RANDOM FOREST MODEL SAMPLING METHOD OF NON-BENGGANG POINTS TO PREDICT BENGGANG EROSION SUSCEPTIBILITY

XU, X. M.* – ZHANG, X. Y. – QIN, L. H. – LI, R. – DU, J. – GUO, Y. H.

School of Geography and Environmental Engineering, Gannan Normal University, Ganzhou 341000, Jiangxi, China

**Corresponding author*
e-mail: xuxiangming@gnnu.edu.cn

(Received 3rd May 2025; accepted 26th Jun 2025)

Abstract. Benggang is a prevalent and distinct type of soil water erosion found in southern China that causes considerable harm, including severe soil erosion and soil degradation. Previous studies have explored sampling method effects insufficiently on model performance of benggang erosion susceptibility (BES), as well as an in-depth understanding of driving factors of BES. Addressing these gaps is essential to facilitate the establishment of a scientific foundation for developing measures aimed at preventing and controlling benggang erosion. Consequently, three non-benggang points sampling techniques and four machine learning models were employed to examine the impact of sampling methods on the performance of BES. The ensuing findings are as follows: Firstly, the random forest sampling method (RFSM) revealed that the five metrics exhibited significant superiority over the other two sampling methods, namely random sampling method (RSM) and historical data sampling method (HDSM). Furthermore, the Random Forest (RF) model was identified as the optimal model. Secondly, the elevation factor was identified as the predominant key factor of BES in Ganxian County, with Fractional Vegetation Cover (FVC), rainfall erosivity factor, precipitation, and lithology identified as the main factors. Lastly, the number densities of the five grades of BES were 0.02, 0.08, 0.09, 1.11, and 1.34, respectively. The area with relatively high and high grades is of paramount importance for the prevention and control of benggang.

Keywords: *RF model, XGBoost model, SVM model, benggang erosion susceptibility, driving factor*

Introduction

Benggang erosion is a specific type of soil water erosion resulting from the interplay of gravitational and hydraulic forces on the slope of a deep weathering layer (He et al., 2024). A survey conducted in 2005 identified 239,000 benggang sites in the following seven provinces of southern China: Hubei, Hunan, Jiangxi, Anhui, Fujian, Guangdong, and Guangxi (Liao et al., 2022). These sites are distinguished by their substantial erosion, considerable eruption force, and rapid development rate. The erosion modulus in these areas is typically in the range of 30,000 to 50,000 t/km^2·a, which is significantly higher than the permissible soil loss limit of 500 t/km^2·a in the southern China (Chen et al., 2013; Ou et al., 2024). This phenomenon not only exacerbates the risk of geological disasters, such as mudslides and landslides, but also imposes significant constraints on ecological restoration, agricultural production, and sustainable regional socio-economic development. Consequently, research studies focusing on spatial prediction and driving mechanisms are essential. This research is essential for the advancement of early warning systems, the execution of preventive measures, and the management of benggangs. Additionally, the findings should be incorporated into spatial planning initiatives to ensure the effective management of land use and the mitigation of potential risks (Liao et al., 2022; Deng and Cai, 2024).

DOI: http://dx.doi.org/10.15666/aeer/2305_89999016

In recent years, with the rapid advancement of geographic information systems (GIS) and intelligent prediction technology, there has been significant progress in the research on the evaluation of benggang or gully erosion susceptibility. Machine learning models are data-driven approaches that have proven effective in addressing the challenges of overfitting and multicollinearity present in other models (Wei et al., 2021). Machine learning models offer the advantage of high prediction accuracy and have emerged as a prominent area of research in disaster susceptibility assessment models. A significant body of research has been conducted by numerous scholars, who have explored the potential of machine learning models in combination with various environmental factors to assess susceptibility (Nguyen et al., 2021; Senanayake et al., 2022; Berihun et al., 2025). Arabameri et al. (2019) conducted a comparative analysis of the performance of machine learning models in gully erosion susceptibility assessment. Their findings indicated that the boosted regression tree and frequency ratio models exhibited high prediction accuracy (Arabameri et al., 2019). Concurrently, Guo, et al. (2024) performed an investigation of the statistical q-values of 17 environmental factors by employing GeoDetector and selecting distinct combinations of environmental factors based on the percentage of cumulative q-values. Their findings indicated that a more satisfactory model accuracy could be attained by considering the primary environmental factors (Guo et al., 2024).

Density of benggang points typically holds more closely related to the evaluation of BES than benggang impacted sizes. A growing body of research has underscored the significance of the sample set selection method in determining BES, a process that is instrumental in ascertaining the representativeness of the training and test sets of the model (Lana et al., 2022). The dataset of machine learning model involves the designation of benggang points that have occurred in the study area as positive points, while other areas are classified as non-benggang areas by default (Guo et al., 2023, 2024). The conventional sampling methodologies for non-benggang points encompass the random sampling method and the historical data method. The random sampling method involves the random selection of an equal number of points as negative points in non-benggang areas, while the historical data method entails the sampling of an equal number of points as negative points by integrating the findings from historical literature and field investigations (Ji et al., 2019; Liu et al., 2024). The arbitrary selection of negative points may jeopardize the model's efficacy, especially when the disparities between negative and positive points are significant. This may result in a reduction in the model's classification accuracy and stability (Bouramtane et al., 2022; Naceur et al., 2024). The historical data method disregards the concern that an area devoid of historical disaster events does not preclude the possibility of future disasters in that area, thereby compromising the accuracy of model predictions. Machine learning models, such as random forest models, offer several advantages (Khosravi et al., 2023). These include enhanced generalization ability, the ability to capture nonlinear relationships, and robustness to noise. Additionally, they reduce human interference through a data-driven approach (Li et al., 2024a; Gao et al., 2025). The impact of random forest sampling of negative points of non-benggang sites on the prediction of BES remains to be further explored.

This study selected 983 benggang sites in Ganxian county that had been confirmed as positive samples by the Jiangxi Province soil and water conservation planning benggang survey. In order to establish a machine learning sample dataset, three sampling methods were employed to screen an equal number of non-benggang points as negative samples. The optimal machine learning model was employed to evaluate the BES and predict the driving factors. The specific studies are as follows: (1) A comparison of non-benggang sampling methods: This objective involves the comparison of the performance of random sampling method (RSM), historical data sampling methods (HDSM), and random forest sampling method (RFSM) in the evaluation of BES. Additionally, it will explore the impact of different sampling methods on model accuracy. (2) A comparison of model performance: Four machine learning models, namely Random Forest (RF), eXtreme Gradient Boosting (XGBoost), Support Vector Machine (SVM), and Logistic Regression (LR), will be conducted, and the accuracy of the four models in predicting BES will be evaluated. (3) Prediction of BES and analysis of driving factors: The optimal machine learning model will be utilized to predict BES and to analyze the primary driving factors of BES. The objective of this study is to accurately localize potentially benggang-dangered areas. This will provide scientific support for the early warning and prevention of benggang disasters. Moreover, the findings can serve as a valuable reference for evaluation of other geologic hazards and advancement of geologic hazard prediction and prevention technologies.

