Multi-sensor fusion of data for monitoring of Huangtupo landslide in the three Gorges Reservoir (China)

Junqi Liu\textsuperscript{a}, Huiming Tang\textsuperscript{a}, Qi Li\textsuperscript{b}, Aijun Su\textsuperscript{a}, Qianhui Liu\textsuperscript{c} and Cheng Zhong\textsuperscript{a}

\textsuperscript{a}Three Gorges Research Center for Geo-hazard, Ministry of Education, China University of Geosciences, Wuhan, China; \textsuperscript{b}State Key Laboratory of Geomechanics and Geotechnical Engineering, Institute of Rock and Soil Mechanics, Chinese Academy of Sciences, Wuhan, China; \textsuperscript{c}Department of Information Management, Peking University, Beijing, China

ABSTRACT
There hides a certain relationship among various monitoring data in a landslide, and the mining of this relationship is of significance to landslide research. In this paper, we first collect multiple monitoring data of riverside 1\# slump-mass of Huangtupo landslide, the Three Gorges Reservoir Region, China, including Global Positioning System (GPS) monitoring data, inclinometer data, reservoir water level, rainfall, water content, crack width, groundwater level and temperature data, etc. By adopting the combination of quantitative statistics and qualitative simulation method for multi-sensor fusion monitoring data analysis, we overcome the one-sidedness of using a single method or single data type. The result of fusion analysis has indicated that in time periods with low rainfall or when the rainfall is not the major factor, main factors affecting landslide movement are crack development, water content of the landslide and water level of the Three Gorges Reservoir. Compared with the actual monitoring data, the fusion analysis results has a maximum error of 1.9\%, which shows a good effect.

1. Introduction
Since the publication of Big Data: A Revolution That Will Transform How We Live, Work, and Think by Viktor and Kenneth in 2013, Big Data analysis and processing has drawn a great deal of attention from the scholars around the world. Wu et al. (2014) proposed HACE theorem and conducted in-depth studies on data mining (DM) of Big Data. Vitolo et al. (2015) expounded the data processing, simulation and forecast for large heterogeneous datasets. Recently, the Big Data analysis has provided a new way of handling problems – there hides a certain relationship among data, and the mining of this relationship is very meaningful for a comprehensive understanding of things.

CONTACT Junqi Liu liujqg@126.com
© 2018 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
Landslide is a kind of common geological disasters, and many factors are involved in its preparation, development and occurrence, which certainly have an implicit relationship. The searching of the hidden relationship is helpful for landslide forecasting. As early as in 1960s, Saito (1969) proposed the Saito model for landslide forecasting to determine the landslide failure time; Lan et al. (2002) built the multiple regression model of slope stability analysis based on Geographical Information System (GIS); in combination with some factors such as rainfall, groundwater level, etc., Baum et al. (2005) made evaluations on regional landslide disasters through GIS; Guzzetti et al. (2006) assessed the landslide susceptibility models according to the discriminant analysis of various topography factors; Herrera et al. (2009) constructed the landslide forecasting model by combining traditional displacement measurement (total station, inclinometers and differential Global Positioning System (GPS)) and ground-based synthetic aperture radar (SAR) data; based on discriminant analysis and regression analysis of various topographic factors, Rossi et al. (2010) investigated the susceptibility zonation of regional landslides; Doglioni et al. (2012) adopted the multiple regression analysis method to forecast the landslide; using some monitoring instruments such as thermometer, rain gauge and ground-based interferometric radar, Intrieri et al. (2012) constructed a landslide early-warning system; on the basis of artificial immune system and evolutionary algorithms and combined with limit equilibrium analysis, Gao (2014) proposed a forecasting method of landslide disasters; Lombardo et al. (2015) compared the binary logistic regression with stochastic gradient tree boost on landslide susceptibility research; Segoni et al. (2015) performed the early-warning research on the landslide by integrating rainfall thresholds and susceptibility maps; Lian et al. (2015) made the prediction of the displacement in a landslide through neural network; Melillo et al. (2016) conducted an exploration of the effects of rainfall on the sliding of a landslide in detail.

With the combination of the current Big Data analysis method, this paper takes a specific single landslide in the Three Gorges Reservoir Region of central China as an example. By integrating the quantitative analysis of mathematical statistics and the trend simulation of the neural network, we study on multi-sensor monitoring data fusion, and attempt to find the hidden relationship of time and space among various monitoring data.

