mathematics

MDPI

Article

Abnormal Monitoring Data Detection Based on Matrix Manipulation and the Cuckoo Search Algorithm

Zhenzhu Meng [1], Yiren Wang [2,*], Sen Zheng [3], Xiao Wang [4], Dan Liu [1], Jinxin Zhang [1] and Yiting Shao [1]

1 School of Water Conservancy and Environment Engineering & Nanxun Innovation Institute, Zhejiang University of Water Resources and Electric Power, Hangzhou 310018, China; mengzhzh@zjweu.edu.cn (Z.M.); liudan@zjweu.edu.cn (D.L.); zhangjx@zjweu.edu.cn (J.Z.); shaoyt@zjweu.edu.cn (Y.S.)
2 School of Environment and Civil Engineering, Dongguan University of Technology & Guangdong Provincial Key Laboratory of Intelligent Disaster Prevention and Emergency Technologies for Urban Lifeline Engineering, Dongguan 523808, China
3 Laboratory of Environmental Hydraulics, Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland; sen.zheng@epfl.ch
4 Huai'an Hydraulic Surcey and Design Research Institute Co., Ltd., Huaian 223500, China; wangxiao-ha@bewg.net.cn
* Correspondence: wangyr@dgut.edu.cn

Abstract: Structural health monitoring is an effective method to evaluate the safety status of dams. Measurement error is an important factor which affects the accuracy of monitoring data modeling. Processing the abnormal monitoring data before data analysis is a necessary step to ensure the reliability of the analysis. In this paper, we proposed a method to process the abnormal dam displacement monitoring data on the basis of matrix manipulation and Cuckoo Search algorithm. We first generate a scatter plot of the monitoring data and exported the matrix of the image. The scatter plot of monitoring data includes isolate outliers, clusters of outliers, and clusters of normal points. The gray scales of isolated outliers are reduced using Gaussian blur. Then, the isolated outliers are eliminated using Ostu binarization. We then use the Cuckoo Search algorithm to distinguish the clusters of outliers and clusters of normal points to identify the process line. To evaluate the performance of the proposed data processing method, we also fitted the data processed by the proposed method and by the commonly used 3-σ method using a regression model, respectively. Results indicate that the proposed method has a better performance in abnormal detection compared with the 3-σ method.

Keywords: monitoring data; dam displacement; abnormal detection; matrix manipulation; Gaussian blur; Cuckoo Search algorithm

MSC: 68T10; 62P30; 62R99; 86-10; 86-11

check for updates

Citation: Meng, Z.; Wang, Y.; Zheng, S.; Wang, X.; Liu, D.; Zhang, J.; Shao, Y. Abnormal Monitoring Data Detection Based on Matrix Manipulation and the Cuckoo Search Algorithm. *Mathematics* **2024**, *12*, 1345. https://doi.org/10.3390/math12091345

Academic Editor: Radu Tudor Ionescu

Received: 9 April 2024
Revised: 26 April 2024
Accepted: 26 April 2024
Published: 29 April 2024

1. Introduction

Due to environmental and ageing effects, structural behaviours and material properties of dams often have changes compared to initial or designed values after running for years, which may affect the heath status of dams [1,2]. In practical engineering, to evaluate the real time running status of dams, engineers commonly install a series of monitoring devices inside the dam and monitor dam's structural parameters, such as displacement, seepage, and rotation. Once the deformation of the dam exceeds the safety value, the risk of the dam break problem would be very high when the reservoir is operated with high water level. Thus, analyzing the deformation status of the dam is important for evaluating the running status of the dam. Interpolating the displacement data of all monitoring point can provide the deformation field of the dam body. Therefore, displacement is the most crucial parameter [3,4]. Displacement monitoring data modeling is consider as one of the most effective methods to assess a dam's health status.

Previous researchers developed a numbers of displacement prediction models using monitoring data [5–7]. In addition to traditional statistical models, machine learning algorithms such as the neural network method [8], support vector machine method [9,10], and extreme learning machine method [11] were applied to displacement monitoring data modeling in recent years. Most previous studies put emphasis on improving the accuracy of displacement prediction models. As research progressed, the precision of these models has been fairly high.

The reliability of monitoring data modeling not only relies on the performance of prediction model but also depends on the quality of monitoring data [12–14]. However, measurement errors of monitoring device are unavoidable in practical engineering due to technical problems such as false reading [15]. Therefore, detecting the abnormal data of displacement monitoring data is of great importance for improving the reliability of displacement prediction models.

Previous studies have proposed different methods to detect outliers in a dataset [14,16–19]. In early studies, criterion-based methods such as Pauta criterion, Chauvenet criterion, and Grubbs criterion were used to detect outliers [20–22]. Each criterion has different usage conditions. Grubbs criterion is applicable to a dataset with few data, whereas Pauta criterion is applicable for dataset with more data. In recent years, statistical theories such as 3σ criterion have been commonly used to detect abnormal values in monitoring datasets. To increase the rate of outliers being detected, many studies have been conducted to enhance traditional methods [23–25]. For example, Zhao et al. improved the 3σ criterion using the minimum covariance determinant [26]. Song et al. developed an detection method based on the multi-variable panel data model and K-means clustering method [27]. Zhang et al. provided a multi-source information fusion model for outlier detection [28].

These methods exhibited fairly good performance in identifying gross errors. One disadvantage of these methods is that they are mostly computationally complex. Moreover, the outlier detection often depends on time variation tendencies without considering the fluctuations of environmental factors, such as water level and external temperature [29]. In addition, the performance of outliers detection is affected by the setting of threshold, which relies on artificial selection, which may lead to missing judgment and misjudgment problems.

To overcome the shortcomings of these methods, we proposed a outlier detection method which combines matrix manipulation and the Cuckoo Search algorithm to deal with the abnormal dam displacement monitoring data. The principle of the proposed method is that the process line of monitoring data should be continuous while the outliers deviates from the process line [30]. We first generated the scatter plot of the original monitoring dataset, which includes clusters of isolated outliers, normal points, and clusters of outliers. The objective of the proposed method is to identify isolated outliers and clusters of outliers from the scatter plots. For the matrix manipulation method, Gaussian blur and Ostu binarization are used to detect isolated outliers [31–33]. We then applied the Cuckoo Search algorithm, which imitates the habit of brood parasitism, to distinguish clusters of normal points and clusters of outliers [34]. To ensure the efficiency of outliers detection, we implement the process of matrix detection and CS algorithm cyclic until the detection results converges. We also compared the abnormal detection performance of the proposed method with the commonly used 3σ method.

This paper is organized as follows. Section 2 presents the principles of the proposed method, which combines matrix manipulation and the Cuckoo Search algorithm. Section 3 introduces the dataset. The displacement monitoring data of the dam at Jinping-I hydropower station is selected as the dataset. The detection results of the proposed method are presented in Section 4. Comparisons of the proposed model with 3σ criterion are also discussed. Concluding remarks complete the paper in Section 5.

2. Data Processing Method Using Matrix Manipulation and the Cuckoo Search Algorithm

This section presents the mathematical details of the abnormal data processing method on the basis of matrix manipulation and Cuckoo Search algorithm. We first generate a scatter plot of the monitoring data. Once the scatter plot has been drawn, we then consider the scatter plot as an image and export the matrix of the image. Then, the matrix can be pre-processed using Gaussian blur and Ostu binarization. The gray scales of isolated outliers are reduced using Gaussian blur. Then, the isolated outliers are eliminated using Ostu binarization. We then use Cuckoo Search algorithm to distinguish the clusters of outliers and clusters of normal points, so as to identify the process line. Figure 1 shows the flowchart of the proposed abnormal data processing method.

