

Article

Fault Data Detection of Traffic Detector Based on Wavelet Packet in the Residual Subspace Associated with PCA

Xiaolu Li ¹, Xi Zhang ², Peng Zhang ³ and Guangyu Zhu ^{1,*}¹ School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China² Beijing Transport Institute, Beijing 10073, China³ Transport Planning and Research Institute, Ministry of Transport, Beijing 100028, China

* Correspondence: gyzhu@bjtu.edu.cn

Received: 24 July 2019; Accepted: 21 August 2019; Published: 23 August 2019



Abstract: To improve the accuracy and efficiency of fault data identification of traffic detectors is crucial in order to decrease the probability of unexpected failures of the intelligent transportation system (ITS). Since convolutional fault data recognition based on traffic flow three-parameter law has a poor capability for multiscale of fault data, PCA (principal component analysis) is adopted for traffic fault data identification. This paper proposes the fault data detection models based on the PCA model, MSPCA (multiscale principal component analysis) model and improved MSPCA model, respectively. In order to improve the recognition rate of traffic detectors' fault data, the improved MSPCA model combines the wavelet packet energy analysis and PCA to achieve traffic detector data fault identification. On the basis of traditional MSPCA, wavelet packet multi-scale decomposition is used to get detailed information, and principal component analysis models are established on different scale matrices, and fault data are separated by wavelet packet energy difference. Through case analysis, the feasibility verification of traffic flow data identification method is carried out. The results show that the improved method proposed in this paper is effective for identifying traffic fault data.

Keywords: traffic detector; traffic flow data; fault data detection; wavelet packet energy analysis; principal component analysis

1. Introduction

With the development of sensors, computers and communication technologies, more and more high-speed and high-precision sensors are applied in the field of traffic detection. As a result, the amount of data in traffic database is also growing rapidly [1]. At the same time, due to the inconsistency of the front-end sensors and the changeability of the natural environment, the traffic data obtained contains not only normal traffic information, but also abnormal data caused by electromagnetic interference, detector failure, transmission network failure and traffic accidents—thus affecting the reliability of the intelligent transportation system (ITS) [2]. Therefore, it is of great practical significance to carry out research on the fault data diagnosis technology of traffic detectors to effectively improve the reliability of ITS operations [3].

After combing through the previous work in detail, the methods of data process monitoring can be divided into three categories: mathematical model-based, knowledge-based and data-driven methods. In the urban traffic system, the data acquired by the detector is huge and the system dynamics are complex. It is impossible to establish an accurate mathematical model for such a system. Moreover, the knowledge or experience of experts is also subjective. However, the monitoring process data

contains abundant process information, and the operation status of the system can be characterized by the normal or abnormal data. It is pointed out in the literature [4] that research is gradually changing from a model-based approach to a data-driven approach. Literature [5] divides data-driven monitoring methods into machine learning, signal processing, and multivariate statistics. Machine learning methods aim at the correct rate of fault diagnosis [6]. They train machine learning algorithms such as neural networks [7] or support vector machines [8] to diagnose faults by using historical data of objects in normal and various fault situations. It has a wide range of applications, but, because of the need for sample data in case of process faults, and the accuracy of the algorithm being closely related to the completeness and representativeness of the samples, it is difficult to be used in those processes that can not obtain a large amount of fault data. Signal processing methods use signal processing methods such as spectral analysis and wavelet transform to analyze the time-frequency characteristics of measured signals, and then diagnose faults [9]. The multivariate statistical analysis method is represented by principal component analysis (PCA). PCA is the core of data-driven process control and fault diagnosis [10]. It constructs a new set of latent variables and forms a new mapping space based on the original data space. The main change information is extracted from it and the statistical features are extracted, so that the spatial characteristics of the original data can be tabulated and the spatial dimension of the original data can be reduced. At the same time, the projection space statistical eigenvectors are orthogonal to each other, which eliminates the correlation between variables and greatly reduces the complexity of the original process characteristic analysis.

Given their importance, the fault data detection of traffic detectors is always a research focus and has attracted the attention of numerous scholars for years [11]. At present, the research of traffic detector fault data recognition mainly focuses on data fault identification based on a traffic flow three-parameter rule [12], data fault identification based on statistical analysis and data fault identification based on artificial intelligence [13]. Xu [14], by analyzing the influence of inner relationship of data acquisition interval and three traffic flow parameters, then designed a four-step data sieving method. In the literature [15], for the non-stationary traffic flow data, wavelet analysis is used to separate the high-frequency and low-frequency components of the original traffic flow data, and then the least squares method is used to find the abnormal points in the reconstructed signal data, which effectively reduces the false positive rate and the false negative rate. Ngan [16] presents a Dirichlet process mixture model (DPMM) for detecting abnormal data in large-scale traffic data, and the experimental results show that the method has good robustness. In literature [17,18], a clustering method is proposed to detect hardware data errors and abnormal traffic behavior in collected large-scale traffic data. The abnormal values are detected by the relationship between neighborhoods of data points, and the ST (space-time) signal of traffic video is transformed into a two-dimensional vector by principal component analysis (PCA). When applied to data fault detection, PCA is only applicable to the analysis of fault samples that exist on a single scale or frequency segment [19–21]. However, the fault data often occur in different time-frequency ranges in the actually collected sample data. That is, the sample data are multi-scale in nature, so the traditional PCA cannot be applied to diagnose the fault data in this situation [22]. For this reason, Bakshi [22] first proposed a multi-scale principal component analysis (MSPCA) model, and the basic principle of MSPCA is to combine wavelet and principal component analysis, extend the single-scale model to a multi-scale model, combine the ability of principal component analysis to reduce the dimension linearly with the ability of removing the autocorrelation of process variables by wavelet and the advantage of decisive feature of process variables by wavelet. Then, the wavelet coefficients are analyzed by PCA on various scales. In the study of traffic detector data, it is found that normal traffic data are regular and its signal energy range is relatively fixed. When abnormal data appear in the data, it will be directly reflected in the signal energy of the collected data. This anomaly may be caused by noise and may also be a traffic detector failure. Therefore, this article only detects where there is an exception in the detected data, and does not calculate the size of the exception.

Aiming at the problem of fault in dynamic traffic data, the fault data detection models based on PCA model, MSPCA model and improved MSPCA model were proposed respectively in this paper. First, the wavelet packet is used to decompose the original data with multi-scale, and the individual variable was decomposed into approximation coefficients and detail coefficients of multiple scales and the corresponding principal component analysis models in various scale matrices were established. Then, using the model statistical magnitude as the threshold value, the comprehensive principal component analysis model was obtained by reconstructing wavelet coefficients and the fault data were separated.

The innovative contributions of this paper can be summarized as follows:

- (1) Aiming at the fault problem of dynamic traffic data, a fault data detection model based on principal component analysis (PCA) is proposed. The principal component subspace and residual subspace are established by historical data. The obtained traffic detector data are projected into the space, and the data to be checked are described by parameters.
- (2) In order to solve the problem of fixed MSPCA modeling and single problem with principal component subspace and T^2 , SPE parameters. Using the idea of adaptive PCA principal component recursion for reference, the wavelet decomposition is replaced by the wavelet packet decomposition, and the fault data are detected by the energy difference of the wavelet packet, which improves the resolution of the model.

The structure of this paper is as follows: Section 2 briefly introduces the fundamental theories of wavelet packet energy analysis and the principal component analysis theory. The third section, the core idea and detailed process of the fault data recognition model are expounded. The fourth section tests the fault data recognition model based on the actual detector data and validates its effectiveness; Section 5 concludes the paper.

