Rainfall feature extraction using cluster analysis and its application on displacement prediction for a cleavage-parallel landslide in the Three Gorges Reservoir area

Y. Liu[1], L. Liu[2]

[1]School of Mechanical Engineering and Electronic Information, China University of Geosciences, Wuhan, 430074, China
[2]Department of Civil & Environmental Engineering, University of Connecticut, Storrs, CT 06269-3037, USA

Correspondence to: L. Liu (Lanbo.Liu@UConn.edu)

Abstract. Rainfall is one of the most important factors controlling landslide deformation and failures. State-of-art rainfall data collection is a common practice in modern landslide research worldwide. Nevertheless, in spite of the availability of high-accuracy rainfall data, it is not a trivial process to diligently incorporate rainfall data in predicting landslide stability due to large quantity, tremendous variety, and wealth multiplicity of rainfall data. Up to date, most of the pre-process procedure of rainfall data only use mean value, maxima and minima to characterize the rainfall feature. This practice significantly overlooks many important and intrinsic features contained in the rainfall data. In this paper, a feature extraction method using a cluster analysis (CA) is employed for the analysis of rainfall data. With this approach we effectively revealed the most significant features contained in a rainfall sequence and greatly reduced the burden for processing large amount of rainfall data. Meanwhile, it greatly improves the spectrum of usefulness of rainfall data.

For showing the efficiency of using the CA characterized rainfall data input, we present three schemes to input rainfall data in back propagation (BP) neural network to forecast landslide displacement. These three schemes are: the original daily rainfall, monthly rainfall, and CA extracted rainfall features. Based on the examination of the root mean square error (RMSE) of the landslide displacement prediction, it is clear that using the CA extracted rainfall features input significantly improve the ability of accurate landslide prediction.

1 Introduction

Landslides are one of the major geological hazards causing major life loss and socio-economic disruption each year world-widely. An early warning system for potential landslides in steep mountainous area with landslide-prone segments is an effective approach to avoid property damage and casualties. To make the early warning system functions effectively and reliably, information about the behaviour of the landslides, including the sliding mechanics, the potential triggering mechanism and the critical precursors of slope stability for issuing emergency warnings, is the major parameter to be sought. The most critical parameters for early warning output are creep velocity, displacement and instability prediction (Sassa et al. 2009).

Rainfall is not only a crucial index of landslide analysis but also a significant factor in triggering landslides. At present, rainfall data collected are very accurate and we can perform statistical analysis based on daily or even real-time data. However, considering that rainfall data becomes more accurate and thus data volume becomes larger, it is difficult to use them directly in landslide analysis. In previous research work, Cepeda et al. (2010) applied rainfall data to estimation of landslides probability in spatial prediction. Rossi et al. (2012) discussed the rainfall threshold of regional landslide in spatial prediction. Melillo et al. (2015) proposed an algorithm calculating rainfall threshold for different landslides. Many people have also studied triggering mechanism between rainfall and landslide, (e.g., Lee et al. 2016; Li and He 2012). These researches are aimed at the rainfall in a particular landslide, and the relationship between rainfall threshold and the probability of landslide, or rainfall probability in regional landslides and the probability of landslide, with no processing of data. With the recognition of the importance of rainfall data growing, attaching greater significance to the information contained in data, some scholars have begun to study the rainfall data itself. Saito et al. (2010) divided rainfall of landslide in shallow condition into two types: short-cycle, high-intensity (SH) and long-time, low-intensity (LL), putting forward the fact that different rainfall types influence landslides differently. These studies have shown that rainfall data is worthy of digging deeply the information they contain to disclose the effects of different rainfall types on landslides.

More innovative data-processing and information fusion methods such as Feature Analysis, Feature Extraction etc., have emerged and been applied in the processing of landslide monitoring data in recent years. These new approaches can be classified into two categories. The first one is to use the feature extraction of radar detection data to analyze and forecast landslide. For example, Wang et al. (2010) applied airborne-radar data to topographic patterns extraction, and predictions about geological disasters such as landslides. The other category is to acquire relevant information and deformation of landslide through feature extraction of remote sensing images of landslide (Lee et al., 2001; Marcelino et al., 2009). Through studies like this, we can draw a conclusion that the feature extraction methods of landslide are mainly concentrated on the processing of radar data and remote sensing data. Very few studies involved analysis of rainfall data in monitoring landslide.

According to previous studies (Finlay et al., 1997; Hu et al., 2011; Gariano et al, 2015), rainfall data plays a very crucial role in landslide deformation and failures, especially in the cases of rainfall-landslide type. Utilizing some methods processing data, such as quantitative and extreme methods, are not capable to dig out the important information contained in data. Although recently scholars have started to categorize data and conduct information mining for rainfall data, there is a lack of substantial researches in this direction. In this paper, we performed a feature extraction method to the rainfall data which is categorized as clustering analysis. With this approach the computation stress is greatly reduced; in the meanwhile, critical information can be extracted from the data. Finally, this approach is applied and validated to a data set acquired at a cleavage-parallel landslide in the Three Gorges Reservoir area.

The rest of this paper is organized as follows. First, the feature analysis of rainfall data, the relationship between rainfall and evaporation capacity, as well as their influences on rainfall and landslides are discussed. Characteristic indices of rainfall, such as rainfall quantity, duration, and the number of raining days in a given period of time will be introduced. The explanations of how to use clustering analysis to categorize rainfall data, including selection of feature and weight analysis of data are followed. Then, the basis of Clustering Analysis is briefly introduced. Finally, application of feature analysis and feature extraction of rainfall data for a bedding landslide monitoring in The Three Gorges area between June 2003 and December 2008 is presented as a case study. Landslide displacement prediction using back propagation (BP) neural network with the rainfall input in the form of raw data, monthly rainfall, and feature extracted rainfall are compared. The final results demonstrated that the one using featured rainfall has the best forecasting with root mean square error (RMSE).

2. Methodology

For the cleavage-parallel landslides, i.e., the landslides whose formation cleavage plane and therefore the slip surface is parallel to topographic slope, rainfall is a very important factor controlling the onset of slipping. In the existing studies, simple numerical methods, such as using the cumulative rainfall (P (mm)) (Bi et al. 2004), the average monthly rainfall (MMP (mm), the average annual rainfall (MAP (mm)) (Liao et al. 2011), or the 1-day, 3-day, or 7-day maximum rainfall method (Huang 2011) were proposed for extracting rainfall features. These works overlooked some of the important information contained in the rainfall data. It is usually admitted that continuous and heavy rainfalls are necessary conditions in triggering landslides in qualitative analysis; however, intermittent rainfall or sporadic rainfall can also generate certain non-negligible influence on the stability of landslides. There are other factors in rainfall affecting landslides, including evaporation, volume, number of times and duration. These factors are discussed below in details.

2.1 The relationship between rainfall and evaporation

In the studies of rainfall effect on landslides, evaporation is a factor that cannot be simply ignored. The monthly average of evaporation is highly variable, and the changes can be very dramatic. For example, in the Three Gorges Reservoir area, the evaporation is only about 1 mmd^{-1} in winter and spring, but may reach 7 mmd^{-1} in hot summer days. When the evaporation is high while the rainfall is low, rainfall has very little effect on landslides (Wu, **2014**). Usually we would deem rainfall volume less than evaporation invalid in this study. In other words, we cannot talk about rainfall alone without taking evaporation into account.

In this study, we will calculate the average daily evaporation in every month. If the daily rainfall is greater than the average daily evaporation in the month, we would consider it valid. Or the actual rainfall data for that day will be deemed zero. Nevertheless, we would like to point out that the daily evaporation value is calculated by simple division of the monthly value with the number of days in that month. This is the most practical way we can do, due to the lack of more detailed supplementary meteorological observations in this area.

