Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 7157844, 10 pages
https://doi.org/10.1155/2022/7157844

Research Article

Outlier Detection of Gravity Dam Deformation Monitoring Data Based on the Multiple Local Outlier Coefficient Method

Bin Li,[1,2] **Xingping Bai**,[1] **Jun Li**,[1] **and Lirong Wang**[1]

[1]*Northwest Engineering Corporation Limited, Power China Water Conservancy & Hydropower Engineering Div, Pumped Storage Engineering Div, Xi'an, Shannxi, China*
[2]*Xi'an University of Technology State Key Laboratory of Eco-hydraulics in Northwest Arid Region of China, Xi'an University of Technology, Xi'an, Shannxi, China*

Correspondence should be addressed to Bin Li; libin@nwh.cn

Received 25 August 2022; Accepted 3 October 2022; Published 19 October 2022

Academic Editor: Junwei Ma

In the deformation monitoring data of gravity dams, a few outliers are often found, which will have adverse effects on the monitoring of model building and other data analysis work. In this study, the multiple local outlier coefficient method was proposed to quickly detect the outlier data in real time and to provide high-quality data for subsequent data analysis. This method was based on the idea of distance in the outlier detection algorithm, aiming at laws and characteristics of gravity dam deformation monitoring data. First, the basic principle, calculation steps, and basic features of the multiple local outlier coefficient method were studied. Then, for the two important parameters of the algorithm, the appropriate window length was selected using autocorrelation and partial autocorrelation analysis, and the appropriate threshold values were selected using the 3σ criterion, maximum method, and empirical method. Finally, an engineering example was used to verify that the algorithm could accurately detect the outliers in the gravity dam deformation monitoring data, and the deviation degree and meaning of the outliers were understood according to the calculated outlier coefficient. The multiple local outlier coefficient method has the advantages of a simple calculation principle, fast calculation speed, real-time detection, and a clear meaning of calculation results. By selecting the appropriate parameters, the method could satisfy the outlier detection of different types of data, offering an advantage in adapting to the computing demand of massive monitoring data and improving the intelligence and real-time monitoring of dam safety.

1. Introduction

Deformation monitoring data contain important information on dam or slope deformation. These data are an important basis for understanding the deformation mechanism, predicting deformation, and evaluating the safety state [1, 2]. However, in the process of data collection, it is inevitable that error information will appear in monitoring data. For random errors, the relevant methods of empirical mode decomposition are often used to reduce their proportion in the original data, so as to improve the calculation accuracy of the prediction model [3, 4]. For outliers, the causes of their generation usually include instrument failure, structural damage, human factors, or other uncertain factors. The situation is relatively complex. So it is necessary to detect them first. In the analysis of gravity dam deformation monitoring data, these outliers often contain important information about whether the dam is abnormal. Therefore, timely and accurate detection of outliers is very important to ensure the safe operation of the dam.

At present, the methods for detecting outliers mainly include the expert experience method, statistical probability method, multiscale decomposition, and data mining in gravity dam deformation monitoring data analysis. The expert experience method detects outliers by manually analyzing the data process line. This method relies too heavily on the individual levels of experts, and its efficiency is low. The statistical probability method detects outliers according to statistical eigenvalues of data. Commonly used methods include the 3σ criterion [5] Chauvenet criterion [6],

Grubbs criterion [7], Dixon criterion [8], and small probability theory [9]. These methods are simple to apply but usually have certain requirements regarding quantity and quality of the data, with certain assumptions. As a result, these methods have many inconveniences. Wavelet analysis is a representative method of multiscale decomposition, where after multi-scale analysis of data, it can detect outliers by identifying the modulus maximum points of decomposed coefficients. Studies have shown that outliers can be clearly detected in high frequency parts after decomposition [10]. Wavelet analysis is based on the internal relationship of the data for detection, without involving the influence of external causes. Data mining is a fast developing subject in recent years and has been widely used in various fields [11, 12]. Outlier detection is an important research direction in data mining [13]. The representative algorithms include the local outlier factor (LOF) and K-nearest neighbor (KNN) algorithm [14, 15]. These methods usually determine whether there are outliers based on the distance between data. According to this principle, some scholars have proposed the DE-LOF outlier detection algorithm [16]. DE-LOF first performs first-order difference on the original data. For the differenced sequence, according to the distance between each value and the first ten values, an LOF algorithm is used to determine the local outlier coefficient of this value. Finally, a set of sequences composed of local outlier coefficients is obtained, and then, the small probability method is used to determine the threshold value to determine which values are outliers. By default, this method determines whether the value is an outlier according to the first ten values to be detected. Whether the range of these ten values is reasonable needs further study.

In addition, many scholars have attempted a variety of methods for dam data outlier detection. In the Bayesian dynamic linear model framework with the switched Kalman filtering method [17], the observed structure data are decomposed into a group of hidden components and different components are described in different ways. If any changes occur in the local trend, a local acceleration component must be added to model its rate of change. At the same time, this method combines the advantages of dynamic adjustment of the Kalman filter. This method has the advantages of high robustness, high timeliness, and no marking of training data. This method needs to build a model first. When building the model, how to deal with outliers in the training data also exists in the enhanced regression tree method [18], dynamic mutation blind spot [19], SSA-NAR model [20], etc. The above methods for detecting outliers need to establish a prediction model first, and then judge whether there are outliers according to the error between the predicted data and the measured data. However, it fails to perform outlier detection when there are outliers in the data of the training model before the establishment of the model.

