Treatment of Errors in Dam Safety Monitoring Data

Xuan Zhang1*, Junjie Li1

1 School of Water Resources and Environment, Zhengzhou University, Zhengzhou, Henan Province, 450001, China
*Corresponding author’s e-mail: 1013122808@qq.com

Abstract. The error processing of dam safety monitoring data is the basic work of the whole data analysis, and it is the guarantee for objectively evaluating the safety of dams. This paper summarizes the sources of error, analysis methods and research status of dam monitoring data, lists several commonly used error processing methods and the latest research progress, and finally analyzes through engineering examples.

1. Introduction

In the dam safety monitoring, due to the complexity of the dam, the working environment, the differences of observers, the performance of the instrument and the monitoring methods, there are inevitably errors in the monitoring data[1]. According to the nature of the error, it can be generally divided into three categories: random error, systematic error and coarse error. The random error is mainly caused by various random and accidental factors, and it conforms to the normal distribution with zero mean[2]. It is common in the continuous monitoring data of large samples, and generally does not affect the normal statistical and timing analysis. Systematic error refers to the error caused by the instability of the monitoring instrument caused by the independent uncertain factors or the displacement of the monitoring base point. It seriously deviates from the true value (or the established statistical model) and often manifests as abnormal fluctuation of the single-sided point data, and may have a certain continuity and stage. Gross errors are data that contain large errors and deviate significantly from the true value (or a given statistical model), often caused by operational negligence during the observation process and errors in the recording, copying, and calculation of data[2]. In the in-situ data of dam safety monitoring, the above three kinds of errors may exist. The existence of error affects the accuracy of model analysis. Therefore, the error analysis and processing of dam monitoring data are carried out to obtain more accurate and effective basic analysis. Data is the primary task of monitoring data analysis and dam safety assessment.

There are two methods for discriminating gross errors, physical discrimination and statistical discrimination. Physical discrimination is often used to judge particularly significant gross errors caused by factors such as monitoring equipment failure. The statistical discriminant method is to define a distribution interval under the condition of given significant level, and analyze whether the monitoring data exceeds the critical value and becomes a gross difference.

2. Research Status of Gross Errors Discrimination Method

At present, the errors of dam monitoring data are widely processed by the least square method[3]. Since G.F.Gauss proposed the least square method in 1794, a large number of scholars have done a lot of research on measurement adjustment theory and method. In 1947, T.M.Tienstre proposed a correlation
adjustment method to extend the requirement for independent observation to random correlation. In 1962, G. Meiissl et al proposed a rank-deficient free network adjustment, and extended the full rank matrix in the measurement adjustment to the singular array. Kiman et al. proposed a recursive filtering method that has been successfully applied to aerospace and industrial automation. In 1969, Krarup proposed least square filtering, estimation and configuration, and adjusted parameters from non-random variables are generalized to random parameters. In terms of specific research work, US S.G.K and Y.F quantitatively analyzed the influence of the measurement error of the ARMA. Jianping Yue evaluated the accuracy of the measurement system and model more objectively by separating the observation error and model error in the analysis of the cause. Wenbao Liu et al. proposed a dam displacement back analysis method that takes into account prior information, and discusses the influence of measurement errors on the displacement inversion results.

The least square method assumes that the observations only contain accidental errors, which is impossible in practice, thus created new theories for studying systematic errors and gross errors. In the late 1960s, W. B proposed data detection and reliability theory, which laid the foundation for gross errors research.

At present, there are two methods for dealing with gross errors. One is the data detection method which still belongs to the category of least square method, and the other is the difference estimation or robust estimation method which is different from the least square method. In China, Jiangwen Zhou proposed some practical anti-difference algorithms; Dongjian Zheng used the average leverage theory to identify the abnormal values of measured data, and realized the gray box diagnosis of the errors; Bin Zhao introduced the concepts of pollution distribution and observation rights in the dam observation data processing, in order to gradually eliminate the influence of gross errors, a reasonable parameter estimate is finally obtained[4].

3. Several commonly used gross errors statistical discriminant

(1) 3σ criterion

The 3σ criterion, also known as the Pauta criterion, is one of the most common and convenient methods. It is based on the premise that the monitoring data sample is large enough and there is no systematic error. If the value of a data residual is found to be greater than 3 times the standard deviation in the monitoring data, as

\[ |v| > 3\sigma \] (3.1)

The measured value can be considered to contain a gross error.

