Research on Automatic Preprocessing Method of Railway Catenary Geometric Inspection Data

Jianfeng Guo, Zaitian Ke, Jinzhao Liu, Wenxuan Zhang, Zhipeng Yang and Liming Ning

China Academy of Railway Sciences Corporation Limited, Room JC321, Daliushu Road, 2th, Beijing, China
Email: guojf@rails.cn

Abstract. In order to solve the problems of low efficiency and variety of results in preprocessing massive railway catenary geometric inspection data by manual interaction, this paper proposes an automatic preprocessing method for railway catenary geometric inspection data. A mileage jump point identification method is proposed based on the modified threshold variation algorithm; a data outlier removal method is proposed based on the repeatability inspection algorithm; a mileage automatic calibration method is proposed based on the principle of maximizing data similarity; a calibration evaluation method is proposed based on the principle of minimizing relative error. The above mentioned automatic preprocessing methods are used to preprocess the field test data. The results show that the method can not only identify the abnormal jump points in the inspection data, but also achieve efficient mileage calibration and calibration evaluation functions, which improves the efficiency of preprocessing railway catenary geometric inspection data.

1. Introduction

In order to improve the train operation safety, it is necessary to pay attention to the impact of the catenary irregularities on trains. Research on the catenary irregularities are mainly based on the analysis of catenary geometric inspection data. The comprehensive inspection trains and catenary inspection trains are used to inspect the catenary geometric state [1]. A large amount of dynamic inspection data was generated during the inspection process. The inspection data contains noise interference. It is necessary to research the preprocessing technology before data analysis.

The catenary geometric inspection data mainly includes mileage, inspection speed, stagger and the contact wire height. The method of comparative analysis of multiple inspection data is used to study the change of the catenary geometric state. Therefore, it is necessary to preprocess multiple inspection data before comparative analysis, which mainly includes jump points identification, invalid removal and mileage calibration. At present, the preprocessing of catenary geometry inspection data is usually carried out through manual interaction. It has the problems of low processing efficiency and high cost.

In recent years, many scholars have researched on automatic data preprocessing methods. In 2019, Gong Xiaoli in Tianjin University proposed an automatic data jump point identification method based on the modified threshold variation algorithm [2]. In 2018, Wang Haiying in China Academy of Railway Sciences proposed an abnormal value recognition technique for test data of contact wire height based on the absolute mean method and quantile approach [3]. In 2019, Wang Zhenhui in Nanchang University proposed a mileage deviation correction method based on the track irregularity matching irregularity data mileage correction model [4]. In 2020, Wang Xin in Southwest Jiaotong...
University proposed a mileage error evaluation method for track inspection data based on multiple waveform matching [5]. Xu Guiyang and Liu Jinzhao in China Academy of Railway Sciences proposed an automatic preprocessing method for track geometric inspection data and successfully applied it to engineering practice [6], which improved the efficiency of preprocessing.

This paper researched an automatic preprocessing method of the catenary geometric inspection data based on the modified threshold variation algorithm, the repeatability inspection algorithm, the principle of maximizing data similarity and the principle of minimizing relative error. The automatic preprocessing method was verified by field test data. The flow chart of automatic preprocessing method is shown in the following figure 1.

![Flow chart of automatic preprocessing method of the catenary geometric inspection data.](image)

**Figure 1.** Flow chart of automatic preprocessing method of the catenary geometric inspection data.

### 2. Mileage Jump Point Identification Method Based on Modified Threshold Variation Algorithm

The inspection data mileage jump phenomenon may occur during the running of train due to the reasons of partial rerouting or measurement formed long or short chains, electromagnetic interference and mileage positioning. Mileage jump can be divided into three situations: start jump, middle jump and end jump. The start jump and end jump usually occur in railway station. Therefore, there are some mileage jump points at the start or end of the data, which cannot be used for data analysis. The middle jump is caused by mileage positioning, electromagnetic interference and long or short chains. The middle jump usually results in repeated and abnormal mileage of local data. The start jump and end jump often cause abnormal mileage of section data. These three situations have a great impact on subsequent research. It should be able to automatically recognize in the beginning of preprocessing.

Mileage jump point identification method based on modified threshold variation algorithm is shown as follows. Assume that the mileages of two adjacent inspection points at time t are $x_{t-1}$ and $x_t$, given a threshold $w$, which can be modified. If the distance between two points is greater than the threshold, the point $x_t$ is determined as the jump point as shown in the following formula:

$$|x_{t-1} - x_t| > w$$

(1)

In the formula, the threshold value $w$ is determined by the spatial sampling frequency and the threshold is usually modified according to the spatial sampling frequency. Mileage jump points automatic recognition results are shown in the following figures 2 and 3.