2. Monitoring data and data pre-processing

The monitoring data of riverside 1# slump-mass of Huangtupo landslide (in the Three Gorges Reservoir Region, China) are used in this study. Figure 1 shows the layout of the monitoring points, in which the upper part denotes the Three Gorges Reservoir, LJ1# denotes riverside 1# slump-mass of Huangtupo Landslide, GPS1 and GPS2 are two surface GPS monitoring points, CX denotes the drilling inclinometer and underground water level monitoring point, HSL denotes water content monitoring point which was installed in the sliding body, LXJ denotes crack width monitoring point which was installed in the crack position of the test hole, and WDJ denotes temperature measuring point which was installed in the sliding body.
The used 11 types of data were collected from June 6, 2014 to September 30, 2014, including the following:

1. the Three Gorges Reservoir water level (reservoir water level);
2. rainfall;
3. rainfall in the past 48 h;
4. crack width;
5. temperature;
6. groundwater level;
7. water content;
8. GPS1 displacement;
9. GPS2 displacement;
10. the distance between GPS1 and GPS2 monitoring point;
11. inclinometer data (Table 1).

The basic requirements of data pre-processing are described as below. These 11 types of data except the monitoring time should be distributed uniformly on the time scale, and a unified coordinate system was built on the space scale so that a set of data was included in every unit of time in this coordinate system. In the present study, one value was selected for each type of data each day, as shown in Table 1. To be specific, if more than one value was collected actually in a day, we should use the average of the measured value; if only a value was collected actually in several days, we should make the following calculation: the used value in the present study = \( \frac{\text{the previous value} + (\text{the latter value} - \text{the previous value})}{\text{(the number of days)}} \).

![Arrangement diagram of monitoring point.](image)

**Table 1.** The data used (part of the data).

| Date     | Reservoir water level (m) | Rainfall (mm) | Rainfall in the past 48 h (mm) | Crack width (cm) | Temperature (°C) |
|----------|---------------------------|---------------|-------------------------------|------------------|-----------------|
| 2014.6.6 | 153.9                     | 6             | 6                            | 4.10             | 19.1            |
| 2014.6.7 | 153.6                     | 0             | 6                            | 4.11             | 19.1            |
| 2014.6.8 | 153.4                     | 0             | 0                            | 4.11             | 19.1            |
| 2014.6.9 | 153.1                     | 0             | 0                            | 4.12             | 19.1            |
| 2014.6.10 | 152.9                    | 0             | 0                            | 4.13             | 19.2            |

(1) the Three Gorges Reservoir water level (reservoir water level); (2) rainfall; (3) rainfall in the past 48 h; (4) crack width; (5) temperature; (6) groundwater level; (7) water content; (8) GPS1 displacement; (9) GPS2 displacement; (10) the distance between GPS1 and GPS2 monitoring point; (11) inclinometer data (Table 1).
Besides, for the inclinometer data and the GPS data, the components along the main sliding direction were used to substitute for the measured data.

3. Methods

As stated above, the present study was conducted based on the Big Data analysis method. In other words, the collected 11 types of monitoring data were used as the main basis for analysis. The inclinometer data in the landslide were adopted as the sign for the overall displacement of the landslide (since the inclinometer data were less affected than GPS data), and the hidden relationships between the inclinometer data and the other data were then determined by means of variance calculation, correlation analysis, cluster analysis, regression analysis and back propagation (BP) neural network as shown in Figure 2.

The specific implementation procedure is described as follows:

1. Collect the 11 types of data of 1# slump-mass, and select one value for each type of data each day for further calculations and simulations;
2. Eliminate the singular data point in each type of data (i.e. the data point with great deviations) based on the variation calculation results;
3. Conduct correlation analysis on the data and determine the correlation coefficients among these 11 types of data;
4. Based on correlation coefficients, conduct cluster analysis and classify the factors according to the effects on the landslide displacement; the main aim of correlation analysis and cluster analysis is to determine the variables that would be involved in further analyses.
5. Conduct regression analysis on the selected variables (if there are many variables exhibiting great effects, the set of the variables with the greatest effects is selected for analysis; if there are few variables exhibiting great effects, the variables with large or moderate effects are selected), and obtain the regression coefficient and the fitting effect of the regression analysis;
6. Conduct BP neural network simulation on the variables that are involved in regression analysis, in which two-thirds of the data are adopted as training data,
and one-third of the data is adopted as verification (it should be noted that the actual ratio can be adjusted in accordance with the data volume), and acquire the optimal iteration model and the evaluation data for the expected fitting degree;

