Advances in Computer, Signals and Systems (2022)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2022.060112
ISSN 2371-8838 Vol. 6 Num. 1

Study on Abnormal Data Preprocessing and Preliminary Analysis Method in Landslide Monitoring System

Mingyang Guo

Hunan University of Science & Technology, Xiangtan, Hunan, 411201, China

Keywords: Landslide monitoring, Abnormal data, Noise filtering, Outlier elimination

Abstract: In the process of landslide monitoring, abnormal sensors will lead to abnormal alarm at the monitoring points, and the occurrence of this behavior will reduce the accuracy and adaptability of the early warning system. Since the data collected by the automatic monitoring equipment is an electronic signal, it needs to be transformed into the actual physical measurement value, and there is usually a certain noise in the measurement data in this process. At the same time, the monitoring equipment may have abnormal values or noise due to certain interference due to the influence of other external factors, or the monitoring data may be missing due to equipment failure. In most cases, when the original data is directly used to predict the real evolution trend of the slope, the deviation between the prediction results and the actual situation is too large or the model can not be used to predict at all. Therefore, the original measurement data needs to be processed before using the prediction model to analyze and estimate the landslide state. Furthermore, the existence of outliers will have a great impact on the estimation of sample autocorrelation, partial correlation and prediction model parameters, resulting in prediction failure. Since the occurrence of outliers is often unknown, it is very important to detect outliers and estimate their possible impact.

1. Introduction

The monitoring work carried out in different deformation stages of landslide occurrence and development is also different in terms of monitoring scope, monitoring methods and monitoring objects. Therefore, it can be divided into three levels: Level I monitoring, level II Monitoring and level II monitoring. Generally, the lower level monitoring is the supplement and improvement of the upper level monitoring, and basically includes the contents of the upper level monitoring, which are described as follows:

Level I monitoring: it mainly detects whether the deformation of the slope is within the expected stability range, which belongs to the nature of evaluation, and detects the abnormal information in the initial stage of the slope. At this stage, a small number of high-precision instruments can be selected, and the monitoring instruments can be arranged according to the principle of controlling key parts according to the engineering geological conditions and main engineering geological problems concerned.

Level II Monitoring: it mainly observes the displacement development process and trend of abnormal slope, which is of monitoring nature. At this stage, the deformation and failure mechanism of slope body shall be analyzed according to the monitoring results, and the slope body shall be strengthened and reinforced pertinently; the monitoring instrument shall be supplemented and improved according to the deformation development process and deformation range of the slope. If necessary, the combination of appearance method and interior method can be considered.

Level III monitoring: when the safety problem is very prominent in the later stage of slope instability, the monitoring means suitable for this stage shall be arranged for continuous observation to provide guarantee for construction safety, which belongs to the nature of early warning and prediction. The internal instruments at this stage may be out of range and fail, and the appearance monitoring shall be the main method [1].

In the above hierarchical monitoring process, smoothing is one of the basic preprocessing methods in the time series research process. For a group of data X and Y obtained from real-time sampling and measurement, first remove the “noise” in the data through smoothing, and then effectively solve the linear parameters of the corresponding fitting polynomial. In scientific research, smoothing is widely used. Smoothing can reduce the impact caused by statistical errors in the measurement process. In cases where the average value cannot be obtained by multiple repeated measurements, and in those measurement segments where y changes abruptly with X, smoothing is fully used. The main models used for smoothing time series include moving average model, weighted moving average model and exponential smoothing model [2] by applying smoothing model to preprocess, we can reduce the variation amplitude of fluctuation in time series, so as to obtain the variation trend of time series from time series.

For some special observation data series, if the fluctuation of the series is relatively violent and the length of the series is large, polynomial smoothing or orthogonal polynomial smoothing algorithm is generally used to smooth them. In practical sense, the orthogonal polynomial smoothing algorithm can maintain the characteristics of the original data and facilitate the identification of landslide stage and landslide prediction in the subsequent fusion processing. Although piecewise polynomial smoothing can improve the correlation between observation time and deformation, due to its smoothing results, the identification of landslide stage will not be obvious, which will have an adverse impact on landslide time prediction.