2. Preliminaries

In this section, firstly, the basic theory of wavelet packet is elaborated. Then, the basic theory of principal component analysis is elaborated.

2.1. The Basic Theory of Wavelet Packet Energy Analysis

In the process of data failure identification, the fault data are realized by calculating the energy difference between the wavelet packet component of normal data and the corresponding component containing the fault data. Some of the fault information is sometimes hidden in high frequency information and cannot be found in wavelet decomposition. The difference between wavelet packet decomposition and wavelet is that the former decomposes the wavelet decomposition into the same low-frequency part of the same layer in the high-frequency part of each layer. Therefore, in the fault diagnosis process, the wavelet packet algorithm is more suitable than the wavelet decomposition algorithm.

In the Mallat algorithm, there are impulse response functions $h(n)$ and $g(n)$. Therefore, the scale function $\varphi(t)$ and the wavelet function $\psi(t)$ are defined as follows [23]:

$$\begin{cases} \varphi(t) = \sum_n h(n)\varphi(2t - n), \\ \psi(t) = \sum_n g(n)\psi(2t - n). \end{cases} \quad (1)$$

In Formula (1), $g(n) = (-1)^{1-n}h(1-n)$, n is the wavelet packet decomposition channel number. Set a finite energy signal, using the Formula (1), the scale function and the wavelet function can be obtained, and the wavelet packet decomposition is obtained as follows [24]:

$$\begin{cases} u_{2j}(t) = \sqrt{2} \sum_{k \in \mathbb{Z}} h(k) u_j(2t-k), \\ u_{2j+1}(t) = \sqrt{2} \sum_{k \in \mathbb{Z}} g(k) u_j(2t-k), \quad 1 \leq j \leq J, \\ u_0(t) = x(t). \end{cases} \quad (2)$$

In Formula (2), $u_j(t)$ is a wavelet packet coefficient, and j is the decomposition scale, $k \in \mathbb{Z}, k \geq 2$, means the scale factor k is an integer not less than 2, $h(k)$ is the low pass filter coefficient, $g(k)$ is the high pass filter coefficient, $x(t)$ is a finite energy signal.

Wavelet packet energy formula is defined as:

$$E = \int_{-\infty}^{+\infty} x^2(t) dt = \sum_{a=1}^{2^j} \sum_{b=1}^{2^{j-1}} \int_{-\infty}^{+\infty} x_j^a(t) x_j^b(t) dt = \sum_{n=0}^{2^j-1} E_j^n. \quad (3)$$

In the form: $x(t)$ is a finite energy signal, scale parameter j , scale factor a and displacement factor b of basis function $\psi(t)$. $E_j^n = \int_{-\infty}^{+\infty} x_j^n(t)^2 dt$, and is the data energy of the n -th band in the j -th scale decomposition of the decomposed signal. Thus, the wavelet packet energy difference is as follows: $\Delta E_j^n = E_j^n - E_j^{n(d)}$, E_j^n is the normal data of wavelet packet energy, and $E_j^{n(d)}$ is the data to be seized wavelet packet energy.

In any interval, when a fault occurs, the energy difference in the fault point and its vicinity will be abruptly changed, and the energy difference between the data points without fault occurring will be relatively gentle. Therefore, the wavelet packet energy difference is sensitive to the fault data, which can be used to analyze and identify the fault on multiple scales.

2.2. Principal Component Analysis Theory

A traffic detector has many characteristics, such as a large number of sensors, a large dimension of data collected, and a large amount of data. Taking the coil sensor as an example, a section contains only two intersections. If we need to collect enough abundant data, we need several or even more coils, which makes it possible to obtain a large amount of data in a sampling period. Principal component analysis (PCA) is a branch of multivariate statistics, which is especially suitable for multivariate data processing [25].

Principal Component Analysis (PCA) is one of the projection models in multivariate statistical analysis. The principle is to assume that there is a measurement sample containing m sensors: $x \in R^m$. In the measurement sample, each sensor has n independent sampling data, and the data are constructed into a data matrix of size $m \times n$:

$$X = [x_1, x_2, \dots, x_n]^T \in R^{n \times m}. \quad (4)$$

Each column of X represents a measurement variable, and each row represents a sample. The PCA model first standardizes each sample of X , that is, it calculates the covariance matrix of X , and the formula is as follows:

$$\text{cov}(x) \approx \frac{X^T \cdot X}{n-1}. \quad (5)$$

After the feature decomposition of X is completed, the size of the feature values is sorted from large to small. The PCA model decomposes X as follows:

$$\begin{cases} X = \hat{X} + E = T \cdot P^T + E, \\ T = X \cdot P. \end{cases} \quad (6)$$

After separating the fault data, the variable with a larger contribution graph is considered as the cause variable that may generate fault data.

In Equation (6), $P \in \mathbb{R}^{m \times A}$ is the load matrix and is composed of the first A feature vectors of $cov(x)$. $T \in \mathbb{R}^{n \times A}$ is a score matrix, each column of T is called a principal element variable, A is the number of principal elements, and is also the number of columns of the score matrix. It can be known from the properties of the eigenvalues and the eigenvectors that the scoring matrices are mutually orthogonal, so these principal elements are also independent of each other. The covariance of the principal element can be calculated from Equation (7):

$$\Lambda = \frac{X^T \cdot X}{n-1} = \begin{bmatrix} \lambda_1 & & \\ \dots & \dots & \\ & & \lambda_n \end{bmatrix}, \quad (7)$$

where $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$ represents the top A larger eigenvalues of covariance matrix of X . For the selection of principal A , this paper uses the method of calculating the cumulative variance contribution rate to determine. The cumulative variance contribution rate represents the ratio of the amount of data that the first A principals can interpret to the total data [26]. This can be calculated as follows:

$$CPV = \frac{\sum_{i=1}^A \lambda_i}{\sum_{i=1}^m \lambda_i}. \quad (8)$$

Usually, when the CPV (cumulative contribution percentage of variance) reaches more than 85%, it can be assumed that the former A principal component can explain most of the data changes in X . As shown in Figure 1, the red dashed line represents the value of CPV, which is 85% in Figure 1, when the lower limit of CPV is set at 85%, the number of principals is at least two. Therefore, when three principals are selected, the first three principals can contain enough information.

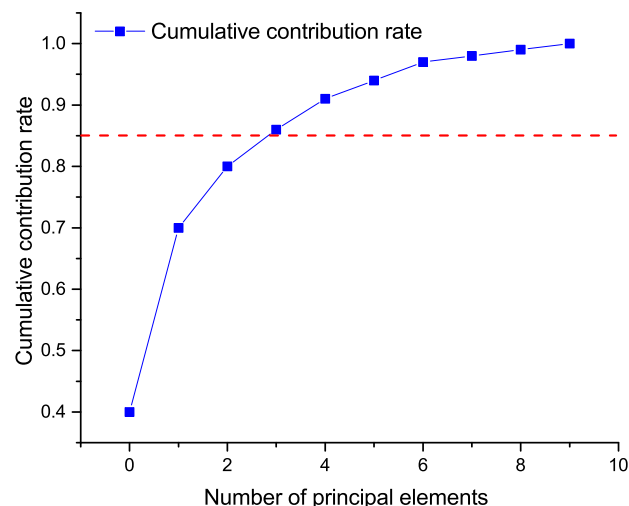


Figure 1. Cumulative contribution rate method to determine the number of principal elements.