In order to improve the accuracy and reliability of the analysis and calculation of gravity dam deformation monitoring data, it is necessary to detect outliers in the original data before data analysis. To solve this problem, this paper proposed a multiple local outlier coefficient (MLOC) method based on the aforementioned research methods and ideas [21, 22]. Different from the DE-LOF method, the MLOC method explored the correlation between the value to be detected and the data before k times. This ensures that highly correlated data were used as a basis for judging whether the data to be detected are an outlier. At the same time, the MLOC method has no restrictions on the selection method of its key parameters, which enables the method to flexibly adjust parameters according to the change rules and characteristics of different types of data, and has stronger adaptability. In addition, this method can independently detect outliers of real-time monitoring data to meet the demand of real-time monitoring of dams.

2. Multiple Local Outlier Coefficient Method

2.1. Basic Conception

2.1.1. Basic Principles. The deformation process of gravity dams is usually stable in the short term but will be periodic and monotonous in the long term. According to deformation characteristics, the MLOC was proposed in this work for outlier detection in gravity dam deformation monitoring data. This method is based on the distance between the data at a certain time and the data before k times, to assess whether it is an outlier. Specifically, at the first moment, a set of window data and a value to be detected are selected from the original monitoring data. They are entered into the MLOC method to judge whether the value to be detected is an outlier. Then, the value that has been detected is added to the window data, and the first value in the window data is removed to complete the update of the window data. After that, the data after the value that has been detected can be detected, that is, the outlier detection at the second moment. By analogy, the method can continuously detect outliers for the latest data. The workflow is shown in Figure 1.

In this work, we assumed three assumptions for this method as follows:

(1) There was a high correlation between the data at a certain time and the data before k times

(2) The changes between data and the data before k times were relatively smooth

(3) Outliers occur in a small amount and at a low rate and should not exceed half of the total amount of data in the window data

The parameters in this method include the window data length k and threshold B. The first k data of the data to be detected were called the window data, where k was the window length. The criterion for determining whether the data to be detected were outliers was called threshold B.

Let the gravity dam deformation monitoring data be expressed as $X = \{x_1, x_2, \ldots, x_t\}$, and the difference sequence of lagging order i of X is expressed as $\Delta_i X = \{x_{i,1}, x_{i,2}, \ldots, x_{i,j}\}$:

$$\Delta x_{i,j} = x_{i,j} - x_{i,j-i}, \tag{1}$$

where $i = 1,2,\ldots, k$ represents the lag order, $j = 1 + i, 2 + i, \ldots, t$ represents the value at the j^{th} moment, and $\Delta x_{i,j}$ represents the difference between x_j and the data at the previous moment.

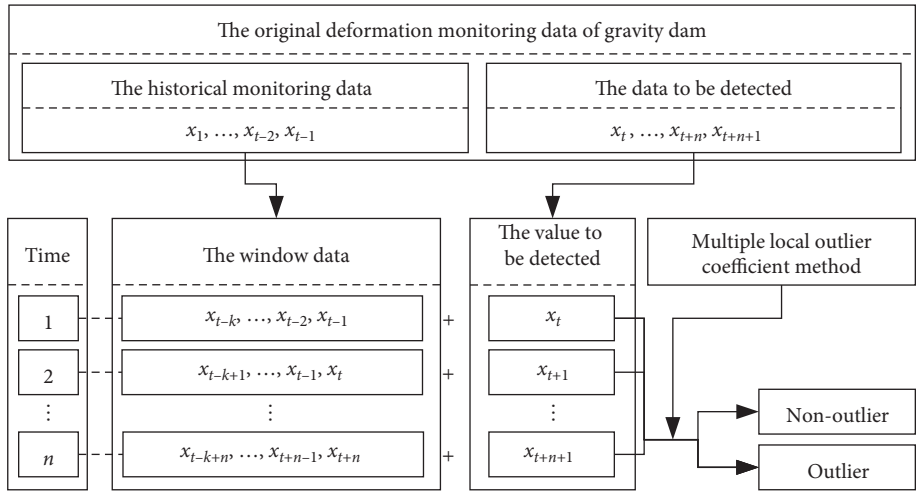

Figure 1: The workflow of the multiple local outlier coefficient method.

If the window length was k, there were k difference sequences $\Delta_i X(i = 1,2,\ldots, k)$. Each difference sequence had two thresholds, maximum and minimum, while the k difference sequences had a total of $2k$ thresholds. The MLOC method was calculated as the ratio between $\Delta x_{i,j}$ and the thresholds of $\Delta_i X$. If the ratio was less than or equal to 1, it indicated that the value to be detected was within the threshold range and was normal, while the outlier coefficient was defined as 0. Otherwise, it meant that the value to be detected was outside the threshold range, and the ratio consisted of the multiple outlier coefficient, where the larger the ratio, the greater the outlier degree.

Because the gravity dam deformation data had a slow monotonic trend in the long run, the positive and negative changes in the data were slightly different. Therefore, there were two thresholds for $\Delta_i X$: maximum and minimum, while k difference sequences had $2k$ thresholds. The data to be detected could obtain k outlier coefficients at most, where the largest absolute value among the outlier coefficients was the final outlier coefficient.

2.1.2. Calculation Steps. Let the deformation monitoring data of the gravity dam be expressed as $\{x_1, x_2,\ldots, x_{t-1}, x_t, x_{t+1}\ldots, x_{t+i}\}$. The calculation steps for the MLOC method are as follows:

(1) Determine the parameters. According to the original monitoring data, the k value and threshold B were determined, where threshold B contained the maximum threshold $B_1 = \{b_{11}, b_{12},\ldots, b_{1k}\}$ and the minimum threshold $B_2 = \{b_{21}, b_{22},\ldots, b_{2k}\}$.

(2) Extract the initial window sequence. The first window data were denoted as $W = \{x_{t-k},\ldots, x_{t-2}, x_{t-1}\}$, and the data to be detected were $\{x_t, x_{t+1},\ldots, x_{t+i}\}$.