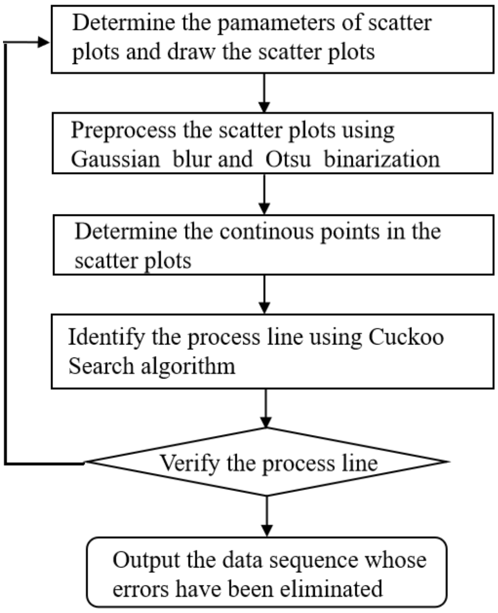

Figure 1. The flowchart of the proposed method.

2.1. Data Pre-Processing Using Gaussian Blur and Ostu Binarization

The Principe of the pre-processing method is to consider the plotted data sequence as an image (i.e., matrix), and identify the outliers in the plot using filters (Gaussian blur and Ostu binarization). Figure 2 shows the linear plot and scatter plot of an example data sequence. It can see from the figure that outliers in scatter plot is easier to be separated from the process line as compared with the linear plot. Thus, the first step of data pre-processing is to generate the scatter plot of the monitoring data sequence. Once the scatter plot has been drawn, we then consider the scatter plot as an image and export the matrix of the image. Then, noises can be reduced using various image-processing techniques, such as Gaussian blur and binary threshold. Scatter plot of monitoring data include isolate outliers, clusters of outliers, and clusters of normal points. At the pre-processing stage, most of the isolate outliers in the matrix can be detected and eliminated using Gaussian blur and Ostu binarization.

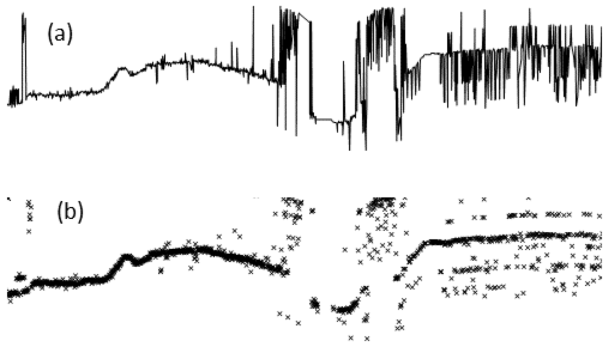

Figure 2. Plots of a data sequence: (**a**) linear plot, (**b**) scatter plot.

The Gaussian blur feature is obtained by smoothing an image using a Gaussian function to reduce the noise level. It can be considered as a nonuniform low-pass filter that preserves low spatial frequency and reduces image noise and negligible details in an image. From a mathematical perspective, the Gaussian blur process is the convolution of a matrix with a normal distribution. Convolving an image with a circular box blur will generate a more precise out-of-focus rendering effect. Since the Fourier transform of a Gaussian function is another Gaussian function, Gaussian blur is a low-pass filter for images. It is typically achieved by convolving an image with a Gaussian kernel. The Gaussian kernel filtering function $G(x, y)$ follows a two-dimension Gaussian distribution:

$$G(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{1}$$

where σ is the standard deviation of the Gaussian distribution. It controls the variance around a mean value of the Gaussian distribution, which determines the extent of the blurring effect around a pixel. With an increase in σ, the high-frequency information content reduces around the pixel. x, y are the coordinates of neighbor pixels.

The Gaussian weighted matrix W_i is:

$$W_i = \begin{bmatrix} \frac{G(-1,1)}{\sum_{j=1}^{n} G_j} & \frac{G(0,1)}{\sum_{j=1}^{n} G_j} & \frac{G(1,1)}{\sum_{j=1}^{n} G_j} \\ \frac{G(-1,0)}{\sum_{j=1}^{n} G_j} & \frac{G(0,0)}{\sum_{j=1}^{n} G_j} & \frac{G(0,1)}{\sum_{j=1}^{n} G_j} \\ \frac{G(-1,-1)}{\sum_{j=1}^{n} G_j} & \frac{G(-1,0)}{\sum_{j=1}^{n} G_j} & \frac{G(-1,1)}{\sum_{j=1}^{n} G_j} \end{bmatrix} \tag{2}$$

where the central pixel is considered as the origin of the coordinate and n is the sum of surrounding pixels and the central pixel. The relation between the gray scale of the central pixel $Gray_c'$ and the gray scale of the surrounding pixels $Gray_i$ can be written as:

$$Gray_c' = \sum_{i=1}^{n} W_i \cdot Gray_i \tag{3}$$

The Gaussian blur filter provides gradual smoothing and preserves the edges better than any other mean filter. We have used Gaussian blur to reduce the high-frequency components. The size of the Gaussian kernel depends on the noise level in the image. If the kernel size is too large, small features within the image may get suppressed, and the image may look blurred. If the kernel size is too small, eliminating the noises within the image will be compromised.

Ostu binarization is often used to separate intra-image pixels into two parts and determine the threshold of the two parts. This algorithm generate a binary image that helped in displaying the desired scatter areas. The binarization is performed on the mask using an elliptical structuring element to smooth the contour of the scatters, break narrow isthmuses that connected the scatters, remove the outlier pixels, and eliminate thin protrusions from the scatters. For a gray image, $G = \{i | i = 1, 2, \ldots, 255\}$ represents the possible set of gray scales, $P = \{n | n = 1, 2, \ldots, N\}$ denotes the set of all pixels. The threshold of the gray scale T_G can be expressed as:

$$T_G = \left\{ T_G | max\left(\sigma^2\right) \right\} \tag{4}$$

where σ^2 denotes the variance between each part and can be written as:

$$\sigma^2 = q_1[1 - q_1][\mu_1 - \mu_2]^2 \tag{5}$$

with:

$$q_1 = \sum_{i=0}^{t} P_i, \quad \mu_1 = \sum_{i=0}^{t} iP_i/q_1, \quad \mu_2 = \sum_{i=t+1}^{255} iP_i/(1 - q_1) \tag{6}$$

where N_i has a number of pixels with gray scale i, P_i is the ratio of N_i to N, q_1 is the ratio of number of pixels with gray scale lower than T_G and N, μ_1 the mean value of the gray scale of pixels in $P_1 = \{n | G_n \leq T_G\}$, μ_2 is the mean value of gray scale of pixels in $P_2 = \{n | G_n > T_G\}$, and G_n is the gray scale of n^{th} pixel.

2.2. Process Line Identification Using Cuckoo Search Algorithm

Using Gaussian blur and Ostu binarization, the scatter plot is processed to be a matrix with clusters of pixel. We define the set including all clusters of pixel as C whose expression can be written as:

$$C = \{C_1, C_2, \ldots, C_n\}(n = N) \tag{7}$$

The set of all possible combinations of the clusters of pixels ξ can be expressed as:

$$\xi = \{Comb_i | \forall C_i \in Comb_i, C_i \in C\} \tag{8}$$

The objective function of process line is:

$$\begin{aligned} F(\zeta) = a_1 \sum_{j=1}^{s} \left| x_j^l - x_j^f \right| + a_2 \sum_{j=1}^{s} \left| y_j^{max} - y_j^{min} \right| \\ - b \sum_{j=1}^{s-1} \left| y_{j+1}^l - y_j^f \right| \end{aligned} \tag{9}$$

where a_1, a_2 are gain coefficients, b is the loss coefficient, (x_j^f, y_j^l) denotes the coordinate of the first pixels, (x_j^l, y_j^l) shows the coordinate of the last pixel, and y_j^{max} and y_j^{min} are the maximum and minimum threshold of vertical coordinates. In addition, each C_i in a $Comb_i$ obey:

$$x_{j-1}^b < x_j^b < x_{j+1}^b, j = 2, 3, \ldots, s - 1 \tag{10}$$

For arbitrary two pixel clusters C_a and C_b in $Comb_i$, $x_a^l < x_b^l$ when $x_b^f > x_a^f$. Then, we can consider the process line identification as an optimization question:

$$\begin{aligned} \max \quad & F(\zeta) \\ s.t. \quad & x_{j-1}^b < x_j^b < x_{j+1}^b, \quad j = 2, 3, \ldots, s - 1 \\ & x_a^l < x_b^l, \end{aligned} \tag{11}$$

The Cuckoo Search algorithm is a stochastic optimization model which is developed based on the brood parasitism of cuckoo birds. Figure 3 shows the flowchart of the Cuckoo Search algorithm.

