[1]Hongfeng Yu
[2]Mingchang Shi*

Evaluation of spatial distribution and sensitivity of erosion gullies based on random forests in low-hill areas of Northeast China

Journal of Electrical Systems

Abstract: - This study explored in depth the problem of erosion gully in farmland in the black soil zone of Northeast China. As a major food production area in China, the northeastern black soil region has been facing the challenge of increasing erosion gully problems in recent years, which not only affects soil quality and ecological environment, but also poses a threat to food security. By using high-resolution remote sensing images, the spatial and temporal distribution and morphological characteristics of erosion gullies in the region were comprehensively analysed. In terms of spatial and temporal distribution, comparison of data from 1968 to 2020 reveals that erosion gullies have increased significantly in number, area and length, reflecting the severity of the soil erosion problem. In terms of spatial distribution, the erosion gullies showed a significant aggregation effect, and this aggregation was enhanced over time. In addition, through the cold hotspot analysis, the hotspot areas of erosion gully activities were successfully identified, which provided a scientific basis for the development of prevention and control measures. In the analysis of influencing factors for the formation and development of erosion gullies, natural and anthropogenic factors, including topography, precipitation, soil and human activities, were considered comprehensively, and a comprehensive database of influencing factors was constructed. Logistic regression model and random forest model were introduced for assessment. Comparison of model performance revealed that the random forest model demonstrated higher accuracy in prediction. Therefore, a sensitivity map was produced based on the random forest model and the predicted probabilities were classified into four levels: low, medium, high, and very high sensitivity zones. The results show that the proportion of each sensitive area is relatively balanced, which provides strong support for the subsequent prediction of erosion gully risk and the development of prevention and control measures.

Keywords: Environment, Comparison, Significant, Demonstrated, Subsequent.

1 INTRODUCTION

In recent years, the study of temporal and spatial distribution of erosion gully has become the focus of academic attention. An in-depth study of the temporal and spatial characteristics of erosion gullies can not only provide scientific basis for soil and water conservation planning, but also provide important reference for formulating effective soil and water conservation measures. According to research, erosion gullies mainly occur in topographic watershed areas, especially in areas that have suffered serious degradation[1]. The formation and distribution of erosion gullies are closely related to the regional rock properties and hydrological conditions, especially in the areas more sensitive to these factors, the development of erosion gullies is more significant [2].The formation and evolution of erosion gully is a complex geological process, which is influenced by many natural factors[3]. In this process, various factors interact to determine the shape, scale and development rate of erosion gullies, and the distribution characteristics of erosion gullies show non-uniformity in different spatial scales[4,5]. It is found that there is a significant correlation between the density of erosion gully and the variation of slope and slope direction

[1] School of Soil and Water Conservation ,Beijing Forestry University, Beijing 100083, Beijing , China.
[1*] School of Soil and Water Conservation ,Beijing Forestry University, Beijing 100083, Beijing , China.Email:wenzikuaifei@126.com

of terrain, and topographic factors play a key role in the formation of erosion gully[6-8]. Specifically, the density of the erosion trench increases with the increase of the slope, which is particularly obvious in the area with higher slope[9].

In addition, climatic factors, especially rainfall and rainfall intensity, have important effects on the development of erosion gullies[10,11,12,13]. Rainfall not only affects the formation and accumulation of surface runoff, but also changes the erosion resistance of soil[14]. Through the analysis of multi-temporal satellite image data, researchers can effectively monitor the dynamic changes of erosion gullies, and build erosion gully sensitivity assessment models based on terrain, soil[15,16], vegetation and other data to provide scientific support for soil and water conservation planning[17,18,19].

In summary, this study comprehensively investigated multiple natural and human factors, including terrain, precipitation, soil and human activities, and built a comprehensive database of influencing factors[20,21]. Through linear regression and correlation analysis, the correlation between each factor and the development of erosion trench was analyzed. The accuracy of Logistic regression model and random forest model in predicting the sensitivity of erosion gully was evaluated.