The principal component constitutes a space, which is called the principal subspace, and the orthogonal complementary space is called the residual subspace. Any sample vector in X can be decomposed into projections on principal and residual subspaces [26].

As shown in Figure 2, X is the measurement data matrix, x is the measurement sample, β is the principal subspace direction, and α and γ are the residual subspace directions. X is projected in two directions to form two spaces, namely principal subspace R_p and residual subspace R_r . The projection

result in the R_p direction is $\hat{x}$, that is, the size of the effective data component in the sample. $\tilde{x}$ is a projection on the residual subspace.

The PCA model divides the variable space into two orthogonal and mutually exclusive subspaces. Any sample vector can be decomposed into projections on these two spaces: $x = \hat{x} + \tilde{x}$; $\hat{x} = PP^T x \in R_p$; $\tilde{x} = (I - PP^T)x \in R_r$. $\hat{x}$ is the modeled part and $\tilde{x}$ is the unmodeled part. From the above relationship, we can conclude that: $(I - PP^T) \cdot \hat{x} = 0$, $(I - PP^T) \cdot \tilde{x} = 0$. Because it is orthogonal projection, the two satisfy the orthogonal relationship: $\hat{x}^T \tilde{x} = 0$. In addition, the two are not statistically relevant, so there are:

$$E(\hat{x}\tilde{x}^T) \approx \frac{X^T X}{n-1} = \frac{PT^T \tilde{X}}{n-1} = 0. \quad (9)$$

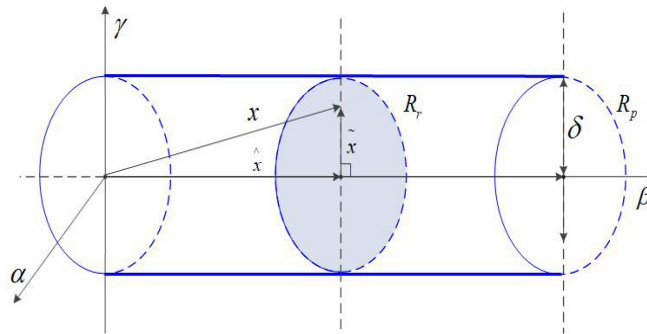


Figure 2. Projection of data on principal and residual subspaces.

From the above analysis, we can find that PCA has natural advantages in process monitoring. It is very suitable for real-time monitoring of traffic detector data.

3. Proposed Fault Diagnosis Method

In this section, the fault data detection models based on the PCA model, MSPCA model and improved MSPCA model are introduced, respectively.

3.1. Diagnostic Model Based on Principal Component Analysis

Based on PCA, the idea of fault separation is to build two spatial models through a large number of normal offline data, and then detect the detected data, separate the correct data and fault data separately. In measuring the normal data, we usually use the Squared prediction error (SPE) and Hotelling's T^2 (hereinafter referred to as statistics T^2) to identify whether there is an abnormality [27].

The SPE indicator measures the variation of the sample vector in the residual space projection:

$$SPE = \|(I - PP^T)x\|^2 \leq \delta_\alpha^2. \quad (10)$$

In Formula (10), δ_α^2 represents the control limit when the confidence level is α , the calculation formula is as follows: $\delta_\alpha^2 = g\chi_{h,\alpha}^2$, $g = \frac{\theta_2}{\theta_1}$, $h = \frac{\theta_1^2}{\theta_2}$, $h_0 = 1 - 2\frac{\theta_1\theta_3}{3\theta_1^2}$, $\theta_i = \sum_{j=A+1}^m \lambda_j^i$ ($i = 1, 2, 3$).

$\chi_{h,\alpha}$ is the distribution critical value of degree of freedom h and confidence α . The occurrence of faults is reflected in the statistical index exceeding the statistical control limit, and the change of correlation between data are reflected in the change of SPE value. Fault detection based on SPE is shown in Figure 3. δ is the radius of control domain, the red dot represents the data point, and the distance from the data point to the β -axis is SPE of residual space. If this distance is maintained in a controllable domain, it indicates that no fault occurs, and, if the data point falls outside the domain, it is considered that a fault occurs.

T^2 statistics measure the change of variables in the principal element space:

$$T^2 = x^T P \Lambda^{-1} P^T x \leq T_\alpha^2. \quad (11)$$

In Formula (11), $\Lambda = \text{diag} \{\lambda_1, \dots, \lambda_A\}$, T_α^2 represents the T^2 control limit when the confidence degree is α .

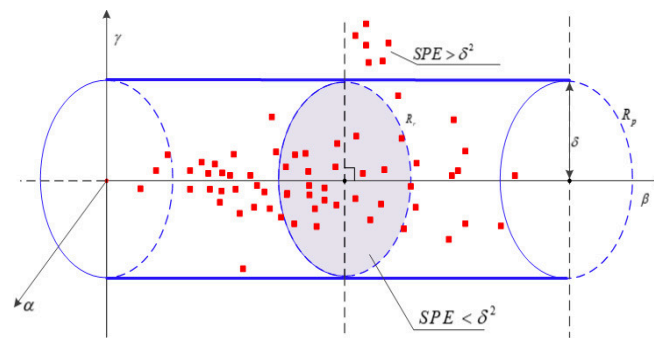


Figure 3. Geometric schematic of fault detection based on SPE (Squared prediction error) index.

Although both SPE and T^2 indicators are used for process fault detection, it is important to point out that the two indicators are complementary. The SPE indicator measures the degree of correlation change between normal process variables and shows abnormal process conditions. The T^2 indicator measures the distance of an existing sample from the origin of the principal subspace. Since the principal subspace contains most of the change in the signal when the process is normal, and the residual subspace mainly represents noise, the control limit defined by T^2 is usually much larger than that defined by SPE [28]. Therefore, smaller faults easily exceed the control limits of the SPE. If the T^2 indicator of a sample exceeds the control limit, its SPE indicator does not exceed its control limit, that is, it does not destroy the relationship between variables. This situation may be a fault, or the process scope may change.

After the completion of the fault data detection, the need to separate the variable that caused the fault. In the process of fault diagnosis of a traffic detector, it is most likely to find out which sensor occurs when the data anomaly is found. The fault separation method used in this paper is a fault separation method based on the SPE contribution graph. The calculation method is as follows [28]:

$$SPE = \|\tilde{C}x\|^2 = \sum_{i=1}^m \text{Cont}_i^{\text{SPE}}. \quad (12)$$

In Formula (12), $\text{Cont}_i^{\text{SPE}} = (\zeta_i^T \tilde{C}x)^2$, $\text{Cont}_i^{\text{SPE}}$ represents the contribution value of each variable to the statistic SPE for the data matrix X to be picked up. $\tilde{C} = I - PP^T$, ζ_i represents the i -th column of the identity matrix I_m . Figure 3 shows the geometric schematic of fault detection based on SPE index. After the fault is separated, the variable with a larger contribution graph is considered to be the cause variable that may cause the fault.

3.2. Fault Data Recognition Model Based on MSPCA

In the long-term study and summary of traffic detectors, it is found that, due to the aging of sensors in detectors and the wear and tear of various devices, traffic flow parameters detected during a certain period of time seem to be at a normal level with little fluctuation of data, but the offset of data are more and more significant, along with this kind of gradual cumulative failure. Ultimately, it will cause systematic false alarm and is not easy to detect.