Materials and methods

Study area

The administrative district of Ganxian is under the jurisdiction of Ganzhou City, which is located in Jiangxi province, China (*Fig. 1*). The coordinates of Ganxian are 114°42'~115°22'E, 25°26'~26°17'N, and its topography is predominantly undulating hills and mountains. The subtropical humid monsoon climate is characteristic of this region, which is situated within the southern periphery of the central subtropics. The frost-free period is protracted, and the average temperature over an extended period is 19.4°C. The region receives abundant rainfall, with an average precipitation volume of 1,438 mm over many years. The precipitation distribution throughout the year exhibits notable seasonal variation (Chen et al., 2013). The Ganxian county is distinguished by its extensive granite deposits, and the benggang disaster is particularly pronounced in this region, which has the highest density, number, and type of benggang in Jiangxi Province. The distribution data of 983 benggang points in this study were primarily derived from the updated data results of the 2015 Jiangxi Province soil and water conservation planning benggang survey (*Fig. 1*). These data were interpreted using a combination of field surveys of benggang points and visual interpretation of remote sensing imagery. To ensure the accuracy of the data, the interpretation process underwent a dual validation process.

Data description

Comprehensive details concerning the data sources in this study are delineated in *Table 1* (Xu, 2025).

Non-benggang points sampling methods

The present study employed three distinct non-benggang point sampling methods: the random sampling method, the historical data sampling method, and the random forest sampling method (Hitouri et al., 2022; Huang et al., 2022).

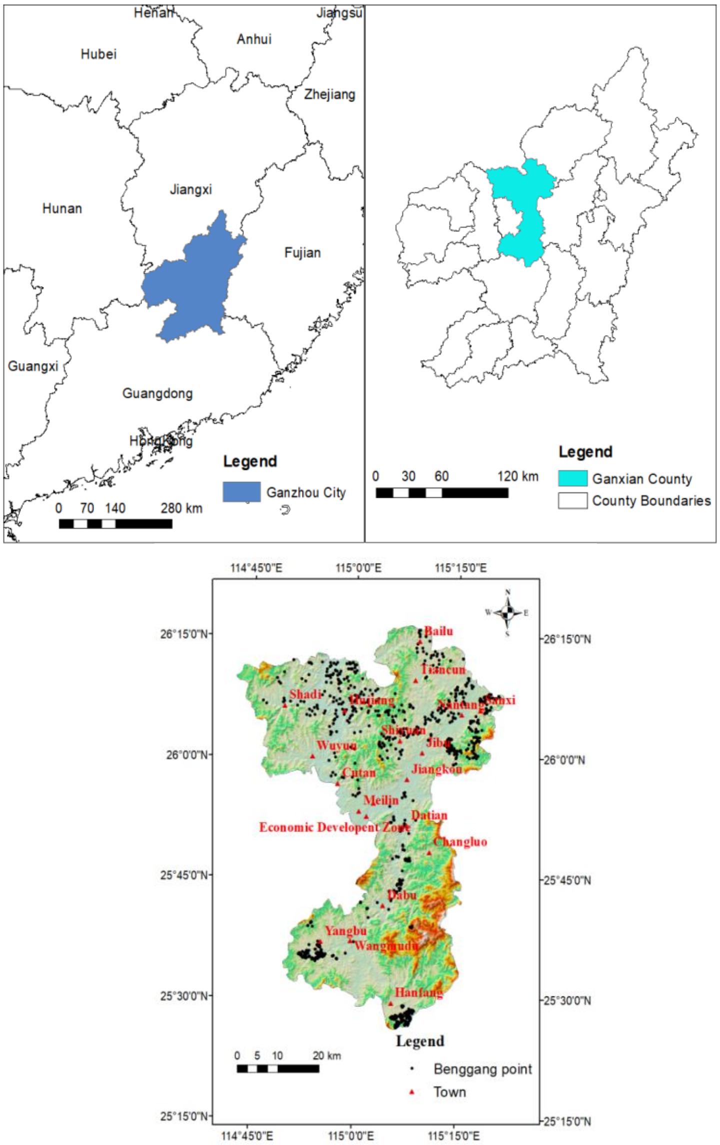

Figure 1. *The location of study area and the distribution of benggang points*

Table 1. *Data sources*

Data Type	Data Sources	Data Accuracy	Year
Normalized Difference Vegetation Index	Resource and Environmental Science Data Platform, Chinese Academy of Sciences (https://www.resdc.cn/)	1 km	2020
Digital Elevation Model	Geospatial Data Cloud (https://www.gscloud.cn/)	30 m	2024
Monthly mean temperature	National Meteorological Science Data Centre (http://data.cma.cn/)	1 km	2000-2020
Monthly mean precipitation	National Meteorological Science Data Centre (http://data.cma.cn/)	1 km	2000-2020
Soil texture (Content of sand, silt,clay)	National Tibetan Plateau Data Center (https://data.tpdc.ac.cn/)	1 km	2009
Land use	Resource and Environmental Science Data Platform, Chinese Academy of Sciences (https://www.resdc.cn/)	30 m	2020

Random sampling method

The RSM is a conventional negative points selection method. The specific steps involved are as follows:

Initially, a regular grid is to be generated, with a dimension of 800 × 800 m in the designated study area. The data at the center point of each grid is then to be extracted, yielding a total of 4,664 non-benggang points.

Subsequently, using Python 3.10, the random seed is established, and 983 points are randomly chosen from the 4,664 non-benggang points as negative points, as depicted in *Figure 2(a)*.

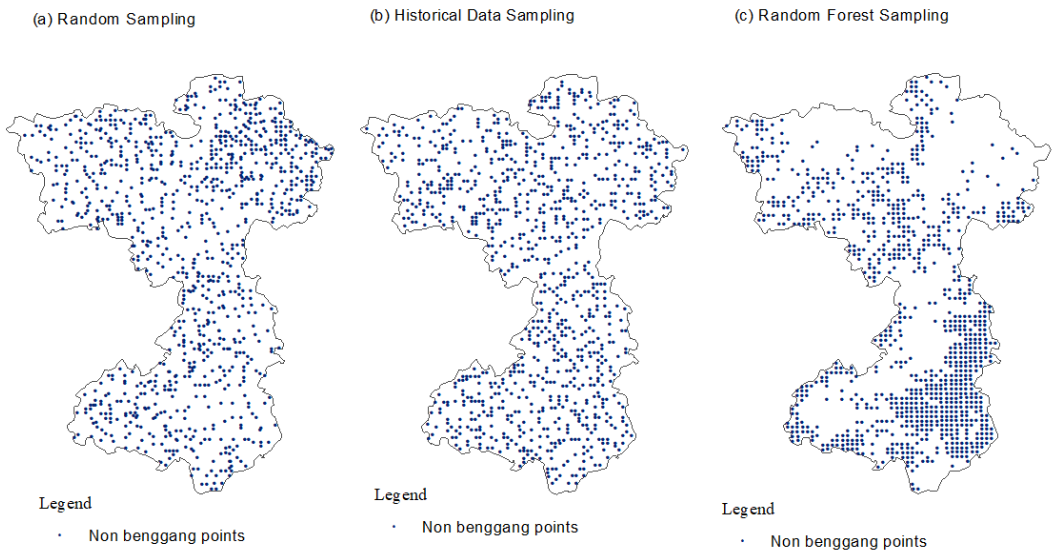

Figure 2. *The three non-benggang points sampling methods*

Historical data sampling method

A regular grid of 800 × 800 m was generated in the study area, and the data at the center point of each grid were extracted, generating a total of 4,664 non-benggang points. By integrating historical literature with field surveys, 983 locations that have not been documented as benggang events and currently exhibit no benggang occurrences were identified as negative spots, as seen in *Figure 2(b)*.