2.2 Statistics of rainfall by times

Up to date, most studies carry out statistical analysis of rainfall based on precipitation per month or per day, or select extreme values in a month or in a few days for landslide analysis (Bui et al, 2012; Du et al., 2013). For example, Crozier and Eyles (1980) used daily rainfall and established thresholds to compare terrain sensitivity and to assess the occurrence probability of landslide. Using daily rainfall data from Kuala Kenderong and Kg. Jeli along the Gerik-Jeli Highway, Lateh et al. (2013) analysed the correlation of landslide events and rainfall precipitation. The rainfall induced landslides was investigated by applying the cumulative rainfall method which comprises the reconstruction of absolute antecedent rainfall for 20 landslide events. However, such statistics did not consider the differences within rainfall types, and it is hard to show the features of rainfalls. In our study, we calculate the number of raining days. When the rainfall volume is less than or equal to the evaporation, we set effective rainfall to zero. Given the rainfall conditions, we set a threshold N (with N=1, 2, 3). If a rainfall ends with more than N non-rainy days followed, it can be considered one rainfall event. We use clustering algorithm to categorize all the data after counting the rainfall events. By doing so, we are able to extract the features of each type, and conduct analysis in a more accurate way.

2.3 The features of the rainfall data: rainfall volume, rainfall duration and rainfall time

The effects of different rainfall types of landslides need to be taken into consideration when determining factors in categorization. Saito et al. (1965) considered the amount of rainfall and rainfall time for the purpose of categorization. In a qualitative analysis, these two factors are usually considered. Based on previous statistics, we emphasize on rainfall duration which can distinguish continuous rainfall from intermittent rainfall. These two different types of rainfall could cause different effects on landslides. In this study, we categorize rainfall based on three factors: rainfall volume, rainfall duration and rainfall time.

Rainfall volume is an important index in categorization. In this research, we select the average daily rainfall volume, which is the rainfall volume divided by the number of days the particular rainfall lasts. Based on our comparative study, the average daily rainfall volume represents the rainfall intensity better and thus differentiates the strong rainfall from continuous rainfall with less ambiguity. The second index we have chosen is rainfall lasting days. It is an important index as it represents both the rainfall volume and the rainfall duration. The third index is the proportion of the raining days in the total number of rain days, which is a crucial index to distinguish continuous rainfall from intermittent rainfall. In addition, since we use millimetre as the measurement unit, the range of rainfall volume data will be (0, 80), the scope of raining days being (0, 6) and (raining time, duration) we need to scale the data to warrant that they are on the same quantitative level, through multiplication by particular coefficients. Based on our numerical tests, the choice of the above values is capable to secure high cohesion and low coupling among the data after categorization. After categorization, we select each kind of rainfall as a particular feature and extract the data, using the BP neural network to demonstrate the effectiveness of feature extraction.

Undoubtedly, similar rainfall events tend to generate similar effects on the stability of landslides, which is consistent with the basic connotation of cluster algorithm. Therefore, this paper employs a widely validated cluster algorithm, K-means, to categorize rainfall data with the purpose of digging out revealing the hidden information (Steinley, 2006). The K-means method is the most matured method in clustering analysis (Steinley, 2006; Hartigan and Wong, 2013). A brief introduction of the K-means clustering algorithm is presented below.

2.4 Clustering Analysis using the K-means clustering algorithm

Like other cluster algorithms, K-means also shares the basic idea that to search for K clusters through iteration which can minimize intra-class distance while maximize inter-class distance. Details of K-means procedure can be found in the flow chart shown in Fig. 1. As for the stopping criteria, it is usually set as that the center of each cluster does not move significantly after several iterations. The mathematical principle of K-means is expressed as Eq. (1).

$$L = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n}||x_i^{(j)}-c_j||^2}{\sum_{j=1}^{k}\sum_{i=1}^{k}||c_i-c_j||^2} \tag{1}$$

In Eq. (1), the numerator corresponds to the overall intra-class distance, and the denominator is the overall inter-class distance. The purpose of K-means is to search for a set of centers c which minimize the cost function L.

The algorithm is simple, fast to converge, as shown in the flow chart of Fig. 1 below. However, the selection of initial centers greatly affects the algorithm's performance. The select strategy for initial centers not only has an impact on the accuracy of categorization, but also contributes significant to the converge rate. While selecting the initial centers, we need to follow the principle of the largest dissimilarity, i.e., the least similarity the initial centers should share. In this paper, a select strategy as Eq. (2) is chosen for determining the K initial centers.

$$D = \sum_{i=1}^{k}\sum_{j=1}^{k}(x_i - x_j)^2 = \max \tag{2}$$

In Eq. 2, D is the total distance (dissimilarity) the k data points share. The larger the D is, the higher probability that these k data points are in the different classes.

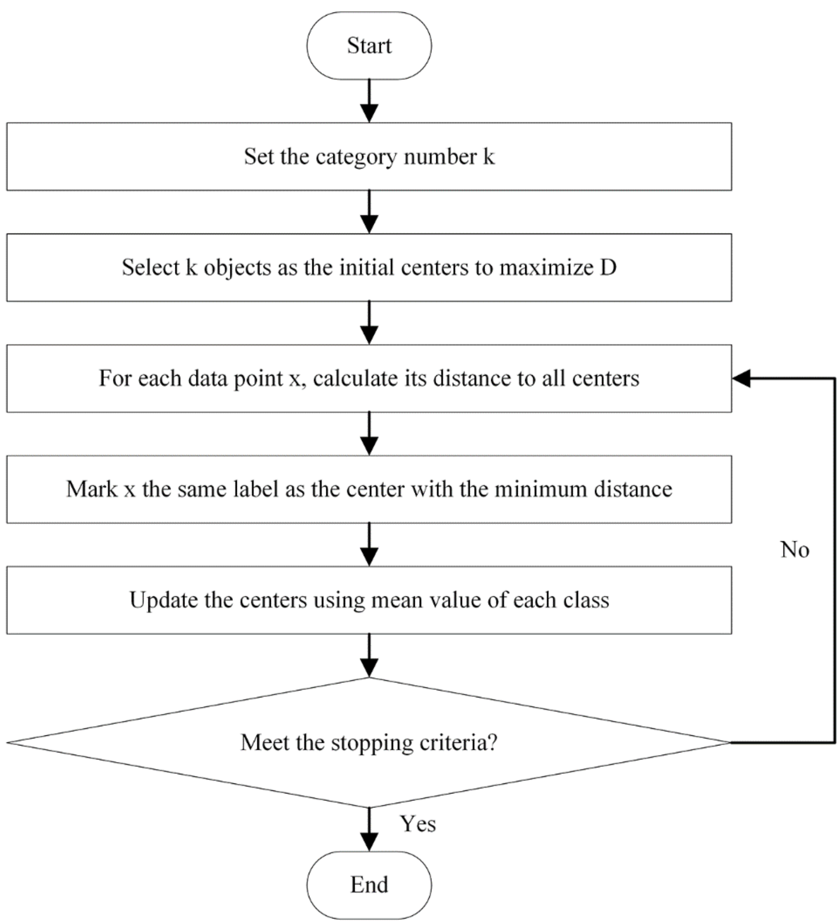

Figure 1: The flow chart of the K-means algorithm.

3 Application to the Baishuihe Landslide field data in the Three Gorges Reservoir

3.1 Geological background and data collection

The Baishuihe Landslide is located in the south bank of the Yangtze River, 56 km away from the Three Gorges Dam (Fig. 2). The landslide is located in the relatively wide open area of Yangtze River valley. It is a single, north-facing, inclined cleavage-parallel slope on the Yangtze River terrace. The rear edge (crown) is about 410 m high from the front edge (toe). The toe is at the 140 m water level of the Yangtze River. Both the left and right flank sides are surrounded by a bedrock ridge and the dip angle is about 30 degrees. It is about 600 m long in sliding direction and 700 m wide laterally. The average depth of the landslide is about 30 m, and the total volume is about 12.6 million cubic meters.