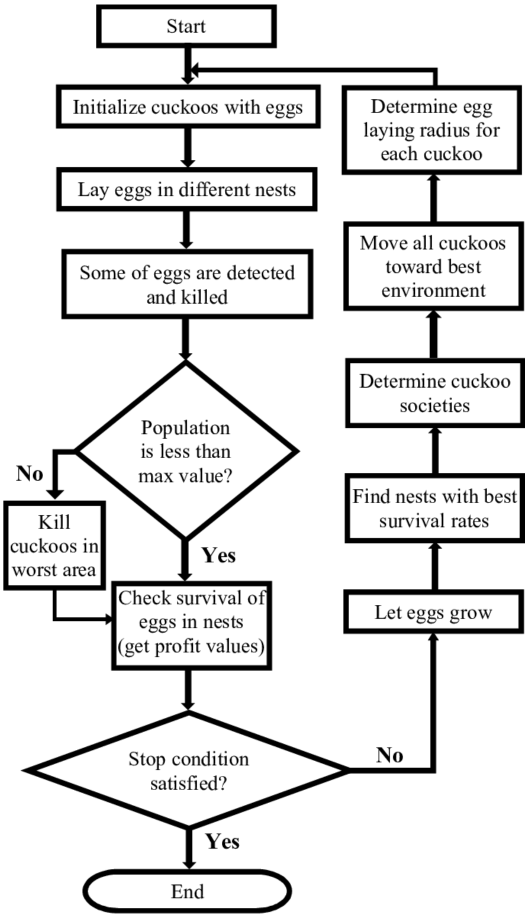

Figure 3. The flowchart of the Cuckoo Search algorithm.

The algorithm follows three principal rules: (a) Each cuckoo lays only one egg each time and picks one nest to place the egg randomly; (b) The best host nests with the highest-quality eggs is kept in the following generation; (c) The number of available host nests is fixed. Details of the algorithm are as follows:

Step 1: Develop the objective function and determine inputs including the threshold, number of iterations, objective accuracy, etc.

Step 2: Establish the initial generation $x_1, x_2, \ldots, x_N$ randomly. Each cuckoo denotes a dataset of attribute values of continuous point, whose expression can be written as:

$$\begin{aligned} N &= \left[\begin{array}{c|c} x_1 & F(x_1) \\ x_2 & F(x_2) \\ \cdots & \cdots \\ x_N & F(x_N) \end{array}\right] \\ &= \left[\begin{array}{cccccc|c} x_1^e & x_1^b & y_1^{min} & y_1^{max} & y_1^e & y_1^b & F(x_1) \\ x_2^e & x_2^b & y_2^{min} & y_2^{max} & y_2^e & y_2^b & F(x_2) \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ x_N^e & x_N^b & y_N^{min} & y_N^{max} & y_N^e & y_N^b & F(x_N) \end{array}\right] \end{aligned} \tag{12}$$

where N is the number of cuckoos. Each cuckoo represents a set of attribute values of continuous points. The best cuckoo x_b^t and the objective function for each cuckoo are determined in this step.

Step 3: Implement the Levy flight. The expression of Levy flight is:

$$x_i^{t+1} = x_i^t + \alpha \oplus Levy(\lambda)(i = 1, 2, \cdots, n) \tag{13}$$

where α denotes the step size, $\oplus$ denotes the entry-wise multiplications, x_i^t and x_i^{t+1} are the positions of t^{th} and $t+1^{th}$ generations of cuckoos, respectively, and $Levy(\lambda)$ is a random searching vector which follows Levy distribution:

$$\begin{aligned} Levy(\lambda) &\sim \frac{\phi u}{|v|^{1/\beta}} \phi \\ &= \left\{ \frac{\Gamma(1+\beta) \times \sin(\pi \times \beta/2)}{\Gamma\{[(1+\beta)/2] \times \beta \times 2^{(\beta-1)/2}\}} \right\}^{1/\beta} \end{aligned} \tag{14}$$

where Γ is the Gamma function, u and v are random numbers which follow Gaussian distributions, and β is the parameter of Levy flight. The nests location for the next generation of cuckoos x_i^{t+1} is given by:

$$x_i^{t+1} = x_i^t + alpha_0 \frac{\phi \times u}{|v|^{1/\beta}} \left(x_i^t - x_b^t\right) \tag{15}$$

where x_b^t is the best nest in t^{th} generation and α_0 is the scaling factor.

Step 4: Eliminate the alien eggs with a probability of $P_a \in [0, 1]$. The mathematical expressions can be written as:

$$x_i^{t+1} = \begin{cases} x_i^t + r \cdot \left(x_{r_1}^t - x_{r_2}^t\right), & if\ r < P_a \\ x_i^t, & otherwise \end{cases} \tag{16}$$

where r is a random number in the range of 0 to 1 and $x_{r_1}^t$ and $x_{r_2}^t$ are two random nest locations in the t^{th} generation, respectively.

Step 5: Determine the objective function of renewal nest locations as well as the optimal cuckoo of $t+1^{th}$ generation x_b^{t+1}. Here, the smaller value between x_b^{t+1} and x_b^t is kept as the $t+1^{th}$ optimal cuckoo.

Step 6: Repeat Step 3 and 4 until the number of iteration other termination criteria reach the set values.

Step 7: To enhance the effect of detecting outliers, the procedure combination needs to be implemented cyclic until the result satisfies the requirement. The threshold of R_y and R'_y are:

$$R_{ymax} = max\{y_1, y_2, \ldots, y_m\}, \quad R_{ymin} = min\{y_1, y_2, \ldots, y_m\} \tag{17}$$

$$R'_{ymax} = max\{y_1, y_2, \ldots, y_n\}, \quad R'_{ymin} = min\{y_1, y_2, \ldots, y_n\} \tag{18}$$

where m counts the number of the pixel clusters in raw data and n counts the number of pixel clusters in processed data. The processed data can be validated once $R_{ymax} = R'_{ymax}$ and $R_{ymin} = R'_{ymin}$.

Figure 4 shows the example of process line detection.

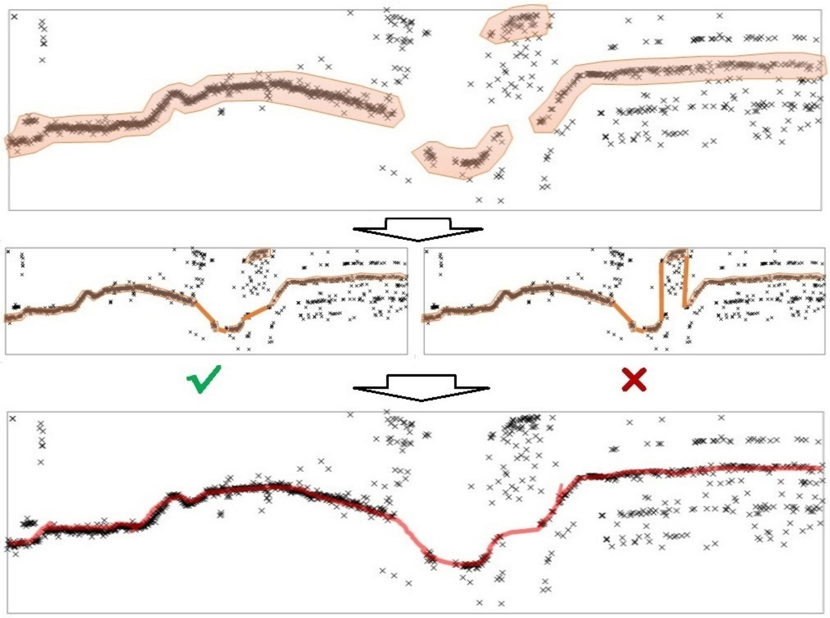

Figure 4. The procedure of process line detection.