![Start-middle jump identification result.](image)

**Figure 2.** Start-middle jump identification result.

![Middle-end jump identification result.](image)

**Figure 3.** Middle-end jump identification result.
3. Data Outlier Removal Method Based on Repeatability Inspection Algorithm

Most of the abnormal and repetitive values in the catenary geometric inspection data are caused by mileage jump points, which have a great influence on the subsequent mileage calibration.

The outliers in the inspection data can be found based on the repeatability inspection algorithm. There are three steps. First of all, the catenary geometric inspection data are arranged in chronological order to find repetitive data. Then the first occurrence value is retained. Finally, all subsequent duplicate values will be deleted. According to the statistical analysis of a large number of catenary inspection data, it is found that the mileage jump will cause a mileage change within a certain range. The abnormal value mostly occurs in the middle jump. In this case, we can modify the selected threshold to find the abnormal mileage value in the inspection data and delete the duplicate samples in which it is located. At this time, the mileage guarantees uniqueness and validity. It can be used for the subsequent mileage calibration research.

4. Automatic Mileage Calibration Technology Based on Maximum Similarity Principle

4.1. The Principle of Maximum Similarity

The catenary geometric inspection data within a certain time have a high similarity by comparing a large number of historical data. The similarity of the waveforms can be quantified and compared by the correlation coefficient. After finding the section with the largest similarity, mileage correction is performed.

The mileage relative error of two inspection data is usually within 100 meters according to the experience of manual mileage calibration. Assume that the first inspection data is standard data. The number of data is N. The second inspection data is to be calibrated. In order to find the waveform with the highest similarity, 100 meters of data are added back and forth in the second inspection. Therefore, the first inspection data can be expressed as \( X_1 = \{x_1(i), i = 1,2,\cdots,N\} \). The second inspection data can be expressed as \( X_2 = \{x_2(j), j = 1,2,\cdots,N+800\} \). The similarity between the two sets of data is measured by the correlation coefficient. The second inspection data needs to be divided into subsequences according to the length of the first inspection data. The subsequence can be expressed as \( X_2^k = \{x_2(k+i-1), i = 1,2,\cdots,N, k = 1,2,\cdots,800\} \).

The correlation coefficient between the two data can be obtained by the following formula:

\[
 r(k) = \frac{N \sum \{x_2(i)\}x_1(x_2(k+i-1)) - \sum \{x_2(i)\} \sum x_1(x_2(k+i-1))}{\sqrt{[N \sum x_1(i)^2 - (\sum x_1(i))^2][N \sum x_2(k+i-1)^2 - (\sum x_2(k+i-1))^2]}}
\]

4.2. Automatic Mileage Calibration Process

The process of automatic mileage calibration is shown in figure 4. In the beginning two group inspection data of the same section should be found from the database. One group is set as the standard group. The other is set to the data to be calibrated. The group to be calibrated adds an additional 100 meters of data back and forth on the basis of the divided sections. The new sections are divided after adding data into equal length subsequences. Then the subsequence corresponding to the maximum correlation coefficient is found and the mileage of this subsequence is replaced by the mileage of the corresponding standard group segment. Finally, the program will draw graphics comparison before and after calibration. We can observe the effect of calibration clearly.
5. Calibration Evaluation Technology Based on Relative Error Minimization Algorithm

Mileage automatic calibration results need to be measured by quantitative indicators. The correlation quantitative index mentioned above cannot be used directly because the data length of the standard group, the group to be calibrated and calibration group are generally different. In order to quantitatively describe calibration result, the concept of relative error is introduced. Taking the stagger in catenary geometric inspection data as an example, two previous inspection data and the mileage section are determined based on a group with a shorter data length. We can compare the inspection data near same mileage point in same section. The two waveforms are closer with the smaller error.

The inspection data of the standard group, the group to be calibrated and the calibration group in the same mileage section are expressed as $x_i'(i=1,2,...,n_1)$, $x_i''(i=1,2,...,n_2)$, and $x_i'''(i=1,2,...,n_3)$. The relative error between standard group and the group to be calibrated is expressed as follows:

$$
\delta_{1-2} = \frac{1}{n_{12}} \sum_{i=1}^{n_{12}} (x_i' - x_i'')^2, i = 1,2,...,n_{12}
$$

(3)

In the formula: $n_{12}$ is a shorter data length between same mileage section of standard group and the group to be calibrated. The relative error of standard group and calibration group is expressed as:

$$
\delta_{1-3} = \frac{1}{n_{13}} \sum_{i=1}^{n_{13}} (x_i' - x_i''')^2, i = 1,2,...,n_{13}
$$

(4)

In the formula: $n_{13}$ is a shorter set of data length between same mileage section of standard group and calibration group. If $\delta_{1-2}$ is smaller than $\delta_{1-3}$, the waveform before calibration is closer and calibration effect is not good. Otherwise, the waveform after calibration is closer and calibration effect is good.