Random forest sampling method

The RFSM is a data mining technique that employs a RF model to identify non-benggang points with the most significant background differences. This objective is realized through the integration of benggang and non-benggang points. The following steps are involved (Biau, 2012):

Initially, a regular grid of 800 × 800 m was generated in the study area, and the data at the center point of each grid were extracted, yielding a total of 4,664 non-benggang points. These 4,664 non-benggang points were amalgamated with 983 benggang points to remediate the category imbalance issue through the implementation of the Synthetic Minority Oversampling Technique (SMOTE).

Subsequently, the dataset was randomly partitioned into training and test sets, with a proportion of 80% and 20%, respectively. The RF model was then trained with the following parameters: n_estimators=200, max_depth=10, random_state=42. The probability of non-benggang points is predicted using the trained model.

The third step involves calculating the Euclidean distance between the non-benggang sample and the mean of the benggang sample.

Finally, the predicted probability of the RF is combined with the Euclidean distance, and the principal component analysis (PCA) is employed for reduction of dimensionality to calculate the comprehensive background difference indicator. The PCA results demonstrated a Euclidean distance of 0.98 and a non-benggang prediction weight of 0.16 in this study. The initial 983 points exhibiting elevated comprehensive background difference indicators were identified as non-benggang points. *Figure 2(c)* illustrates the results of RFSM.

Benggang erosion susceptibility mapping models

Four machine learning models will be utilized to assess BES: RF, XGBoost, SVM, and LR. The specific steps involved are as follows:

Initially, the following fourteen environmental factors were selected as inputs to the model: elevation (m), slope (°), aspect (°), slope length and steepness factor (LS factor, dimensionless), land use (categorical data), Fractional Vegetation Cover (FVC, %), precipitation (mm), rainfall erosivity factor (R factor, $MJ\cdot hm^{-2}\cdot h^{-1}\cdot a^{-1}$), soil texture (categorical data), lithology (categorical data), soil erodibility factor (K factor, $Mg\cdot km^{2}\cdot h\cdot km^{-2}\cdot MJ^{-1}\cdot mm^{-1}$), soil type (categorical data), Topographic Wetness Index (TWI, dimensionless), river network ($km\cdot km^{-2}$). Standardizing all environmental factors ensured that each factor had consistent units of measurement. The extraction of elevation, slope and aspect data was conducted using DEM, while the calculations of the R factor, K factor, LS factor and FVC index were informed by relevant literature, and the calculations were shown in *Equation 1-6*. In this study, land use types were divided into five categories: cultivated land, forestland, grassland, water body, built-up land. Soil textures were also divided into five categories: clay, silt loam, loam, sandy clay loam, sandy loam. Soil types were divided into eleven categories: Rendzic Leptosols (RGc), Rendzic Leptosols, dark phase (RGd), Luvisols (FLe), Orthic Anthrosols (ATc), Greyic Luvisols (Gle), Chromic Acrisols (ACh), Haplic Acrisols (ACu), Calcic Chernozems (CMo), Haplic Alisols (ALh), Hydragric Anthrosols (WR), Arenosols (DS). The main lithology in the study area was gray sandstone, purple conglomerate, slate, quartz sandstone, granite, dolomite, silty clay, diabase, quartz conglomerate, silt stone, fine sandstone, limestone. The *Fig. 3* presents a visual representation of the distribution of environmental factors in Ganxian County.

Subsequently, the four machine learning models were trained using the training data (Cristianint and Shawe-Taylor, 2000; Maalouf et al., 2011; Biau, 2012; Chen and Guestrin, 2016).

Finally, the performance metrics of the four models are to be assessed using the data of test set.

R factor

The R factor was estimated using the Wischmeier equation (Kite, 2001) based on monthly and annual average rainfall data. The calculations are shown in *Equation 1.*

$$R=\sum_{i=1}^{12}1.735\times10^{(1.5\lg\frac{p_i^2}{p})-0.8188} \qquad \text{(Eq.1)}$$

In the equation, P_i represents the monthly average rainfall (mm), and P represents the annual average rainfall (mm).