3. Dataset

In this study, we used the monitoring data of the dam at Jinping-I hydropower station as the dataset. The dam is located at Yalong River, Sichuan Province, China, and famous as being one of the highest concrete arch dam worldwide. The elevation of the dam's crest is 1885 m and that of the dam's foundation is 1580 m. The normal impounded water level is 1880 m. Figure 5a,b show the geological location and the photo of the dam, respectively.

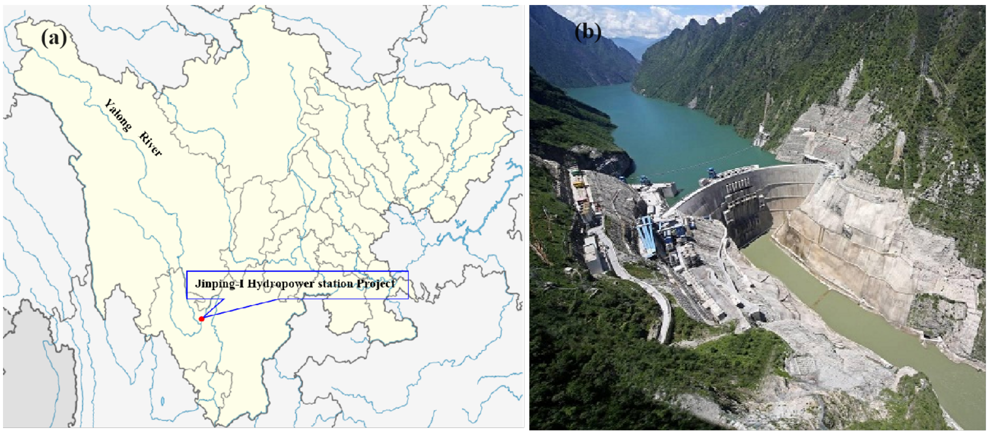

Figure 5. (**a**) Location of the Jinping-I hydroppower station; (**b**) Photo of the Jinping-I arch dam.

As shown in Figure 6, the monitoring points were well distributed on the cross section of the dam. As the dataset, we used displacement monitoring data of six selected monitoring points which located on three different plumb lines. The selected monitoring points PL11-1, PL11-3, PL13-1, PL13-3, PL16-1, and PL16-3 are highlighted by red square in Figure 6. The data sequence was separated into two parts, where 80% of the dataset was used to test the detection ability (data during the period 1 July 2017 to 30 February 2018) and 20% of the dataset was used to evaluate the prediction performance (from 1 March 2018 to 30 May 2018).

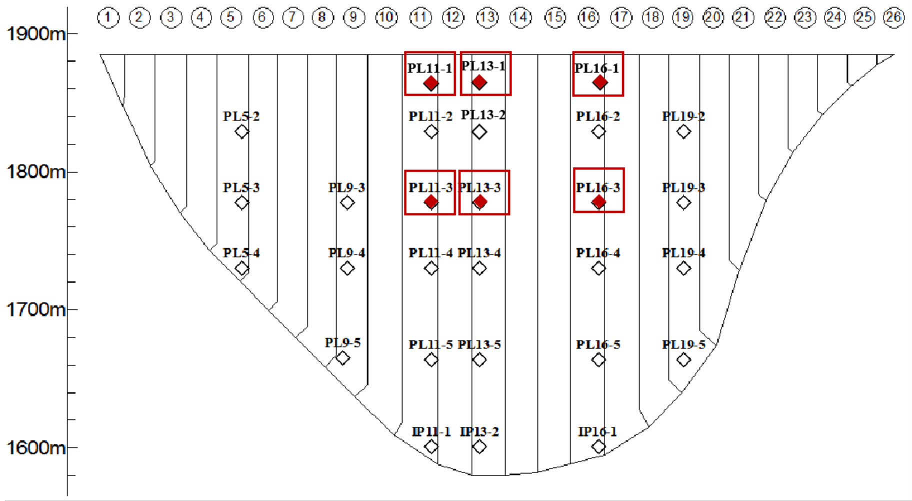

Figure 6. The distribution of monitoring points (red boxes denote the selected monitoring points).

4. Results

4.1. Optimal Settings of the Scatter Plot of the Original Data

For the data processing method based on matrix manipulation and Cuckoo Search algorithm, the first step is to generate a scatter plot of the original data. Then, the scatter plot is considered an image and the matrix of the image is exported. Attributes of scatter such as shape and size affect the performance of matrix manipulation including Gaussian blur and Ostu binarization. Thus, we first determine the optimal settings of attributes of scatters. Figure 7 shows the stack of patterns with different shapes including square, cross, and isscross. All these three patterns have nine pixels. It can be seen from the figure that when the patterns stacks together, the cross pattern could keep more information as compared with the square pattern and isscross pattern.

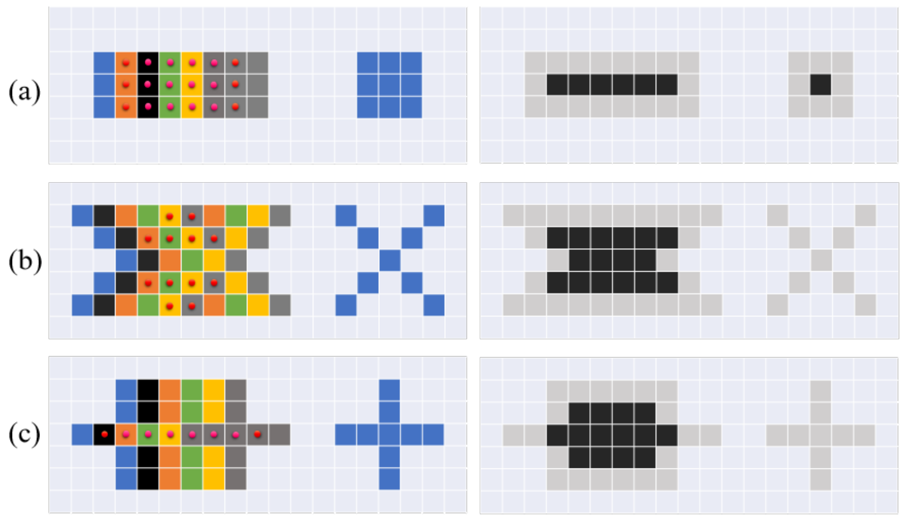

Figure 7. Stack of patterns with different shapes: (**a**) square, (**b**) cross, and (**c**) isscross.

To obtain the optimal settings of the attributes of scatters, we constructed a scatter plot of a sample data sequence using four different shapes of scatters (circle, square, cross, and isscross) with the same size, and compared the filtering performance of Gaussian blur and Ostu binarization. Figure 8a–d exhibit the plots processed by Gaussian blur and Ostu binarization using a circle, square, cross, and isscross as scatters, respectively.

It can be seen from the figure that the data processing using cross as the scatter shape has the best performance in outlier detection. Using cross as the scatter shape, more outliers are eliminated, and the clusters of continuous points are identified. Comparing with the square and isscross, cross patterns have more dispersing distributed pixels. Using cross as the scatter shape, the gray scale of isolated outliers can be reduced more intensely by Gaussian blur, and thus, the outlier can be easier eliminated by the filter. In addition, using a circle and square as scatter shapes, the outliers detection performance is significantly worse than that of cross and isscross. The detection performance of scatter plots using a cross is slightly higher than using isscross. Therefore, we selected cross as the shape of scatter in the data preprocessing using Gaussian blur and Ostu binarization.

