Universe Scientific Publishing

Study on the correlation between deforestation and flood occurrence: an analysis based on Lasso regression

Hanyue Liu [1] Tengfei Meng [2] Junchao Zhang [2] Zhengtao Liao [3] Fuyun Chen [4]*

1 Cyberspace Security College of Qufu Normal University, Jining Shandong 272000
2 Guilin University of Technology at Nanning, Chongzuo Guangxi 532100
3 Sanya University, Sanya Hainan 572000
4 Chengdu University of Information Technology, Chengdu Sichuan 610225

Abstract: Flood is a natural phenomenon in which the water volume of rivers and lakes increases rapidly or the water level rises suddenly due to natural factors such as heavy rain, melting ice and snow, and storm surges. As a natural disaster with strong suddenness and wide impact, it poses a serious threat to human society. Based on historical flood event data, this paper focuses on the study of the multi-level risk warning system of flood disasters, aiming to improve the response capability and disaster reduction efficiency of flood events. Through Lasso regression and cross-validation technology, five key factors closely related to flood occurrence are screened out-"deforestation, climate change, silt deposition, agricultural practice and insufficient planning". The role mechanism of these factors in causing floods is further analyzed, and scientific suggestions and optimization measures for early prevention of floods are proposed, in order to provide data support and theoretical basis for the formulation of disaster prevention and mitigation policies.

Keywords: flood disaster; risk warning; Lasso regression; deforestation; disaster prevention and mitigation

1 Introduction

Flood is an extremely destructive natural disaster. It is caused by natural factors such as heavy rain, melting ice and snow, and storm surge. It causes the rapid increase or sudden rise of water levels in rivers, lakes and other water bodies, which has a great impact on life safety, infrastructure, agricultural production and ecosystems. pose a serious threat. In the context of climate change and intensified human activities, the frequency and intensity of flood disasters have increased significantly [1]. How to understand the causes of floods, identify key influencing factors and build an effective early warning system has become a core topic in disaster prevention and reduction research.

Traditional research mostly focuses on hydrological and meteorological factors, but the disturbing effects of human activities such as deforestation, land development, and agricultural practices on hydrological systems are increasingly apparent. Deforestation reduces the surface's ability to absorb water, while climate change further amplifies flood risks by changing precipitation patterns and increasing the frequency of extreme weather. Unreasonable agricultural practices and land planning exacerbate surface water loss and soil erosion [2]. Based on historical flood event data, this article uses Lasso regression combined with ten-fold cross-validation to screen out five key variables: deforestation, climate change, silt deposition, agricultural practices and insufficient planning, and conducts an in-depth analysis of their role in the causes of floods. Scientific flood early warning systems and disaster prevention and reduction policies provide theoretical basis and practical support.

Lasso regression is widely used in high-dimensional data analysis due to its variable selection and shrinkage capabilities. By introducing the L1 regularization term, Lasso regression can effectively prevent overfitting and eliminate redundant variables, thereby improving the model's explanatory power and prediction accuracy. Based on historical flood event data, this study uses Lasso regression combined with ten-fold cross-validation to screen out five key variables: deforestation, climate change, silt deposition, agricultural practices, and insufficient planning, and deeply analyzes their role in flood causes [3]. At the same time, this paper proposes targeted risk prevention and management measures, aiming to provide theoretical and practical support for building a scientific flood warning system and formulating disaster prevention and mitigation policies.

2 Related work

In recent years, flood risk research has become an important topic in the field of environmental science and disaster management, fo-

cusing on flood causes, risk prediction, and early warning system construction. A large number of studies have shown that climate change has significantly increased the risk of floods by increasing extreme rainfall events and changing precipitation patterns, exacerbating glacier melting and river water level fluctuations [4]. At the same time, deforestation, as an important means of human activity intervention, further increases flood risks by weakening the ability to regulate surface water and exacerbating soil erosion. Some studies have also linked land use changes, urbanization and unreasonable agricultural practices to flood occurrence, revealing the profound impact of human activities on the hydrological cycle.