2 MATERIALS AND METHODS

2.1 Study area

The study area is located in the south of Dongliao County, Liaoyuan City, Jilin Province, with coordinates ranging from 124°57‘~125°8’E and 42°36‘~42°51’N (Fig. 1), covering an area of about 223.6 km2, with elevations ranging from 287 to 603 m, and with obvious changes in topography. The study area is situated in the northeastern low-mountainous hilly area, with complex and varied topography, which is conducive to the formation and development of erosional gullies. The climate of the study area is temperate monsoon climate, and the average precipitation is 662mm per year, which is concentrated in June-September, and reaches the peak in July. Strong precipitation events increase surface runoff in a short time, accelerate the process of soil erosion, and then promote the formation and expansion of erosion gullies. The study area is dominated by Quaternary loose sediments and Paleozoic to Cenozoic sedimentary rock formations with complex rock types, including sandstone, shale and volcanic rocks. This geological background promotes a diversity of soils, with the main soil types being brown and sandy brown soils, which are susceptible to erosion under certain climatic conditions. This topographic diversity provides the physical basis for the formation of erosion gullies. The river system in the study area is dominated by the Dongliao River and its tributaries, which not only provide important local water resources, but also play a key role in terrain formation and erosion. The erosive action of water flow deepens the river valleys and promotes the development of erosion gullies along the rivers. Combined with the topographic relief, geological structure, soil type and climatic conditions of the area, the formation and development of erosion gullies in the study area show a complex dynamic process.

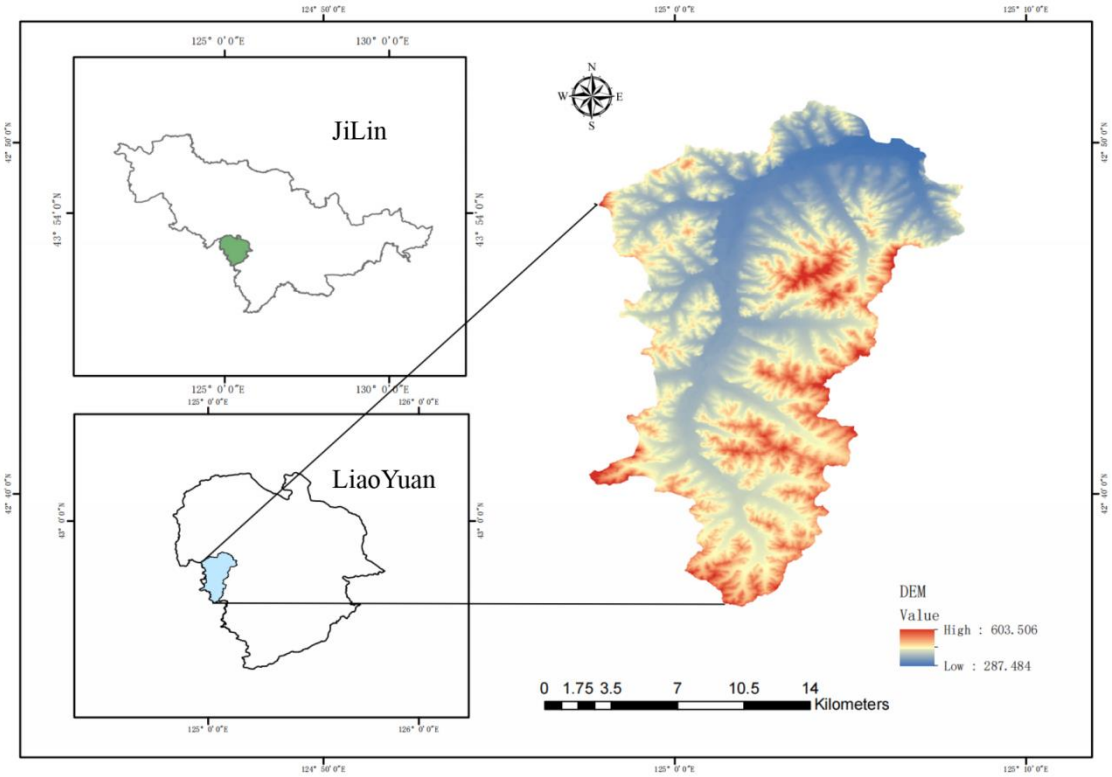


Figure 2.1 Overview of the study area

2.2 Erosion gully distribution and change

Taking the erosion gully in typical low mountain hilly area of the Northeast Black Soil Region as the research object, the data of erosion gully channels were visually deciphered from remote sensing images at three time points of 1968, 2010 and 2020 using ArcGIS 10.2 software to construct the data system. And the quantitative analysis of the indicators of the basic morphological parameters of erosion gully (erosion gully area, length, density and severity) was calculated and analysed based on the DEM-divided sub-watersheds as a unit.