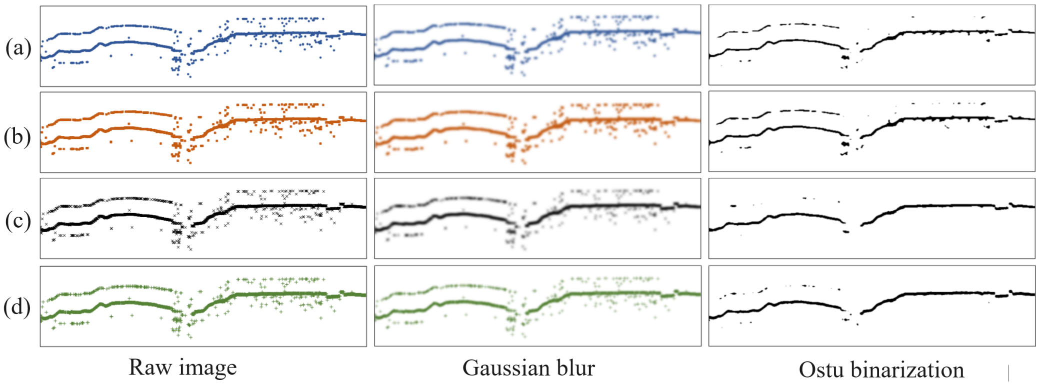

Figure 8. Gaussian blur and Ostu binarization of a plot using different shapes of scatters: (**a**) circle, (**b**) square, (**c**) cross, (**d**) isscross.

As the shape of scatters have been selected, the size of the scatters should be determined. To determine the optimal set of scatter size, we compare the outlier detection performance of scatter plot using cross scatter with four different sizes. As shown in Figure 9, the number of pixels of these four sizes are 5, 9, 13, and 17, respectively. Comparing Figure 9 with Figure 8a,b, using a cross with five pixels, the detection performance is similar to those using a circle and square as scatter shapes. This is because when the size of the cross is small, the pixels are centrally distributed, that is, the micro shape is similar to a circle and square. When the size of the cross is increased, the scatter is less centralized distributed. The cross with nine pixels has the best performance in outlier detection, that is, more outliers are detected and eliminated after the data processing. Therefore, a nine-pixels cross is the optimal size and shape.

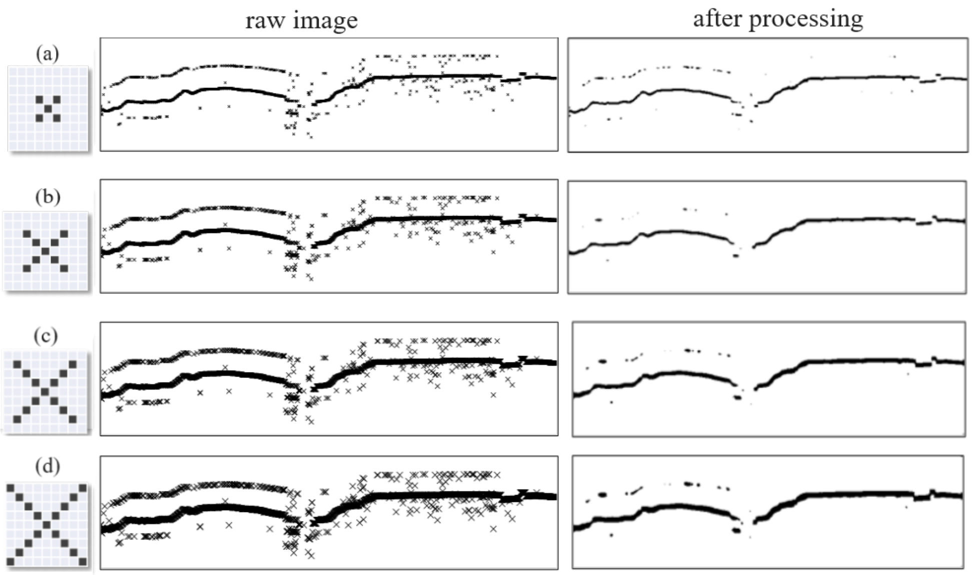

Figure 9. Gaussian blur and Ostu binarization processing scatter plots using a cross with different sizes: (**a**) 5 pixels, (**b**) 9 pixels, (**c**) 13 pixels, and (**d**) 17 pixels.

4.2. Results of Abnormal Data Detection Based on the Proposed Method

According to the analysis in Section 4.1, a cross of nine pixels is selected as the scatter in the pre-processing procedure using Gaussian blur and Ostu binarization. For the process line identification using the Cuckoo Search algorithm, Levy flight β is set to 2.0 and the discard probability P_a is set to 0.2.

Using the sample data sequence as an example, Figure 10 shows the whole procedure of the proposed method, which combines matrix manipulation and the Cuckoo Search algorithm. As presented in Section 2, the processing procedure of the proposed method consists of three steps: (1) Gaussian blur, (2) Ostu binarization, and (3) process line identification using the Cuckoo Search algorithm.

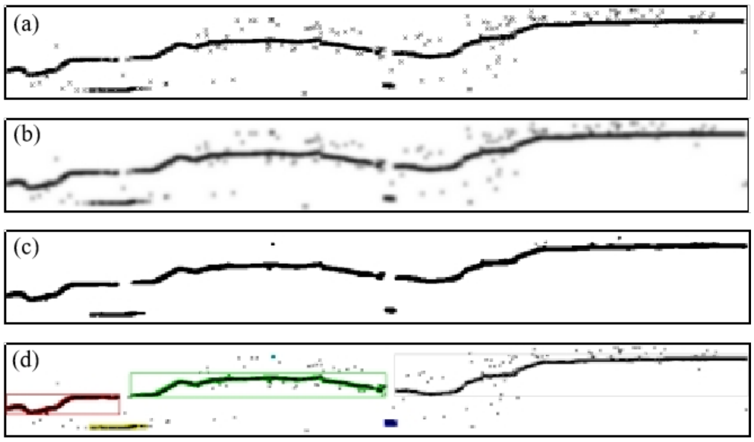

Figure 10. The error processing of the sample data sequence using the proposed method: (**a**) raw image, (**b**) Gaussian blur, (**c**) Ostu binarization, and (**d**) process line identification.

Displacement monitoring datasets of the six selected monitoring points installed in the dam at Jinping-I hydropower station are used to validate the propose method. Here, in order to verify the processing ability of the proposed method, we added the numbers of outliers in the original dataset, so as to increase the detection difficulty. We then use the proposed method to detect and eliminate outliers in the data sequence. Table 1 shows the total data number N_t and number of outliers N_d detected by the proposed method for each monitoring point.

Table 1. The total data number N_t and number of outliers N_o detected by the proposed method.

Monitoring Points	PL11-1	PL11-3	PL13-1	PL13-3	PL16-1	PL16-3
N_t	830	788	872	820	860	860
N_o	86	91	75	76	70	89

4.3. Comparison of the Proposed Method with 3-σ Method

To evaluate the efficiency of the proposed method, we process the same dataset using the 3-σ method, which is a classical method in outlier detection, combining multidimensional regression and 3-σ criterion.

The factors dominating the displacement of dam includes three components: the hydrostatic component δ_H, the temperature component δ_T, and the aging component δ_θ. The expressions of δ_H, δ_T, and δ_θ are as follows:

$$\delta_H = \sum_{i=0}^{4} a_i H^i \tag{19}$$

$$\delta_T = \sum_{j=1}^{4} b_j T_j \tag{20}$$

$$\delta_\theta = c_1\theta + c_2 \ln\theta \tag{21}$$

where H is the upstream water level, T_j is the external temperature, and $\theta = \frac{t}{100}$ and t is time. The displacement δ can be written as:

$$\delta = a_0 + \sum_{i=1}^{4} a_i H^i + \sum_{j=1}^{4} b_j T_j + c_1\theta + c_2 \ln\theta \tag{22}$$

where a_0, a_i, b_j, c_1, and c_2 are the coefficients of explanatory variables and can be solved using the least square regression method.