With the maturity of wavelet technology, the multi-scale characteristics of fault data have made remarkable progress. In order to provide more detailed and high-resolution analysis data and to overcome the shortcomings of customer service PCA, Bakshi [22] proposed a multi-scale principal component analysis fault diagnosis method. This model does not depend on the analytical model of the target system, but proceeds from the process data, combines wavelet analysis with PCA, and uses

PCA to process more offline data. The multiscale characteristic of wavelet is used to overcome the shortcomings of conventional PCA in meta-statistical analysis.

Figure 4 is the flow chart of MSPCA. It can be seen from the graph that MSPCA combines wavelet analysis with PCA organically. The model uses offline normal data to establish principal component subspace, then decomposes the sample data by wavelet, and models the coefficient matrix of wavelet analysis coefficients separately in principal component space. After reconstructing the decomposed signal, the principal component will be modeled again.

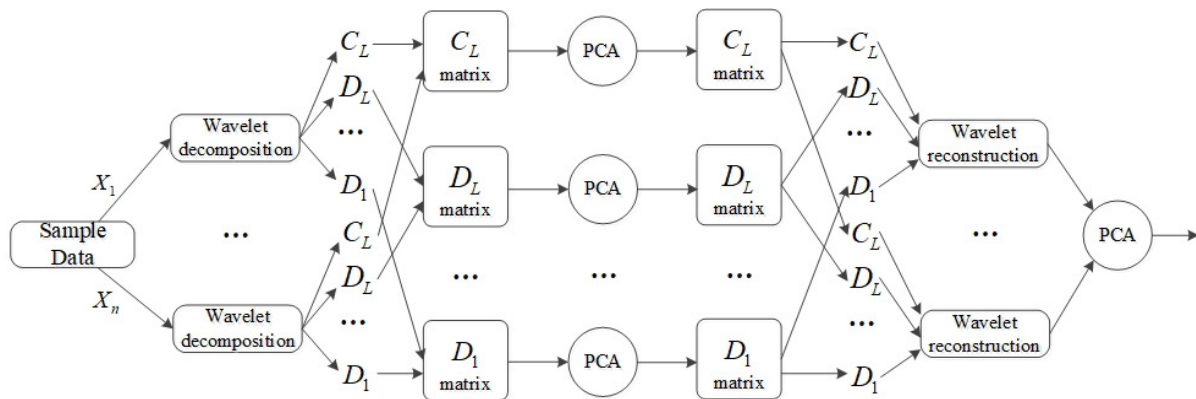


Figure 4. The flow chart of the calculation of MSPCA (multiscale principal component analysis) mode.

3.3. Fault Data Recognition Model Based on Improved MSPCA

When monitoring the actual dynamic traffic data, the noise distribution of the data are random and its intensity is time-varying. However, when dealing with the wavelet coefficients beyond the control threshold, the traditional MSPCA uses the unified threshold under the decomposition scale to reconstruct the wavelet, lacking the consideration of the time-varying noise. Therefore, part of the noise will be mistaken as fault data and separated into residual space. In addition, part of the fault data covered by noise will be expanded, resulting in an increased false alarm rate.

In order to solve the problems of a fixed model, principal component subspace, SPE, T^2 and a single parameter in traditional MSPCA, based on the recursive idea of the adaptive PCA, the traditional MSPCA is adjusted. The modifications are as follows:

- (i) Segmental processing of collected traffic flow data;
- (ii) In order to improve the resolution of the model, the wavelet decomposition is adjusted to the wavelet packet decomposition;
- (iii) The energy difference of nodes in each decomposition scale is calculated by the energy difference of wavelet packet to locate the fault data location.

Figure 5 shows the flow chart of the improved method.

The model is mainly divided into two modules: offline module and online fault diagnosis module. The offline module preprocesses the model with normal data. Through this process, the scale parameters of wavelet packet and the number of principal components of principal component analysis can be determined. The online fault detection module uses the parameters calculated by the offline module to monitor the online data in real time, remove the noise interference, and finally complete the data fault diagnosis.

As shown in Figure 5, in offline mode, the model firstly preprocesses the normal historical data, uses the wavelet packet to decompose the data at J - level, calculates the wavelet packet energy of each component on each scale, and models the component data on each scale by PCA, and chooses the appropriate number of principal components. The data space is decomposed into principal component space and residual space, and the control limits of T^2 - SPE statistics for each component on each scale are calculated.

In the online fault diagnosis mode, the data are denoised by wavelet threshold. Referring to the scale parameters calculated in the offline mode, the data are decomposed into J -scales. The wavelet packet energy of each component on the scale is calculated, and the difference of wavelet packet energy on each scale is obtained. According to the set threshold, the fault can be judged. The components with faults are selected for principal component analysis, and the location and size of data faults are determined according to $T^2 - SPE$ control limits obtained from offline mode. Finally, all the data are analyzed by PCA to determine which sensors are responsible for the fault and the severity of the fault.

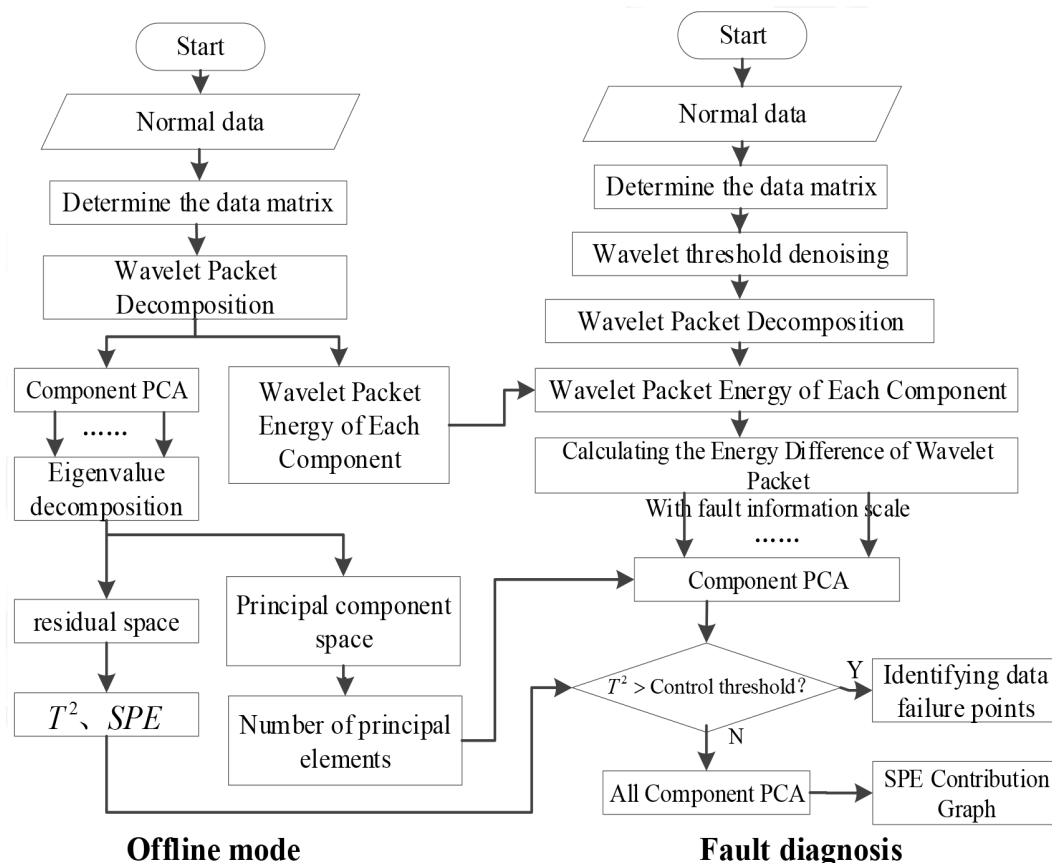


Figure 5. The flow chart of the improved method.