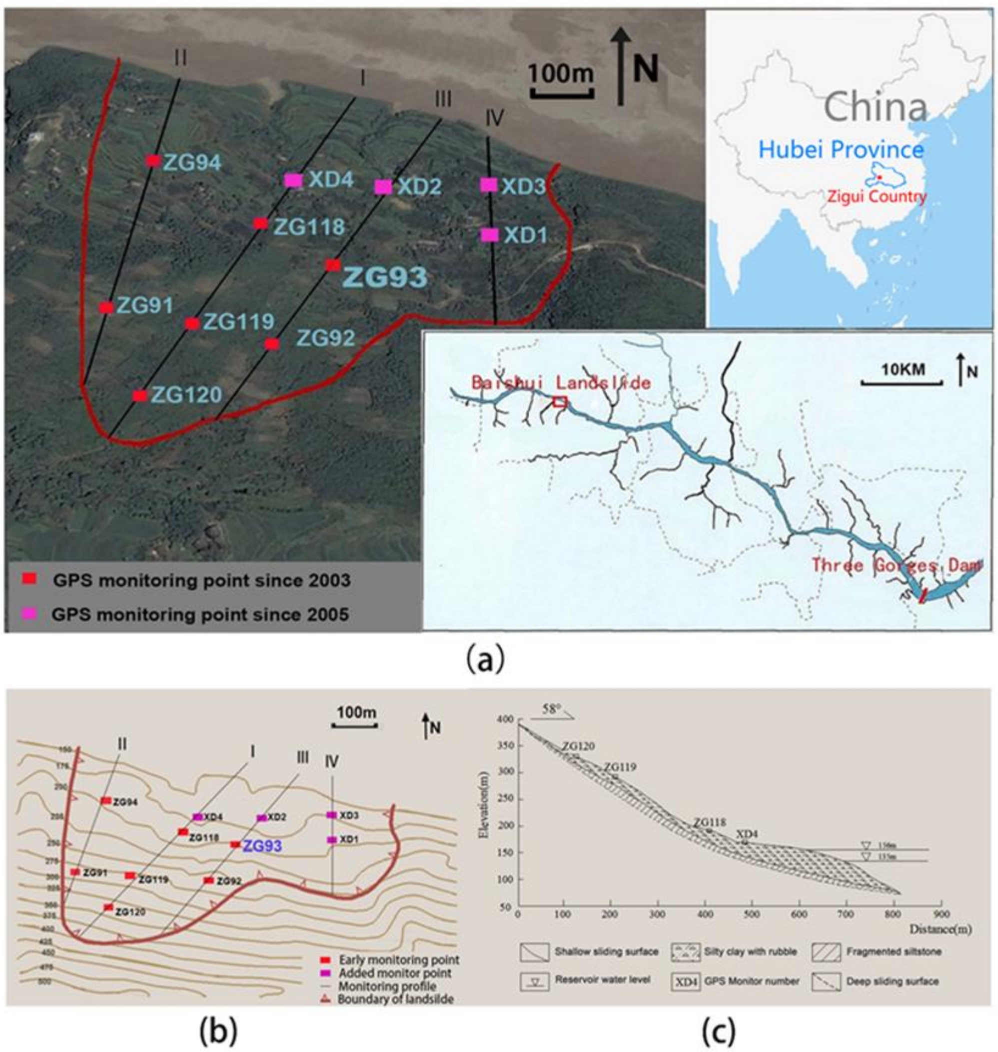

Figure 2. (a) The location of the Baishuihe Landslide (the west most red open square) in the Three Gorges Reservoir area; (b) The locations of the GPS benchmarks (the red and magenta solid squares) for displacement monitoring in the Baishuihe Landslide; (c) The vertical geological cross-section of the Baishuihe Landslide along Profile I.

For monitoring the deformation of this landslide, seven GPS monitoring benchmarks were built along three longitudinal profiles in the Yangtze River in June 2003 (the solid red squares with labels initiated by ZG). Later on in June 2005 four more GPS monitoring benchmarks were added to the right part of the landslide. There is a GPS reference point on each side of the flanks in the rock ridge. In order to better represent the landslide sliding incidence and verify our processing method of rainfall data, we select the monitoring point ZG93 as the experimental object for training and prediction of the neural network algorithm. The selection of ZG93 is based on: 1) It is roughly located at the center of the Baishuihe Landslide so that it is the most unlikely point to be contaminated by false alarm or local signals generated by boundary effect in those monitoring points close to landslide flanks; 2) Observational facts, as shown as the red curve and triangles in Fig. 3 below, support our selection for the fact that it is sensitive enough to catch the subtle displacement in the early stage of the monitoring period (prior to the end of 2007) on one hand; and behaved as the average of all the point after rapid change occurred in May 2007 on the other hand.

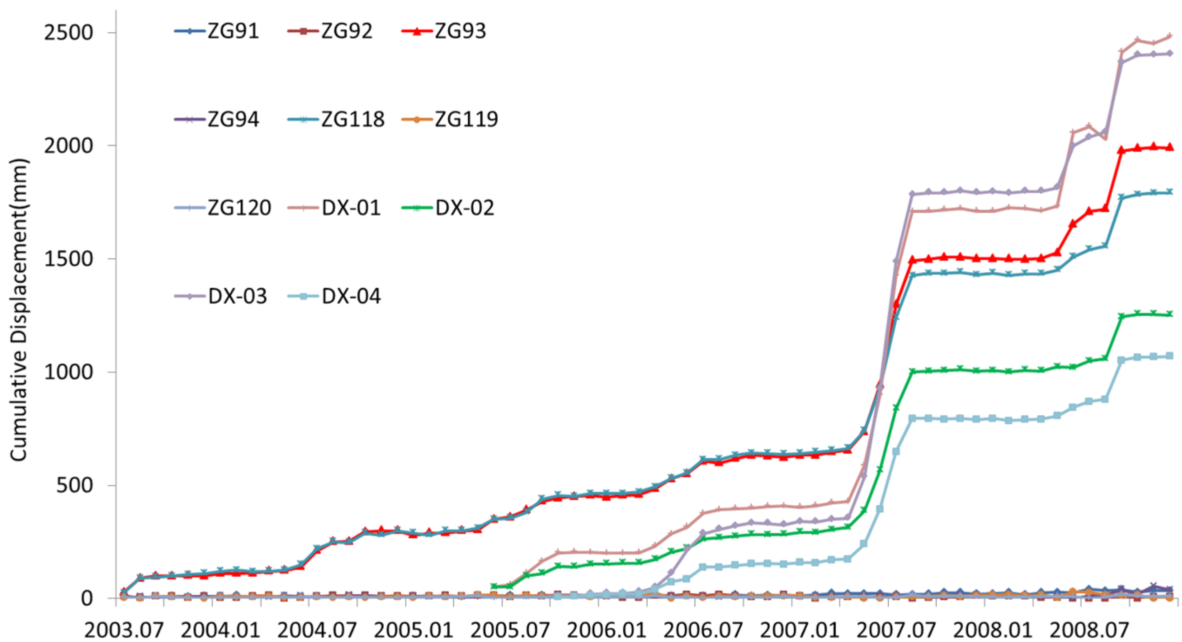

Figure 3: The cumulative displacement of monitoring points in the Baishuihe Landslide.

3.2 Feature analysis of rainfall data

We use the daily rainfall data from June 2003 to December 2008 (Fig. 4) in Zigui County, Hubei Province, China to conduct the analysis to seek the effects of rainfalls to landslide displacement of the Baishuihe Landslide.

