Flood Occurrence Probability Modeling and Evaluation Based on Spearman Analysis and MLP Algorithm

Zijun Lin*, Jianan Lin

School of Mathematics and Statistics, Guangdong University of Technology, Guangzhou, China, 510520

*Corresponding author: lzj17817277905@163.com

Abstract. Global climate change and human activities have intensified flood risk. Therefore, accurate prediction of flood probability and disaster loss reduction has become the focus of disaster prevention and mitigation research. At the same time, traditional statistical methods are difficult to meet the needs of flood prediction in complex environments. Given this, this study starts with many index factors that affect the probability of flood occurrence and combines the multi-layer perceptron model to predict the likelihood of flood occurrence. This model has a good effect in the field of prediction and can make up for the shortcomings of traditional statistical methods. Specifically, this paper first conducts Spearman correlation analysis for different indicators, aiming to select 11 indicators with high correlation. The multi-layer perceptron model was further used to modify the model parameters and data set division. Finally, the flood probability prediction model was obtained with a 60% training set and 40% verification set, the number of hidden layer neurons was 64 and 32 respectively, and the root mean square error RMSE was 0.038242. The model was considered reliable. Based on this, it is applied to the event prediction of unknown flood occurrence probability, and the results of the event prediction are 0.534, 0.470, 0.452, etc., and most of them fall between 0.45 and 0.55. At the same time, this study deeply explored the distribution of flood probability, combined with the Q-Q diagram and probability distribution, and found that the data set conforms to the normal distribution, which is consistent with the real world.

Keywords: Flood Disaster, Spearman Correlation Analysis, Multilayer Perceptron, Flood Prediction Model.

1. Introduction

With the intensification of global climate change and the impact of human activities, the frequency and intensity of extreme weather events are increasing, and natural disasters such as floods have brought severe challenges to human society and economic development. Flood disasters not only pose a threat to human life and property but also cause long-term damage to the ecological environment and infrastructure. Therefore, how to accurately predict the probability of flood occurrence [1] and reduce the risk of flood has become a key research direction in the field of disaster prevention and mitigation [2].

Traditional flood prediction methods [3] are usually based on historical hydrological data, using statistical models and meteorological forecasting tools for analysis. However, the effectiveness of these methods is often limited when dealing with high-dimensional data and nonlinear relationships, and it is difficult to meet the needs of flood prediction in complex environments. In recent years, with the development of data science [4, 5] and artificial intelligence technology [6], neural networks [7, 8] have gradually become a research hotspot and have been widely used in many fields, and multi-layer perceptron (a branch of neural networks) has played a significant role in the field of prediction. For example, Zhang et al. [9] predicted the good mud content in a uranium mine by using a multi-layer perceptron, and Su et al. [10] predicted species diversity by using a multi-layer perceptron, which proved the feasibility of multi-layer perceptron in the prediction field. Therefore, to make up for the limitations of traditional methods to deal with complex systems, this paper adopts a multi-layer perceptron algorithm to predict flood problems.

First, this paper collected the index data set of related factors affecting the probability of flood occurrence and pre-processed the data. Then Spearman correlation analysis [11] was used to analyze

the correlation between each index and the probability of flood occurrence. Several indicators with strong correlation were selected and a multi-layer perceptron was used to build a flood prediction model, and the model was used to predict events with unknown flood occurrence probability. The distribution characteristics of the predicted probability are analyzed.

2. Data Preprocessing and Spearman Correlation Analysis

The data sets used in this study are all from http://www.apmcm.org/

2.1. Data preprocessing

1) Missing value detection

For the data given in the data set, first of all, directly use the positioning function of Excel table, select the null value for positioning, and get the prompt that no cell is found, indicating that the table data has no missing value and does not need to be processed.

2) Outlier test

The Z-score method was used to test the outlier values of the data detected by missing values.

The Z-score method can standardize the original score, converting it into a relative position relative to the mean, so that it can be compared between different data distributions. The value of the Z-score can be positive, negative, or zero. If the Z-score is positive, the original score is higher than the mean; If the Z-score is negative, the original score is lower than the average. If the Z-score is zero, the original score is equal to the average. The Z-score is the deviation from the mean in standard deviation, where both standard deviation and variance are 1 and the mean is 0.

The calculation formula is:

$$Z = \frac{X-\mu}{\sigma} \tag{1}$$

Where X is the original fraction, μ is the mean, and σ is the standard deviation.