(3) Determine the initial outlier coefficient. We made sure that the initial window data were all normal values, where the outlier coefficient λ of each value was 0.

(4) Mark the normal data. The MLOC method defaulted to the majority of data in the window data as normal data, according to the outlier coefficient. This step is needed to mark the normal data in the window data and group them into a set Z, $Z \subseteq W$.

(5) Calculate the outlier coefficient.

(i) We calculated the difference value d_i between the value x_t to be detected and each value in Z, $d_i = x_t - x_i (x_i \in Z)$, where the set comprising d_i was D.

(ii) We calculated the multiple outlier coefficients u_i for x_t, where the set comprising u_i was U:

$U_i = d_i/b_{1i}$ ($d_i \in$ D 且 $d_i \geq 0$); $u_i = d_i/b_{2i}$ ($d_i \in$ D 且 $d_i < 0$).

(6) Determine the outlier coefficient. If $|u_i| > 1$ is in U, we selected the $u_{\max}$ with the largest absolute value in U as the outlier coefficient λ_t of x_t.

If $|u_i| \leq f$, the value x_t to be detected was normal data, and its outlier coefficient $\lambda_t = 0$ was defined.

(7) Update the window data. We added x_t to the window data W and eliminated x_{t-k}. Then, we used the data x_{t+1} at the next moment as the data to be detected. Steps 4–7 were repeated to continuously detect real-time data.

2.1.3. Basic Features. According to the above introduction for the MLOC method, this method has the following four features.

(1) High efficiency. Faced with a considerable amount of streaming data, the MLOC method could extract, store, and calculate a small amount of data, and the calculation process was simple. Therefore, this method was highly efficient.

(2) Clear meaning. The outlier coefficient represented the ratio between the distance of the data to be detected, deviating from the first k data and the corresponding threshold, and the meaning of the outlier degree was clear.

(3) Real time. The MLOC method only had to measure the value to be detected and the first k data and did not need to know the data after the value. Therefore, the data acquired in real time could be detected.

(4) Not affected by outliers. The MLOC method defaulted to the minority of outliers in the window data, and this method judged the value to be detected based on the majority of normal values in the window data. Therefore, even if there were outliers in the window data, it did not affect the calculation process.

(5) Wide range of applications. The traditional outlier detection method could only detect spike outliers, while the MLOC method could detect not only the spike outlier but also the step outlier, as shown in Figure 2. However, spike outlier did not exceed half of the window length, and the step outlier only appeared once in the window data.

2.2. Window Length Selection. To determine the window length, it was necessary to know whether there was a high correlation between the data at a certain moment and the data at a previous i moment. In this work, autocorrelation and partial autocorrelation were used to analyze the correlation of the gravity dam deformation monitoring data.

2.2.1. Autocorrelation. Autocorrelation refers to the correlation between a sequence and its lag i-order sequence. When a sequence changes in the same direction as its lag i-order sequence, this indicates that the sequence is a positive autocorrelation. Otherwise, it will be a negative autocorrelation. The deformation monitoring data of a gravity dam consisted of a one-dimensional time series, where the first-order linear autoregressive form is as follows [23]:

$$x_i = \rho x_{i-1} + \varepsilon_i, \tag{2}$$

where x_i is a random variable, ρ is the self-covariance coefficient or the first-order autocorrelation coefficient ($-1 < \rho < 1$), and ε_i is the random interference term.

Equation (1) is actually the unitary linear regression model, where x_i is the dependent variable, x_{i-1} is the independent variable, and ρ is the regression coefficient. Calculating the autocorrelation of x_i actually consisted of the process of calculating the autocorrelation coefficient, and the least square method is used to solve the autocorrelation coefficient as follows [23]:

$$\widehat{\rho} = \frac{\sum_{i=2}^{n} x_i x_{i-1}}{\sum_{i=2}^{n} x_{i-1}^2}. \tag{3}$$

With a large sample size, the autocorrelation coefficient can usually be estimated according to the following formula [23]:

$$\widehat{\rho} = \frac{\sum_{i=2}^{n} x_i x_{i-1}}{\sqrt{\sum_{i=2}^{n} x_i^2 \sum_{i=2}^{n} x_{i-1}^2}}, \tag{4}$$

where the correlation coefficient is a function of time t. Calculating the correlation served as the process for solving the autocorrelation function.

2.2.2. Partial Autocorrelation. Usually, the correlation between random variables x_i and x_{i-k} will not be completely independent, as they will also be affected by x_{i-1}, x_{i-2},. . ., x_{i-k-1}. If only the correlation between x_i and x_{i-k} needs to be considered, the partial autocorrelation between them should be calculated. Therefore, the high-order autoregressive model is given as follows [24]:

$$x_i = \phi_1 x_{i-1} + \phi_2 x_{i-2} + \cdots + \phi_k x_{i-k} + \varepsilon_i, \tag{5}$$

where ϕ_k is the partial autocorrelation coefficient and ε_i is the random interference term.