2. Data Smoothing Method

In the process of using multi-source sensors to monitor landslide mass, due to the influence of various factors, landslide monitoring data always inevitably contain noise. Among them, random noise is accompanied by the whole monitoring process. The performance of these noises on the graph is such as “burrs and spikes”. If the unsmoothed data are directly analyzed in the next step, these noises may be amplified and directly affect the results of multi-source data fusion. In order to obtain high-quality monitoring data, it is necessary to smooth the monitoring data before multi-source data fusion analysis. Through data smoothing, the interference components can be eliminated and the variation characteristics of the original curve can be maintained. Liu Jinsen and others eliminated the high-frequency noise in the original observation data by using the linear sliding smoothing principle, and obtained high-quality data. Weiqing uses window moving polynomial smoothing and wavelet denoising methods to denoise slope monitoring data, and discusses the applicability of these two methods in specific cases [3] At present, there are some research results and practical experience in monitoring data denoising in various fields and disciplines. This section mainly discusses four widely used data smoothing methods in the field of landslide deformation monitoring: simple moving average method, weighted moving average method, exponential moving average method and savitzky Golay smoothing method [4].

(1) Simple moving average method

The core idea of the algorithm is to solve the average of two or more adjacent items in the original sequence of landslide monitoring, as a smooth value, and re form a new sequence.

(2) Weighted moving average method

Simple moving average is equal to treating each data in the average period equally and giving them the same weight, which is 1 / m. However, according to experience, when predicting the data change trend, the closer the data is, the greater the reference significance is. Therefore, different weights should be given to the data at different times.

(3) Exponential moving average method

The exponential moving average method is a weighted average of all past data. All weights are uniquely determined by the smoothing coefficient, and the weighted index decreases with the passage of time, so it is called exponential moving average.

4) Savitzky Golay smoothing method

Savitzky Golay smoothing method (also known as moving window polynomial quasi smoothing method) was proposed by savitzky and Golay in 1964. It is a method of best fitting by moving window using least square method.

3. Monitoring Data Faqs

Because the rain gauge data represents the accumulated rainfall within the time of two adjacent data and is the accumulated amount within a time period, it is different from the processing method of displacement data and needs to be processed separately [5] Through the data analysis of various models of multiple equipment manufacturers, it is found that there are the following common problems:

① The same data is sent repeatedly, resulting in errors in the calculated cumulative rainfall and rainfall intensity. This problem is often caused by the data transmission mechanism of the monitoring equipment. The duplicate data can be eliminated through data traversal on the server side.

② The returned rainfall value is cumulative rather than the rainfall in two adjacent times, which will also lead to the calculation of wrong cumulative rainfall and rainfall intensity.

In addition to rainfall, in the process of other data acquisition and transmission, due to the influence of equipment error, human factors and environmental factors, a small part of data will be different from the overall behavior characteristics, structure or correlation. This part of data is called outliers, also known as outliers or outliers [6]. The occurrence of outliers will bring false information and affect the accuracy and accuracy of subsequent multi-source data fusion results. Therefore, outliers in the original data must be detected and eliminated.

4. Overview of Abnormal Data Processing Methods in Monitoring Process

In time series, observations are sometimes affected by abnormal events. The influence of uncertain interference or error will cause false prediction results. These abnormal observations can be called outliers. The existence of outliers in time series will have a great impact on the estimation of sample autocorrelation, partial correlation and prediction model parameters, resulting in prediction failure. Since the occurrence of outliers is often unknown, it is very important to detect outliers and estimate their possible impact [7]. In this section, we mainly analyze and discuss some methods of outlier detection and processing, and determine the scheme of outlier processing of monitoring data series suitable for general types of systems. During the collection and transmission of the data obtained through the monitoring of measuring instruments, due to the objective existence of environmental interference or human factors, the value of individual data may not conform to the actual monitoring situation, or even the data may be missing. The type of sample data is what we call abnormal value. Outliers are not only the extreme performance caused by the inherent random error in the data, but also may be caused by negligent error [8] Through analysis, in general, we can deal with outliers in the following four ways:

1) Keep the outliers in the sample and participate in the subsequent data unified calculation.
2) It is allowed to eliminate outliers and exclude outliers from the sample.
3) It is allowed to eliminate outliers and add appropriate observations to the sample.
4) Correct the abnormal value after finding the actual cause.

In various monitoring systems based on landslide prediction, the authenticity of data is particularly important. In order to restore the objective authenticity of data and obtain better prediction and analysis results after later fusion, it is necessary to eliminate the abnormal values of the original data first; In addition, for the manually observed data and the data obtained through the data acquisition system, the interference of “noise” will inevitably be superimposed (in the fitting process, it is reflected in the output graph that some “burrs and spikes” appear on the curve). In order to improve the quality of data, the data must be smoothed to remove noise interference. However, in the actual operation process, it is difficult to distinguish the systematic error, machine error and fault error encountered in the measurement. In the statistical test, we make full use of mathematical statistics to analyze the error, so as to correctly evaluate the measurement data and provide useful information.