As can be seen from Fig. 4, rainfall in this region mainly concentrates in the summer months from April to September, and the heaviest rainfalls happen in July. During this period of 5 years and 7 months, the highest daily rainfall volume is 81.8 mm, occurred in June 2006; while the longest continuous rainfall occurred in July 2008, lasted for more than 11 days.

Han et al. (2012) discussed the evaporation of Zigui area from 2001 to 2010. The average annual evaporation in Zigui County in the last decade is 937.0 mm, and the total evaporation between May and September is 668.5 mm, the monthly maximum is 187.8 mm, occurred in July. From October to next year's April, the total evaporation is only 269.1 mm. We take 4.0 mmd^{-1} as the daily average of evaporation for May, June, August, September, 6.3 mmd^{-1} for July, and 1.3 mmd^{-1} for October to April. By taking evaporation into account, when processing the rainfall data, if the daily average of evaporation is greater than the rainfall volume of a particular day, we consider the rainfall volume of that day to be zero. In other words, if rainfall volume is less than evaporation, it is deemed invalid rainfall. Only when the rainfall volume is higher than daily evaporation, the actual rainfall volume is used for data processing.

In this analysis, we set interval threshold of rainfall N = 2; that is to say, if there is no effective rainfall for 2 days, we consider the rainfall ends. If there is only one day without rainfall since the first raining day, we consider the rainfall is not over yet. The rain is over until there is no effective rainfall for 2 days. Based on this premise, the total number of rainfalls is 211 from June 2003 to December 2008, most of which are single rain-day, accounting for 112 events. More-than-one-day rainfalls account for 99 events.

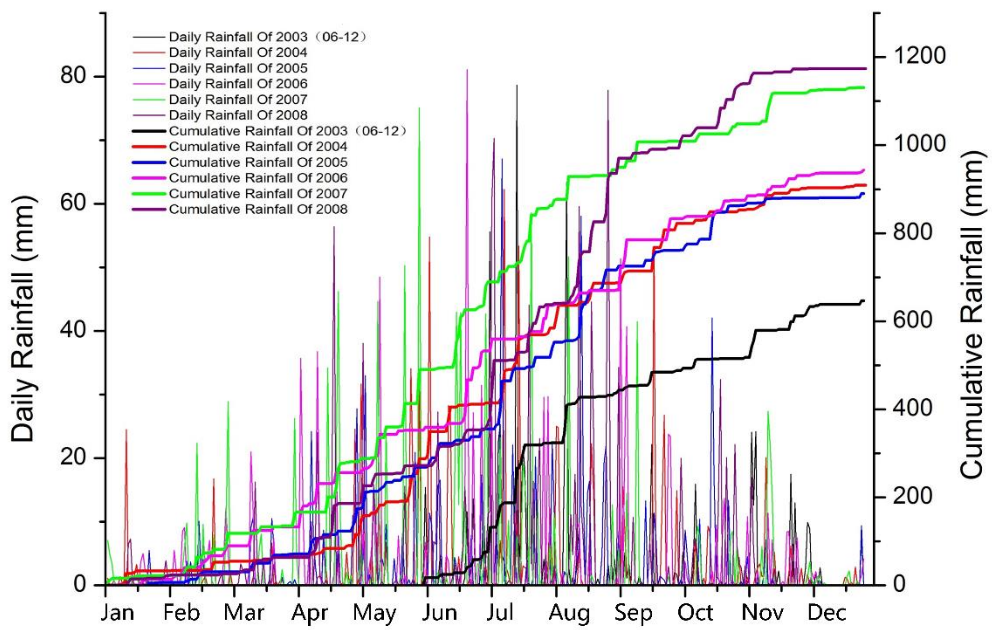

Figure 4. Annual rainfall data in terms of daily and cumulative precipitations for the period from June 2003 to December 2008.

We have further analysed the rainfall data by getting the average monthly column chart as shown in Fig. 5 below, along with the Empirical Cumulative Distribution Function (ECDF) for the duration and cumulative rainfall as Fig. 6 (a and b). The results confirmed that the duration is basically 1-2 days, and the rainfall amount is 2-15 mm for each rainfall event.

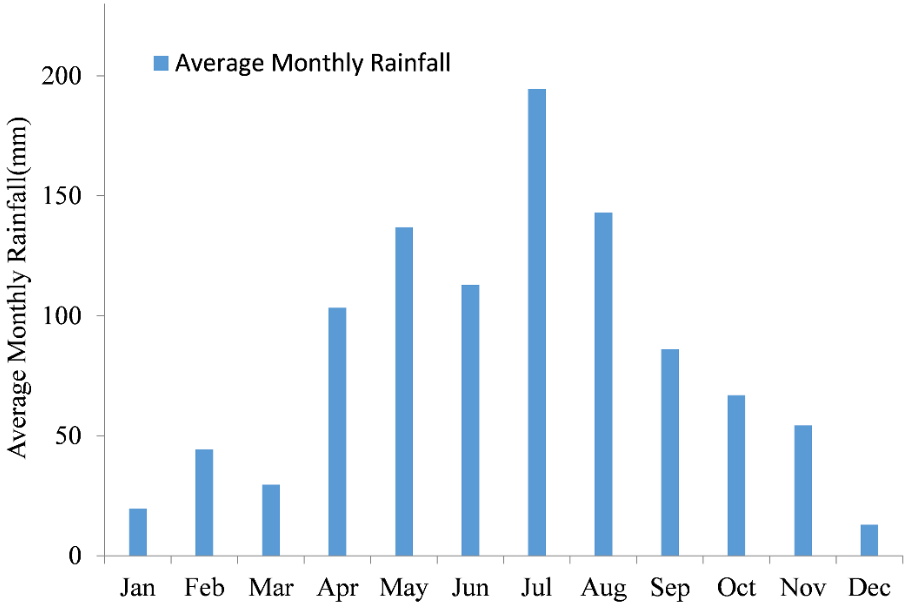

Figure 5. The average monthly rainfall column chart for the period of 2003-2008.

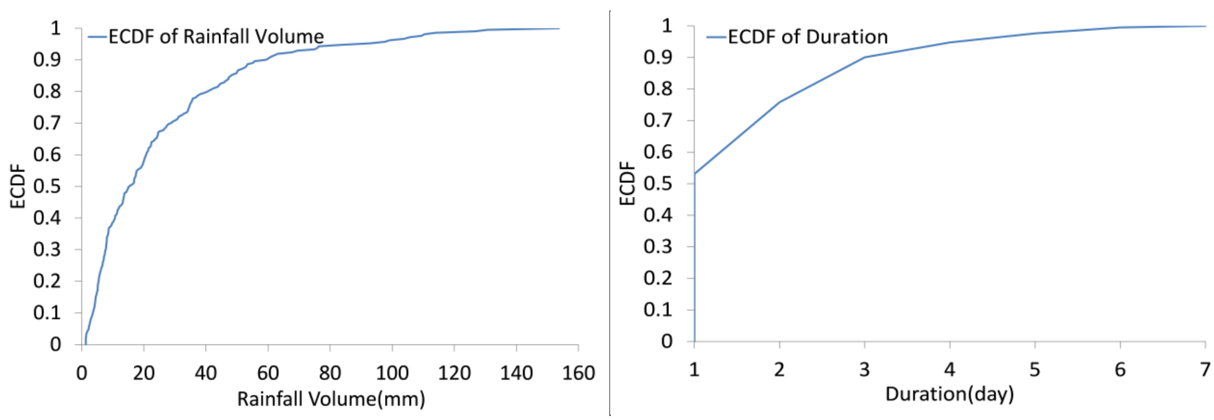

Figure 6. The ECDF plots of the cumulative rainfall and the duration for rainfall events.

Table 1. Cumulative Rainfall from 2003 to 2008.