The following formula can be obtained by multiplying both sides of (5) by x_{-1}, x_{i-2},. . ., x_{i-k} [24]:

$$\begin{cases} \rho_1 = \phi_1 \rho_0 + \phi_2 \rho_1 + \cdots + \phi_k \rho_{k-1}, \\ \rho_2 = \phi_1 \rho_1 + \phi_2 \rho_0 + \cdots + \phi_k \rho_{k-2}, \\ \quad \vdots \\ \rho_k = \phi_1 \rho_{k-1} + \phi_2 \rho_{k-2} + \cdots + \phi_k \rho_0. \end{cases} \tag{6}$$

This equation gives the Yule–Walker equation. Cramer's method can be used to solve ϕ_k as follows [24]:

$$\begin{aligned} \phi_1 &= \rho_1, \\ \phi_2 &= \frac{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & \rho_2 \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & \rho_1 \end{vmatrix}}, \\ \% \quad \phi_3 &= \frac{\begin{bmatrix} 1 & \rho_1 & \rho_1 \\ \rho_1 & 1 & \rho_2 \\ \rho_2 & \rho_1 & \rho_3 \end{bmatrix}}{\begin{bmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_1 \\ \rho_2 & \rho_1 & 1 \end{bmatrix}}. \ldots \end{aligned} \tag{7}$$

2.2.3. Other Instructions. Through correlation analysis, we could determine that the data at a certain time had a significant correlation with the data at the previous K time. When a step outlier was encountered, the data sequence was split into two parts, as shown in Figure 1. When two parts had the same amount of data, it was impossible to test the following data. Therefore, to ensure that the following data could be detected based on the two parts of the data and improve the range of data correlation, the $k = 2K + 1$ method was used to determine the final window length.

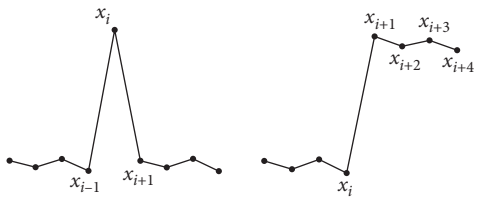

Figure 2: Spike outlier and step outlier.

2.3. Threshold Selection Method. Threshold is a key criterion that can be used to determine whether the data to be detected are outliers or not. In the MLOC method, different thresholds should be set for different lag sequences. Threshold selection is complicated, and the commonly used method involves the 3σ criterion. A reasonable threshold can be finally determined by considering the calculation results of various methods.

2.3.1. 3σ Criteria. The 3σ criterion is considered the most representative statistical method for detecting outliers. It assumes that data contain only random noise and then calculates the mean value μ and standard deviation σ of data, selecting an appropriate interval according to a certain probability. If data exceed this interval, they will not be considered random noise but an outlier. Otherwise, data will be considered normal data.

Let the data sequence be $\{x_1, x_2, \ldots, x_j, x_{j+1}, \ldots, x_t\}$, where the mean value of this data group is as follows:

$$\mu = \frac{1}{t}\sum_{i=1}^{t} x_i. \tag{8}$$

The standard deviation is as follows:

$$\sigma = \sqrt{\frac{1}{t}\sum_{i=1}^{t}(x_i - \mu)^2}. \tag{9}$$

A reasonable threshold could be determined based on μ and σ, where 3σ has usually been used to determine whether data are outliers. The 3σ criterion has usually been required to ensure that historical monitoring data will be sufficient and include data for extreme conditions. Only following this approach, data could represent the normal deformation of the dam body under various conditions.

2.3.2. Other Methods. Threshold determination can be accomplished in many ways. Specifically, the extreme value method adopts the maximum and minimum values of historical normal data as the threshold, while the finite element method can be used to calculate the safe deformation model of the dam body as the threshold value. The finite element method can also be used to calculate the safe deformation range of the dam body, and the results can be used as the threshold. This process is typically considered complex for determining the threshold, and it is usually necessary to determine the threshold based on expert experience and a comparison of similar projects [25, 26].

Figure 3: Photo of the gravity dam.

3. Example Calculation

3.1. Project Overview. In this study, a power station hub consists of a barrage, a water discharge structure, a water delivery system, an underground powerhouse and a ground switch station. The barrage is a roller-compacted concrete gravity dam. The maximum dam height of a concrete gravity dam was set to 72.4 m, the maximum length of the dam crest was 206 m, and the maximum width of the dam crest width was 7.5 m. The total reservoir capacity was 47 million m^3, the total installed capacity was 250 MW, and the dam was divided into 9 sections, as shown in Figure 3. Except for dam section no. 9, a displacement monitoring point was set at the top of the other eight dam sections. The monitoring points were numbered EX1~EX8, as shown in Figure 4.

In this work, the data of EX4 point in dam section no. 5 were selected for analysis, and the data are shown in Figure 5. The EX4 point was located at the dam top in the middle of the riverbed. Its variation range and law were representatives to some extent, which could reflect the general law of horizontal displacement of the dam top of the concrete gravity dam. The data for EX4 were measured once a day, including 869 data from June 2, 2016, to October 22, 2018, and 720 data were used as training data, while 149 subsequent data were used as test data.

3.2. Selection of Window Length. In this work, autocorrelation and partial autocorrelation analysis were used to analyze the correlation of the deformation monitoring data of the gravity dam. Then, the window length k was determined according to the calculation results.

Before the correlation calculation of the EX4 point, the first-order difference calculation was first conducted to ensure that the data consisted of the stationary sequence, namely, the Δ_1EX4 sequence. The calculation results of the autocorrelation coefficient and partial autocorrelation coefficient of the Δ_1EX4 sequence are shown in Figure 6. The results showed that Δ_1EX4 sequences were significantly correlated within the three orders. This also indicated that certain data in the EX4 sequence had a significant correlation with the data in the first three moments.

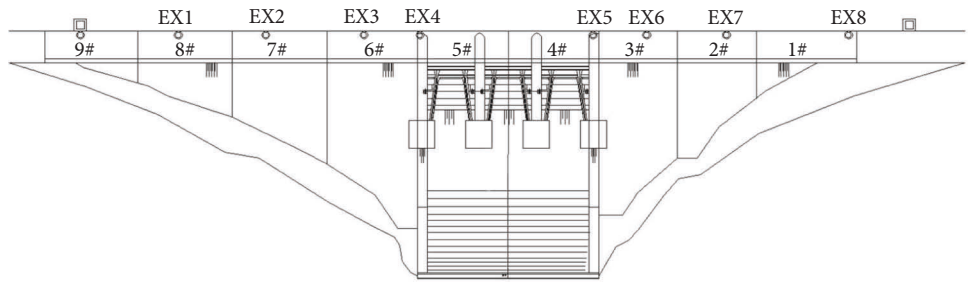

Figure 4: Downstream elevation view of the gravity dam.