In terms of the application of data-driven models, machine learning methods such as random forests, support vector machines, and statistical regression models such as ridge regression are widely used in flood risk assessment, while Lasso regression is particularly prominent in high-dimensional data analysis due to its variable selection ability [5]. By introducing L1 regularization, Lasso regression can eliminate redundant variables, solve feature collinearity problems, and achieve sparse modeling. In addition, research has gradually shifted from single-factor models to multi-source data fusion, combining multidimensional data such as meteorology, topography, land use, and socio-economics to build a regionalized flood assessment system. However, multivariate models still face challenges such as data quality dependence and regional difference adaptability. This paper combines Lasso regression with cross-validation technology to screen out the key factors of flood occurrence, providing new technical paths and practical suggestions for risk assessment and management.

3 Screening of characteristic variables of flood outbreaks

Lasso regression (Least Absolute Selection and Shrinkage Operator) is a compression estimation method based on linear regression. Its full name is “least absolute selection and shrinkage operator”. Lasso regression can effectively achieve variable selection and prevent model overfitting and collinearity by adding L1 regularization terms to the loss function [6]. Its optimization goal is to minimize the following loss function:

$$L(\beta) = \frac{1}{2N}\sum_{i=1}^{N}(\gamma_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\sum_{|j=1}^{p}|\beta_j| \quad (1)$$

Among them, the first term is the sum of squares of the residuals, which measures the degree of fit of the model to the data; the second term is the L1 regularization term, which promotes the unimportant variable coefficients to shrink to zero by constraining the regression coefficients to achieve sparse modeling. The regularization parameter λ controls the regularization strength. The larger its value, the more variable coefficients are shrunk to zero, thereby simplifying the model.

Lasso regression is widely used in high-dimensional data analysis, especially in scenarios with a large number of features and strong correlation between variables. Compared with traditional least squares regression, Lasso regression can automatically screen out key features that have a significant impact on the target variable, while improving the generalization ability of the model [7].

In the study of this article, Lasso regression analysis was performed on the variables (x_1,x_2,x_3,…,x_{19},x_{20}), and the cross-validation method was used to evaluate the model performance. First, by gradually adjusting the regularization parameter λ, a regression coefficient path diagram is drawn. Figure 1 shows the regression coefficient path diagram, in which the horizontal axis is the logarithm of the regularization parameter log λ, and the vertical axis is the regression coefficient of different variables. The path diagram intuitively reflects the shrinkage process of the coefficients of each variable as the λ value changes. When λ is small, the regularization intensity is low, the coefficients of most variables are non-zero, and the model is more complex and contains most feature variables. As λ increases (log λ moves to the right), the regularization intensity gradually increases, the coefficients of most variables gradually shrink to zero, and the model begins to simplify. When λ increases to a certain extent, only a few variables have non-zero coefficients, indicating that these variables are important influencing factors of the target variable, and the remaining variables are eliminated. It can be observed from the figure that the curves of different colors represent the regression coefficients of each variable. As λ increases, the curves of only a few variables continue to the larger λ region, indicating that these variables make significant contributions to the model and are key characteristic variables [8]. Specifically, the variables whose curves remain significantly non-zero in the path plot are the most important and have a higher correlation with flood occurrence, while the coefficients of the remaining variables gradually shrink to zero, showing a weak impact on the target variable. When λ is small, the mod-

el retains a large number of variables, and has strong explanatory power but may have the risk of over-fitting; when λ is large, only a small number of variables are retained, the model is simpler, the interpretability is enhanced, and the risk of over-fitting is reduced.