Year	Cumulative Rainfall (mm)
2003	646.30
2004	909.22
2005	890.50
2006	943.20
2007	1130.97
2008	1172.80

3.3 Feature extraction of Rainfall data and Categorization results

After we sample the rainfall data based on the total number of rainfall events, we now can characterize the average daily rainfall by using three indices for each rainfall event. These three indices are the average daily rainfall volume r, the number of days of rainfall d, and the ratio of rainfall days over the contiguous days T. To ensure the data of these three indices be on the same magnitude, the three features extracted will be multiplied separately by some coefficients p.

The first index (the average daily rainfall volume r) is defined as:

$$\mathrm{r} = \frac{R}{d} p_1 \tag{3}$$

where R is the total volume of a rainfall event, d is the number of raining days in this rainfall event, p_1 is a scaling coefficient close of 0.1. The measuring unit is millimeter.

The second index (the number of days of rainfall d) has a range of 1-6 in our sample. The original data can be used without any scaling.

The third index (the ratio of rainfall days over the continuous days T) can be defined as:

$$\mathrm{T} = \frac{d}{D} p_2 \tag{4}$$

where d is the number of raining days, D is the total number of days during the particular rainfall event, and p_2 is another scaling coefficient. According to our test, we can reach an optimal point of maximum cohesion and minimum coupling effect by setting p_2 = 11.

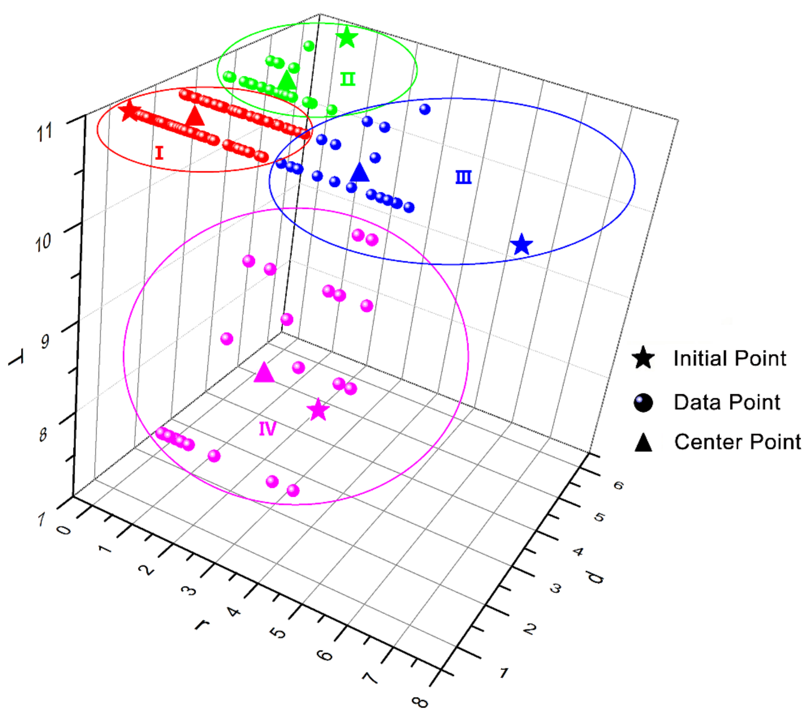

Figure 7. The classified rainfall types based on cluster analysis: I: sporadic rainfall; II: long-duration rainfall; III: short-duration storms; and IV: long-duration intermittent rainfall.

Using the K-means clustering algorithm to calculate the parameters r, d and T for each of these 211 rainfall events, we can characterize the rainfall events into four clusters.

To reduce the number of iterations and improve the clustering performance, four points are selected as the initial clustering centers based on Eq. (2). These four initial points are C_1=(0.13, 1, 11), C_2=(0.52, 6, 11), C_3=(7.51, 1, 11), C_4=(2.45, 4, 7.33), respectively (Fig. 7). We represent these points in the form of (r, d, T).

In the clustering process a new sample x_i is added each time and use $\mathrm{M} = \sqrt{\sum_{j=1}^{3}(x_{ij} - C_{ij})}$ to calculate the distance between this point and the four cluster centers. Based on the minimum distance principle, this new sample is assigned to the closest cluster. Add the samples sequentially to exhaust these 211 samples; and each of rainfall samples must belong to one of these four clusters. Next, update the cluster centers by the equation $\mathrm{C}_i = \frac{1}{n}\sum_{x \in C_i} x$, calculate distance between each sample and the new cluster centers, and re-cluster it according to the distances. Repeat this process until the stopping criteria is met. Finally, four cluster centers: (1.00, 1.31, 11), (1.06, 3.42, 11), (4.23, 1.40, 11), and (1.69, 3.09, 7.99) are obtained. The above data are rounded to two decimals. There are 142 samples in the first cluster, 25 in the second cluster, 21 in the third cluster and 23 in the fourth cluster, as shown in Fig. 7. The first cluster (category) is characterized by low-rainfall, duration of 1-2 days, which are mainly the sporadic rainfalls (the red cluster in Fig. 7). The characteristics of the second type of rain are comparatively less volume, but with long duration and no interruption (the green cluster in Fig. 7). The third type of rainfall is characterized by short duration, usually 1-2 days, but the rainfall volume is very big (storms, the blue cluster in Fig. 7). Finally, the fourth type of rainfall is long duration with moderate rainfall volume and intermittent rainfall (the magenta cluster in Fig. 7).

Rainfall volume is the most important factor in causing the variations of displacement in the cleavage-parallel landslide (Gariano et al., 2015). Therefore, after categorization, we use rainfall volume as the feature for extraction, taking the total rainfall volume in the same category as the feature of that particular category. In displacement prediction as described later in this paper, we conduct statistics on the rainfall volume per month of each type of rainfalls. For example, in the period between August 16 and September 15, 2008, there were five events of effective rainfall. The five samples were measured using our (r, d, T) set at (0.98, 2, 11), (3.39, 2, 11), (3.68, 3, 11), (3.43, 1, 11) and (1.07, 1, 11), respectively. Among this small sample set, there were 2 first-type rainfalls, 0 second-type rainfalls, 3 third-type rainfalls, and 0 fourth-type rainfalls. The total rainfall volumes were 30.3, 0, 212.5, and 0 mm for each type of the rainfall respectively. Using feature extraction, the feature vector for rainfall in that month would be (30.3, 0, 212.5, 0).

3.4 Prediction of landslide displacement with BP neural network

After the discussion of rainfall feature characterization and extraction with the clustering algorithm, we are ready to touch the major topic of the effect of rainfalls on landslides displacement. Using simple correlation just shown as Fig. 8 below, one can find that the connection between rainfall and landslide displacement at the Baishuihe site is quite obvious. Nevertheless, more closed and quantitative examination is needed to enable us reach more definitive conclusion of this causality.

The back propagation (BP) network is a kind of multilayer feedforward neural network. It is a widely tested and validated error back propagation algorithm. The network consists of an input layer, a number of hidden (middle) layers and an output layer. Based on Kolmogorov's theorem, a three layer BP neural network can achieve approximation for any arbitrary nonlinear functions, so that we choose BP neural network to carry out this quantitative examination.

To verify the effectiveness of feature extraction after using cluster analysis, we utilize BP neural network to predict displacement with the following three treatments of the rainfall data: 1) original daily rainfall (mm); 2) monthly total rainfall (mm); and 3) the extracted rainfall feature processed through cluster analysis and feature extraction. We use the rainfall data of the current month and the last month, along with the displacement of last month as the input to predict the displacement in the current month with BP neural network. We use only one hidden layer. And the node number is n_1 with: $n_1 = \sqrt{n+m} + a$; where n is the input layer node number, m is the output layer node number; and a is a constant, which is set to be 2 in this work.