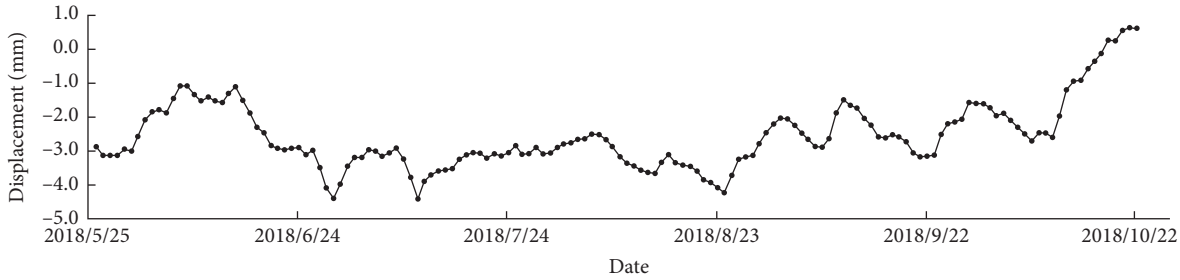

Figure 5: Data process line of the EM4 point.

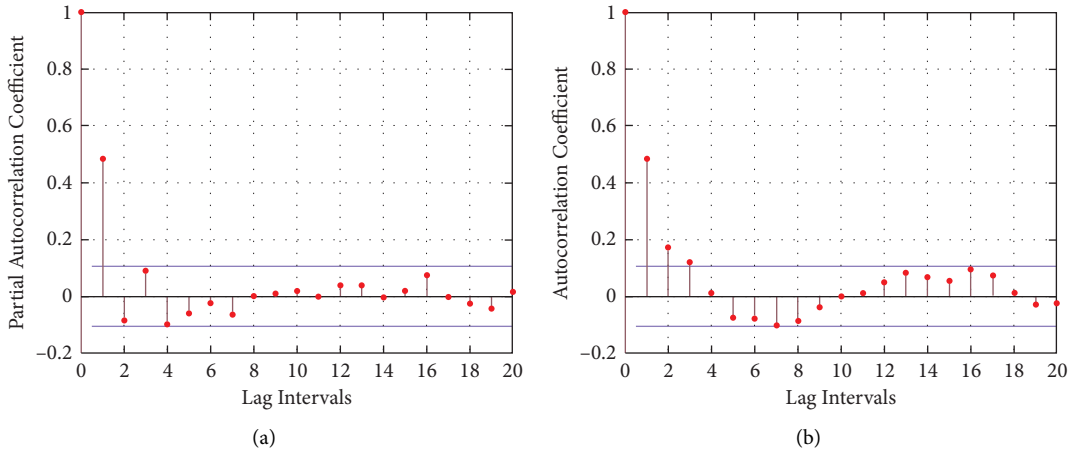

Figure 6: The autocorrelation coefficient and the partial autocorrelation coefficient of $\Delta_1 EX4$.

According to the above calculation results, K was 3. According to section 2.2, the window length $k = 2K + 1$ is 7.

3.3. Threshold Selection. In this work, the 3σ criterion was first used to calculate the thresholds of the $\Delta_1 EX4 \sim \Delta_7 EX4$ sequence. Then, the calculation results were analyzed by expert experience. Finally, the maximum and minimum values were used as thresholds.

According to equations (8) and (9), 730 historical data for EX4 points were calculated. The 3σ calculation results of the $\Delta_1 EX4 \sim \Delta_7 EX4$ data sequences are shown in Table 1 and Figure 7.

As shown in the above results, some of the values in the $\Delta_1 EX4 \sim \Delta_7 EX4$ sequences were not in the range of 3σ. However, after expert analysis, these values were not outliers, because there were only more than two years of data available for EX4. Compared to the lifetime of gravity dams over decades, the total amount of data was small and there were not representative enough data. Therefore, the calculation results were too conservative.

The MLOC method did not limit the selection method of thresholds. Multiple methods could be used to determine the threshold value. The purpose of this work was to verify the feasibility of the MLOC method and whether it could achieve the expected purpose. Considering this, we adopted

Table 1: 3σ calculation results of the $\Delta_1 EX4 \sim \Delta_7 EX4$ sequences.

Data sequences	$\Delta_1 EX4$	$\Delta_2 EX4$	$\Delta_3 EX4$	$\Delta_4 EX4$	$\Delta_5 EX4$	$\Delta_6 EX4$	$\Delta_7 EX4$
μ	−0.0043	−0.0087	−0.0130	−0.0176	−0.0224	−0.0269	−0.0310
σ	0.1912	0.3269	0.4368	0.5276	0.6054	0.6742	0.7343
3σ	0.5737	0.9806	1.3104	1.5828	1.8162	2.0225	2.2030
$\mu - 3\sigma$	−0.5780	−0.9893	−1.3234	−1.6004	−1.8386	−2.0494	−2.2340
$\mu + 3\sigma$	0.5694	0.9719	1.2974	1.5653	1.7938	1.9956	2.1720