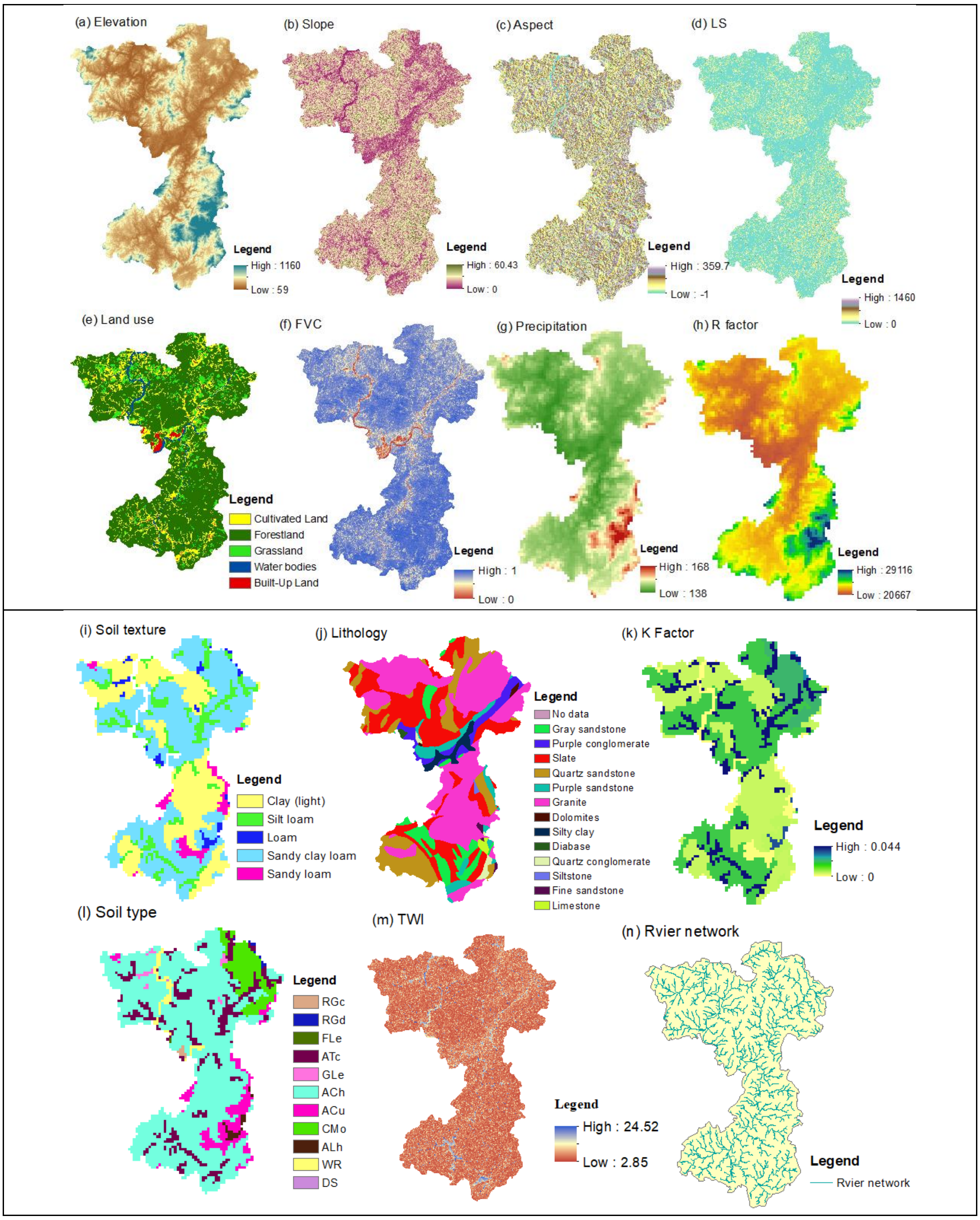

Figure 3. *Distribution of environmental factors of benggang erosion susceptibility: a Elevation map, b Slope map, c Aspect map, d LS map, e Land Use map, f FVC map, g Precipitation map, h R factor map, i Soil Texture map, j Lithology map, k K factor map, l Soil Type map, m TWI map, n River Network map*

K factor

Soil erosion is primarily influenced by the medium of soil, and the soil erodibility factor (K) is employed to evaluate the soil susceptibility to erosion. In this study, the EPIC model proposed by Williams (1990) was utilized to simulate and calculate the soil erodibility factor based on a soil database. The calculation is predominantly contingent on the content of silt, sand, clay, and organic matter in the soil. The calculations are delineated in *Equation 2*.

$$K = 0.1317\left\{0.2 + 0.3\exp\left[-0.0256SAN\left(1-\frac{SIL}{100}\right)\right]\right\}\times\left(\frac{SIL}{SIL+CLA}\right)^{0.3}\times\left[1-\frac{0.25C}{C+\exp\left(3.72-2.95C\right)}\right]\times\left[1-\frac{0.7SN1}{SN1+\exp\left(-5.51+22.9SN1\right)}\right] \quad \text{(Eq.2)}$$

In the formula, 0.1317 is the coefficient for converting US units to international units. SAN, SIL, CLA, and C represent the sand, silt, clay, and organic matter content (%) of soils, respectively. SN1 =1- SAN/100.

LS factor

The slope length and steepness factor (LS) is a metric that reflects the impact of terrain features on soil erosion (El Jarjini et al., 2023). The precise relationship between these factors remains to be elucidated. However, within a certain range, it has been observed that the longer the slope, the greater the accumulation of flow, and the steeper the slope, the faster the runoff velocity. However, once a threshold is attained, the rate of soil erosion will no longer increase. The calculations for slope (S) are delineated in *Equation 3*.

$$S = \begin{cases} 10.8\sin\theta + 0.03, \theta < 9\% \\ 16.8\sin\theta - 0.05, 18\% > \theta \geq 9\% \\ 21.91\sin\theta - 0.96, \theta \geq 18\% \end{cases} \quad \text{(Eq.3)}$$

In the equation, L denotes the slope length factor, S signifies the slope steepness factor, and θ represents the slope value derived from DEM data (El Jarjini et al., 2023; Lai et al., 2024).

Using ArcGIS 10.8 software, the slope length factor (L) was calculated using the *Equations 4 and 5* (Qiu et al., 2018).

$$L = \left(\frac{\lambda}{22.13}\right)^{m} \quad \text{(Eq.4)}$$

$$\begin{cases} m = 0.5 & \theta \geq 9\% \\ m = 0.4 & 9\% > \theta \geq 3\% \\ m = 0.3 & 3\% > \theta \geq 1\% \\ m = 0.2 & \theta < 1\% \end{cases} \quad \text{(Eq.5)}$$

In the equation, L denotes the slope length factor; λ signifies the slope length (m); m indicates the slope length exponent; and θ represents the slope steepness (%).

FVC index

FVC is defined as the percentage of the vertical projection area of vegetation in a given unit area. These calculations are shown in *Equation 6* (Cai et al., 2000; Li et al., 2021):

$$FVC = \frac{NDVI - NDVI_{\min}}{NDVI_{\max} - NDVI_{\min}} \tag{Eq.6}$$

In the formula, FVC represents the vegetation cover (%) , NDVI is the normalized difference vegetation index, with the multi-year monthly average used in this study. $\mathrm{NDVI_{max}}$ refers to the NDVI value of pixels completely covered by vegetation, and $\mathrm{NDVI_{min}}$ refers to the NDVI value of bare soil or areas without vegetation cover (Li et al., 2024b).

Validation

The following five evaluation metrics are employed in this study: Accuracy, Precision, Recall, F1-Score, and Area Under Curve (AUC) (Zabihi et al., 2018).

The accuracy of models is determined by the proportion of points that are correctly classified. It is calculated as follows (Shen et al., 2024):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{Eq.7}$$

In this formula, TP (true positive) denotes true cases, TN (true negative) denotes true negative cases, FP (false positive) denotes false positive cases, and FN (false negative) denotes false negative cases.

The precision rate is calculated as follows:

$$\mathrm{Precision} = \frac{TP}{TP + FP} \tag{Eq.8}$$

The calculation of recall is as follows:

$$Recall = \frac{TP}{TP + FN} \tag{Eq.9}$$

The F1-Score is the reconciled average of precision and recall, calculated as follows:

$$F1 - Score = 2 \times \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{Eq.10}$$

The area under the curve (AUC) of the ROC curve is a metric employed to evaluate the classification performance of a model. A model with optimal classification performance will have an AUC value that is close to 1 (Laraib et al., 2024). The AUC is computed as follows:

$$TPR = \frac{TP}{TP + FN} \tag{Eq.11}$$

DOI: http://dx.doi.org/10.15666/aeer/2305_89999016

$$FPR = \frac{FP}{FP + TN} \qquad \text{(Eq.12)}$$

Feature importance of driving factors

The optimal model was employed to assess the importance of each environmental factor on the BES. The optimal model quantifies the degree of influence of each environmental factor on the occurrence of benggang by calculating its characteristic importance. The extent of this influence is measured by the characteristic significance, with elevated values signifying a more substantial effect of the component on benggang occurrence.

The optimal model employs a multifaceted approach to assess the importance of each environmental factor. This multifaceted approach involves the calculation of the number of splits in the decision tree and the information gain from these splits. The feature importance is calculated as Rajbahadur et al. (2021):

$$\text{Importance}(f) = \frac{1}{N}\sum_{i=1}^{N}\{Gain(f,i) \times Split(f,i)\} \qquad \text{(Eq.13)}$$

In this formula, *Gain(f,i)* denotes the information gain brought by factor f in the ith decision tree, *Split(f,i)* denotes the number of splits of factor f in the ith decision tree, and N denotes the total number of decision trees.

Results

Non-benggang points sampling methods effects on model performance

The following five metrics are utilized to assess the accuracy performance of the models: Accuracy, Precision, Recall, F1-Score, and AUC (*Table 2 and Fig. 4*).