Generally, for data preprocessing, algorithms are selected and designed based on the following principles:

1) Whether the algorithm is stable;
2) Whether the logical structure of the algorithm is simple;
3) Whether the number of operations of the algorithm and the amount of storage space required by the algorithm are as small as possible.

Based on the above principles, the basic steps of eliminating outliers in the system are to first set a confidence level value, and then determine a confidence limit. When there is an error value exceeding this limit, it will be determined that it is an outlier, so as to eliminate it. Under the premise of normal distribution of monitoring data W, the common methods to eliminate outliers include laida method, shoville method and first-order difference method [9].

5. Laida Method (Unequal Confidence Probability)

Let the measured be measured with equal precision, x1, X2..., xn are obtained independently, the arithmetic mean X and residual error VI = xi-x (I = 1,2,..., n) are calculated, and the standard deviation is calculated according to Bessel formula σ, If the residual error VB (1 < = B < = n) of a measured value XB satisfies | VB | = | xb-x | > 3 σ It is considered that XB is a bad value with gross error value and should be eliminated. When sorting out test data, we often encounter such a situation, that is, a few suspicious data with large deviation are found in a group of test data. Such data are called outlier or exceptional data, which are often caused by negligence error.

6. Schweiler Method (Equal Confidence Probability)

In the N measurement results, if the number of possible errors is less than half, the abnormal value will be eliminated. Firstly, the current confidence probability is set as L-L / 2n. Through the confidence probability value, we can calculate the schoville coefficient. At the same time, in engineering application, in order to improve the calculation efficiency, it can also be processed directly from the corresponding coefficient table by looking up the table. Where mean = sum (V) /

N, SD = sqrt (sum (V (I) * V (I) - mean * mean, I = 1 to n) / (n - 1)). If a sample data d satisfies | V (d) | > W (n) * SD, V (d) is considered to be an abnormal data and is eliminated

Table 1 Schoville coefficient table

n	3	4	5	6	7	8	9	10	11	12
W(n)	1.38	1.53	1.65	1.73	1.80	1.86	1.92	1.96	2.00	2.03
n	13	14	15	20	30	40	50	100	200	500
W(n)	2.07	2.10	2.13	2.24	2.39	2.49	2.58	2.81	3.02	3.20

The characteristics of this method are:

(1) Suitable for real-time data acquisition and processing;

(2) The accuracy is not only related to the allowable error limit, but also related to the accuracy of the measured values of the first two points;

(3) If the variation law of the measured physical quantity is not monotonically increasing or monotonically decreasing, this method will produce large errors at the inflection point of the function, which will not be used in serious cases.

7. First Order Difference Method (Prediction Comparison Method)

The algorithm flow of this method is as follows: first, the first two measured values are used to estimate the new measured value, and then the estimated value is compared with the actual measured value. If the result is greater than the preset allowable difference limit, the measured value is eliminated.

The characteristics of this method are:

(1) Suitable for real-time data acquisition and processing;

(2) The accuracy is not only related to the allowable error limit, but also related to the accuracy of the measured values of the first two points;

(3) If the variation law of the measured physical quantity is not a monotonic increasing or decreasing function, this method will produce large errors at the inflection point of the function, which will not be used in serious cases.

8. Conclusion

Generally, the landslide displacement observation data shall reflect the stable and continuous deformation process of the slope. Once an abnormal value appears in the data, it may be the gross error caused by the inaccurate reading of the observer, the error caused by equipment failure or instability, or the real anomaly of the sudden deformation of the rock and soil mass before the slope instability and failure. Chen Zhijian divides the abnormal values in geotechnical engineering monitoring into false anomalies, apparent anomalies and symptomatic anomalies, and expounds the identification principles and methods of abnormal values in slope monitoring. Based on the research of Chen Zhijian et al. [10], the identification principles of deformation outliers can be summarized into the following five points;

(1) Non single point principle. The deformation and failure of slope is not limited to isolated displacement data points, and the deformation of monitoring points on potential sliding body should have a certain correlation within a certain range. Therefore, when detecting outliers, it is necessary to compare the displacement value and deformation direction of adjacent dirty measuring points. The greater the correlation of outliers between adjacent monitoring points, the greater the possibility of true outliers. If it is an isolated large anomaly, it is likely to be a false anomaly [11].