6. Example Verification of Automatic Data Preprocessing

6.1. Automatic Recognition of Mileage Jump Point

Using the automatic preprocessing method proposed in this paper to preprocess the inspection data set, the automatically recognized mileage jump points result is as follows in figure 5.
6.2. Abnormal Value Removal

Using the automatic preprocessing method proposed in this paper to preprocess two inspection data, the results are shown in the following figures 6 and 7. The mileage correction results in the straight line section and curve section are performed respectively.

The blue solid line represents the first inspection data waveform and the dotted red line represents the second inspection data waveform. It can be seen from figures 6 and 7 that the waveforms of two inspection data in same mileage range are different. There are horizontal lines throughout the figures because there are mileage jump points in the data. In order to ensure the accuracy of subsequent analysis, it is necessary to eliminate the influence of horizontal lines then correct mileage of inspection data. The results of automatic preprocessing using outlier and duplicate value removal techniques based on the repeatability inspection algorithm proposed in this paper are as shown in figures 8 and 9.

It can be seen from figures 8 and 9 that after eliminating the effects of repeated and abnormal values, there are no horizontal lines. Then the second inspection data can be automatically calibrated.

6.3. Automatic Mileage Calibration

After removing the influence of the repeated and abnormal values, the waveforms of the straight line
and curve sections are calibrated according to the above mentioned automatic calibration technology based on maximization of similarity. The results are shown as in figures 10 and 11.

![Figure 10. Calibration data waveform in straight section.](image)
![Figure 11. Calibration data in curve section.](image)

It can be seen from figures 10 and 11 that the waveforms calibrated according to the above mentioned mileage automatic calibration technique based on similarity maximization can achieve automatic calibration and alignment to a certain extent.

6.4. Quantitative Evaluation of Calibration Results

According to the calibration evaluation standard based on the minimization of relative error proposed in this paper, two data in the range of 100 kilometer are selected and evaluated by dividing the unit per kilometer. It is found that the relative error becomes smaller after calibration in 79 units. It can be defined that the score is 79 of the calibration procedure. The calibration result on each kilometer unit is shown in figure 12.

![Figure 12. Result of calibration quantitative evaluation.](image)

7. Conclusion

This paper proposes an automatic preprocessing method for railway catenary geometric inspection data. It is verified by field test data. The results show that the modified threshold variation method in the preprocessing process can accurately identify the mileage jump points of the catenary geometric inspection data. The repeatability inspection algorithm can effectively eliminate repeated outliers of data. The automatic calibration method for catenary geometric inspection data based on the principle of maximizing data similarity can quickly calibrate multiple inspection data automatically. The catenary geometric inspection data mileage calibration evaluation method based on the relative error minimization algorithm can accurately quantify the calibration results. The railway catenary geometric
inspection data automatic preprocessing method proposed in this paper has a good application effect.

Acknowledgment
This paper was supported by China Railway corporation science and technology research and development projects (J2018G006, P2018G051, P2018G047).

References
[1] Weidong W, Shiping G, Chao Y, et al. 2012 High-speed comprehensive inspection train. *Railway Technology Innovation* 2 (1) 12-15.
[2] Xiaoli G and Xiong X 2019 Research on stock market jump and volatility dynamics based on modified realized threshold power variance. *Operations Research and Management* 28 (5) 124-133.
[3] Haiying W, Wei Z and Gang S 2018 Abnormal value recognition technique for test data of contact wire height. *China Railway* 1 (6) 93-97.
[4] Zhenhui W, Hongtao Z 2019 Research on correction of mileage deviation of historical data of track irregularity. *Railway Standard Design* 63 (12) 49-53.
[5] Xin W, Ping W, et al. 2020 Mileage error estimation and correction for high-speed railway track inspection data based on multiple data waveform. *Journal of Railway* 42 (2) 110-116.
[6] Guiyang X, Tianyun S, Jinchao L, et al. 2013 Research on automatic pre-processing methods for track geometry inspection data of high speed railway. *China Railway Science* 34 (6) 8-12.