2.2.1 Spatial autocorrelation calculations

Spatial autocorrelation is a more commonly used method to assess the spatial aggregation pattern, and this analysis method effectively reflects the degree of spatial correlation between the value of an attribute of a spatial unit and its neighbouring units [22] . In this paper, Moran's I was chosen to assess the global spatial autocorrelation of erosion gully data [23], and the formula is as follows.

$$\text{Moran's I} = \frac{n}{s_0}\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j}z_i z_j}{\sum_{i=1}^{n} z_i^2} \quad (2.1)$$

where z_i is the deviation of an element attribute of i from its mean value $(x_i - \bar{X})$, $w_{i,j}$ is the spatial weight between elements i and j, n is the total number of elements, and s_0 is the set of all spatial weights.

$$s_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} W_{i,j} \quad (2.2)$$

The final tally of Z scores:

$$z_i = \frac{I - E[I]}{\sqrt{V[i]}} \quad (2.3)$$

2.2.2 Standard deviation of ellipse

Erosion gullies generally show a relatively obvious directional distribution in geographical space, and the analysis of the directional characteristics of erosion gullies in the study area can understand the overall development trend of erosion gullies from a macro perspective. In this study, the method of standard deviation ellipse [24]was adopted, which firstly determined a circle center and calculated by means of arithmetic average. After determining the center coordinates of the ellipse (SDEx, SDEy), the square of the fitting ellipse is further calculated, the true north direction is 0°, and the length of the X and Y axes of the ellipse is finally determined (equation 2.4-2.6).

the calculation formula is as follows：

$$\tan\theta = \frac{\sum_{i=1}^{n} \Delta x_i^2 - \sum_{i=1}^{n} \Delta y_i^2 + \sqrt{(\sum_{i=1}^{n} \Delta x_i^2 - \sum_{i=1}^{n} \Delta y_i^2) + 4(\sum_{i=1}^{n} \Delta x_i \Delta y_i)^2}}{2\sum_{i=1}^{n} \Delta x_i \Delta y_i} \quad (2.4)$$

$$\varphi_x = \sqrt{\frac{1}{n}\left[\sum_{i=1}^{n} (\Delta x_i \cos\theta - \Delta y_i \sin\theta)^2\right]} \quad (2.5)$$

$$\varphi_x = \sqrt{\frac{1}{n}\left[\sum_{i=1}^{n} (\Delta x_i \cos\theta - \Delta y_i \sin\theta)^2\right]} \quad (2.6)$$

Where: φ_x and φ_y represent the standard deviation of the X and Y axes respectively, Δx_i and Δy_i represent the deviation of the coordinate point of each point element from its mean center respectively, θ represents the Angle of rotation of the ellipse, and n represents the total number of risk units.