(2) Principle of consistency. The displacement monitoring data reflecting the slope instability and failure process usually follow a certain regularity and are disturbed as much as possible, but the

overall direction of deformation should be consistent and continuous in time. For the displacement outliers with large, small and disordered directions, it is necessary to further analyze the causes, perhaps due to interference or technical and equipment failure.

(3) Progressive principle. The deformation and failure of slope usually has a process of occurrence and development, and its deformation response should show gradual and cumulative. If accidental abnormal information is found, it may be gross error or error.

(4) Principle of rationality. For different types, manufacturers and models of monitoring instruments, the measurement range and accuracy may be different, but the data measured by each instrument shall be within its allowable normal range, and the abnormal values beyond its normal measurement range may be errors.

(5) Relevance principle. Deformation displacement is the direct reflection of slope deformation and failure process, and the slope deformation process is often related to atmospheric precipitation, groundwater level and human activities. Therefore, there is often a certain correlation between the monitoring information of different elements. If the occurrence of displacement anomaly is related to rainfall process or groundwater level change, it is more likely to be a true anomaly.

(6) Visibility principle. When corresponding abnormal phenomena are observed near the monitoring point of abnormal data, such as cracks in nearby slope, cracks in house wall, ground bulging, sudden change of spring water and other visual phenomena, the abnormal value is more likely to be true [12].

According to the above principles, it can be seen that the traditional anomaly detection method can only simply find the anomaly value from the observation data, but can not analyze and explain the cause of the anomaly. However, the anomaly value that truly reflects the sudden change of slope rock ± body deformation needs to be retained and deeply studied [13]. Therefore, the anomaly detection of monitoring data cannot completely rely on relevant algorithms, It is necessary to use the combination of anomaly detection algorithm and professional experience analysis to detect and eliminate outliers.

References

[1] Chaoyang He. Research on key technology and application of landslide real-time monitoring and early warning system.2020.Chengdu University of Technology,PhD dissertation.

[2] Junqing Fan. Research on multi-source heterogeneous sensor information fusion method for landslide monitoring.2015.China University of Geosciences, PhD dissertation.

[3] Zhiwei Wang. Research on multi-source heterogeneous monitoring data fusion algorithm of loess landslide.2020.Chang'an University,MA thesis.

[4] Chuowen Feng, et al.”Comparative study on detection methods of abnormal wind power data.” New technology of electrical energy 40.07(2021):55-61. doi:CNKI:SUN:DGDN.0.2021-07-007.

[5] Bo Yu, et al.”Sensitive data identification and abnormal behavior analysis of unstructured documents.” Journal of Intelligent Systems: doi:10.11992/tis.202104028.

[6] Deping Gao.”Anomaly detection of mobile terminal network data based on isolated forest.” information technology .06(2021):125-129. doi:10.13274/j.cnki.hdzj.2021.06.023.

[7] Ge Yang, et al.”Research on abnormal value identification technology of dam safety monitoring data based on singular spectrum analysis.” Hydropower generation .

[8] Jiaxu Huang, Xianhui Zeng,and Chenjun Shi.”Research on equipment energy consumption anomaly recognition algorithm based on real-time data stream feature extraction.” Information technology and network security 40.05(2021):45-50. doi:10.19358/j.issn.2096-5133.2021.05.008.

[9] Li Hui, et al.”Toward data anomaly detection for automated structural health monitoring: Exploiting generative adversarial nets and autoencoders.” Structural Health Monitoring 20.4(2021): doi:10.1177/1475921720924601.

[10] Zhang Ting, et al.”A new method of data missing estimation with FNN-based tensor heterogeneous ensemble learning for internet of vehicle.” Neurocomputing 420.(2021): doi:10.1016/J.NEUCOM.2020.09.042.

[11] “Alibaba Group Holding Limited; Patent Application Titled “Method And Device For Determining Data Anomaly” Published Online (USPTO 20200329063).” Internet Business Newsweekly .(2020).

[12] Boukela Lynda, et al.”An outlier ensemble for unsupervised anomaly detection in honeypots data.” Intelligent Data Analysis 24.4(2020): doi:10.3233/IDA-194656.

[13] Yufei Guo.Study on prediction and early warning system of single landslide.2013.China University of Geosciences (Beijing),PhD dissertation.