7. Adjust the number of the variables involved in analysis (e.g. the variables with great effects are selected first, and then the variables with great and moderate effects are selected, and then all of the variables are selected), return to (5) and (6); repeat this process for several times until the regression coefficient with the highest fitting degree (i.e. the minimum error with the BP neural network simulation result) is found;

8. Obtain the multi-sensor fusion analysis results.

Among various monitoring data of landslide, there inevitably exists certain objective relationship, but this relationship is not clear at present. Compared with other related research, this method has several features:

1. To classify and filter all kinds of data through correlation analysis and cluster analysis, provided classification data for regression analysis and neural network analysis. The selection of sample data is very important for BP neural network analysis;
2. Through circular process of different data assembly, and mutual authentication between regression analysis and BP neural network analysis (the above steps 5–7), we get the optimized results;
3. This method combined mathematical statistics-based quantitative analysis and BP neural network-based trend analysis, and fused a variety of monitoring data, thereby overcoming the one-sidedness using a single method or a few data.

4. Results and discussion

Using the method described above, the 11 types of data are analysed and the following conclusions can be drawn.

1) Through correlation analysis on the 11 types of data, it can be found that the inclinometer data are significantly correlated with two GPS displacements, the distance between two GPS monitoring points, reservoir water level, water content, crack width, groundwater level and temperature at the level of 0.01.

2) Based on the correlation coefficients among various data, cluster analysis is conducted on the other variables through inclinometer data and GPS data, and the factors (or variables) that affect the landslide’s sliding can be classified as four types – the factors with the greatest effect (I-type), reservoir water level, groundwater level, water content and crack width; the factor with the relatively large effect (II-type), temperature; the factor with the moderate effect (III-type), the distances between two GPS monitoring points; the factors with the smallest effect (IV-type), rainfall and rainfall in the past 48 h (Table 2; Figure 3).

Since the inclinometer data and GPS data are direct factors that can reflect whether a landslide slides or not, the cluster analysis was performed based on the inclination data and GPS data. Also, the measured inclination data are stable and the
slippage of the inclination data along the main sliding direction is used for the final landslide forecasting. Therefore, two GPS displacements are regarded as the factors that impose relatively large effect on the landslide’s sliding.

According to the correlation analysis and cluster analysis results, rainfall and rainfall in the past 48 h impose little effect on the landslide’s sliding during the period from June 6, 2014 to September 30, 2014, even the influence extent is smaller than that of the distance between two GPS measuring points. The distance between the two GPS points is meaningless to a certain extent, so these two factors would be ignored in further analyses.

(3) Conduct regression analysis on inclinometer data, GPS1 displacement, GPS2 displacement, the distance between GPS1 and GPS2 monitoring points, the Three Gorges Reservoir water level, crack width, temperature, groundwater level and water content. The results are shown in Tables 3–5.

From Tables 3 and 4, we can see that independent variables (GPS1 displacement, GPS2 displacement, the distance between GPS1 and GPS2 monitoring point, the Three Gorges Reservoir water level, crack width, temperature, groundwater level and water content) can accurately predict dependent variable (the inclination data), with the coefficient of determination of 0.997, that is, the predictive fitting results are favourable. Table 5 lists the regression coefficients.
By comparing the inclination data calculated by the regression coefficient with the measured data, the maximum error of regression analysis is 1.9%.

(4) Conduct BP neural network simulation on inclinometer data, GPS1 displacement, GPS2 displacement, the distance between GPS1 and GPS2 monitoring point, the Three Gorges Reservoir water level, crack width, temperature, groundwater level and water content. A set of data is selected from every five sets of data as the verification data, and other data are selected as the training data.

The simulation results demonstrate that when 16 neurons are used in the hidden layer, the simulation could obtain the optimal results. The optimal model exhibits favourable fitting performances, with the maximum error of 0.51% (Table 6; Figure 4).

The BP neural network simulation results can further verify the regression analysis results. Since both two fitting results are desirable (with the maximum error below 2%), the hidden relationship among various monitoring data of LJ1# landslide can be written as follows:

The inclination displacement \(=0.492 \times \text{crack width} + 0.278 \times \text{water content} + 0.248 \times \text{the Three Gorges Reservoir water level} + 0.102 \times \text{groundwater level} - 0.072 \times \text{GPS1 displacement} + 0.070 \times \text{the distance between two GPS monitoring points} - 0.026 \times \text{temperature} + 0.003 \times \text{GPS2 displacement}\)

where the inclination displacement denotes the component of inclination along the main sliding direction, and the GPS displacement denotes the component of GPS displacement along the main sliding direction.