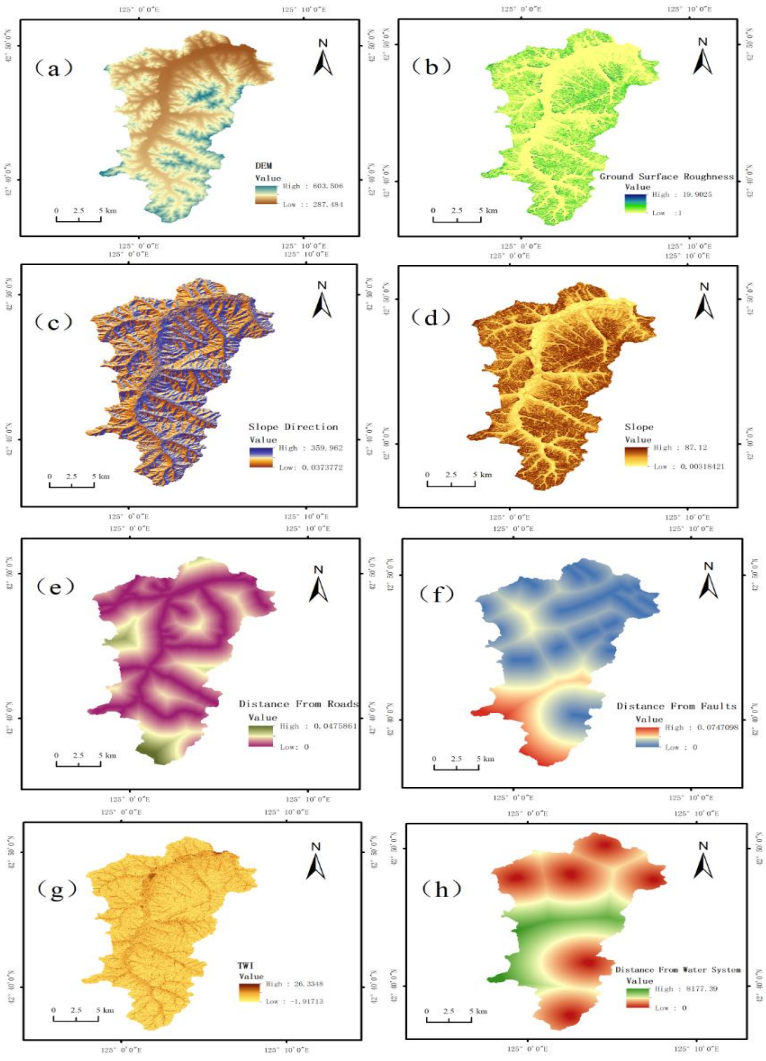
In this study, the directional distribution tool integrated in ArcGIS 10.2 software was used to determine the overall trend of the distribution of erosion gully. However, the calculated inclination of the fitted ellipse ranges from 0 to 180°.

2.3 Influencing factors and data sources

Gully erosion is a complex process driven by both natural and human factor[25]. Therefore, reasonable selection of influencing factors will improve the performance of the model. Based on the influence factors of previous studies and the characteristics of the northeast black soil area, the factors affecting gully erosion mainly include topography, hydrology, soil, vegetation, geology and human activities. Topographic factors affect gully erosion by influencing runoff generation and accumulation[26,27,28]. Topographic factors include elevation, slope, aspect, topographic wetness index (TWI), ground roughness, hydrodynamic index (SPI), plane curvature and profile curvature. They are calculated based on a 3.2m resolution digital elevation model (DEM) produced by GF-7 images. Euclidean distance tool in ArcGIS is used to calculate the distance to water system, road and fault. Hydrological factors include precipitation erosivity and distance from water system. Precipitation is the main source of runoff and the driving force of gully erosion[29,30]. In this study, the average annual precipitation from

1968 to 2020 was selected to calculate the rainfall erosivity. The data are from "China Surface Climatological Data Daily Dataset (V30)", provided by the National Data Center for Meteorological Sciences (http://data.cma.cn/). Soil factors, including soil erodibility factors, were calculated using the HSWD China Soil data set provided by the Institute of Soil Science, Chinese Academy of Sciences, Nanjing. Vegetation factors include NDVI. NDVI data is derived from monthly NDVI raster data from 2000-2022 of the MOD13A3 dataset, with a spatial resolution of 1km*1km. Factors of human activity include land use and distance from roads. Land use affects gully development by influencing material erodibility and surface runoff. So using 1985-2020, wuhan university CLCD land cover classification data set (https:/doi.org 10.5281 / zenodo. 5816591), the spatial resolution of 30 m * 30 m; ArcGIS was used to calculate the distance between the erosion trench and the road. The road data were derived from the vector road network data of OpenStreetMap(OSM) (www.openstreetmap.org), and the Euclidene distance tool was used to calculate the distance. Geological factors include distance from faults. The distance from faults is calculated using the national 1：200000 digital geological map spatial database (http://dcc.cgs.gov.cn) and Euclidean distance tool.