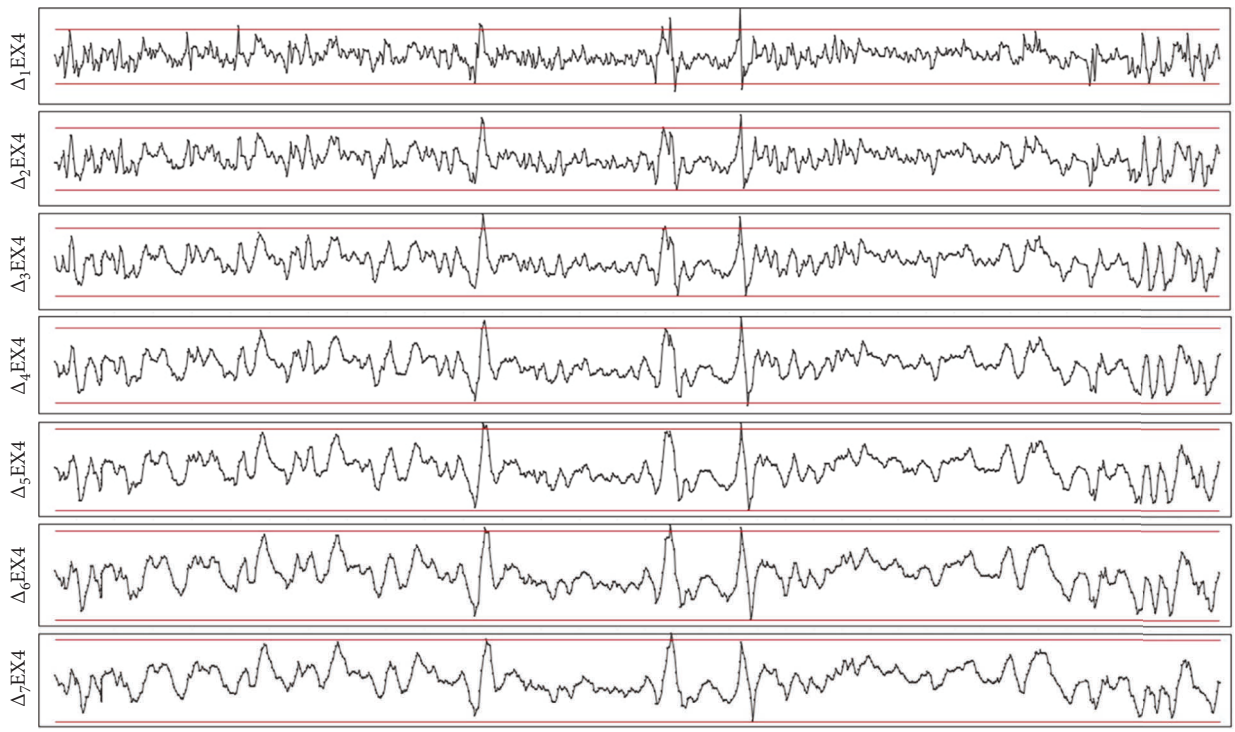

Figure 7: $\Delta_1 EX4 \sim \Delta_7 EX4$ sequences and their 3σ interval.

Table 2: Thresholds of the $\Delta_1 EX4 \sim \Delta_7 EX4$ sequences.

Data sequence	$\Delta_1 EX4$	$\Delta_2 EX4$	$\Delta_3 EX4$	$\Delta_4 EX4$	$\Delta_5 EX4$	$\Delta_6 EX4$	$\Delta_7 EX4$
Maximum threshold	1.0275	1.3599	1.8079	2.0180	2.0699	2.2350	2.4682
Minimum threshold	−0.7412	−0.9674	−1.3599	−1.6994	−1.8610	−2.0226	−2.1796

the maximum method to finally determine the threshold. According to the extreme value method, the threshold calculation results of the maximum method are shown in Table 2.

In Table 2, the absolute values of the minimum thresholds were all smaller than the absolute values of the maximum thresholds. This was because dam deformation not only achieved relatively stable periodicity but also underwent creep with time. Thus, the forward displacement of dam deformation was greater than the backward displacement, and the cumulative forward creep displacement was generated.

3.4. Calculation Results of Multiple Local Outlier Coefficients. According to the determined window length k and threshold set B, outliers could be detected for the test data of EX4. To better illustrate the outlier detection effect of the MLOC method, three spike outliers and step outliers were added to the test data. The calculation results of the outlier coefficient for the test data are shown in Figure 8, and the information about outliers detected is shown in Table 3.

According to the calculation results of the outlier coefficient, three spike outliers, step outliers, and four outliers afterward were detected. This showed that the MLOC method met the expected detection target. In addition, four outliers were also detected in the original test data. This