Table 2. *Metrics for evaluating the performance of machine learning models (* represents optimal values)*

Non-benggang point sampling method	Machine learning model	Accuracy	Precision	Recall	F1-Score	AUC
RSM	LR	0.5964	0.5845	0.6237	0.6035	0.6317
	RF	0.6904	0.7045	0.6392	0.6703	0.7502
	SVM	0.5939	0.7073	0.2990	0.4203	0.6112
	XGBoost	0.6675	0.6479	0.7113	0.6781	0.7399
HDSM	LR	0.6294	0.6176	0.6495	0.6332	0.6793
	RF	0.7487	0.7363	0.7629	0.7494	0.8202
	SVM	0.5457	1.0000	0.0773	0.1435	0.6844
	XGBoost	0.7360	0.7250	0.7474	0.7360	0.7897
RFSM	LR	0.7893	0.7734	0.8093	0.7909	0.8642
	RF	0.9619(*)	0.9891(*)	0.9330(*)	0.9602(*)	0.9830
	SVM	0.6878	0.6564	0.7680	0.7078	0.7083
	XGBoost	0.9467	0.9676	0.9227	0.9446	0.9836(*)

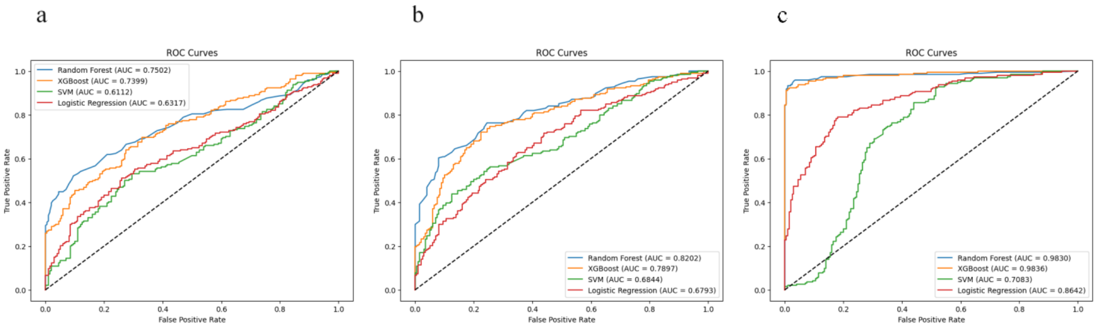

Figure 4. *The three sampling methods effects on the AUC metric. a RSM b HDSM c RFSM*

The performance of the models is delineated in *Table 2*. A comparison of the performance of the four models of the RSM method reveals that the Accuracy, Precision, Recall, F1-Score, and AUC metrics of the RF model are 0.6904, 0.7045, 0.6392, 0.6703, and 0.7502, respectively. The RF model demonstrates the highest Accuracy and AUC. The XGBoost model exhibited the highest Recall and F1-Score, at 0.7113 and 0.6781, respectively. The SVM model exhibited the highest precision at 0.7073. The LR model, conversely, exhibited a lower overall performance. A comparison of the performance of the four models of the HDSM method reveals that the RF model exhibits the highest Accuracy, Recall, F1-Score, and AUC at 0.7487, 0.7629, 0.7494, and 0.8202, respectively. The XGBoost model demonstrated the second-highest values, while the SVM model exhibited the highest precision metric. A similar comparison of the performance metrics of the four models from the RFSM method reveals that the RF model exhibits optimal values for Accuracy, Precision, Recall, and F1-Score at 0.9619, 0.9891, 0.9330, and 0.9602, respectively. The XGBoost model, on the other hand, demonstrates the optimal value for AUC at 0.9836. The remaining two models demonstrate suboptimal performance metrics.

The effect of the three non-benggang points sampling methods on model performance (*Fig. 5*) reveals that the RFSM method demonstrates optimal performance, with mean values of Accuracy, Precision, Recall, F1-Score, and AUC of 0.8464, 0.8466, 0.858, 0.8508, and 0.8847, respectively. The results indicate that the RFSM exhibit significant superiority over the other two sampling methods, namely RSM and HDSM ($p<0.01$).

A thorough evaluation of the model's performance across all 12 combinations indicates that the RFSM, employed by the sampling method, in conjunction with the RF model, exhibits optimal performance. This combination attains the highest accolades in terms of accuracy, precision, recall, and F1-score. The AUC value of 0.9830 for this model is marginally lower than the 0.9836 achieved by the XGBoost model in the RFSM method. The RFSM method exhibits a significant improvement in predictive accuracy of the model relative to the other two sampling techniques. The RF model performs optimally, with the XGBoost model ranking second and the LR and SVM models exhibiting comparatively lower performance. Consequently, in this study, the RFSM was selected as the non-benggang points sampling method, and the RF model was selected as the optimal model for the BES prediction.

DOI: http://dx.doi.org/10.15666/aeer/2305_89999016

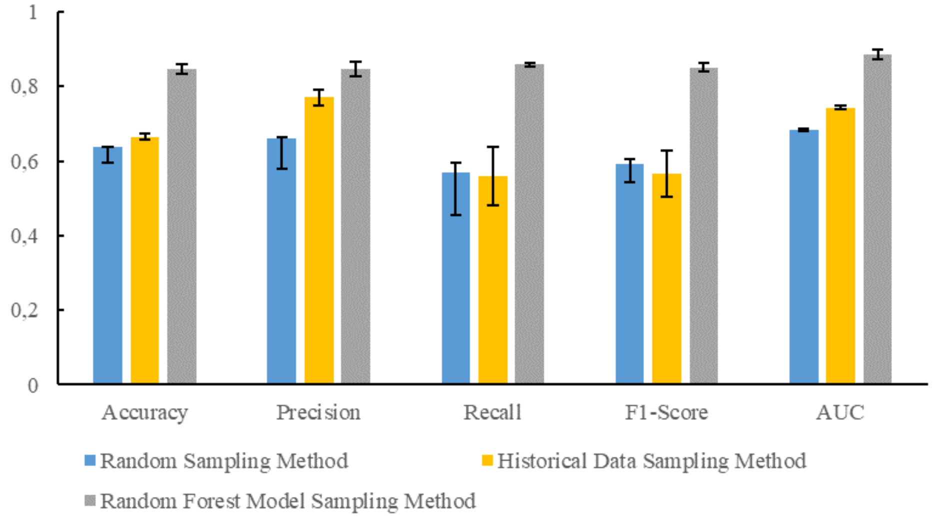

***Figure 5.** The three sampling methods effects on model performance metrics*

Variable importance analysis

The RF model was utilized to assess the significance of each environmental component on BES. The RF model is a statistical learning method that assesses the feature importance of each environmental component to ascertain its impact on the occurrence of benggang.

The outcomes of the impact factor detection are illustrated in *Figure 6*. The 14 impact factors were then classified into four distinct grades: key factors, main factors, relatively important factors, and low-impact factors, respectively. Key factors are defined as those with feature importance ≥ 0.2, and in this study, the feature importance of elevation is 0.2092, thus designating it as the most significant factor contributing to BES. The main factors are those with feature importance of 0.1 ≤ importance value < 0.2, and in this study, the main factors include R factor, FVC, precipitation, and lithology, with feature importance of 0.1223, 0.1199, 0.1177, and 0.1155, respectively. The relatively important factors are those with feature importance of 0.025 ≤ importance value < 0.1, and include land use, slope, soil texture, K factor, distance to river, and soil type. The low-impact factor, which is equivalent to the importance value of <0.025, encompasses the LS factor, TWI, and aspect in this study.