According to the principle of the 3-σ method, the probability of the absolute error between modeling data and original data $|\varepsilon|$ less than 3σ is 99.7%. Here, σ is the residual standard deviation whose expression can be written as:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}\left[Y(i) - \widehat{Y}(i)\right]^2}{n-k-1}} \tag{23}$$

where n is the data number of the dataset, k is the degree of freedom of the model, and $Y(i)$ and $\widehat{Y}(i)$ denote the monitoring data sequence and modeling data sequence, respectively.

We suppose that outliers should have a large deviation from the modeling displacement. Thus, 3σ can be used as the threshold for outlier detection. That is, the monitoring data are regarded as outliers once $|\varepsilon|$ exceeds 3σ. The mathematical descriptions can be written as:

$$\begin{cases} |\varepsilon| > 3\sigma & Y(i) \text{ is outlier} \\ |\varepsilon| \leq 3\sigma & Y(i) \text{ is valid value} \end{cases} \tag{24}$$

where:

$$|\varepsilon| = \left|Y(i) - \widehat{Y}(i)\right| \tag{25}$$

Figure 11 compares the outlier detection results obtained by the 3-σ method and the proposed method. Black dots present the original data sequence without outlier detection, and the black dots without marks present outliers detected by the 3-σ method and the proposed method. Red cross and yellow square denote outliers detected by 3-σ method and the proposed method, respectively. For the 3-σ method, only outliers located in the area between monitoring data and modeling data exceeds 3-σ can be detected. Outliers located in the surrounding areas of the process line can not be eliminated. Compared with the 3-σ method, the proposed method has a better performance. It can detect most outliers existing in all these six data sequences.

We defined the ratio of number of detected outliers N_d to number of outliers N_o as detection ratio r_d:

$$r_d = \frac{N_d}{N_o} \tag{26}$$

Table 2 exhibits the number of detected outliers detected N_d and the detection ratio r_d of each monitoring point for the proposed method and 3σ method. The average of r_d is 32.22% for the 3-σ method and 9% for the proposed method. Using the 3-σ method, r_d ranges between 26.37% and 51.42% for all monitoring points. Using the proposed method, r_d ranges between 87.20% and 100% for all monitoring points. In general, the proposed method has a significantly higher performance in outlier detection than the 3-σ method.

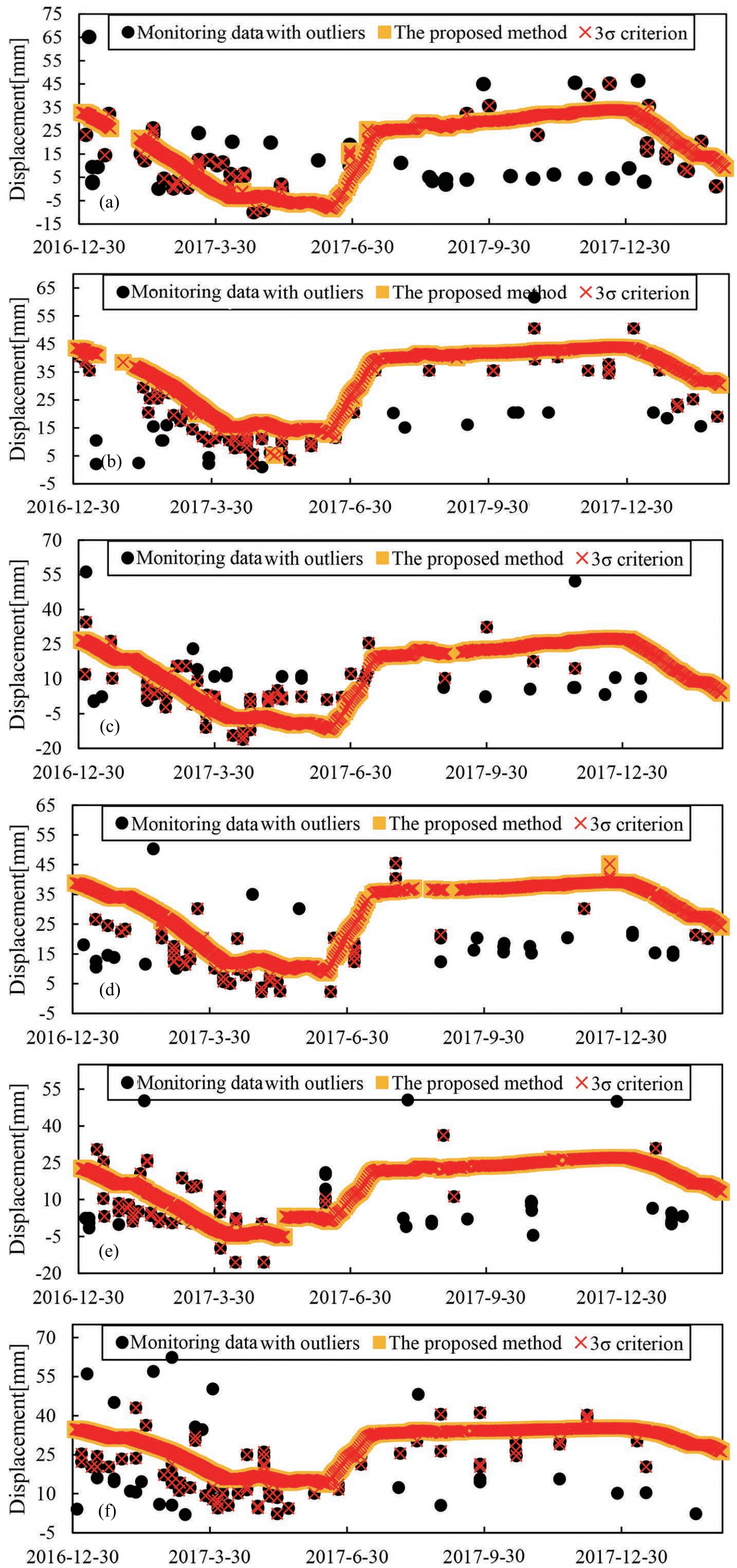

Figure 11. The results of outlier detection of the proposed method and 3σ method of: (**a**) PL11-1, (**b**) PL11-3, (**c**) PL13-1, (**d**) PL13-3, (**e**) PL16-1, and (**f**) PL16-3.

Table 2. N_d and r_d of the 3-σ method and proposed method for each monitoring point.

Monitoring Points	The Proposed Method		3σ Method	
	N_d	r_d(%)	N_d	r_d(%)
PL11-1	75	87.20	31	36.04
PL11-3	84	92.31	24	26.37
PL13-1	72	96.00	29	38.66
PL13-3	72	94.73	28	36.84
PL16-1	70	100.00	36	51.42
PL16-3	88	98.87	31	34.83

4.4. Regression Model Development Using Processed Data

The regression models are developed using data processed by the 3-σ method and proposed method, to verify the efficiency of outlier detection for monitoring data modeling. The principal expression of the regression model is:

$$\hat{y} = a_0 + a_1x_1 + a_2x_2 + \cdots + a_ix_i \tag{27}$$

where $\hat{y}$ is the modeling data, x_i is the explanatory variables, and a_i is the coefficients of explanatory variables. x_i consists of the three components introduced in Section 4.3: the hydrostatic component δ_H, the temperature component δ_T, and the aging component δ_θ. The coefficients of explanatory variables can be solved using the ordinary least square method.

Figure 12 exhibits the fitting results using data processed by the 3-σ method and proposed method. The displacement modeled using data processed by both these two methods are fitted well with the monitoring data. In general, the prediction results obtained using both datasets are quite similar to the observed data.

The coefficient of determination R^2 and standard deviation RMSE are selected as indicators, to quantify the fitting performance using these two dataset and the predicting accuracy of the model. The equations of R^2 and RMSE are as follows:

$$R^2 = \frac{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y}_i)^2} \tag{28}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{n}} \tag{29}$$

where $\overline{y}_i$ is the average of the monitoring data, $\widehat{y}_i$ is the modeling data, y_i is the monitoring data, and n is the total data.