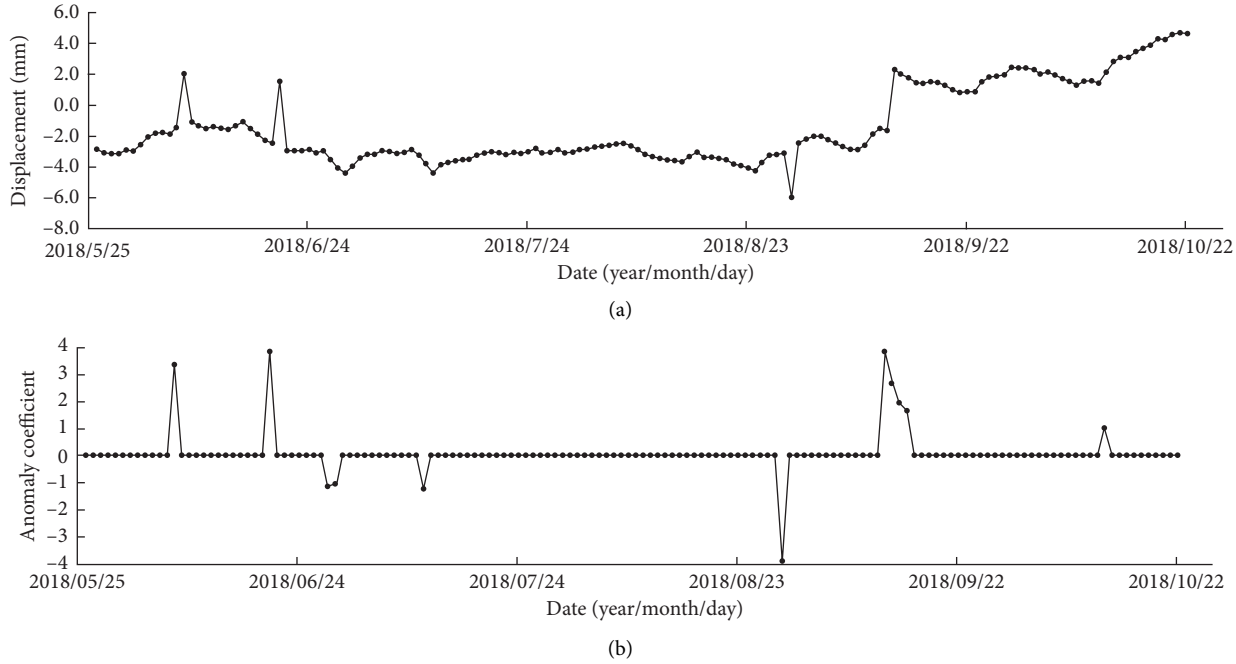

Figure 8: The calculation results of the multiple local outlier coefficient algorithm. (a) Test data process line-added outliers. (b) Outlier coefficient for test data-added outliers.

Table 3: Information for the outliers detected.

Date	Data	Outlier coefficient	Differential order	Difference value	Threshold
2018.6.7	2.0000	3.3637	1	3.4562	1.0275
2018.6.20	1.5000	3.8597	1	3.9658	1.0275
2018.6.28	−4.0808	−1.1307	2	−1.0940	−0.9674
2018.6.29	−4.4181	−1.0525	3	−1.4314	−1.3599
2018.7.11	−4.4256	−1.2230	2	−1.1833	−0.9674
2018.8.29	−6.0000	−3.8779	1	−2.8743	−0.7412
2018.9.12	2.2536	3.8326	1	3.9380	1.0275
2018.9.13	1.9609	2.6803	2	3.6453	1.3599
2018.9.14	1.7351	1.9628	6	4.3870	2.2350
2018.9.15	1.4027	1.6427	7	4.0546	2.4682
2018.10.12	2.7894	1.0233	2	1.3917	1.3599

indicated that the difference between them and the data at a certain time beforehand exceeded the maximum or minimum values in the historical data. For the data at 2018.6.28, the historical minimum value (the minimum value in the 730 training data) of the lag 2 order difference sequence was −0.9674. However, in the test data, the difference value of lag 2 order was −1.0940, which was less than −0.9674. This indicated that the deformation of the dam on June 28, 2018, was less than the historical minimum, and its outlier coefficient was in the ratio of −1.0940 to −0.9674, which was −1.1307. The minus sign indicated that the dam was moving upstream.

Step outliers were generated from September 12, 2018, and the following three data were also judged as outliers, which was the principle of the MLOC method. When step outliers occurred, the following three data appeared to meet the normal law; however, they were still a small number of outliers compared to the general law of the previous data. The discriminant range set by this method was greater than half of the window data, that is, greater than 7/2. Therefore, the step value and the three values behind it were judged as outliers, and the values after that were judged as normal.

The added outliers tended to deviate greatly from the original data, and this was easy to detect. The four outliers detected in the original test data were not outliers according to their changing trend. This was because this work determined the threshold based on the maximum value of historical data. The 730 historical data sequence was short; thus, the threshold setting was more cautious. However, this had no impact on the overall analysis of the data, and the MLOC

method did not limit the method of determining the threshold. Thus, the threshold could be determined in a more reasonable way, and then, outliers could be detected.

4. Conclusions

According to the law and characteristics of gravity dam deformation monitoring data, this work put forward the MLOC method for the real-time detection of outliers. Through systematic research and example calculation of this method, the following conclusions were obtained:

(1) The values before and after the deformation monitoring data of the gravity dam showed a significant correlation within a certain range. According to this feature, whether the value to be detected was an outlier could be judged according to its value at a time before k.

(2) When calculating the thresholds based on the historical deformation monitoring data, the historical data required sufficient capacity and representativeness. The MLOC method did not limit the method of determining the threshold; therefore, a variety of methods and expert experience were fully considered to determine the threshold.

(3) The MLOC method has the advantages of fast calculation speed and real-time detection. The example calculation results showed that this method could accurately detect outliers in the monitoring data, and the deviation degree and meaning of the outliers could be understood by the outlier coefficient.

The MLOC method is based on the distance between the data at a certain time and the data before k times to assess whether it is an outlier. The key parameters include the window length and the threshold. As long as the values before and after a set of data have a certain correlation, the outlier detection can be carried out by adjusting parameters. However, the selection of the threshold is always a very complex research problem in engineering, which usually needs to be determined by combining a variety of methods and expert experience. The MLOC method has no restrictions on the method of parameter selection, so it has strong flexibility and can be applied to real-time outlier detection of various types of data.

Data Availability

The data can be obtained from the corresponding author upon request. The calculated data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This research was supported by the Natural Science Basic Research Plan in Shaanxi Province(2021JM-331, 2019JQ-318).

References

[1] J. W. Ma, D. Xia, Y. K. Wang et al., "A comprehensive comparison among metaheuristics (MHs) for geohazard modeling using machine learning: insights from a case study of landslide displacement prediction," *Engineering Applications of Artificial Intelligence*, vol. 114, pp. 0952–1976, Article ID 105150, 2022.