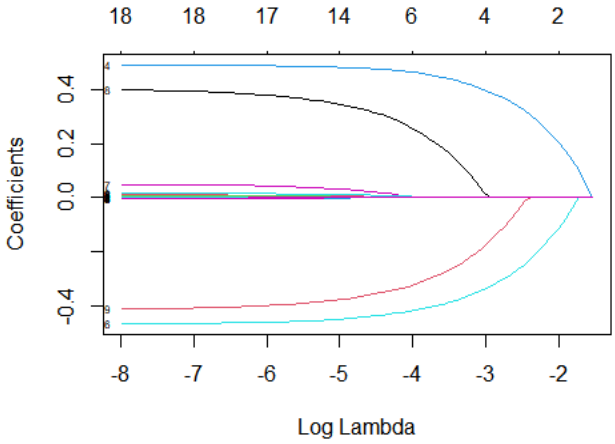

Figure 1 Regression coefficient path diagram

Figure 2 shows the ten-fold cross validation results of Lasso regression, which is used to select the optimal regularization parameter λ. The horizontal axis is the logarithm of the regularization parameter log λ, and the vertical axis is the mean squared error (MSE). The red dot represents the mean MSE corresponding to each λ value, and the gray error bar shows its standard error range. Through this figure, we can intuitively observe how the generalization ability of the model changes with the change of λ, and finally determine the λ value that makes the model perform best.

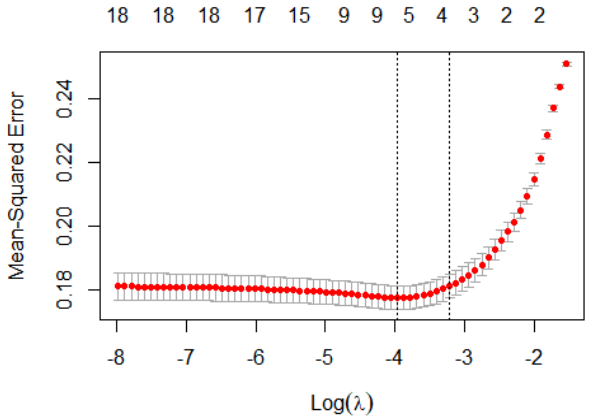

Figure 2 Ten-fold cross validation graph

As λ increases monotonically, the regression coefficients of more and more indicator variables gradually shrink to zero, which indicates that the increase in regularization strength effectively eliminates the redundant and collinear variable characteristics in the model. Lasso regression automatically realizes variable selection through this process, while simplifying the model structure and improving the generalization ability.

Combined with the results of ten-fold cross-validation, we can further analyze that when $\ln(\lambda_1) \approx -4$, the mean square error (MSE) reaches the minimum value, at which time the model fits best, retains the five most important variables, and better balances the complexity and fitting ability of the model. However, considering the need for model simplicity and stability in practical applications, the "one standard error criterion" can be used for optimization. When $\ln(\lambda_2) = -3.5$, although the MSE increases slightly (close to one standard error of the minimum error), the remaining indicator variables screened out are reduced to 4 as shown in Table 1.

Table 1 λ parameter comparison table

Main factors		
Intercept value	0.4702	0.4849
Deforestation	0.4648	0.4196
Climate change	-0.4155	-0.3631
Silt	0.2504	0.0865
Agricultural practices	-0.3207	-0.2215
Inadequate planning	0.0041	0
Other	0	0

The regression effects under both λ parameters are very good. In order to improve the accuracy of the model as much as possible, we finally chose to include five indicator variables (deforestation X_4, climate change X_6, silt X_8, agricultural practice X_9, and insufficient planning X_{19}) when $\ln(\lambda_1) \approx -4$ to solve subsequent problems.

4 Discussions

This study used Lasso regression combined with ten-fold cross-validation technology to screen out five key factors: deforestation, climate change, silt deposition, agricultural practices and insufficient planning, providing a scientific basis for early warning and prevention of flood risks. Deforestation, as the most important indicator, has a profound impact on flood formation. It weakens the surface water regulation ability, intensifies soil erosion and sedimentation, significantly reduces river capacity, and reduces precipitation interception capacity, resulting in a significant increase in flood risks. In addition, climate change directly promotes the formation of floods by changing rainfall patterns, increasing the frequency of extreme precipitation events, and accelerating the melting of ice and snow and rising river levels, and interacts with other factors to amplify the complexity of disasters. These findings provide important scientific basis for strengthening forest protection and addressing climate change policies.