Outlier recognition is divided into three steps. First, the Z-score is calculated and the outlier threshold is set. Here, ±3 is selected as the threshold. For outliers, due to the huge sample data, it is considered to directly delete outliers, and use the positioning function of Excel tables to delete the row where the deleted data is located, and only the row with complete data is retained for subsequent analysis.

2.2. Data analysis

Since it is uncertain whether the data obey the normal distribution, and the assumption of Pearson correlation analysis is that the data conform to the normal distribution, this paper considers using Spearman correlation analysis, which is not affected by the distribution, to analyze the correlation of the processed data.

Spearman correlation analysis is a non-parametric statistical method that uses the rank (or order) of the observed values of two variables to calculate the correlation between them. This method does not depend on the specific distribution of the data and is not affected by outliers, so it is particularly suitable for cases where the data does not conform to a normal distribution or there are outliers.

1) The algorithm is divided into four steps. The first is ranking conversion, by converting the data for each variable into a rank (or sort). If there are parallel data, take their average rank. For example, if two data points are both the smallest, their rank is the average of the number of these two data points (i.e., if the parallel smallest is two data points, then they are both 1.5). The ranking difference is further calculated by calculating the difference in the rank of each pair of variables. That is, for each pair of observations, the difference between its rank on one variable and its rank on another variable is calculated. The sum of the squares of the ranking differences is then calculated by squaring the ranking differences of all observations and summing them. This sum of squares reflects the degree of inconsistency in the ordering of the two variables. Finally, the Spearman correlation coefficient is calculated according to formula (2)

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \quad (2)$$

In practice, the connection between the variables is irrelevant, so ρ can be calculated using the difference between the grades of the two observed variables in a simple step:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad (3)$$

2) Algorithm solution

The built-in function in MATLAB was used to calculate the Spielman correlation coefficient, and the correlation coefficient matrix and P value of the 20 indicators and the probability of flood occurrence were obtained in descending order, as shown in Table 1:

Table 1 Ranking table of Spearman coefficients

The Spearman correlation coefficient matrix	P value matrix	Name of index
0.1812	0	Infrastructure deterioration
0.1804	0	Terrain drainage
0.1795	0	Monsoon intensity
0.1791	0	Dam quality
0.1787	0	River management
0.1781	0	Silting-up
0.1776	0	Population score
0.1768	0	Deforestation
0.1766	0	Climate change
0.1766	0	Landslip
0.1759	0	Ineffective disaster prevention
0.1749	0	Agricultural practice
0.1748	0	Wetland loss
0.1746	0	Drainage area
0.1736	0	Policy factors
0.1731	0	Insufficient planning
0.1719	0	Urbanization
0.1714	0	Corrosion
0.1695	0	Drainage system
0.1692	0	Coastal vulnerability

According to the P-value matrix is all are less than 0.05, it is 95% sure that these 20 indicators are related to the probability of flood occurrence. According to the Spearman correlation coefficient matrix, the correlation between each index and the probability of flood occurrence is not strong, and all of them are relatively close. Considering that there are too many factors, to facilitate subsequent model training and reduce model complexity, this study sets ineffective disaster prevention as the dividing line to screen factors.

3) Correlation coefficient interpretation

This paper chooses ineffective disaster prevention as the dividing line and considers that ineffective disaster prevention and indicators higher than it: Infrastructure deterioration, terrain drainage, monsoon intensity, dam quality, river management, silting-up, population score, deforestation, climate change, and landslip are closely related to flood probability, among which infrastructure deterioration and terrain drainage are most closely related to flood probability. Agricultural practices, wetland loss, drainage area, policy factors, insufficient planning, urbanization, corrosion, drainage systems, and coastal vulnerability are less strongly associated with flood probability.

3. Flood Probability Prediction Using Multilayer Perceptron

3.1. Construction of prediction model of occurrence probability

According to the correlation coefficient between each index and the probability of flood occurrence obtained above, the first eleven indicators are selected here, and the multi-layer perceptron in the neural network algorithm is considered for prediction.

Multilayer perceptron (MLP) is a feedforward neural network model whose basic structure includes an input layer, an output layer, and at least one or more hidden layers. Each layer is composed of multiple neurons, each of which receives input from neurons in the previous layer and, after processing with weighted summation and activation functions, generates output and passes it to the next layer.

This paper divides the processed data into a training set and a validation set. Due to the large number of samples, through continuous parameter adjustment, the data set was finally divided into about sixty-four, that is, 600,000 events were randomly selected as the training set, and the remaining 385,308 data were selected as the verification set. The training cycle was set to one round, and the number of neurons in the first hidden layer was 64 and the number of neurons in the second hidden layer was 32. After a round of training, the training result is obtained: the root mean square error RMSE of the verification set is 0.038242, indicating that the training effect is good and the model is reliable.