The detailed algorithm process is as follows:

- Step1 Input the normal data matrix X , set the size of the data window is m , and N is a multivariate variable dimension;
- Step2 The wavelet packet transform is used to decompose each row of the data matrix X , that is, each variable to a scale J , and to calculate the wavelet packet energy value of each component under the scale;
- Step3 Perform the principal component analysis on each wavelet packet component to calculate the number of each principal component and the two control-limit parameters of the component;
- Step4 Input the online monitoring data, carry out the wavelet threshold diagnosing for each component, and reconstruct the wavelet packet with the same scale according to the corresponding components in Step2;
- Step5 Calculate the wavelet packet energy of each component. In addition, subtract from the corresponding offline data wavelet packet energy, to obtain the components of the wavelet packet energy difference curve;

- Step6 Select the faulty component or the component of interest, and perform the principal component modeling of the same parameters as in Step3, and use the previously obtained SPE control parameter and control parameter to separate the fault;
- Step7 Repeat Step4–Step6, give statistics of the detailed location of the data failure under each component;
- Step8 Conduct principal component analysis on all the data and draw 2D contribution map based on SPE.

3.4. Definition of Error Index

Define the false alarm rate γ_f , the missing reporting rate γ_m , and the algorithm accuracy rate γ_A . The false alarm rate describes a data that is not a fault but is misreported as a fault. The missing reporting rate describes data that should be a fault but has not been detected. The calculation method is as follows:

$$\begin{cases} \gamma_f = \frac{n}{N}, \\ \gamma_m = \frac{m}{N}, \\ \gamma_A = \sqrt{\frac{\theta_{T^2}^2 + \theta_{SPE}^2}{2}}. \end{cases} \quad (13)$$

In Equation (13), n is the number of false alarm points, m is the number of missing reporting points, and N is the total number of data. θ_{T^2} and θ_{SPE} respectively indicate the accuracy of fault diagnosis when using T^2 and SPE , among which:

$$\begin{cases} \theta_{T^2} = 1 - (\gamma_{f,T^2} + \gamma_{m,T^2}), \\ \theta_{SPE} = 1 - (\gamma_{f,SPE} + \gamma_{m,SPE}). \end{cases} \quad (14)$$

In Equation (14), γ_{f,T^2} and $\gamma_{f,SPE}$ respectively indicate the accuracy of false alarm rate when using T^2 and SPE , γ_{m,T^2} and $\gamma_{m,SPE}$ respectively indicate the accuracy of missing reporting rate when using T^2 and SPE .

4. Case Study

The data of the coil detector in the intersection of Beijing is selected as the experimental data source during the week (13–17 August 2018). There is a coil in front of the stop line in each lane of the intersection. There are eight coils in total. Each detector generates about 555 data a day. Figure 6 is a working day inspection data. In coil 3 and coil 7, faults occur between data points 150 to 200, 300 to 350. After adding high frequency random white noise to all data, the data are used as the data to be checked with fault and noise.

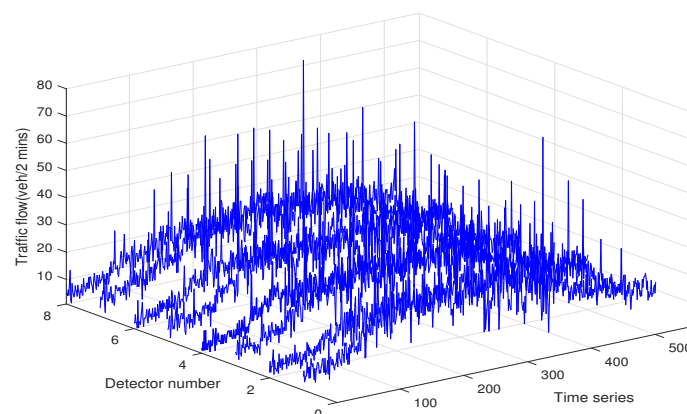


Figure 6. Traffic flow data of eight detectors.

In the parameter setting of simulation, the DB3 wavelet is used as the basis function of wavelet decomposition and wavelet packet decomposition in traditional MSPCA and the improved MAPCA model proposed in this paper. The confidence level is 99% when calculating the parameters of statistics and SPE statistics. The cumulative contribution rate of principal component is not less than 85%.

4.1. Data Fault Recognition and Analysis

From Figure 7a, it can be found that the data near the sampling point 300–350 appear to be of a more serious fault type. Other sampling points, such as 500, have a different degree of fault data in the vicinity of the sampling point. Compared to Figure 7b, it is known that the No. 3 detector and No. 7 detector have the greatest contribution to the fault data, of which the number 3 detector has a relatively more serious fault. Among them, the real line represents the T^2 control limit of 99% of confidence.

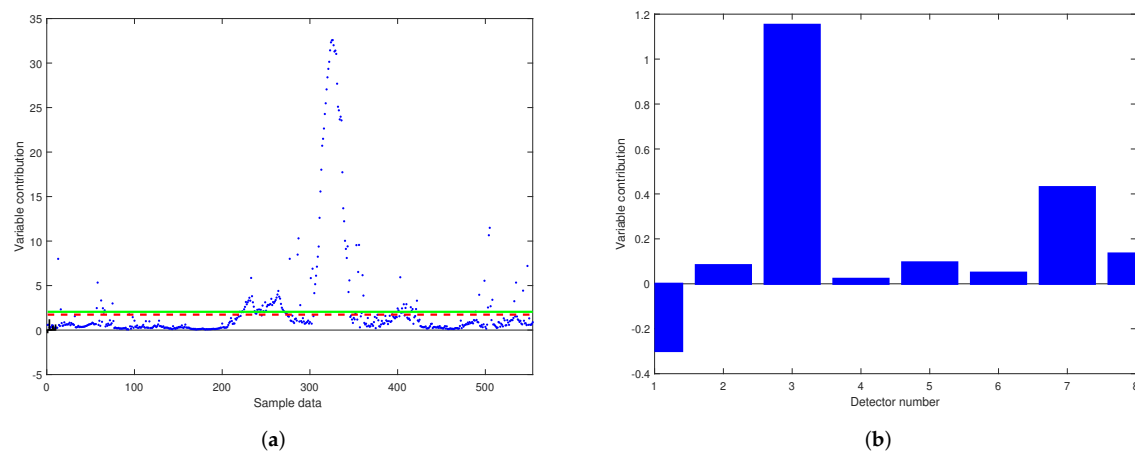


Figure 7. The SPE threshold and contribution graph for matrix X: (a) the SPE control limit of X; (b) each variable contribution graph of X.

In order to verify the advantages of the proposed fault data detection model, each vector of the data to be detected is first decomposed into a wavelet packet, and the principal component analysis is used to model the matrix composed of each node of the wavelet packet decomposition. The wavelet packet energy difference of each node is calculated, and the component of the fault is found to be analyzed by the PCA model, and the location where the data fails under the component can be determined.

4.2. PCA Fault Data Recognition Results

In the T^2 and SPE monitoring diagrams, the dotted line indicates that the confidence level is 99% or the SPE control limit parameter. As can be seen from Figure 8, a large number of data points have been misreported. In the monitoring chart, there have been more serious error alarms in the 100 data points at the beginning of the data, and similar problems occurred near 200. Although noise interference is less affected by the SPE monitoring graph, there is a serious misreporting phenomenon.