Benggang erosion susceptibility map

The RFSM was selected for the non-benggang points, and the RF model was chosen as the optimal model for predicting BES maps. This study utilized the natural breakpoint approach to classify the BES into five grades: low, relatively low, moderate, relatively high, and high, respectively. The spatial distribution of these grades within Ganxian county is depicted in *Figure 7*. It is imperative to emphasize that areas exhibiting relatively high or high grades necessitate particular attention for the prevention and control of benggang. It is significant that 86.17% of the benggang points were concentrated in these regions. The ratio of projected benggang points to area was utilized to calculate the density of benggang. The densities of the five grades of benggang were 0.02, 0.08, 0.09, 1.11, and 1.34, respectively (*Table 3*). The results show that relatively high and high grades are primarily concentrated in Bailu, Tiancun, Nantang, Sanxi, and

Hujiang towns in northern Ganxian County, as well as Yangbu and Hanfang towns in southern Ganxian County. The monitoring and prevention of benggang in these areas has been identified as a key priority.

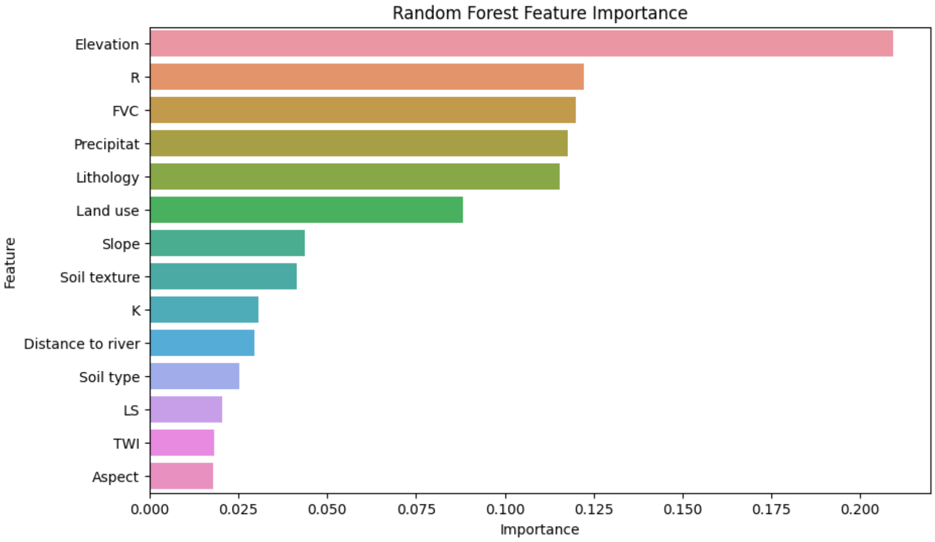

Figure 6. *Feature importance of different environmental factors*

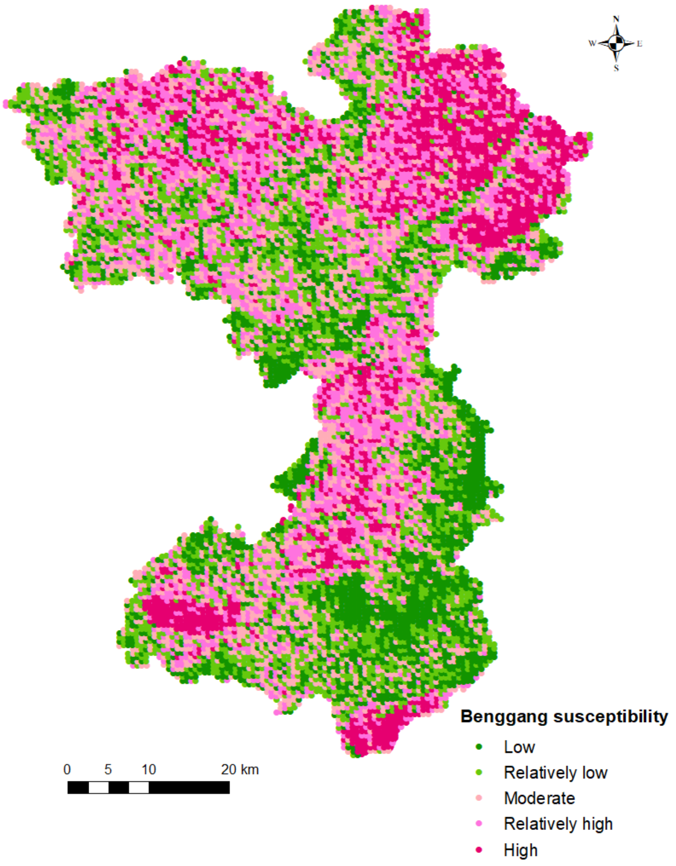

Figure 7. *Benggang erosion susceptibility map*

DOI: http://dx.doi.org/10.15666/aeer/2305_89999016

Table 3. *Percentage of benggang erosion susceptibility in Ganxian county*

Grades of susceptibility	Area (km^2)	Percentage of Area	Predicted number	Percentage of predicted number	Density of benggang (point/km^2)
low	947.80	31.70%	23	2.34%	0.02
Relatively low	734.76	24.57%	60	6.10%	0.08
Moderate	594.51	19.88%	53	5.39%	0.09
Relatively high	480.83	16.08%	536	54.53%	1.11
High	232.21	7.77%	311	31.64%	1.34

Discussion

The three sampling methods to predict benggang erosion susceptibility

The enhancement of the performance of benggang prediction models constitutes a pivotal scientific imperative in the realm of benggang control. The extant methods to enhance model prediction accuracy primarily encompass multi-model coupling, high-precision data acquisition, machine learning optimization, and spatial heterogeneity analysis (Hu et al., 2024; Fagbohun et al., 2024; Shen et al., 2024). It is noteworthy that the field has rarely investigated the potential of enhancing prediction accuracy through the optimization of benggang and non-benggang points dataset composition. This subject merits rigorous investigation.

The accuracy of data-driven models is contingent upon the quality of positive and negative points. The objective of the BES evaluation is to ascertain the risk level of the region, not to predict continuous values, which is a classification problem, and its dependent variable is the benggang and non-benggang points (1=exist, 0=don't exist). The sample dataset for the evaluation of BES comprises benggang points that have occurred in the study area, designated as a positive sample, and the same number of non-benggang points selected as the positive sample, designated as a negative sample. The sampling methods for the non-benggang point dataset have been employed in previous studies using the RSM, HDSM and the frequency ratio methods. In this study, the RFSM was adopted for the non-benggang dataset, and metrics of Accuracy, Precision, Recall, F1-Score, and AUC were found to be significantly superior to traditional sampling methods ($P < 0.01$). The RFSM has been demonstrated to possess the capacity to capture non-linear relationships and demonstrate robustness to noise. It has also been shown to accurately capture the difference between the environmental factors of benggang and non-benggang points. This capability effectively addresses the limitations of traditional sampling methods, thereby optimizing the composition of the non-benggang point dataset, enhancing the accuracy of the prediction model. The present study aims to predict the susceptibility of benggang with a high degree of accuracy (Chen et al., 2022; Berihun et al., 2025).