(2) Grubbs criterion

Unlike the 3σ criterion, the Grubbs criterion is based on small-volume sample and has extremely high precision for finding a suspicious value mixed in the sample. It is rearranged by updating the monitoring data \( x_i \) in ascending order:

\[
(3) \quad x(1) \leq x(2) \leq \ldots \leq x(n)
\]

Grubbs calculated the distribution of \( g(1) = \frac{x(1) - x(\bar{x})}{\sigma} \) and \( g(n) = \frac{x(n) - x(\bar{x})}{\sigma} \), determined the significance level \( \alpha \) (usually 0.01 or 0.05), and looked up the table to obtain the Grubbs threshold \( g_0(n, \alpha) \).

If you think \( x(i) \) is suspicious, then

\[ g(i) \geq g_0(n, \alpha) \] (3.2)

(3) Dixon criterion

The Dixon criterion is similar to the Grubbs criterion and has higher accuracy in small-volume sample. The difference from the Grubbs criterion is that the Dixon criterion does not need to calculate the standard deviation \( \sigma \), and the accuracy ratio can be used to obtain more accurate results. Monitor the data in ascending order:

\[
(4) \quad x(1) \leq x(2) \leq \ldots \leq x(n)
\]

The statistic distribution of \( x(1) \) is calculated:

\[ r_{10} = \frac{x(n) - x(n-1)}{x(n) - x(1)} \] (3.3)
\[ r_{11} = \frac{x(n) - x(n-1)}{x(n) - x(2)} \]
\[ r_{21} = \frac{x(n) - x(n-2)}{x(n) - x(2)} \]
\[ r_{22} = \frac{x(n) - x(n-3)}{x(n) - x(3)} \]

Determine the significant level \( \alpha \), check the table to determine the critical value \( r_0(n, \alpha) \), if the statistic of the monitoring data \( x(i) \) is greater than the critical value:

\[ r_{ij} > r_0(n, \alpha) \]  (3.4)

Then \( x(i) \) may be a gross error.

Dixon gave the statistic selection reference. When \( n \leq 7 \), the effect of using \( r_{10} \) is good; when \( 8 \leq n \leq 10 \), the effect of using \( r_{11} \) is good; when \( 11 \leq n \leq 13 \), the effect of using \( r_{21} \) is good; when \( n \geq 14 \), using \( r_{22} \) works well.

(4) Romanov criterion

The Romanov criterion, also known as the t-test, is characterized by first rejecting a suspicious value and then examining the suspected value that is rejected by the t-distribution for gross error.

Suppose \( x_j \) is a suspicious value in a set of monitoring data, remove it and then calculate the average:

\[ \bar{x} = \frac{1}{n-1} \sum_{i \neq j} x_i \]  (3.5)

And calculate the standard deviation:

\[ \sigma = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n-2}} \]  (3.6)

Check the table to determine the t-distribution test coefficient \( K(n, \alpha) \)

If

\[ |x_j - \bar{x}| > K\sigma \]  (3.7)

Then the monitoring data \( x_j \) may be a gross error and it is correct to reject it.

If the sample size of the monitoring data is particularly small, it is very reasonable to use the t-distribution test to discriminate the coarse error. In addition, there are several mathematical methods for discriminating gross errors, such as the Chovini criterion, the Nair criterion and the fine criterion, which are not introduced here, as detailed in the literature[4].

4. Latest research progress

(1) Mathematical model

According to the comprehensive influence of the amount of dam effect on the environmental quantity, a multi-factor statistical model is constructed to test whether there is gross error in the effect quantity data.

Let the statistical model of a monitored quantity be:

\[ \delta = f(x_1, x_2, x_3) \]  (4.1)

Where “\( \delta \)” : the estimated value of the monitoring amount;
“\( x \)” : is the impact factor affecting the monitoring amount;
“\( f \)” : represents the functional relationship between the effect size and the predictor.

If

\[ |\delta - Y| > K\sigma_\delta \]  (4.2)

It is considered that there is a gross error in the monitoring data.

Where “\( Y \)” : represents the monitored value;
“\( K \)” : indicates the coefficient, generally takes 2;
\( \sigma \): represents the standard deviation corresponding to the model.