Table 3 exhibits the R^2 and RMSE of these two dataset for each monitoring point. R^2 exceeded 0.9 for both datasets, which indicates that the regression model can be validated. R^2 ranges between 0.9474 and 0.9854 using the dataset processed by the 3-σ method, ranges between 0.9933 and 0.9998 using dataset processed by the proposed method. It can be noted that the regression model has a better fitting performance using the dataset processed by the proposed method. The regression model developed using the dataset with fewer outliers performs better in prediction accuracy.

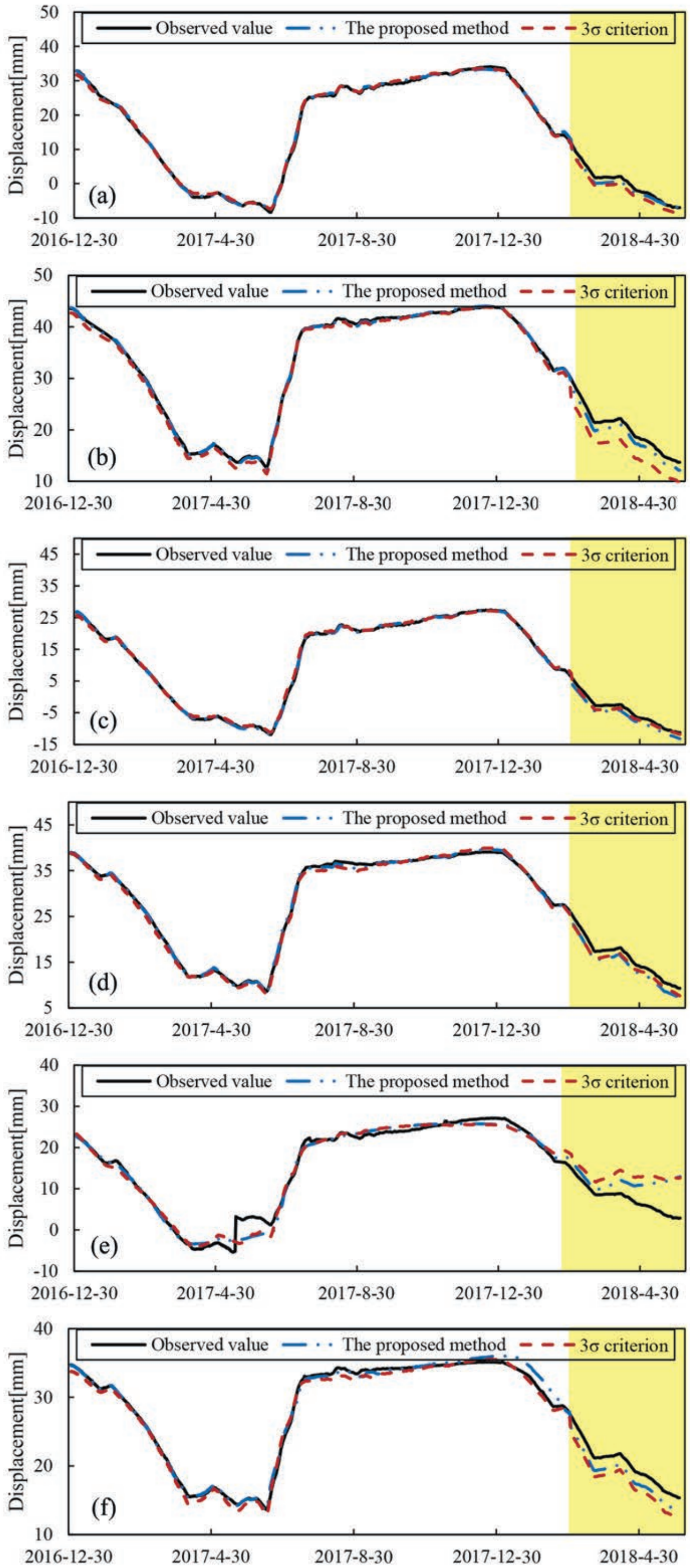

Figure 12. The regression model developed using dataset processed by the 3-σ method and proposed method: (**a**) PL11-1, (**b**) PL11-3, (**c**) PL13-1, (**d**) PL13-3, (**e**) PL16-1, and (**f**) PL16-3.

Table 3. R^2 and RMSE of the regression models using the dataset processed by the 3-σ method and proposed method.

Monitoring Points	R^2		RMSE	
	The Proposed Method	**3-σ Method**	**The Proposed Method**	**3-σ Method**
PL11-1	0.982	0.954	0.943	2.371
PL11-3	0.983	0.959	0.538	2.228
PL13-1	0.998	0.941	0.228	2.274
PL13-3	0.993	0.974	0.393	2.213
PL16-1	0.933	0.962	1.236	2.734
PL16-3	0.992	0.947	0.304	2.561

5. Conclusions

Displacement monitoring data analysis is an effective method to evaluate the running status of dams. Measurement error greatly affects the accuracy of monitoring data modeling. In order to process the abnormal dam displacement monitoring data, we proposed a data processing method by combining matrix manipulation and the Cuckoo Search algorithm.

In this paper, we first generate a scatter plot of the original monitoring data. Once the scatter plot has been drawn, we then consider the scatter plot as an image and export the matrix of the images. The matrix consists of isolated outliers, clusters of outliers, and clusters of normal data. At the pre-processing stage, the isolated outliers are detected and eliminated using Gaussian blur and Ostu binarization. Gaussian blur reduce the gray scales of isolated outliers, and Ostu binarization eliminate these isolated outliers from the matrix. Using the Cuckoo Search algorithm, we search the optimal series of clusters for determining the process line.

The proposed method is validated using the displacement monitoring data of the dam at Jinping-I hydropower station. By comparing the pre-processed results obtained by different sets of scatter plot, the scatter plots of nine-pixels cross is used in this study. The ratio of outlier detected r_d using the proposed method is over 85% for each monitoring point, and it is significantly higher than that of the 3-σ method. In addition, we regress the processed dataset and original dataset using a statistical model, respectively. The results indicate that the regression model fitted with data pre-processed by the proposed model has a better performance compared with the regression model using the original dataset and dataset pre-processed using the 3-σ method.

The proposed method provides a novel solution for detecting outliers in time series data with continuous characteristics. Engineering application of the method in this article is to detect abnormal data in monitoring data of dam displacement. The proposed method is not applicable for datasets without continuous characteristics, i.e., time series data without varying patterns or time-invariance data. One future direction of this study is to increase the engineering applications of the proposed method. The application can be extended to other structures, such as bridges, slopes, etc. In addition, this study introduced the image processing method into abnormal data detection. For both the image processing part and detection part, we selected mutual and routine methods; the detection accuracy can be further improved if we use other high performance methods. Therefore, future studies will need to improve the abnormal detecting performance by introducing a high-ability image processing method and searching algorithm.

Author Contributions: Conceptualization, S.Z.; methodology, S.Z. and Z.M.; validation, Z.M. and Y.W.; formal analysis, X.W., D.L., J.Z. and Y.S.; writing–original draft preparation, Z.M.; writing—review and editing, Y.W.; project administration, Z.M.; funding acquisition, Z.M. and D.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Joint Funds of the Zhejiang Provincial Natural Science Foundation of China (Grants Nos. LZJWY24E090005, LZJWY22E090003), University-Level Key Course of Zhejiang University of Water Resources and Electric Power (Grant No. ZDKC202319), and Nanxun Scholars Program for Young Scholars of ZJWEU (Grant No. RC2023021192).

Data Availability Statement: Data are available upon request.

