represents the set of valid observation points within \([t_i - 3, t_i + 3]\) time period. If \( \text{neighbor}(t_i) \) is not null, \( \mu_{\text{nei}} \) represents the mean of all data; If \( \text{neighbor}(t_i) \) is a null set, \( \mu_{\text{nei}} \) represents the mean of all valid observation points. \( m_i \) is the number of days that there are no effective observation points in \([t_i - 3, t_i + 3]\) time period, i.e. \( m_i = 7 - |\text{neighbor}(x_i)| \), \( \text{validdata} \) represents the set of all effective observation points. Then \( p(t_i) \) can be estimated according to the following formula:

\[
p(t_i) = \begin{cases} 
\int_{\mu_i - 0.05}^{\mu_i + 0.05} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(x - \mu_i)^2}{2\sigma_i^2}\right)dx, & \sigma_i \neq 0 \\
1, & \sigma_i = 0
\end{cases} \tag{15}
\]

The relative missing information is defined as:

\[
\text{relI} = \begin{cases} 
\frac{\sum_{x_j \in \text{miss}} I(t_i)}{\sum_{x_j \in \text{validdata}} I(x_j)}, & \sum_{x_j \in \text{validdata}} I(x_j) \neq 0 \\
0, & \sum_{x_j \in \text{validdata}} I(x_j) = 0
\end{cases} \tag{16}
\]

Where \( \text{miss} \) is the index set of missing points and \( \text{validdata} \) is the set of all effective observation points. The score of data on the integrity index can be given according to the relative missing information. Similar to the information index, the following model can be used to give the relationship between the relative missing information and the integrity score:
Score(Completeness) = 5.5 - \frac{2.75 \cdot e^{r_{relI}}}{5.5 - 0.5 + 0.5 \cdot e^{r_{relI}}} \quad (17)

\(r\) is the parameter, and the selection method is the same as that in the validity model.

2.5. Accuracy
Unwashed data often contains errors that degrade the quality of the data, and accuracy measures how much and how severe the errors are.

For a given data set \(S\), set the number of valid data points as \(n\), use the weighted local anomaly factor algorithm to identify the error data, and record the corresponding \(\text{wlof}\) value of these error data, denoted as \(\{l_1, l_2, \ldots, l_k\}\). Define the degree of anomaly as:

\[
\text{abn}(S) = \frac{\sum_{l=1}^{k} \min(50, l_i)}{n} \quad (18)
\]

Generally, it can be considered that the degree of anomaly is linearly correlated with the accuracy score, and the higher the degree of anomaly, the lower the accuracy score. Therefore, the score of the data on the accuracy index can be obtained according to the following equation:

\[
\text{Score}(\text{Accuracy}) = \max(0, 5 - \text{abn}(S)) \quad (19)
\]

3. Experimental results
Based on the above five evaluation indexes, the data quality evaluation tool is constructed to evaluate the data quality of transmission and transformation equipment. At the same time, the data quality was evaluated respectively. Quality assessment results of different data of transmission and transformation are shown in table 1.

| Data source               | Validity | Redundancy | Information Quantity | Completeness | Accuracy |
|---------------------------|----------|------------|----------------------|--------------|----------|
| Online monitoring data    | 4.73     | 4.44       | 4.15                 | 4.73         | 4.21     |
| Meteorological data      | 4.83     | 4.89       | 4.86                 | 5.00         | 4.98     |
| Equipment operation data | 5.00     | 5.00       | 4.92                 | 5.00         | 4.97     |

4. Conclusion
This paper introduces the specific methods of data quality assessment, and gives the indicators of data performance in the five aspects of validity, redundancy, information quantity, completeness and accuracy. The proportion of valid data to total data is evaluated. Redundancy assesses the proportion of duplicate data. By introducing the concept of information quantity, there is a unified standard for evaluating the amount of information contained in the data. The completeness index evaluates the impact of missing data on the overall quality of data. Based on the concept of information quantity, the integrity index can more accurately describe the impact of continuous missing and scattered missing on the overall quality of data. The accuracy index is calculated by calculating the \(\text{wlof}\) value of the wrong data, and the quantity of the wrong data and the degree of anomaly are estimated. This paper shows the overall effect of data quality assessment based on the above five indicators, which can effectively guide data pre-processing.
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] Li Yuxuan, Gao Peng, Zhao Qinglei. Application of Wavelet Analysis in Power Load Data Processing [J]. Shandong electric power, 2018, 45 (5): 5-9.

[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.