(2) Wavelet analysis

The data sequence of dam safety monitoring can be regarded as a digital signal sequence composed of different frequency components, and the suspicious value in the data sequence appears as a mutation under the normal signal. When the wavelet analysis method is used to diagnose the suspect value, the digital signal is firstly decomposed by multi-scale, and the detailed coefficients of each layer are reconstructed. At the signal mutation, the coefficient of the digital signal after wavelet transform has a modulus maximum. Therefore, the reconstructed detail coefficient has a point of modulus maximum, and there may be a gross error\(^5\).

(3) Random fuzzy diagnosis

Since the size and position of the suspicious value cannot be predicted and controlled in advance, it is determined that the suspicious value shows randomness in quality and ambiguity in quantity, that is, the wild value has random fuzzy duality. Therefore, random fuzzy processing can be used to diagnose outliers\(^5\).

Assume that the observed value of a monitored amount is \( \{x_1, x_2, \ldots, x_n\} \), its corresponding forecast value is \( \{x'_1, x'_2, \ldots, x'_n\} \).

Suppose

\[
y_i = x_i - x'_i
\]

Assume its scope is

\[
U = \{y_1, y_2, \ldots, y_n\}
\]

\( A \) is a fuzzy subset (F subset) on \( U \). In the subset \( A \), the values \( x_1, x_2, \ldots, x_n \) are taken as probabilities \( p_1, p_2, \ldots, p_n \), respectively, and their membership functions for \( A \) is:

\[
\mu_i = e^{-D_{i1}(y_i)}
\]

\[
D_{i1}(y_i) = \frac{y_i^2 + \omega_i^2}{2y_{imax}^2 - y_{jmax}^2}
\]

\( D_{i1} \) is the Mahalanobis distance of \( y_i \) with respect to the core point \( A_0 \) of the subset \( A \). The random fuzzy probability density function that \( y_i \) satisfies is:

\[
p_i = \lambda_i e^{-\lambda_i(a-u_i)}
\]

\[
\lambda_0 = \ln \sum_{i=1}^{n} e^{-\lambda_i(a-u_i)} - 1
\]

\[
\sum_{i=1}^{n} (a-u_i)e^{-\lambda_i u_i} = 0
\]

Where \( "a" \) is a constant; \( "n" \) is the sample size.

After a given confidence factor \( \beta \), the random fuzzy probability distribution function that \( y_i \) satisfies is:

\[
F(u_\beta) = \int_{-\infty}^{u_\beta} f(u) du = \beta
\]

\[
f(u) = -k\lambda_i e^{(l+\lambda_i) a - \lambda_i u}
\]

In the formula, \( "k" \) is the adjustment coefficient, and \( "k" \) is set to consider the influence of the model error of the membership function, the error of the weight factor, and the non-random factor in the observation, \( "k" \) is determined by the following formula:

\[
\int_{-\infty}^{+\infty} f(u) du = 1
\]

\( u_\beta \) is the test threshold. If \( y_i \) exists, its membership is \( u_i \), and \( u_i < u_\beta \), then \( y_i \) is considered to be a
outlier.

(4) Multi-source information fusion method

Different from the mathematical model of single point monitoring with a single monitoring point as the research object, multi-source information fusion method is combined with modern computer technology, and the observation data of multiple sensors in large complex systems are reasonable according to certain criteria. Dominance, combination, comprehensive analysis and utilization, linking the mechanisms between the points to complete the required decision-making and evaluation tasks. Jinping He\(^{[6]}\) adopts the Bayes theory in data fusion technology, taking multiple monitoring points of the same kind of monitoring effect as the research object, taking the deformation monitoring effect amount as the research focus, using variance as the fusion characteristic parameter, and measuring the single effect quantity. The point monitoring data is combined, correlated and combined to establish a multi-measurement abnormality diagnosis model based on Bayes fusion theory, which effectively overcomes the existing single-point monitoring data analysis and modeling method in analyzing and grasping the overall structure of the dam. The limitations of sexuality provide a new and more reasonable and effective way for dam safety assessment and safety monitoring. On this basis, Ziyang Li\(^{[2]}\) and other maritime safety monitoring data characteristics analysis, combined with multi-source information fusion technology, to build a dam safety monitoring data rationality fusion diagnosis system, through the numerical analysis of single point data analysis to achieve gross errors identification; through the feature fusion of the same monitoring type data, to realize the system errors identification; through the decision fusion to achieve the systematic errors caused by the observation base point and the special changes caused by the complex causes such as dam lesions.