Feature importance analysis

Benggang erosion formation and development are controlled by environmental variable evolution and human activities (Zhu et al., 2023; Liao et al., 2023). According to previous studies, the environmental variable vary significantly. Lithology and weathering crust significantly influence the development of benggang, with 41.94% of benggang forming in 25.88% of the granite area (Liao et al., 2019). The study in Ganzhou identified the primary driving factors for benggang erosion as being rainfall erosivity, elevation, and land use (Liao et al., 2022). This indicated that a thorough investigation into the driving elements affecting a particular type of benggang erosion could enhance the comprehension of the benggang erosion process (Liu et al., 2024; Laraib et al., 2024).

This study utilizes the RF model to assess the importance features on the susceptibility of benggangs. The RF model is employed to ascertain the degree of influence on the occurrence of benggangs by calculating the characteristic importance of each environmental factor. The study found that elevation emerged as the predominant key factor of BES in the region, with R, FVC, precipitation, and lithology identified as the primary factors. Meanwhile, land use, slope, soil texture, K, distance to river, and soil type were identified as relatively important factors, while LS, TWI and aspect were found to be low impact factors. Existing studies indicated that the benggang number initially increases and subsequently decreases with rising elevation. Simultaneously, benggang is demonstrated to differ from gully erosion in terms of catchment area, lithological requirements, morphology, predominant erosion type, and erosion modulus. The mechanisms of benggang and gully erosion in the red soil region exhibit significant disparities, and the investigation of their respective regional drivers can facilitate the development of effective prevention and control measures (Xia et al., 2021; Liao et al., 2022).

Conclusion

The present study employed three non-benggang points sampling methods and four machine learning models to analyze the sampling effects on model performance of BES. The optimal machine learning model will be utilized to predict BES and to analyze the primary driving factors. The result reveals that the RFSM method demonstrates optimal performance. The results demonstrate that the five metrics of RFSM exhibit significant superiority over the other two sampling methods. Additionally, the RF model was selected as the optimal model. The elevation factor emerged as the predominant key factor of BES in the region, with R, FVC, precipitation, and lithology identified as the main factors. The densities of the five grades of BES were 0.02, 0.08, 0.09, 1.11, and 1.34, respectively. The area with relatively high and high grades is of paramount importance for the prevention and control of benggang.

Acknowledgements. This study was supported by the National Natural Science Foundation of China (42261015), the Science and Technology Foundation of the Education Department of Jiangxi Province (GJJ211426).

REFERENCES

[1] Arabameri, A., Pradhan, B., Lombardo, L. (2019): Comparative assessment using boosted regression trees, binary logistic regression, frequency ratio and numerical risk factor for gully erosion susceptibility modelling. – Catena 183: 104223.

[2] Berihun, M. L., Tsunekawa, A., Haregeweyn, N., Bayabil, H. K., Fenta, A. A., Meshesha, T. M., Kassa, S. B., Bizuneh, B. B., Hailu, Y. B., Vanmaercke, M. (2025): Unveiling gully erosion susceptibility: A semi-quantitative modeling approach integrated with field data in contrasting landscapes and climate regions. – Geomorphology 468: 109485.

[3] Biau, G. (2012): Analysis of a Random Forests Model. – J. Mach. Learn. Res. 13: 1063-1095.

[4] Bouramtane, T., Hilal, H., Rezende-Filho, A. T., Bouramtane, K., Barbiero, L., Abraham, S., Morarech, M. (2022): Mapping gully erosion variability and susceptibility using remote sensing, multivariate statistical analysis, and machine learning in South Mato Grosso, Brazil. – Geosciences 12: 235.

[5] Cai, C., Ding, S., Shi, Z., Huang, L., Zhang, G., (2000): Study of Applying USLE and Geographical Information System IDRISI to Predict Soil Erosion in Small Watershed. – Journal of Soil and Water Conservation 14: 19-24.

[6] Chen, X., Yang, J., Xiao, S., Song, Y., Zheng, H., Shen, L. (2013): Distribution Characteristics and Causes of Collapse Erosion. – Journal of Mountain Science 31: 716-722.

[7] Chen, T., Guestrin, C. (2016): XGBoost: A Scalable Tree Boosting System. – Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, August, pp. 785-794.

[8] Chen, P., Liu, X., Yu, S., Xu, J., Hong, B., Ma, J., Ding, J., Chen, Y., Chen, Y., Lu, C. (2022): Stability assessment of the restored Benggang units in a weathered granite crust region of South China. – Ecological Engineering 182: 106709.

[9] Cristianint, N., Shawe-Taylor, J. (2000): An introduction to support vector machines and other kernel-based learning methods. – Cambridge University Press.

[10] Deng, Y., Cai, C. (2024): Progress of Survey and Monitoring and Control Technology of Benggang Erosion in Red Soil Hilly Area. – Acta Pedologica Sinica 5.

[11] El Jarjini, Y., Morarech, M., Valles, V., Touiouine, A., Touzani, M., Arjdal, Y., Barry, A. A., Barbiero, L. (2023): Surface Formations Salinity Survey in an Estuarine Area of Northern Morocco, by Crossing Satellite Imagery, Discriminant Analysis, and Machine Learning. – Soil Systems 7: 33.

[12] Fagbohun, B. J., Aladejana, O. O., Okonye, I. F., Tobore, A. O. (2024): Assessing gully erosion susceptibility dynamics using information value and hazard index methods: A case study of Agulu-Nanka watershed, Southeast Nigeria. – Catena 241: 108070.

[13] Gao, R., Gao, M., Yao, S., Wen, Y. (2025): Gully erosion susceptibility mapping considering seasonal variations of NDVI using a machine learning approach in the Mollisol region of China. – Soil & Tillage Research 245: 106322.

[14] Guo, F., Wu, D., Wang, X., Dai, Q., Lai, P., Chen, Y., Xia, D. (2023): Susceptibility Assessment of Benggang Based on Random Forests Model and Geodetector in Xingguo County of South Jiangxi. – J. of China Three Gorges Univ. 45: 44-50.

[15] Guo, F., Jiang, G., Huang, X., Wang, X., Xia, D., Chen, Y., Li, X. (2024): Impact of environmental factor combinations and negative sample selection on Benggang susceptibility assessment in granite area. – Transactions of the CSAE 40: 191-200.

[16] He, Y., Huang, Y., Lin, J., Lin, X., Ji, X. (2024): Relationship between benggang erosion and landscape pattern in the southern red soil zone based on path analysis. – Chinese Journal of Applied Ecology 35: 2872-2880.

[17] Hitouri, S., Varasano, A., Mohajane, M., Ijlil, S., Essahlaoui, N., Ali, S. A., Teodoro, A. C. (2022): Hybrid machine learning approach for gully erosion mapping susceptibility at a watershed scale. – ISPRS Int. J. Geo-Inf. 11: 401.