[2] J. W. Ma, D. Xia, H. X. Guo et al., "Metaheuristic-based support vector regression for landslide displacement prediction: a comparative study," *Landslides*, vol. 19, no. 10, pp. 2489–2511, 2022.

[3] J. R. Zhang, H. M. Tang, D. D. Tannant et al., "Combined forecasting model with CEEMD-LCSS reconstruction and the ABC-SVR method for landslide displacement prediction," *Journal of Cleaner Production*, vol. 293, Article ID 126205, 2021.

[4] J. R. Zhang, H. M. Tang, T. Wen et al., "A hybrid landslide displacement prediction method based on CEEMD and DTW-ACO-SVR—cases studied in the three gorges reservoir area," *Sensors*, vol. 20, no. 15, p. 4287, 2020.

[5] B. B. Xu and C. F. Cui, "Research of outlier detection on automatic monitoring data of dam," *Journal of Geomatics*, vol. 40, no. 2, pp. 59–61, 2015.

[6] H. X. Zhao, S. N. Zhou, and H. Xiao, "The comparison and discussion of four criterions of gross-error detection," *PHYSICAL EXPERIMENTOFCOLLEGE*, vol. 30, no. 5, pp. 105–108, 2017.

[7] J. X. Tao, H. Y. Xiong, and B. Hu, "Method for judging abnormal values of dam safety monitoring data," *J of Three Gorges Univ (Natural Sciences)*, vol. 38, no. 6, pp. 15–17+41, 2016.

[8] S. Lach, "The application of selected statistical tests in the detection and removal of outliers in water engineering data based on the example of piezometric measurements at the Dobczyce dam over the period 2012-2016," *E3S Web of Conferences*, vol. 45, pp. 1–8, 2018.

[9] Z. Jiang and J. He, "Method of fusion diagnosis for dam service status based on joint distribution function of multiple points," *Mathematical Problems in Engineering*, vol. 2016, no. 9, 10 pages, Article ID 9049260, 2016.

[10] J. Liu and J. J. Lian, "Outliers detection of dam displacement monitoring data based on wavelet transform," *Applied Mechanics and Materials*, vol. 71-78, pp. 4590–4595, 2011.

[11] R. Domingues, M. Filippone, P. Michiardi, and J Zouaoui, "A comparative evaluation of outlier detection algorithms: experiments and analyses," *Pattern Recognition*, vol. 74, pp. 406–421, 2018.

[12] H. Ghallab, H. Fahmy, and M. Nasr, "Detection outliers on internet of things using big data technology," *Egyptian Informatics Journal*, vol. 21, no. 3, pp. 131–138, 2020.

[13] M. Orellana and P. Cedillo, "Outlier Detection with Data Mining Techniques and Statistical Methods," in *Proceedings of the 2019 International Conference on Information Systems and Computer Science (INCISCOS)*, Quito, Ecuador, November 2019.

[14] F. G. Zheng, "Abnormal deformation analysis concrete dam based on local outlier factor [J]," *Water Resources and Power*, vol. 34, no. 6, pp. 103–105, 2016.

[15] B. Wang and Z. Mao, "A Dynamic Ensemble Outlier Detection Model Based on an Adaptive K-Nearest Neighbor rule," *Information Fusion*, vol. 63, no. 3, 2020.

[16] J. Yang, X. D. Qu, D. X. Hu, J. Song, L. Cheng, and B. Li, "Research on singular value detection method of concrete

dam deformation monitoring," *Measurement*, vol. 179, Article ID 109457, 2021.

[17] L. H. Nguyen and J. A. Goulet, "Anomaly detection with the switching kalman filter for structural health monitoring," *Structural Control and Heath Monitoring*, vol. 25, no. 4, pp. 1–18, 2017.

[18] F. Salazar, J. M. González, M. Á. Toledo, and E. Oñate, "A methodology for dam safety evaluation and anomaly detection based on boosted regression trees," in *Proceedings of the 8th European Workshop On Structural Health Monitoring (EWSHM 2016)*, pp. 1–11, Bilbao, Spain, July 2016.

[19] Z. K. Xu, W. Xiong, B. W. Wei, and L. H Li, "A new method of dam safety monitoring for identifying displacement mutation," *Applied Mechanics and Materials*, vol. 687-691, pp. 925–928, 2014.

[20] J. T. Song, Y. C. Chen, and J. Yang, "A novel outlier detection method of long-term dam monitoring data based on SSA-NAR," *Wireless Communications and Mobile Computing*, vol. 2022, p. 1, Article ID 6569367, 2022.

[21] A. Taha and A. S. Hadi, "Anomaly detection methods for categorical data: a review," *ACM Computing Surveys*, vol. 52, no. 2, pp. 1–35, 2019.

[22] Y. Goi and C. W. Kim, "Bayesian outlier detection for health monitoring of bridges," *Procedia Engineering*, vol. 199, pp. 2120–2125, 2017.

[23] C. Q. Tao, *Econometrics*, Fudan University Press, Shanghai, 2012.

[24] J. N. Yu, *Econometrics*, University of International Business Economics Press, Beijing, 2014.

[25] B. Li, J. Yang, and D. X. Hu, "Dam monitoring data analysis method:a literature review," *Structural Control and Health Monitoring*, vol. 27, no. 3, pp. 1–14, 2020.

[26] Y. Xu, H. B. Huang, Y. L. Li, J. Zhou, X. Lu, and Y Wang, "A three-stage online anomaly identification model for monitoring data in dams," *Structural Health Monitoring*, vol. 21, no. 3, pp. 1183–1206, 2022.