All data processing results are shown in Figure 2.2



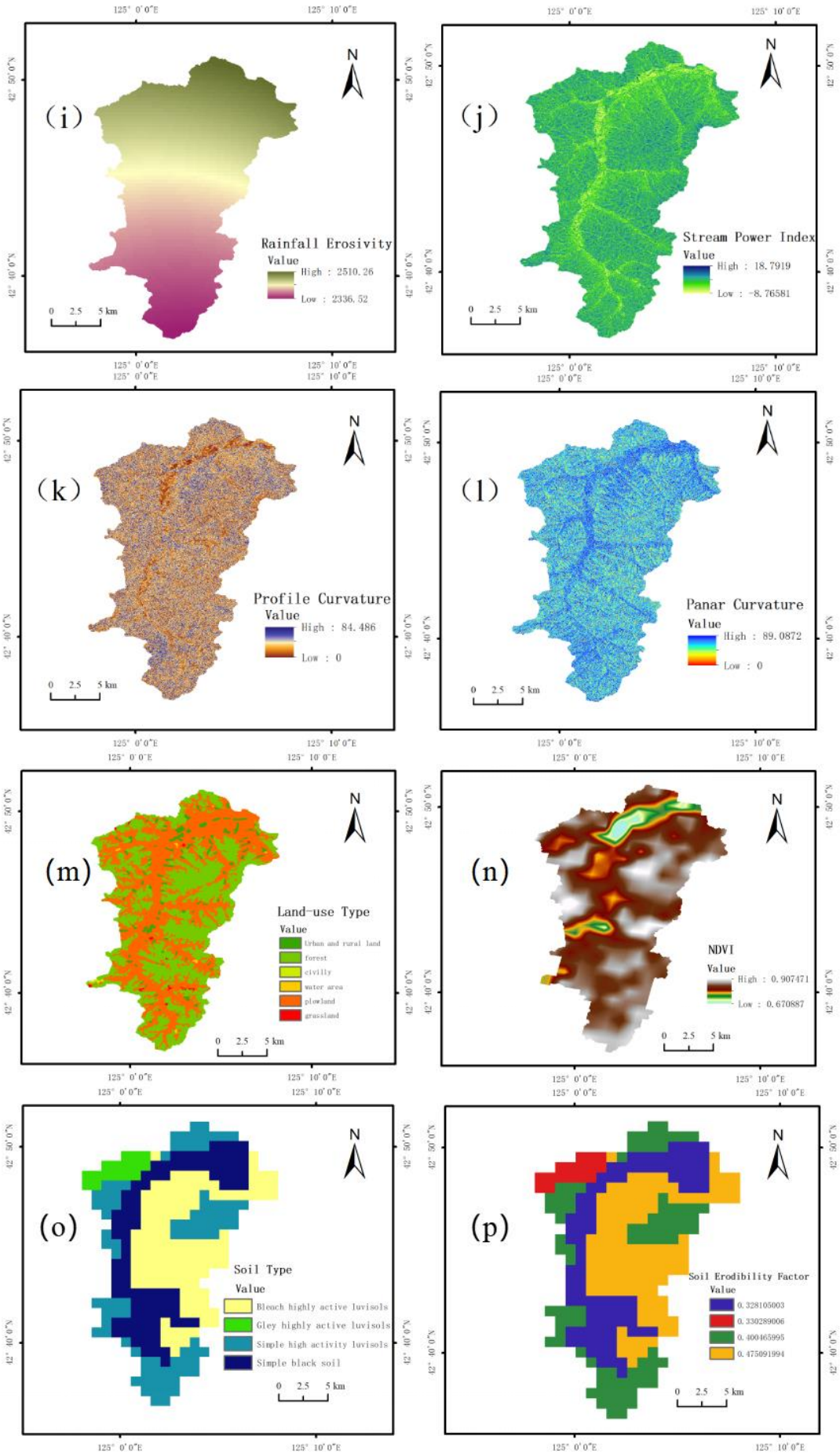

(i)
Rainfall Erosivity
Value
High : 2510.26
Low : 2336.52
(j)
Stream Power Index
Value
High : 18.7919
Low : -8.76581
(k)
Profile Curvature
Value
High : 84.486
Low : 0
(l)
Panar Curvature
Value
High : 89.0872
Low : 0
(m)
Land-use Type
Value
Urban and rural land
forest
civilly
water area
plowland
grassland
(n)
NDVI
Value
High : 0.907471
Low : 0.670887
(o)
Soil Type
Value
Bleach highly active luvisols
Gley highly active luvisols
Simple high activity luvisols
Simple black soil
(p)
Soil Erodibility Factor
Value
0.328105003
0.330289006
0.400465995
0.475091994
N
0 2.5 5 km
125° 0'0"E
125° 10'0"E