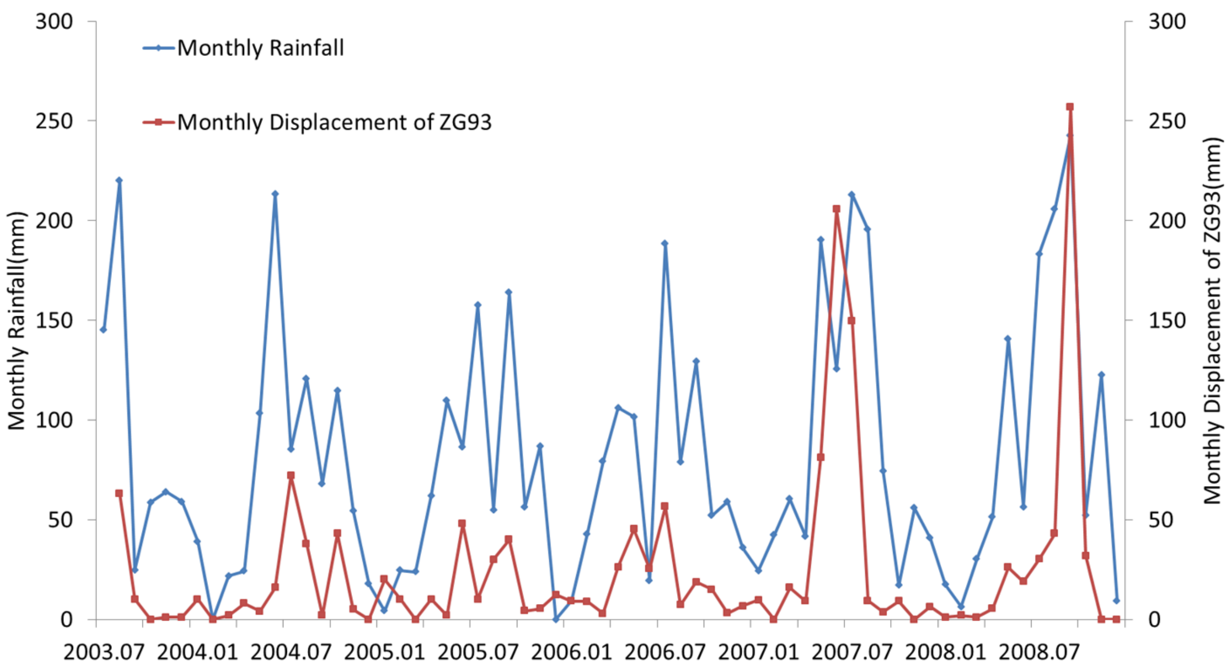

Figure 8. The monthly rainfall in Zigui County and displacement recorded by GPS survey mark ZG93 at the Baishuihe site for the period from June 2003 to December 2008.

First, we use rainfall data and displacement data between June 2003 and December 2005 to train the BP neural network. Then we use the trained neural network to predict displacement between January 2006 and December 2008. In the prediction process, once the prediction of the displacement of each month is finished, we use the newly obtained data to train the neural network again, and use the newly trained network for prediction of the displacement of next month. The prediction results are shown as Figs. 9 and 10; and the network structure; the errors of training; operation times of training and prediction by BP neural network is shown in Table 2.

Table 2. The root mean square error of training of and prediction by BP neural network.

	Network structure [n, n_1, m]	RMSE of training	Operation times of training	RMSE of prediction
Daily rainfall of 60 days	[61,9,1]	4.55E-04	1.21E+09	3.23E+02
Monthly total rainfall	[3,4,1]	1.08E+00	2.33E+07	4.08E+02
Extracted rainfall feature	[9,5,1]	2.43E-03	1.40E+08	2.62E+02

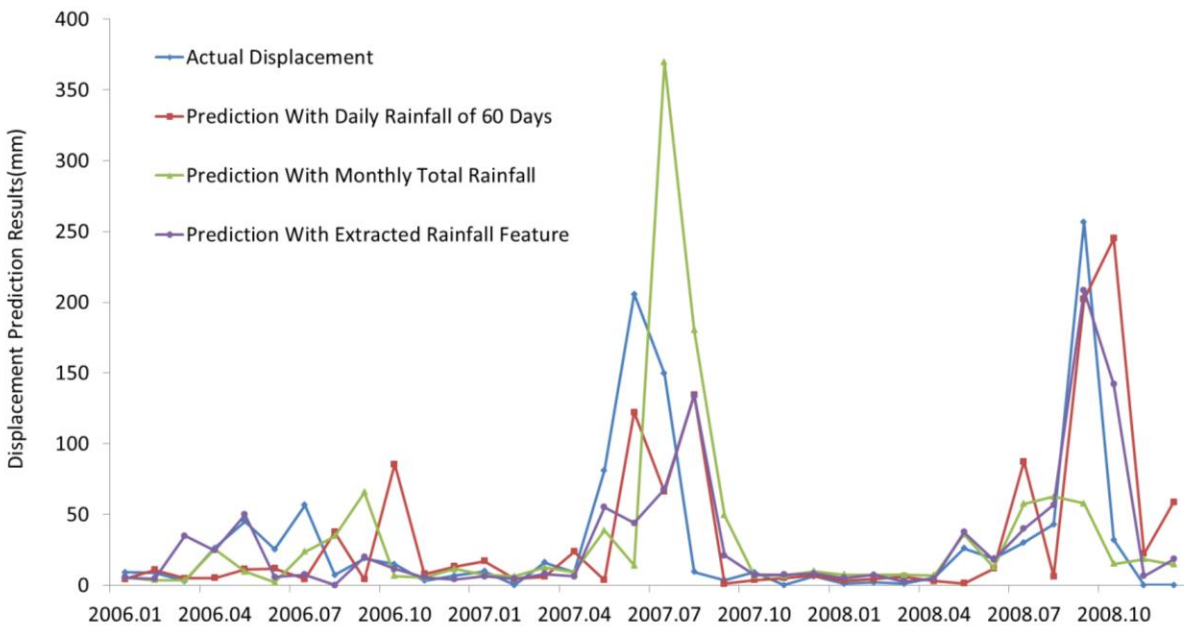

Figure 9. Landslide displacement prediction based on 3 types of rainfall input.

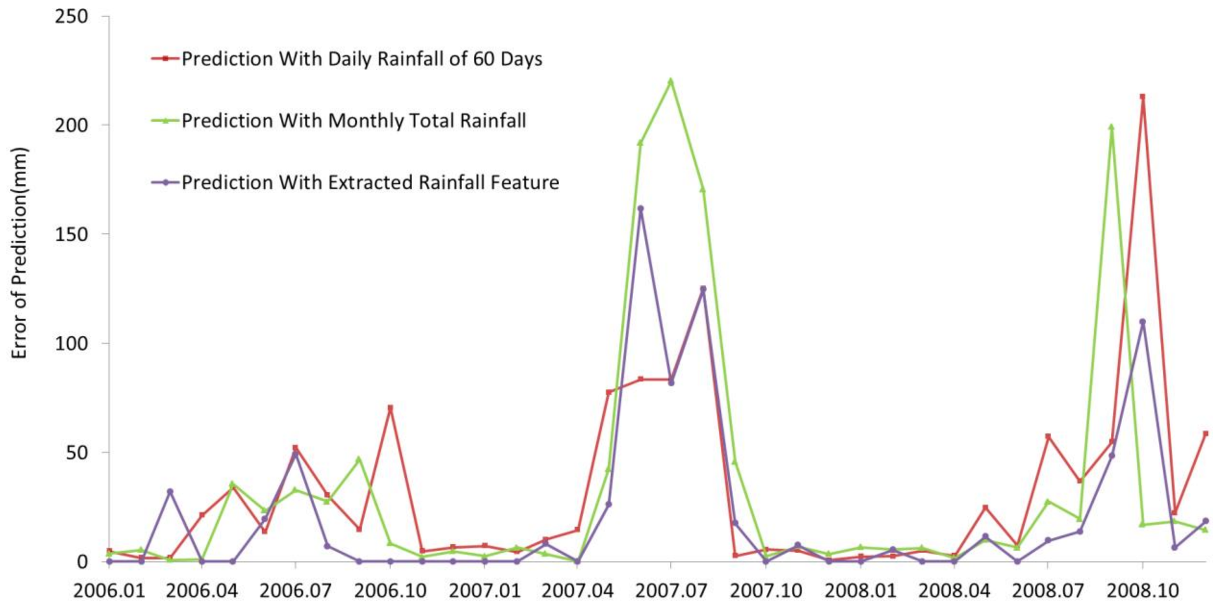

Figure 10. Comparison of the displacement prediction errors based on 3 types of rainfall input.