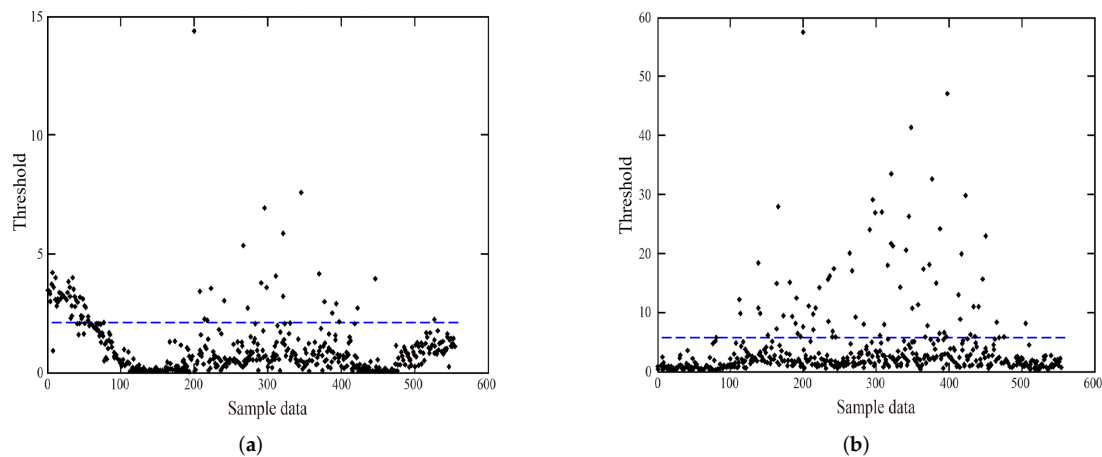


Figure 8. Control limit chart of the PCA (Principal Component Analysis) fault recognition model: (a) PCA- T^2 ; (b) PCA-SPE.

The fault recognition of the PCA model is caused by the fact that the PCA has no noise interference characteristics, so, if the noise is more intense, the noise information will be treated as the fault information. In addition, traffic flow information is characterized by nonlinear and time-varying characteristics, and the possibility of mutation is relatively large. Therefore, the method of “one size fits all” with various control parameters such as PCA will inevitably lead to errors.

4.3. MSPCA Fault Data Recognition Results

MSPCA combines the wavelet analysis method on the basis of PCA, and carries out unified threshold processing after wavelet decomposition, which has a certain anti-jamming ability to noise. The simulation results are shown in Figure 9. MSPCA can more effectively resist the interference of noise, the false positives of fault are also significantly reduced, but, due to the MSPCA in establishing the main space of atoms and the residual subspace being single, in dealing with traffic flow data, many false positives still appeared.

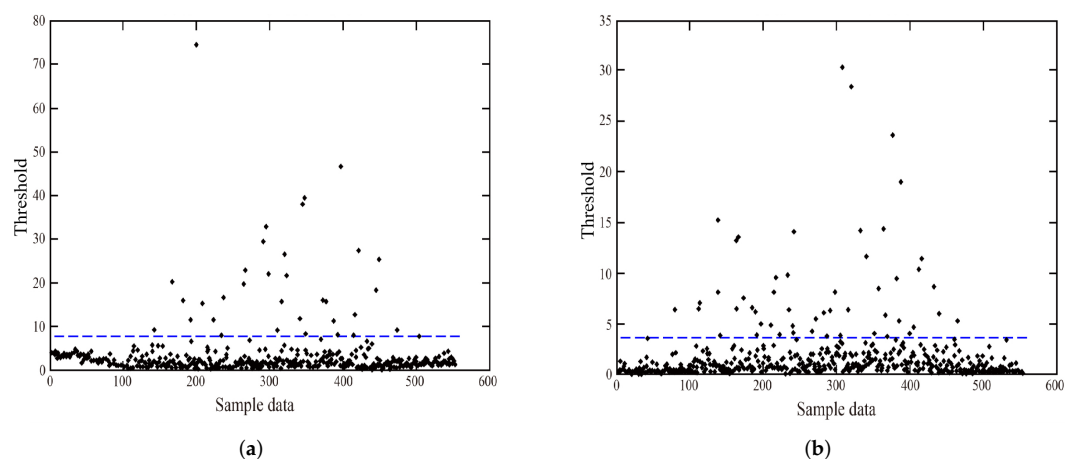


Figure 9. Control limit chart of MSPCA fault diagnosis model: (a) MSPCA- T^2 ; (b) MSPCA-SPE.

4.4. Improved MSPCA Fault Data Recognition Results

After de-noising by wavelet threshold, the reconstructed signal is decomposed into three-scale wavelet packets. The energy of two nodes is obtained in the first scale, four in the second scale and eight in the third scale. The energy difference between the wavelet packet energy and its corresponding normal data at each scale and the same node at the same scale is calculated, respectively. The result of each scale is shown in Figures 10–12.

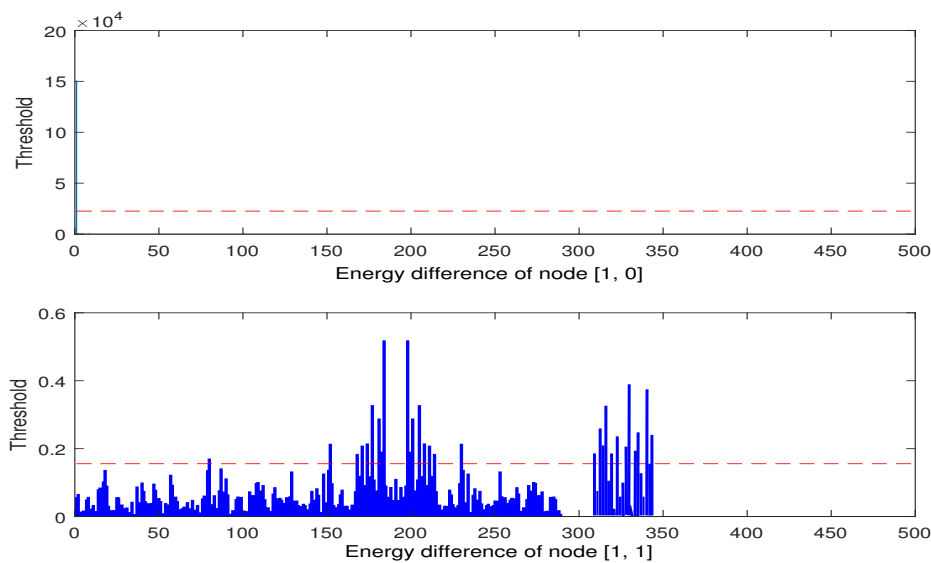


Figure 10. Energy difference result of first layer decomposition.