Figure 2.2 Map of base data: (a) DEM, (b) ground surface roughness, (c) slope direction, (d) slope, (e) distance from roads, (f) distance from faults, (g) TWI (h) distance from water system, (i) rainfall erosivity, (j) SPI, (k) profile curvature, (l) planar curvature, (m) land-use type, (n) NDVI , (o) soil type, (p) soil erodibility factor

2.4 Multi-collinearity analysis

The occurrence of erosion gullies is usually influenced by a variety of factors, and when there are too many influences and the input datasets are highly correlated, the problem of multiple covariance occurs, which may lead to some errors in the prediction process. Prediction of disturbance sensitivity to erosion gullies. The screening of erosion gully influencing factors through the test of covariance, which is an important part of statistical analysis, was carried out using variance inflation factor (VIF) and tolerance (TOL). The formulae are as follows:

$$\mathrm{VIF} = \frac{1}{1 - \mathrm{R}_j^2} \qquad (2.7)$$

Where R^2 is the coefficient of complex determination of the regression on the other independent variables when thought to be the dependent variable. When R is smaller, the correlation between the variables is higher and the VIF is greater.

VIF is concerned with the change in the standard error of the erosion gully impact factor, i.e. the lower the standard error, the lower the risk of multicollinearity. If the VIF value is greater than 10 or the tolerance value is less than 0.1, it indicates that there is a multicollinearity problem between the influencing factors. Selecting influencing factors with low VIF or high TOL values can effectively reduce or even avoid model prediction errors Some statistical software uses TOL instead of VIF, where the tolerance is:

$$\mathrm{TOL} = \frac{1}{\mathrm{VIF}} \qquad (2.8)$$

Multi-collinearity analysis is one of the commonly used factor selection methods to evaluate the "non-independence" of channel erosion induced factors. The main reason is that there is a strong correlation between variables, which leads to inaccurate test results and unreliable prediction results. Multicollinearity in a data set is referred to as a linear relationship between two or more erosion gully influence factors .Tolerance (TOL) and variance inflation factor (VIF) values <0.1 and >10, respectively, indicate good multicollinearity between variables in a data set.

2.5 Model Construction

2.5.1 Random Forest

Random forest (RF) algorithm, as a powerful supervised learning algorithm, is used to solve classification and regression problems. Based on the principle of ensemble learning, this method enhances the prediction ability of the model by integrating the output of multiple decision trees. For classification problems, the random forest model integrates the classification results of multiple decision trees, and takes the most frequent categories as the overall output, following the majority voting principle. In contrast, for regression problems, random forest

algorithms summarize the predicted values of all individual decision trees and calculate their average values as the final prediction result. This method ensures that the model's predictions are consistent and robust across different decision trees. This way of integrating the output of multiple decision trees not only improves the stability and accuracy of the model, but also makes the random forest algorithm have a wide application prospect in the field of data mining and machine learning. The structure and decision mechanism of the random forest model are as follows:

In the process of building a random forest, samples are selected randomly at first. Specifically, N samples were selected from the original training set containing M samples through the sampling method with put back (i.e. bootstrap sampling). Since samples may be selected repeatedly or unselected in the sampling process, the sample sets obtained each time are different, and these sets are used to train each decision tree respectively. Secondly, random selection of features is carried out. Assuming that each sample has K feature attributes, the algorithm will randomly select k (k<K) features among these features. Based on these k features, the most optimized segmentation attributes are selected to construct the CART decision tree. In the whole process of decision tree construction, the number of selected features k remains fixed. Finally, m fully grown and unpruned decision trees are constructed by repeating the above steps of random selection of samples and features. These collections of decision trees form what is called a random forest. In the regression problem, the final result is obtained by averaging the predicted values of each tree. This integrated approach enhanced by randomness and diversity significantly improves the predictive performance and generalization ability of the model. Figure 2.3 shows.