Based on the whole fusion analysis process and its results, in time periods with low rainfall, crack width has the greatest impact on the inclination displacement of 1# slump-mass of Huangtupo landslide, which needs to be monitored specially. Among all the monitoring data, landslide water content, reservoir water level and groundwater level have positive correlation with inclination displacement, and this correlation is consistent with the general knowledge of landslide stability. And landslide water content has the greatest impact among the three. GPS surface displacements have very small impact on the inclination displacement due to external interference factors, such as farming, construction and other human activities. In addition, from the data analysis, the impact of temperature on the displacement is also very small.

| Table 3. F test. Sum of squares | DOF  | MS     | Significance |
|-------------------------------|------|--------|--------------|
| Regression                    | 178.849 | 8 | 22.356 | 0.000* |
| Residual                      | 0.498 | 108 | 0.005 |
| Total                         | 179.347 | 116 |

| Table 4. Summary sheet. | Table Layout. |
|------------------------|----------------|
| \(R\) | \(R^2\) | Adjusted \(R^2\) | Error of estimation |
| 0.999 | 0.997 | 0.997 | 0.0679 |

*Forecasting variable (inclination): (constant), GPS1 displacement, GPS2 displacement, the distance between two GPS monitoring points, temperature, crack width, water content, groundwater level, reservoir water level.
The fusion analysis results are in good agreement with the measured data, with a maximum error within 2%, and the results are close to the results of neural network simulation as well. This indicates that the multi-sensor data fusion method combining multiple methods is reliable to some extent. The fusion results show that when rainfall is not the major factor, main factors affecting landslide movement are crack development, landslide water content and reservoir water level, and crack development is most relevant to the movement.

This paper adopted the Big Data analysis and explored the hidden relationship among various data mainly based on data analysis. The error of fusion analysis results is within 2%, that is, the fusion analysis is favourable. Some further discussions are described as follows.

5. The problem of rainfall

In common sense, rainfall is one of the major factors affecting landslide. However, in this paper, rainfall has little impact which can even be neglected. The reasons are as follows:

1. In the research period, the rainfall in the area is relatively small. There are only three times of heavy rain greater than 35 mm that lasted for 1 day. About two-thirds of the time, there is no rain, and only small rain ranging 3–10 mm occurred in other periods.
2. The landslide studied in this paper is creep landslide and under development, so this kind of rainfall frequency and intensity has little impact on landslide movement.

3. The rainfall factor is integrated into the landslide water content factor to some extent.

6. Multi-sensor data fusion problem

The data collected by monitoring sensors are continuous on time but not continuous on space. The data obtained by each sensor only reflect the local feature of a single measure point at a particular aspect, so the monitoring data from multi-sensor are not consistent, and even can be contradictory. Multi-sensor data fusion can remove its one-sidedness and inconsistency to some extent.

This paper describes a method of combining mathematical statistics and BP neural network to solve data fusion of multi-sensor. The quantitative influence coefficient of multiple monitoring data on the measured oblique displacement is obtained, which achieves the objective of multi-sensor data fusion.

At the same time, the selection of sample data types can greatly affect the result of neural network simulation. Through the method shown in Figure 2, the quantitative mathematical statistics can be effectively combined with the tendency simulation of BP neural network to achieve a better result.

7. Verification using new data and dynamic updating

The result of data fusion discussed in this paper can be updated dynamically with updating data. If there is updated data, then cycle analysis can be performed again based on previous analysis to update the result.

8. Advantages and disadvantages of Big Data analysis

As a recognized resource, data and information have been showing explosive growth over the years. Based on the partial sample data, and the study of many models, the certainty of traditional research results is affected in the face of data with low value density and rapid growth. Big Data studies are based on large amounts of data rather than partial sample data, based on actual data rather than some certain model, based on large volume of data with low value density rather than a set of selected data with high value density. It can quickly extract valuable information from a large number of data with low value density and evaluate some of the trends or probabilities that may arise. Big Data analysis is characterized by the fact that the data set ‘is not a random sample, but the whole data’, the allowable data quality ‘is not accuracy, but the hybrid’, the revealed meaning of the data ‘is not causality, but correlation’ (Viktor and Kenneth 2013). This characteristics can significantly help solve the problem of random sampling and small sample space, the lack of reliable mechanism of action, causality and dynamic model, the limitation of making determination and prediction only based on a small amount of observation data and some of the inherent models.
This method depends heavily on data. For better results, more data should be collected and are needed to be dynamically updated.