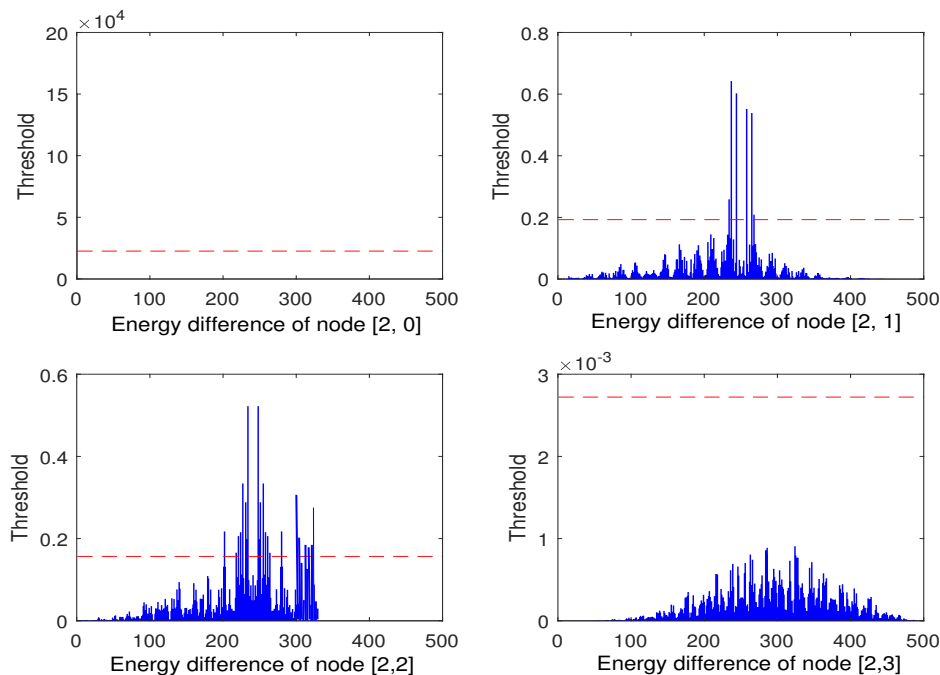


Figure 11. Energy difference result of second layer decomposition.

Figure 11 shows the energy difference of eight vectors decomposed by layer 3 wavelet packet of the detector data. The dashed line represents the energy difference threshold of each component.

In comparison, it can be found that, in Figure 10, no fault information is found in the energy difference of the first layer wavelet packet decomposition signal node [1, 0], and fault information is found in the node [1, 1] near the data point 175–225, 300–350. In Figure 11, the energy of wavelet packet of node [2, 1] detects data faults at about 225–275 data points, and the range of data faults detected by node [2, 2] is about 200–275, 300–350 data points. In the nodes [2, 0] and [2, 3], there is no fault information. At the same time, it can be found that, in the nodes [2, 0], the energy difference is almost zero, which is the same as in the first layer of nodes [1, 0]. Although the energy difference of nodes [2, 3] fluctuates, it does not exceed the set threshold, so it is considered that the change is mainly caused by the noise contained in the data. In Figure 12, node [3, 0] and node [1, 0] have no fault data as

in [2, 0]. However, data faults occurred in the vicinity of 150–200, 300–350 in nodes [3, 1], nodes [3, 2], nodes [3, 3], and nodes [3, 4].

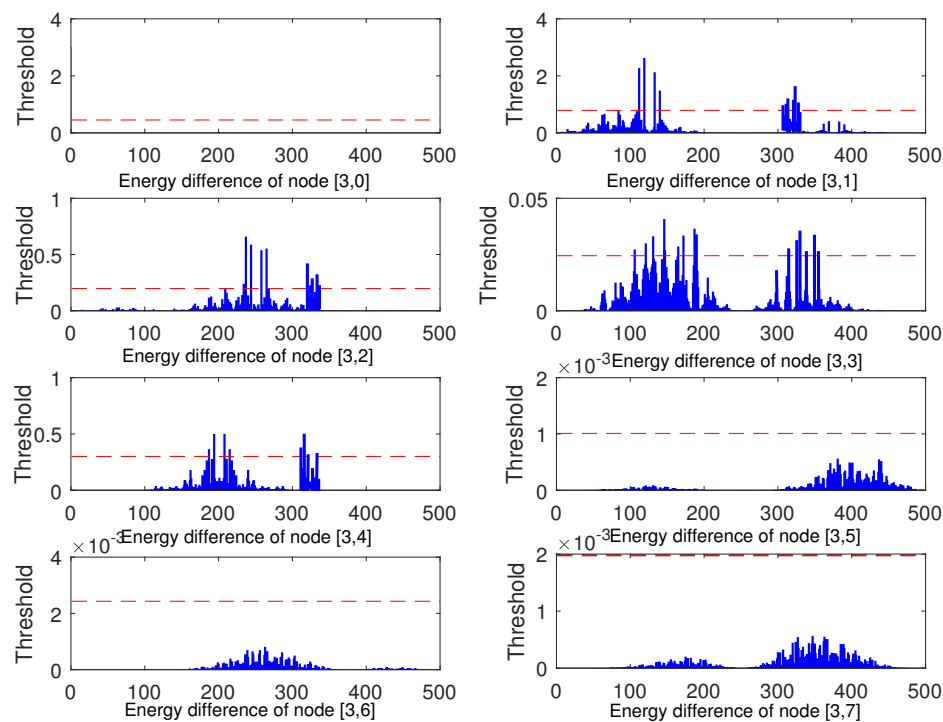


Figure 12. Energy difference result of third layer decomposition.

Through three-level decomposition of the wavelet packet, the suspected fault location of the data to be checked can be obtained between 150–225 and 275–350 data points. Parameters and principal component modeling analysis are carried out for nodes beyond the threshold at each level, principal component space and residual space are established, and T^2 control limit is set. The data with fault information is projected into the principal component space, and the result is shown in Figure 13.

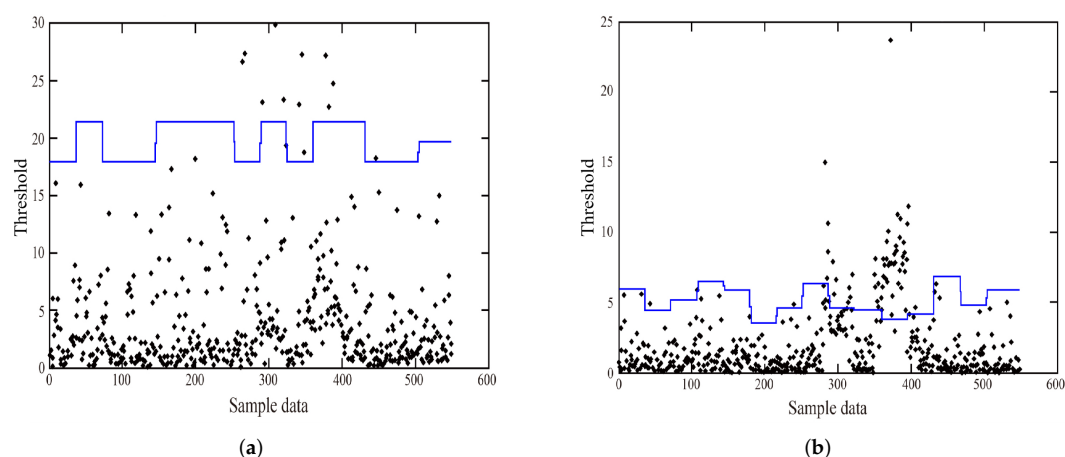


Figure 13. Control chart of improved MSPCA fault diagnosis model: (a) improved MSPCA- T^2 ; (b) improved MSPCA-SPE.

In contrast to the results of the three models, we can find that the improved MSPCA model of this paper can be more effective at making noise reduction in the data, and the error reporting in the malfunction diagnosis is obviously reduced, and it can be obvious from the control limits and the SPE

control chart to find the location of the two points of failure. Evaluate the accuracy of each model, as shown in Table 1.

Table 1. Accuracy comparison of traffic detector fault data identification results.