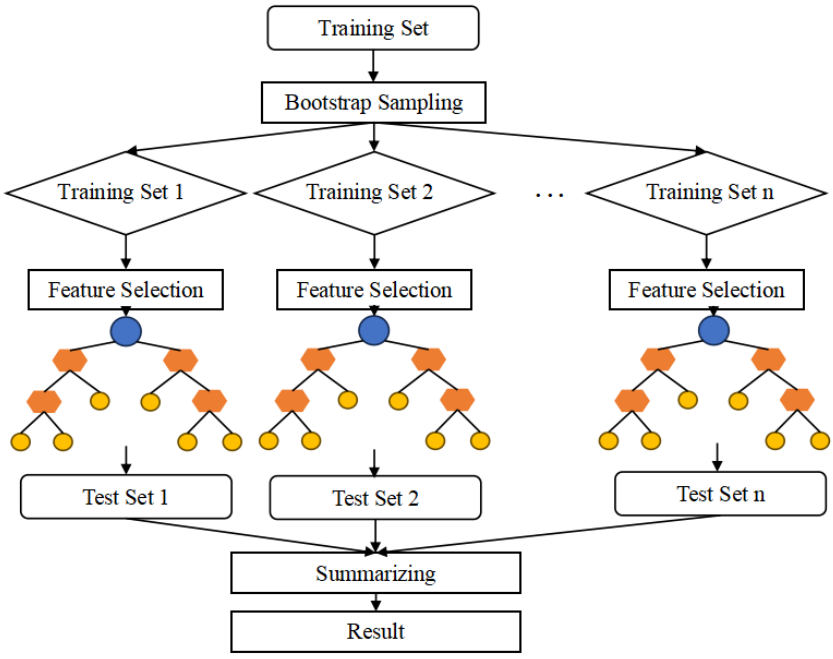


Figure 2.3 Schematic Diagram of Random Forest Model Structure and Decision Mechanism

2.5.2 Logistic Regression

In the study, y_i^* is set as a continuous potential dependent variable, which is usually used as an internal indicator to measure whether an event has occurred. This indicator is often used to indicate the occurrence or non-

occurrence of an event, and its range extends from negative infinity to positive infinity [96]. At the same time, y_i is defined as a binary variable to directly reflect the occurrence of an event, and a constant value threshold $a = 0$ is set as the limit to judge whether the event has occurred. When $y_i^* > 0$, the event is considered to have occurred, and $y_i = 1$; If $y_i^* \leq 0$, the event does not occur, in which case$y_i = 0$.

Assume the following linear relationship between y_i^* and x_i (Equation 2.9).

$$y_i^* = \alpha + \beta x_i + \varepsilon_i \quad (2.9)$$

The

deformation results in formula (2.10)

$$P(y_i = 1|x_i) = P[(\alpha + \beta x_i + \varepsilon_i) > 0] = P[\varepsilon_i > (-\alpha - \beta x_i)] \quad (2.10)$$

Where: P stands for probability.

It is usually assumed that the error term ε follows the Logistic distribution. Since the Logistic distribution has symmetry, the equation (2.11) is obtained by changing the direction of the inequality.

$$P(y_i = 1|x_i) = P[\varepsilon_i \leq (\alpha + \beta x_i)] = F(\alpha + \beta x_i) \quad (2.11)$$

If F in equation (2.11) is the cumulative distribution function of Logistic distribution, then the model is a Logistic regression model.

2.6 Accuracy evaluation

2.6.1 Confusion matrix

Confusion Matrix is a common tool for evaluating the performance of a classification model, which can show the classification ability of the model for different classes. The confusion matrix is a square matrix with rows representing the true class, columns representing the class predicted by the model, and elements in the matrix representing the number of samples[31,32].According to the classification task, the real label situation and the model predicted the label situation are divided into four types:

Figure 2.1 Confusion matrix

True Value	**Predicted Value**	
	1	**0**
1	TP	FN
0	FP	TN

Where true positive TP represents the number of samples that are actually positive and correctly predicted by the model; False negative FN indicates that it is actually a positive class. However, the number of samples incorrectly predicted by the model to be negative; False positive FP indicates the number of samples that are actually negative, but the model incorrectly predicts positive; True negative TN: represents the number of samples that are actually

negative and correctly predicted by the model to be negative.