5. Project example

Taking the monitoring data of the dam top line deformation monitoring point JC6 of a dam from 2003 to 2006 as an example, the gross error recognition results of the $3\sigma$ criterion, the Grubbs criterion and the mathematical model method are compared and analyzed.

Establish a statistical model of dam crest displacement\(^{[7]}\)

$$\delta = \delta_H + \delta_T + \delta_\theta$$

Stepwise regression analysis to determine specific parameters:

$$\delta = 272.681 - 8.946H + 0.099H^2 - 0.000351H^3 + 1.633 \sin \frac{2\pi t}{365} + 1.576 \cos \frac{2\pi t}{365} - 0.478 \sin \frac{4\pi t}{365} + 0.501 \cos \frac{4\pi t}{365} + 2.4730 - 44.250 \ln \theta$$

(4.9)

Where “$\delta$” is the horizontal displacement of the dam crest;
“$\delta_H$” is the water pressure component;
“$\delta_T$” is a temperature component;
“$\delta_\theta$” aging component;
“$H$” is the upstream head;
“$t$” is the cumulative number of days on the monitoring day;
“$\theta$” is the cumulative number of days divided by 100.

The correlation coefficient of the model is 0.9, indicating that the fitting effect is better.
Figure 1. JC6 regression process line of the dam crest line measurement point.

| Gross number | 3σ criterion | Grubbs criterion | Mathematical model |
|--------------|--------------|------------------|--------------------|

The sample size of the monitoring data is 51, and the discrimination result is as shown in the above table. It can be seen from the statistical model expression that the water pressure component is the main factor affecting the displacement of the dam crest. It can be seen from the above figure that the process of the displacement fitting curve and the hydraulic component are basically the same, and the measured value process line is in 2003. On March 16, 2003, May 9, 2003 and November 5, 2004, the three nodes and the water pressure component process line changed significantly. Therefore, it can be judged that there are coarse errors in the data of the three nodes.

The calculation of the 3σ criterion is extremely simple, but the accuracy is relatively low, which is suitable for error analysis of dam safety monitoring data with low precision requirements. The Grubbs criterion is applicable to small-capacity sample data. For large-capacity sample data, the accuracy is low, and it is difficult to promote the application in long-sequence dam safety monitoring data error analysis. The mathematical model method is a multi-factor equation based on the effect quantity and the environmental quantity, and has high discriminative precision. The improved probability identification method and multi-source information fusion method have large computational complexity and complicated process, and the utility is small.

6. Conclusion
This paper makes an in-depth summary of the errors in the dam safety monitoring data and its discovery methods, and analyzes the new methods that have emerged in recent years. The following conclusions were obtained:

1) The mathematical model method considers the influence of the independent variables of the measuring points, so that the gross error identification of the effect quantity is no longer a simple mathematical calculation, which makes it more objective and more suitable for the preprocessing of dam safety monitoring data.

2) The multi-factor statistical model method of constructing independent variables and dependent variables is based on the premise that there is no error in the dependent variable data. Although the statistical model method can find some errors in the dependent variable data, it can not judge the independent variable data.
References
[1] Li, Z.Y., Guo, L., Ma, H.F., et al. (2018) Rationality test of dam monitoring data based on statistical diagnosis. Advances in Science and Technology of Water Resources., 38:71-75.
[2] Li, Z.Y., Ma, H.F., Hua, W.N. (2013) Diagnosis of dam safety monitoring data rationality based on multiple-source information fusion. HYDRO-SCIENCE AND ENGINEERING., 1:41-45.
[3] Yang, J., Li, Z.K., Li, Z.X., et al. (2012) Hydraulic building safety monitoring and control. Yellow River Water Conservancy Press, Zhengzhou.
[4] Jiang, P. Zhao, J.Y. Wei, J., (2014) Error theory and data processing. National Defense Industry Press, Beijing.
[5] Xu, B. Jing, K. (2010) A Diagnostic Process of Outliers in Dam Safety Monitoring Data Analysis. Water Resources and Power., 28:64-66.
[6] He, J.P. Tang, Y.Y. Shi,Y.Q., et al. (2012) Model of Diagnosing Abnormal Behavior of Dam Based on Multi-monitoring Points and Bayes Fusion Theory. Journal of Yangtze River Scientific Research Institute., 29:63-66.
[7] Wu, Z.R. (2003) Hydraulic building safety monitoring theory and its application. Higher education press, Beijing.