9. Conclusion

Landslide is a kind of geological disaster that has frequently occurred around the world. In recent 5 years, over 10,000 landslides occurred in China each year, causing serious life and property loss.

This paper proposes a multi-sensor monitoring data fusion method based on Big Data: using the various monitoring data as the main bases, and integrating multiple methods to determine the hidden relationship in various monitoring data. Moreover, this method combines mathematical statistics-based quantitative analysis and BP neural network-based trend analysis, that is, determine the data types involved in fusion analysis based on mathematical statistics results, so as to adjust the data types involved in fusion analysis and minimize the errors. The integration of multiple methods can overcome the one-sidedness using a single method to a certain extent and then provide reliable results.

The fusion analysis results of Huangtupo landslide in the Three Gorges Reservoir show that in areas or time periods with relatively low rainfall, three main factors affecting landslide movements are crack development, landslide water content and reservoir water level, which need to be monitored specially. Crack width is most relevant to the landslide movement.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

Support for this research is partly provided by the National Natural Science Foundation of China through Grant Nos. 41572336, 41502280, 41502290, 41102210.

References

Baum RL, Coe JA, Godt JW, Harp EL, Reid ME, Savage WZ, Schulz WH, Brien DL, Chleborad AF, McKenna JP, Michael JA. 2005. Regional landslide-hazard assessment for Seattle, Washington, USA. Landslides 2(4):266–279.

Doglioni A, Fiorillo F, Guadagno F, Simeone V. 2012. Evolutionary polynomial regression to alert rainfall-triggered landslide reactivation. Landslides 9(1):53–62.

Gao W. 2014. Forecasting of landslide disasters based on bionics algorithm (Part 1: Critical slip surface searching). Comp Geotech. 61:370–377.

Guzzetti F, Reichenbach P, Ardizzone F, Cardinali M, Galli M 2006. Estimating the quality of landslide susceptibility models. Geomorphology. 81(1–2):166–184.

Herrera G, Fernandez-Merodo JA, Mulas J, Pastor M, Luzi G, Monserrat O. 2009. A landslide forecasting model using ground based SAR data: the Portalet case study. Eng. Geol. 105(3–4):220–230.

Intrieri E, Gigli G, Mugnai F, Fanti R, Casagli N. 2012. Design and implementation of a landslide early warning system. Eng Geol. 147:124–136.
Lan HX, Wu FQ, Wang SJ. 2002. GIS based landslide CF multi-variable regression model and its application. *J Mount Sci*. 20(6):732–737.
Lian C, Zeng ZG, Yao W, Tang HM. 2015. Multiple neural networks switched prediction for landslide displacement. *Eng Geol*. 186(SI):91–99.
Lombardo L, Cama M, Conoscenti C. 2015. Binary logistic regression versus stochastic gradient boosted decision trees in assessing landslide susceptibility for multiple-occurring landslide events: application to the 2009 storm event in Messina (Sicily, southern Italy). *Nat Hazard*. 79(3):1621–1648.
Melillo M, Brunetti MT, Peruccacci S, Gariano SL, Guzzetti F. 2016. Rainfall thresholds for the possible landslide occurrence in Sicily (Southern Italy) based on the automatic reconstruction of rainfall events. *Landslides* 13(1):165–172.
Rossi M, Guzzetti F, Reichenbach P, Mondini AC, Peruccacci S. 2010. Optimal landslide susceptibility zonation based on multiple forecasts. *Geomorphology* 114(3):129–142.
Saito M. 1969. Forecasting time of slope failure by tertiary creep. Proceedings of the 7th International Conference on Soil Mechanics and Foundation Engineering, Mexico City, 2. p. 677–683.
Segoni S, Lagomarsino D, Fanti R, Moretti S, Casagli N. 2015. Integration of rainfall thresholds and susceptibility maps in the Emilia Romagna (Italy) regional-scale landslide warning system. *Landslides* 12(4):773–785.
Viktor M-S, Kenneth C, 2013. Big Data: a revolution that will transform how we live, work, and think. *Eamon Dolan/Houghton Mifflin Harcourt*. p. 242.
Vitolo C, Elkhatib Y, Reusser D, Macleod CJA, Buytaert, W. 2015. Web technologies for environmental Big Data. *Environ Model Software* 63:185–198.
Wu XD, Zhu XQ, Wu GQ, Ding W. 2014. Data mining with Big Data. *IEEE Trans Knowl Data Eng*. 26(1):97–107.