Conflicts of Interest: Author Xiao Wang was employed by the company Huai'an Hydraulic Surcey and Design Research Institute Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. The Huai'an Hydraulic Surcey and Design Research Institute Co., Ltd. had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Kim, Y.S.; Kim, B.T. Prediction of relative crest settlement of concrete-faced rockfill dams analyzed using an artificial neural network model. *Comput. Geotech.* **2008**, *35*, 313–322. [CrossRef]
2. De Sortis, A.; Paoliani, P. Statistical analysis and structural identification in concrete dam monitoring. *Eng. Struct.* **2007**, *29*, 110–120. [CrossRef]
3. Kao, C.Y.; Loh, C.H. Monitoring of long-term static deformation data of Fei-Tsui arch dam using artificial neural network-based approaches. *Struct. Control Health Monit.* **2013**, *20*, 282–303. [CrossRef]
4. Dardanelli, G.; La Loggia, G.; Perfetti, N.; Capodici, F.; Puccio, L.; Maltese, A. Monitoring displacements of an earthen dam using GNSS and remote sensing. In Proceedings of the Remote Sensing for Agriculture, Ecosystems, and Hydrology XVI, Amsterdam, The Netherlands, 22–25 September 2014 ; Volume 9239, pp. 574–589.
5. Wu, Z. *Safety Monitoring Theory and Its Application of Hydraulic Structures*; Higher Education: Beijing, China, 2003.
6. Bukenya, P.; Moyo, P.; Beushausen, H.; Oosthuizen, C. Health monitoring of concrete dams: A literature review. *J. Civ. Struct. Health Monit.* **2014**, *4*, 235–244. [CrossRef]
7. Leger, P.; Leclerc, M. Hydrostatic, temperature, time-displacement model for concrete dams. *J. Eng. Mech.* **2007**, *133*, 267–277. [CrossRef]
8. Mata, J. Interpretation of concrete dam behaviour with artificial neural network and multiple linear regression models. *Eng. Struct.* **2011**, *33*, 903–910. [CrossRef]
9. Hipni, A.; El-shafie, A.; Najah, A.; Karim, O.A.; Hussain, A.; Mukhlisin, M. Daily forecasting of dam water levels: Comparing a support vector machine (SVM) model with adaptive neuro fuzzy inference system (ANFIS). *Water Resour. Manag.* **2013**, *27*, 3803–3823. [CrossRef]
10. Hariri-Ardebili, M.A.; Pourkamali-Anaraki, F. Support vector machine based reliability analysis of concrete dams. *Soil Dyn. Earthq. Eng.* **2018**, *104*, 276–295. [CrossRef]
11. Kang, F.; Liu, J.; Li, J.; Li, S. Concrete dam deformation prediction model for health monitoring based on extreme learning machine. *Struct. Control Health Monit.* **2017**, *24*, e1997. [CrossRef]
12. Avendano-Valencia, L.D.; Fassois, S.D. Gaussian mixture random coefficient model based framework for shm in structures with time–dependent dynamics under uncertainty. *Mech. Syst. Signal Process.* **2017**, *97*, 59–83. [CrossRef]
13. Alimohammadi, H.; Chen, S.N. Performance evaluation of outlier detection techniques in production timeseries: A systematic review and meta-analysis. *Expert Syst. Appl.* **2022**, *191*, 116371. [CrossRef]
14. Samara, M.A.; Bennis, I.; Abouaissa, A.; Lorenz, P. A survey of outlier detection techniques in IoT: Review and classification. *J. Sens. Actuator Netw.* **2022**, *11*, 4. [CrossRef]
15. Chen, L.; Gu, C.; Zheng, S.; Wang, Y. A Method for Identifying Gross Errors in Dam Monitoring Data. *Water* **2024**, *16*, 978. [CrossRef]
16. Bourquin, J.; Schmidli, H.; van Hoogevest, P.; Leuenberger, H. Pitfalls of artificial neural networks (ANN) modelling technique for data sets containing outlier measurements using a study on mixture properties of a direct compressed dosage form. *Eur. J. Pharm. Sci.* **1998**, *7*, 17–28. [CrossRef]
17. Chakravarty, S.; Demirhan, H.; Baser, F. Fuzzy regression functions with a noise cluster and the impact of outliers on mainstream machine learning methods in the regression setting. *Appl. Soft Comput.* **2020**, *96*, 106535. [CrossRef]
18. Zhao, L.; Akoglu, L. On using classification datasets to evaluate graph outlier detection: Peculiar observations and new insights. *Big Data* **2023**, *11*, 151–180. [CrossRef]
19. Chen, H.; Huang, S.; Xu, Y.P.; Teegavarapu, R.S.; Guo, Y.; Nie, H.; Xie, H. Using baseflow ensembles for hydrologic hysteresis characterization in humid basins of Southeastern China. *Water Resour. Res.* **2024**, *60*, e2023WR036195. [CrossRef]
20. Bao, Y.; Tang, Z.; Li, H.; Zhang, Y. Computer vision and deep learning–based data anomaly detection method for structural health monitoring. *Struct. Health Monit.* **2019**, *18*, 401–421. [CrossRef]
21. Domingues, R.; Filippone, M.; Michiardi, P.; Zouaoui, J. A comparative evaluation of outlier detection algorithms: Experiments and analyses. *Pattern Recognit.* **2018**, *74*, 406–421. [CrossRef]
22. Miao, Y.; Su, H.; Xu, O.; Chu, J. Support vector regression approach for simultaneous data reconciliation and gross error or outlier detection. *Ind. Eng. Chem. Res.* **2009**, *48*, 10903–10911. [CrossRef]
23. Malhotra, P.; Ramakrishnan, A.; Anand, G.; Vig, L.; Agarwal, P.; Shroff, G. LSTM-based encoder-decoder for multi-sensor anomaly detection. *arXiv* **2016**, arXiv:1607.00148.

24. Rico, J.; Barateiro, J.; Mata, J.; Antunes, A.; Cardoso, E. Applying advanced data analytics and machine learning to enhance the safety control of dams. In *Machine Learning Paradigms: Applications of Learning and Analytics in Intelligent Systems*; Springer: Cham, Switzerland, 2019; pp. 315–350.
25. Mishra, G.; Kumar, R. An individual fairness based outlier detection ensemble. *Pattern Recognit. Lett.* **2023**, *171*, 76–83. [CrossRef]
26. Zhao, Z.; Chen, K.; Zhang, H.; Li, Y.; Wu, Z. The method of gross error identification of dam monitoring data based on robust estimation. *J. Water Resour. Power* **2018**, *36*, 68–71.
27. Song, J.; Zhang, S.; Tong, F.; Yang, J.; Zeng, Z.; Yuan, S. Outlier detection based on multivariable panel data and K-means clustering for dam deformation monitoring data. *Adv. Civ. Eng.* **2021**, *2021*, 3739551. [CrossRef]
28. Zhang, P.; Li, T.; Wang, G.; Wang, D.; Lai, P.; Zhang, F. A multi-source information fusion model for outlier detection. *Inf. Fusion* **2023**, *93*, 192–208. [CrossRef]
29. Li, M.; Li, M.; Ren, Q.; Li, H.; Song, L. DRLSTM: A dual-stage deep learning approach driven by raw monitoring data for dam displacement prediction. *Adv. Eng. Inform.* **2022**, *51*, 101510. [CrossRef]
30. Petrou, M.M.; Petrou, C. *Image Processing: The Fundamentals*; John Wiley & Sons: Hoboken, NJ, USA, 2010.
31. Flusser, J.; Farokhi, S.; Höschl, C.; Suk, T.; Zitova, B.; Pedone, M. Recognition of images degraded by Gaussian blur. *IEEE Trans. Image Process.* **2015**, *25*, 790–806. [CrossRef]
32. Waltz, F.M.; Miller, J.W. Efficient algorithm for gaussian blur using finite-state machines. In Proceedings of the Machine Vision Systems for Inspection and Metrology VII, Boston, MA, USA, 4–5 November 1998; Volume 3521, pp. 334–341.
33. Russ, J.C. *The Image Processing Handbook*; CRC Press: Boca Raton, FL, USA, 2006.
34. Mareli, M.; Twala, B. An adaptive Cuckoo search algorithm for optimisation. *Appl. Comput. Inform.* **2018**, *14*, 107–115. [CrossRef]