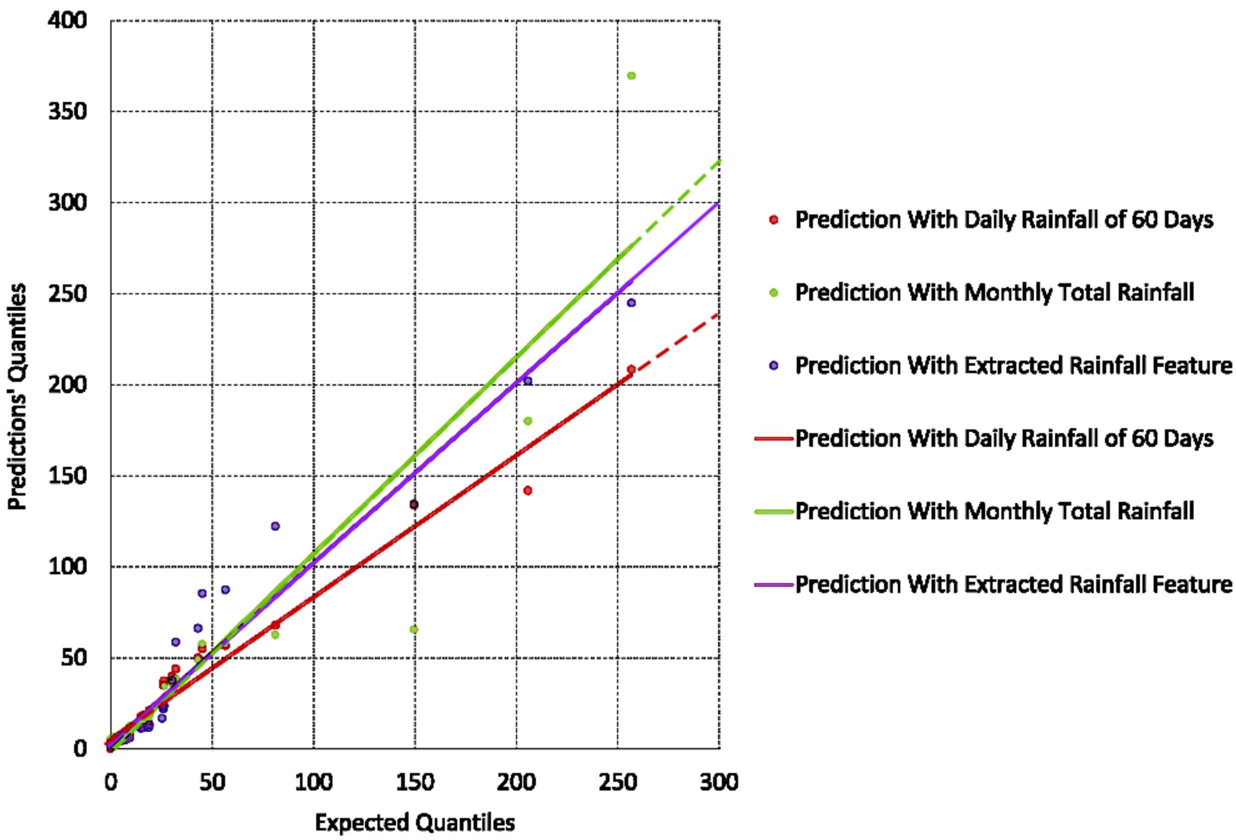

Figure 11: The q-q plot for landslide displacement prediction based on 3 types of rainfall input.

As can be seen from the Figs. 9, 10 and Table 2, when using the original daily rainfall of 60 days, we have too much data for the neural network to process. The operation time of training is as high as 1.21E+09. The neural network, unfortunately, has very limited capability to handle large volume of data. There are too many possible matching internal functions in the training stage. Therefore, we have the smallest mean squared error in the training stage but not the best prediction among the three methods.

In the second approach, when we use monthly total rainfall to forecast displacement, the volume of data to be processed is greatly reduced; but it is at the sacrifice of great reduction in rainfall features. In both the training and prediction stage, the results are the worst among the three approaches.

In the third method, when we the extracted rainfall feature after feature extraction, we also have much less volume of data for the neural network to process, by comparison with using the first rainfall type; meanwhile, it is not at the sacrifice of great reduction in rainfall features when compared with the second approach. Although we have slightly higher mean squared error in the training stage, but the prediction results are the best among the three methods.

The q-q plot shown in Fig. 11 is an exploratory graphical expression used to check the validity of a distributional assumption for data sets. It is employed for analyzing the relationship between observed displacement data and the predictions with three types of rainfall input. If the observed and the predicted data sets have the same distribution, the fitted line in the q-q plot will approach y=x. As can be seen from Fig. 11, the fitted curve of the data points from the prediction with extracted rainfall feature is closer to the line y=x with slope of 1; while the prediction with monthly total rainfall is overestimated and the prediction with daily rainfall of 60 days is underestimated. It indicates that the extracted rainfall feature represents real rainfall better than daily rainfall of 60 days and monthly rainfall in landslide displacement prediction.

4 Result Discussion

After analysing the precipitation and evaporation of the region studied, rainfall data is categorized by times based on three indexes: rainfall volume, rainfall duration and rainfall time. As for different study areas, category numbers should vary accordingly with the geological characteristics, so that features extracted can be more in line with the real situation.

Some scholars mentioned the effect of different rainfall types on landslide before (Brand et al. 1984; Glade et al. 2000; Glade 2000). However few study conducted the analysis and discussion of the comprehensive effect mixed-type rainfall contributes on landslide. In this paper, a tentative research is proposed and reasonable results are obtained. Four cumulative rainfall volume of each category in each month are used as the monthly rainfall feature in the prediction for landslide displacement, by which the influence of category sequence and non-raining days within each rainfall event can be circumvented. Study considering non-raining days within each rainfall event requires a large amount of penetration and evaporation data, which is our further study focus.

We used 2 years and 7 months data to train the BP neural network and 3 years for forecasting of the displacement of landslides (Figs. 9 and 10). The results showed some important features. First, by using the proposed feature extraction approach of the rainfall data, the computational burden for forecasting was greatly reduced. Second, the comparison of the predicted and the observed displacement indicates that using the feature extraction approach has led less forecasting error than using other rainfall reduction methods (e.g., monthly total rainfall or daily rainfall of 60 days). Moreover, one more interesting feature is noteworthy. From the prediction results (Fig. 9) we can see that the forecasting capability has no significant decay with the increase of time accumulation. The prediction of the displacement peak in the summer of 2008 is even more precise than the prediction of the peak in summer 2007. This fact may lead us to suspect that either there are other significant contributing factor(s) to the displacement peak in 2007; or there are more characteristics in the rainfall in summer 2007 that has not been essentially characterized by the current approach. After all, we can confidently state that the feature extraction approach is an important improvement in rainfall-landslide characterization process.