Silt deposits, agricultural practices and inadequate planning largely reflect localized impacts of human activities. Silt deposition reduces the flood discharge capacity of rivers and is closely related to soil erosion caused by unreasonable agricultural practices, while insufficient planning aggravates the problem of surface hardening during urbanization and accelerates the formation and spread of floods. The results of this study suggest that it is necessary to comprehensively improve the refined level of flood risk assessment and management starting from regional land use and infrastructure design. In addition, Lasso regression performs well in automatic variable selection and processing multi-dimensional correlation features, providing an efficient technical means for extracting key variables in complex environmental systems. However, the dependence of model performance on regularization parameters needs to be further optimized to ensure explanatory power. balance with simplicity.

5 Conclusion

In the study of this article, by screening the characteristic variables of flood occurrence and using Lasso regression combined with ten-fold cross-validation method, we successfully identified core indicators closely related to flood events. The findings indicate that deforestation, climate change, silt deposition, agricultural practices and inadequate planning are the five key factors influencing the occurrence of floods. These indicators are not only statistically significant, but also have strong explanatory power in actual environmental mechanisms. However, further analysis also shows that although appropriately increasing the regularization intensity may sacrifice a certain degree of fitting accuracy, it can simplify the model and further reduce the variables to four, making the focus of the analysis more focused, especially in actual policy formulation and intervention. It is of great significance in the implementation of the measures. The research results not only methodologically prove the applicability of Lasso regression in flood risk analysis, but also clarify the significant impact of human activities and environmental changes on flood events in a practical sense. Deforestation and climate change are important components of global environmental problems, and their driving effects on flood events directly reflect the causal link between ecosystem degradation and climate instability. At the same time, regional factors such as silt deposition, agricultural practices, and inadequate planning reveal potential problems in human land use and infrastructure planning. These factors work together to influence the dynamic changes of the hydrological cycle and

the frequency of floods. Therefore, intervention and management of these factors are not only the key to flood risk prevention and control, but also provide a practical path for the sustainable development of regional ecosystems.

References:

[1]Satapathy, D. P., & Mishra, B. P. (2024). Flood susceptibility modeling by integrating tree-based regression with metaheuristic algorithm, BWO. Transactions in GIS, 28(5), 1043-1064.

[2]Gao, B., Shan, Y., Liu, X., Yin, S., Yu, B., Cui, C., & Cao, L. (2024). Prediction and driving factors of forest fire occurrence in Jilin Province, China. Journal of Forestry Research, 35(1), 21.

[3]Shojaeian, A., Shafizadeh-Moghadam, H., Sharafati, A., & Shahabi, H. (2024). Extreme flash flood susceptibility mapping using a novel PCA-based model stacking approach. Advances in Space Research.

[4]Tapias-Rivera, J., & Gutiérrez, J. D. (2023). Environmental and socio-economic determinants of the occurrence of malaria clusters in Colombia. Acta Tropica, 241, 106892.

[5]Khosravi, K., Golkarian, A., Melesse, A. M., & Deo, R. C. (2022). Suspended sediment load modeling using advanced hybrid rotation forest based elastic network approach. Journal of Hydrology, 610, 127963.

[6]Kainthura, P., & Sharma, N. (2022). Hybrid machine learning approach for landslide prediction, Uttarakhand, India. Scientific reports,12(1), 20101.

[7]Kamil, N. N., Xiao, S., Salleh, S. N. S., Xu, H., & Zhuang, C. C. (2024). Nonlinear impacts of climate anomalies on oil palm productivity.Heliyon, 10(15).

[8]Kalu, I., Ndehedehe, C. E., Okwuashi, O., & Eyoh, A. E. (2021). Assessing freshwater changes over Southern and Central Africa (2002–2017). Remote Sensing, 13(13), 2543.