3.2. Predict the probability of flooding

The trained model is used to predict the probability of flood occurrence according to the data of each index in the data set. Part of the forecast results are shown in Table 2:

Table 2 Part of the prediction probability sample table

id	Flood probability
1	0.534
2	0.470
3	0.452
4	0.487
5	0.489

According to the forecast results, a histogram and line chart of flood occurrence probability was drawn, as shown in Figure 1 and Figure 2

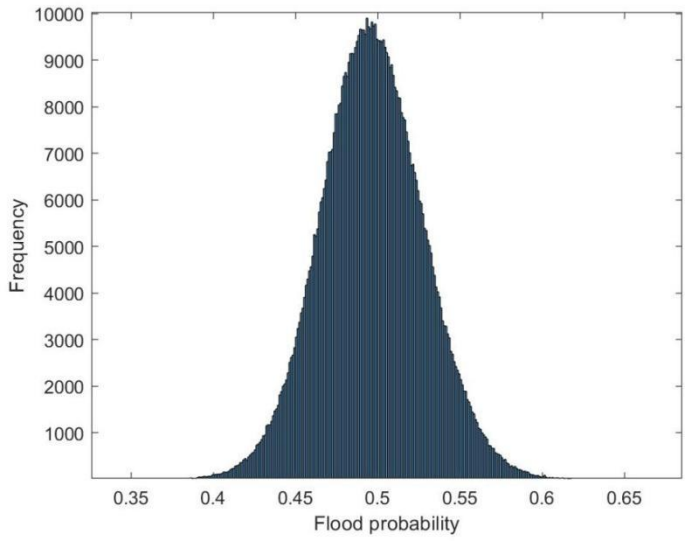

Figure 1. Histogram of forecast flood probability

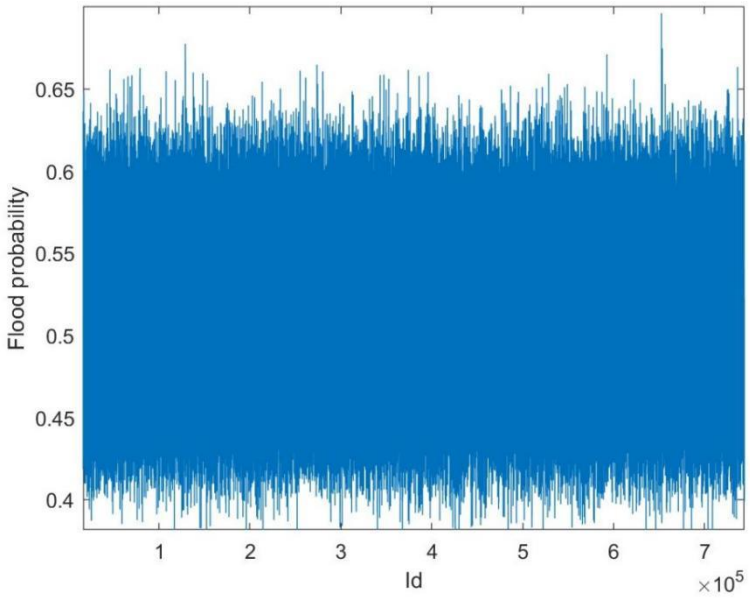

Figure 2. Line chart of forecast flood probability

According to the histogram and line chart, it is found that most of the probability distributions are between 0.45 and 0.55. Therefore, this paper speculated that the prediction probability may follow a normal distribution, so a Q-Q chart is used to judge whether the data obeys normal.

Q-Q plots [12] are often used to evaluate whether a data set conforms to a particular theoretical distribution (such as a normal distribution). The verification results with the Q-Q diagram are shown in Figure 3