[18] Hu, Y., Qi, Z., Zhou, Z., Qin, Y. (2024): Detection of Benggang in Remote Sensing Imagery through Integration of Segmentation Anything Model with Object-Based Classification. – Romote Sensing 16: 428.

[19] Huang, D., Su, L., Fan, H., Zhou, L., Tian, Y. (2022): Identification of topographic factors for gully erosion susceptibility and their spatial modelling using machine learning in the black soil region of Northeast China. – Ecol Indic 143: 109376.

[20] Ji, X., Huang, Y., Lin, J., Jiang, F., Ge, H. (2019): Spatio-temporal Erosion Features and Prediction for the Erosion Gullies on Collapsing Hills. – Mountai Research 37: 86-97.

[21] Khosravi, K., Rezaie, F., Cooper, J. R., Kalantari, Z., Abolfathi, S., Hatamiafkoueieh, J. (2023): Soil water erosion susceptibility assessment using deep learning algorithms. – Journal of Hydrology 618: 129229.

[22] Kite, G. (2001): Modelling the Mekong: hydrological simulation for environmental impact studies. – Journal of Hydrology 253: 1-13.

[23] Lai, J., Li, J., Liu, L. (2024): Predicting Soil Erosion Using RUSLE and GeoSOS-FLUS Models: A Case Study in Kunming, China. – Forests 15(6): 1039.

[24] Lana, J. C., Castro, P. D., Lana, C. E. (2022): Assessing gully erosion susceptibility and its conditioning factors in southeastern Brazil using machine learning algorithms and bivariate statistical methods: a regional approach. – Geomorphology 402: 108159.

[25] Laraib, S., Xiong, D., Zhao, D., Shrestha, B. R., Liu, L., Qin, X., Xie, X., Rai, D. K., Zhang, W. (2024): Assessment of gully influencing factors and susceptibility using remote sensing-based frequency ratio method in Sunshui River Basin, Southwest China. – Environ Monit Assess. 196: 731.

[26] Li, M., Yin, L., Zhang, Y., Su, X., Liu, G., Wang, X., Au, Y., Wu, X. (2021): Spatio-temporal dynamics of fractional vegetation coverage based on MODIS-EVI and its driving factors in Southwest China. – Acta Ecologica Sinica 41: 1138-1147.

[27] Li, H., Jin, J., Dong, F., Zhang, J., Li, L., Zhang, Y. (2024a): Gully Erosion Susceptibility Prediction Using High-Resolution Data: Evaluation, Comparison, and Improvement of Multiple Machine Learning Models. – Remote Sens. 16: 4742.

[28] Li, J., Tian, Y., Wang, D., Zhang, Q., Tao, J., Zhang, Y., Lin, J. (2024b): Matching and driving mechanism analysis of the supply and demand relationships of soil conservation services in karst peak-cluster depression basin in Southwest Guangxi, China. – Catena 246: 108438.

[29] Liao, Y., Yuan, Z., Zheng, M., Li, D., Nie, X., Wu, X., Huang, B., Xie, Z., Tang, C. (2019): The spatial distribution of Benggang and the factors that influence it. – Land Degradation & Development 30: 2323-2335.

[30] Liao, K., Song, Y., Xie, S., Luo, Y., Liu, Q., Lin, H. (2022): Quantitative Analysis of the Factors Influencing the Spatial Distribution of Benggang Landforms Based on a Geographical Detector. – ISPRS Int. J. Geo-Inf. 11: 337.

[31] Liao, Y., Yuan, Z., Li, D., Zheng, M., Huang, B., Xie, Z., Wu, X., Luo, X. (2023): What kind of gully can develop into benggang? – Catena 225: 107024.

[32] Liu, Z., Wei, Y., Cui, T., Lu, H., Cai, C. (2024): Spatial scaling effects of gully erosion in response to driving factors in southern China. – J. Geogr. Sci. 34: 942-962.

[33] Maalouf, M. (2011): Logistic Regression in Data Analysis: An Overview. – Int. J. Data Anal. Tech. Strateg. 3: 281.

[34] Naceur, H. A., Abdo, H. G., Igmoullan, B., Namous, M., Alshehri, F., Albanai, J. A. (2024): Implementation of random forest, adaptive boosting, and gradient boosting decision trees algorithms for gully erosion susceptibility mapping using remote sensing and GIS. – Environmental Earth Sciences 83: 121.

[35] Nguyen, K. A., Chen, W., Lin, B., Seeboonruang, U. (2021): Comparison of Ensemble Machine Learning Methods for Soil Erosion Pin Measurements. – ISPRS Int. J. Geo-Inf. 10: 42.

[36] Ou, H., Mu, X., Yuan, Z., Yang, X., Liao, Y., Nguyen, K. L., Sombatpanit, S. (2024): Mapping benggang erosion susceptibility: an analysis of environmental influencing factors based on the maxent model. – Sustainability 16: 7328.

[37] Qiu, H., Cui, P., Regmi, A. D., Hu, S., Wang, X., Zhang, Y. (2018): The effects of slope length and slope gradient on the size distributions of loess slides: Field observations and simulations. – Geomorphology 300: 69-76.

[38] Rajbahadur, G. K., Wang, S., Oliva, G. A., Kamei, Y., Hassan, A. E. (2021): The Impact of Feature Importance Methods on the Interpretation of Defect Classifiers. – IEEE Transactions on Software Engineering 48: 2245-2261.

[39] Senanayake, S., Pradhan, B., Alamri, A., Park, H. (2022): A new application of deep neural network (LSTM) and RUSLE models in soil erosion prediction. – Science of the Total Environment 845: 157220.

[40] Shen, S., Chen, J., Cheng, D., Liu, H., Zhang, T. (2024): Benggang segmentation via deep exchanging of digital orthophoto map and digital surface model features. – International Soil and Water Conservation Research 12: 589-599.

[41] Wei, Y., Wu, X., Wang, J., Yu, H., Xia, J., Deng, Y., Zhang, Y., Xiang, Y., Cai, C., Guo, Z. (2021): Identification of geo-environmental factors on Benggang susceptibility and its spatial modelling using comparative data-driven methods. – Soil & Tillage Research 208: 104857.

[42] Williams, J. R. (1990): The erosion-productivity impact calculator (EPIC) model: A case history. – Philos. Trans. Biol. Sci. 329: 421-428.

[43] Xia, J., Cai, C., Wei, Y., Zhou, Y., Gu, J., Xiong, Y., Zhou, X. (2021): Variations of soil hydraulic properties along granitic slopes in Benggang erosion areas. – Journal of Soil and Sediments 21: 1177-1189.

[44] Xu, X. (2025): Construction of ecological security patterns in hilly cities based on morphological spatial pattern analysis and minimum cumulative resistance models: a case study of Ganzhou, China. – Applied Ecology and Environmental Research 23(1): 1475-1496.

[45] Zabihi, M., Mirchooli, F., Motevalli, A., Khaledi Darvishan, A., Pourghasemi, H. R., Zakeri, M. A., Sadighi, F. (2018): Spatial modelling of gully erosion in Mazandaran province, northern Iran. – Catena 161: 1-13.

[46] Zhu, X., Gao, L., Wei, X., Li, T., Shao, M. (2023): Progress and prospect of studies of Benggang erosion in southern China. – Geoderma 438: 116656.