Model	Statistic	False Alarm Rate	Rate of Missing Report	Accuracy Rate
PCA	T^2	32.76	8.78	67.21
PCA	SPE	20.26	4.12	67.21
MSPCA	T^2	18.45	5.73	78.23
MSPCA	SPE	12.44	7.87	78.23
ImprovedMSPCA	T^2	9.12	4.17	89.26
ImprovedMSPCA	SPE	7.76	0.08	89.26

5. Conclusions

In this paper, the traffic flow data collected by traffic detector is studied in fault recognition and data repair. First of all, this paper puts forward the improved MSPCA model, on the basis of traditional MSPCA model, to improve the identification precision of the model, wavelet decomposition to the wavelet packet decomposition, using a wavelet packet energy difference method to test the fault information, and to isolate the fault data. Simulation results show that the improved MSPCA fault data diagnosis model can effectively identify abnormal data.

Author Contributions: Data curation, P.Z.; Funding acquisition, G.Z.; Investigation, X.Z.; Methodology, X.L.; Software, X.Z.; Supervision, G.Z.; Validation, P.Z.; Writing—original draft, X.L.

Funding: This work is supported by the National Key Research and Development Plan of the Ministry of Science and Technology (2016YFB1200203-02), the Fundamental Research Funds for the Central Universities (No. 2019YJS103).

Conflicts of Interest: The authors declare no conflict of interest.

References

- Wang, X.Y.; Zhang, J.L.; Yang, X.Y. *Key Theory and Method of Traffic Flow Data Cleaning and State Identification and Optimization Control*; Science Press: Beijing, China, 2011.
- Liu, W.; Guo, P.; Ye, L. A Low-Delay Lightweight Recurrent Neural Network (LLRNN) for Rotating Machinery Fault Diagnosis. *Sensors* **2019**, *19*, 3109.
- Chen, Y.; Zhang, T.; Luo, Z.; Sun, K. A Novel Rolling Bearing Fault Diagnosis and Severity Analysis Method. *Appl. Sci.* **2019**, *9*, 2356.
- Qin, S.J. Statistical process monitoring: Basics and beyond. *J. Chemom.* **2010**, *17*, 480–502.
- Zhou, D.H.; Hu, Y.Y. Fault Diagnosis Techniques for Dynamic Systems. *Acta Autom. Sin.* **2009**, *35*, 748–758.
- Kankar, P.K.; Sharma, S.C.; Harsha, S.P. Fault diagnosis of ball bearings using machine learning methods. *Expert Syst. Appl.* **2011**, *38*, 1876–1886.
- Silva, S.; Costa, P.; Gouvea, M.; Lacerda, A.; Alves, F.; Leite, D. High impedance fault detection in power distribution systems using wavelet transform and evolving neural network. *Electr. Power Syst. Res.* **2018**, *154*, 474–483.
- Salahshoor, K.; Kordestani, M.; Khoshro, M.S. Fault detection and diagnosis of an industrial steam turbine using fusion of svm (support vector machine) and anfis (adaptive neuro-fuzzy inference system) classifiers. *Energy* **2010**, *35*, 5472–5482.
- Gustafsson, F. Statistical signal processing approaches to fault detection. *Annu. Rev. Control.* **2007**, *31*, 41–54.
- Wang, H.Q.; Song, Z.H.; Wang, H. Fault Detection Behavior Analysis of Pca-Based Process Monitoring Approach. *J. Chem. Ind. Eng.* **2002**, *53*, 297–301.
- Bae, B.; Kim, H.; Lim, H.; Liu, Y.; Han, L.D.; Freeze, P.B. Missing data imputation for traffic flow speed using spatio-temporal cokriging. *Transp. Res. Part Emerg. Technol.* **2018**, *88*, 124–139.
- Li, H.; Qin, L.; Chang, X.; Rong, J.; Ran, B.; Jia, L. Sensor layout strategy and sensitivity analysis for macroscopic traffic flow parameter acquisition. *Iet Intell. Transp. Syst.* **2017**, *11*, 212–221.

13. Wen, C.L.; Lv, F.Y.; Bao, Z.J.; Liu, M.Q. A Review of Data Driven-based Incipient Fault Diagnosis. *Acta Autom. Sin.* **2016**, *42*, 285–1299.
14. Xu, C.; Qu, Z.W.; Tao, P.F.; Jin, S. Methods of real-time screening and reconstruction for dynamic traffic abnormal data. *J. Harbin Eng. Univ.* **2016**, *37*, 211–217.
15. Xiao, Q.; Wang, D.J.; Liu, D. Abnormal traffic flow data detection based on wavelet analysis. In Proceedings of the MATEC Web of Conferences, Hong Kong, China, 26–27 April 2016.
16. Ngan, H.Y.T.; Yung, N.H.C.; Yeh, A.G.O. Outlier detection in traffic data based on the Dirichlet process mixture model. *Intell. Transp. Syst. Lett.* **2015**, *9*, 773–781.
17. Wong, C.H.M.; Ngan, H.Y.T.; Yung, N.H.C. Modulo-k Clustering based Outlier Detection for Large-scale Traffic Data. In Proceedings of the Int'l Conference IEEE Information Technology and Application (ICITA), Sydney, Australia, 1–4 July 2016.
18. Dang, T.T.; Ngan, H.Y.T.; Liu, W. Distance-based k-nearest neighbors outlier detection method in large-scale traffic data. In Proceedings of the IEEE International Conference on Digital Signal Processing, Singapore, 21–24 July 2015; pp. 507–510.
19. Cheng, Z.; Gao, X.; Tao, X.; Yuan, L.; Pang, Y. Fault detection and diagnosis strategy based on a weighted and combined index in the residual subspace associated with PCA. *J. Chemom.* **2018**, *32*, e2981.
20. Misra, M.; Yue, H.H.; Qin, S.J.; Cheng, L. Multivariate process monitoring and fault diagnosis by multi-scale PCA. *Comput. Chem. Eng.* **2002**, *26*, 1281–1293.
21. Choi, S.W.; Martin, E.B.; Morris, A.J.; Lee, I.B. Fault detection based on a maximum-likelihood principal component analysis (PCA) mixture. *Ind. Eng. Chem. Res.* **2005**, *44*, 2316–2327.
22. Bakshi, B.R. Multiscale PCA with application to multivariate statistical process monitoring. *Aiche J.* **1998**, *44*, 1596–1610.
23. Reddy, M.J.; Mohanta, D.K. A wavelet-fuzzy combined approach for classification and location of transmission line faults. *Int. J. Electr. Power Energy Syst.* **2007**, *29*, 669–678.
24. Fan, W.; Wang, X. An image compression method based on orthogonal wavelet packet transform. *Bol. Tec. Tech. Bull.* **2017**, *55*, 18–25.
25. Hancock, J.M.; Zvelebil, M.J. Principal-Components Analysis (PCA). In *Dictionary of Bioinformatics and Computational Biology*; John Wiley and Sons, Inc.: Hoboken, NJ, USA, 2004.
26. Kim, K.I.; Jung, K.; Kim, H.J. Face recognition using kernel principal component analysis. *IEEE Signal Process. Lett.* **2002**, *9*, 40–42.
27. Li, S.; Wen, J. A model-based fault detection and diagnostic methodology based on PCA method and wavelet transform. *Energy Build.* **2014**, *68*, 63–71.
28. Zhang, Y.; Ma, C. Fault diagnosis of nonlinear processes using multiscale KPCA and multiscale KPLS. *Chem. Eng. Sci.* **2011**, *66*, 64–72.



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).