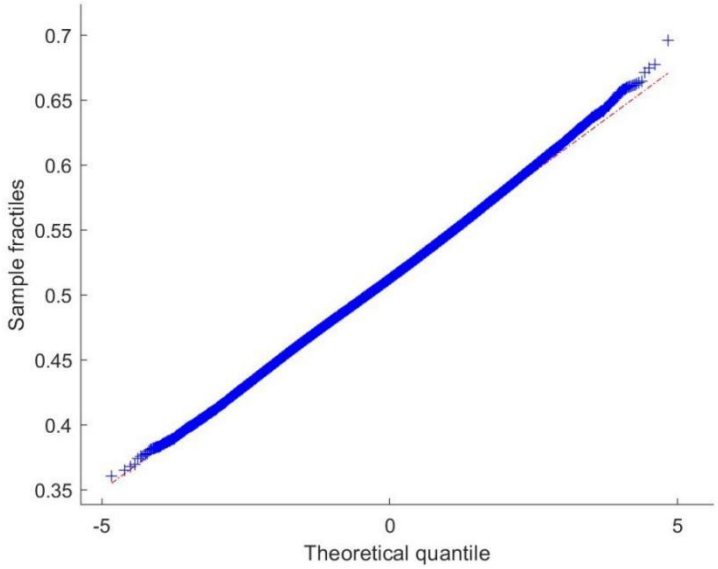

Figure 3. Q-Q graph normal distribution test

As can be seen from the figure, the data points roughly fall on a straight line, so it is considered that the data set conforms to a normal distribution.

4. Conclusions

In this paper, the missing value and outlier data in the data set were first processed, and then Spearman correlation analysis was carried out between these 20 indicators and the probability of flood occurrence to obtain the correlation between each indicator and the probability of flood occurrence.

Then, the first 11 indicators with high correlation were selected, the data set was divided into a training set and verification set, and the prediction model was established by using a multi-layer perceptron. A prediction model with a root-mean-square error of 0.038242 was obtained, and the accuracy of the model was judged. Based on the established prediction model, the required index data is extracted from the predicted data set and put into the model for prediction. The histogram and line chart are drawn according to the predicted flood probability, and the Q-Q chart is combined for verification. It is considered that the distribution of the predicted probability is approximately subject to the normal distribution, thus proving the accuracy of the model.

This paper provides a research idea and framework for flood prediction. Through data processing and analysis, model construction, and accuracy evaluation, the feasibility of Spearman correlation analysis combined with a multi-layer perceptron flood probability prediction model is proved.

References

[1] YANG Luji. Application of Rise Rate Analysis Method in Flood Prediction and Early Warning of Small and Medium-sized Rivers in Guangxi: A Case Study of Flood at Shatou Town Hydrological Station in Dong'an River[J]. Guangxi Water Resources and Hydropower, 2024(02):52-57.

[2] YU Shiming. Discussion on the Value of Flood Prediction Model in Disaster Mitigation [J]. Sichuan Water Resources, 2020,41(03):124-126

[3] LIU Zhiyu. Research and Practice of Key Technologies for Flood Prediction and Forecasting[J]. China Water Resources, 2020(17):7-10

[4] Kou Ye. Research on early warning and forecasting of flood in Langju water based on big data[J]. Hydraulic Science and Cold Region Engineering, 2024,7(01):85-89

[5] ZHANG Shi'an, WANG Qiang, WU Changjie Application of technological innovation and digital tools in river basin governance[C]//: 2023 China Water Conservancy Conference, Zhengzhou, Henan, China, 2023

[6] Yun Aoting, Zhang Jingfang, Zhang Haofei, et al Key Technologies and Applications of Flood Forecasting and Dispatching Model Based on Artificial Intelligence Technology[J]. Inner Mongolia Water Resources, 2021(06):15-17

[7] Chen J, Li Y, Zhang S. Fast Prediction of Urban Flooding Water Depth Based on CNN-LSTM[J]. WATER, 2023,15(7):14.

[8] Chen J, Li Y, Zhang C, et al. Urban Flooding Prediction Method Based on the Combination of LSTM Neural Network and Numerical Model[J]. INTERNATIONAL JOURNAL OF ENVIRONMENTAL RESEARCH AND PUBLIC HEALTH, 2023,20(2):12.

[9] Zhang Zhe'an, Liu Longcheng, Wang Shuli, et al Application of Multiple Linear Regression Model and Multilayer Perceptron Neural Network in Prediction of Argillaceous Content in Uranium Ore Logging[J]. Uranium Geology, 2024,40(05):1007-1013

[10] Su Riguga, Zhang Jintun, Wang Yongxia Species diversity and neural network prediction of forest communities in Songshan Nature Reserve, Beijing[J]. Acta Ecologica Sinica, 2013,33(11):3394-3403

[11] Song H Y, Park S. An Analysis of Correlation between Personality and Visiting Place using Spearman's Rank Correlation Coefficient[J]. KSII TRANSACTIONS ON INTERNET AND INFORMATION SYSTEMS, 2020,14(5):1951-1966.

[12] Idelkun, Sun Kai, Wang Bin, et al Delay Distribution Between Nodes in Clock Synchronization in Wireless Sensor Networks[J]. Journal of Computer Applications, 2020,40(S2):85-89.