5 Conclusions

In this paper, we first analysed the characteristics of rainfall data, extracted the volume, duration and onset time for each single rainfall event. With this process, the amount of rainfall data is greatly reduced and the characteristics of rainfall data are substantially preserved and extracted. As the second step, the featured information of rainfalls was used in landslide displacement prediction. We used the extracted features for the characteristic analysis and prediction to the Baishuihe Landslide in the Three Gorges area on the Yangtze River. The BP neural network method is applied to three types of rainfall data: the characteristic value, the daily rainfall, and the monthly rainfall, as the input into BP neural network to forecast the landslide displacement, respectively. Comparisons of the errors and efficiency for these three approaches are made and the main conclusions are described as follows:

1) We have carried out statistical analysis on original rainfall events. By taking this approach we preserved the rainfall details as much as possible, while reduced the burden of processing large amount of raw data.

2) We introduced the K-means cluster algorithm for those rainfall events sharing maximum similarity.

3) The four cumulative rainfall volumes of K categories in each month are used as the monthly rainfall feature.

4) Finally, our analysis results showed that using the rainfall feature extracted can lead to a better performance in landslide displacement prediction.

Acknowledgements

This research was funded by the National Natural Sciences Foundation of China (Project Nos. 41302278, 41272377, and 41272306), and the Fundamental Research Funds for National Universities, China University of Geosciences-Wuhan (No. CUG120119). The authors are grateful to the Zhangjiachong Soil and Water Conservation Experiment Station in Zigui County for providing the rainfall data. The first author wishes to thank the China Scholarship Council for funding his visit to the University of Connecticut where this study was conducted.

References

Bi, H. X., Nakakita, O., and Abe, K.: Spatial distribution prediction and hazard zonation of landslide based on GIS techniques, Journal of Natural Disasters, 13, 50–57, 2004.

Brand, E. W., Premchitt, J., and Phillipson, H. B.: Relationship between rainfall and landslides in Hong Kong, in: Proceeding 4th International Symposium Landslides, Toronto, 377–384, 1984.

Bui, D. T., Pradhan, B, Lofman, O, Revhaug, I, and Dick, Ø. B.: Regional prediction of landslide hazard using probability analysis of intense rainfall in the Hoa Binh province, Vietnam, Natural Hazards, 66, 707-730, doi:10.1007/s11069-012-0510-0, 2013.

Cepeda, J., Höeg, K., and Nadim, F.: Landslide-triggering rainfall thresholds: a conceptual framework, Quarterly Journal of Engineering Geology and Hydrogeology, 43, 69–84, doi:10.1144/1470-9236/08-066, 2010.

Crozier, M. J., and Eyles, R. J.: Assessing the probability of rapid mass movement, in: 3rd Australia-New Zealand Conference on Geomechanics, Wellington: New Zealand Institution of Engineers, 6, 247-251, 1980.

Du J., Yin K., and Lacasse S.: Displacement prediction in colluvial landslides, Three Gorges Reservoir, China, Landslides, 10, 203-218, doi:10.1007/s10346-012-0326-8, 2013.

Finlay, P. J., Fell, R., and Maguire, P. K.: The relationship between the probability of landslide occurrence and rainfall, Canadian Geotechnical Journal, 34, 811-824, 1997.

Gariano, S. L., Brunetti, M. T., Iovine, G., Melillo, M., Peruccacci, S., Terranova, O., Ven-nari, C., and Guzzetti, F.: Calibration and validation of rainfall thresholds for shallow landslide forecasting in Sicily, southern Italy, Geomorphology, 228, 653-665, doi:10.1016/j.geomorph.2014.10.019, 2015.

Glade, T., Crozier, M., and Smith, P.: Applying probability determination to refine landslide-triggering rainfall thresholds using an empirical “antecedent daily rainfall model”, Pure & Applied Geophysics, 157, 1059–1079, doi:10.1007/s000240050017, 2000.

Glade, T.: Modelling landslide triggering rainfall thresholds at a range of complexities, in: Proc of the VIII International Symposium on Landslides, Cardiff, Telford, London, 2, 633–640, 2000.

Han, Q. Z., Xiang, F., Ma, L., Xia, L. Z., Xiang, L., and Wang, G. M.: The Three Gorges typical district 2001-2010 local meteorological factors changing trend analysis, Soils, 44, 1029–1034, 2012.

Hartigan J. A., and Wong M. A.: A K-means clustering algorithm, Applied Statistics, 28, 100-108, doi:10.2307/2346830, 2013.

Hu, M., Wang, R., and Shen, J. H.: Rainfall, landslide and debris flow intergrowth relationship in Jiangjia Ravine, Journal of Mountain Science, 8, 603-610, 2011.

Huang, G. D: The Research about model and analysis of Landslide Stability on Intelligence Algorithms, China University of Geosciences Doctoral dissertation, 2011.

Lateh, H., Tay L. T., Khan, Y., Kamil, A., and Azizat, N.: Prediction of landslide using Rainfall Intensity-Duration Threshold along East-West Highway, Malaysia, Caspian Journal of Applied Sciences Research, 2, 124-133, 2013.

Lee, M. J., Park, I., Won, J. S., and Lee, S.: Landslide hazard mapping considering rainfall probability in Inje, Korea, Geomatics, Natural Hazards and Risk, 7, 424-446, doi:10.1080/19475705.2014.931307, 2016.

Lee, S., and Min, K.: Statistical analysis of landslide susceptibility at Yongin, Korea, Environmental Geology, 40, 1095-1113, 2001.

Li, D. X., and He, S. M.: The deformation prediction model on rainfall-triggered shallow landslide, Journal of Mountain Science, 30, 342-346, 2012.

Liao, R. X., Wang, G., and Zou, L. C.: GIS-based landslide spatial database system design for Three Gorges Reservoir area, Journal of China Three Gorges University, 33, 24–27, 2011.

Marcelino, E. V., Formaggio, A. R., and Maeda, E. E.: Landslide inventory using image fusion techniques in Brazil, International Journal of Applied Earth Observation and Geoinformation, 11, 181-191, doi:10.1016/j.jag.2009.01.003, 2009.

Melillo, M., Brunetti, M. T., Peruccacci, S., Gariano, S. L., and Guzzetti, F.: An algorithm for the objective reconstruction of rainfall events responsible for landslides, Landslides, 12, 311-320, doi:10.1007/s10346-014-0471-3, 2015.

Rossi, M., Kirschbaum, D., Luciani, S., Mondini, A. C., and Guzzetti, F.: TRMM satellite rainfall estimates for landslide early warning in Italy: preliminary results, Proceedings of the SPIE-The International Society for Optical Engineering, 85230D, doi:10.1117/12.979672, 2012.

Saito, H., Nakayama, D., and Matsuyama, H.: Two types of rainfall conditions associated with shallow landslide initiation in Japan as revealed by Normalized Soil Water Index, Sola, 6, 57–60, doi:10.2151/sola.2010-015, 2010.

Saito, M.: Forecasting the time of occurrence of a slope failure, in: Proceedings of the 6th International Conference on Soil Mechanics and Foundation Engineering, Montreal, 2, 537–541, 1965.

Sassa, K., Picarelli, L., and Yueping Y.: Monitoring, prediction and early warning, in: Landslides-Disaster Risk Reduction, Springer Berlin Heidelberg, 351–375, 2009.

Steinley, D.: K-means clustering: a half-century synthesis, British Journal of Mathematical & Statistical Psychology, 59(Pt 1), 1-34, doi:10.1348/000711005X48266, 2006.

Wang, Z., Li, H. Y., and Wu, L. X.: Geodesics-based topographical feature extraction from airborne Lidar data for disaster management, in: 18th International Conference on Geoinformatics, 1–5, 2010.

Wu H.: Monitoring and theoretical analysis of rainfall infiltration of huangtupo landslide in the three gorges reservoir, China University of Geosciences for the Master Degree of Engineering, 2014.