Through the confusion matrix, some important classification performance indicators can be calculated, such as Accuracy (A), Precision (P), Recall (R), F1-score, AUC, etc. These metrics help to fully evaluate the performance of the model.

$$A = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.12)$$

$$P = \frac{TP}{TP + FP} \quad (2.13)$$

$$R = \frac{TP}{TP + FN} \quad (2.14)$$

$$F1\text{-score} = \frac{2 \times P \times R}{P + R} = \frac{2 \times TP}{\text{Sample count}+TP-TN} \quad (2.15)$$

2.6.2 ROC curve

ROC stands for receiver operating characteristic curve. Is a two-dimensional plane curve with the false positive rate FPR as the horizontal coordinate and the true rate TPR as the vertical coordinate.

In order to effectively evaluate and compare the ROC Curve performance of different models, the index AUC (Area Under the Curve) is introduced, which represents the closed area between the ROC curve and the axis. Ideally, a good model should have an ROC curve above the diagonal and AUC values in the range of 0.5 to 1. When the AUC value is in the range of 0.5 to 0.7, the accuracy of the model is low. An AUC value between 0.7 and 0.9 indicates that the model has average accuracy and good performance. An AUC value of more than 0.9 indicates that the model has high accuracy and good performance. As the AUC value increases and approaches 1, the classification performance of the model is considered to be more efficient. Therefore, the AUC value is an important basis to judge the reliability of the model.

$$TPR = \frac{TP}{(TP + FN)} \quad (2.16)$$

$$FPR = \frac{FP}{(TN + FP)} \quad (2.17)$$

3 RESULTS AND DISCUSSION

3.1 Spatial distribution of erosion gullies

3.1.1 Spatial distribution changes of erosion gully

Based on the analysis of the number of erosion gullies at three time nodes in 1968, 2010 and 2020 in 19 small basins, the number of erosion gullies in 1968 was concentrated in the range of 9~92 (average: 31.21). In 2010, the number of erosion gullies in small basins ranged from 15 to 113 (average: 40.05), and in 2020, the number of erosion gullies in small basins ranged from 18 to 127 (average: 46.58). In the past 50 years, the change rate of

erosion gullies ranged from 0.17 to 0.66m/year(average: 0.03m/y), and the increase rate of erosion gullies from 1968 to 2020 was 38.12% to 100% (average: 49.24%).

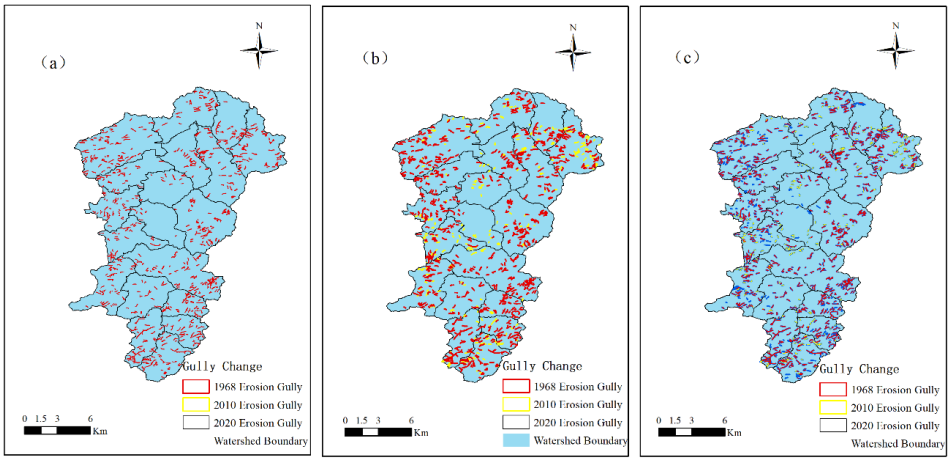


Figure 3.1 Spatial changes in erosion gullies in different watersheds, 1968-2020

3.1.2 Spatial autocorrelation analysis of erosion gully distribution

For the calculation of the Moran index, the spatial autocorrelation in the spatial statistical tool in ArcGlS 10.2 software was adopted, which could calculate the Moran index based on the distribution data of erosion gullies and analyze its confidence interval. Table 3.6 and Table 3.7 show the calculation results of global Moran index of gully density and erosion gully fragmentation in a small watershed, respectively. For this small watershed, the Moran's I values of gully density and erosion gully fragmentation from 1968 to 2020 are significantly greater than 0 ($Z>4.23$, $P<0.001$). In the past 50 years, the spatial distribution and dynamic changes of erosion gullies have significant positive correlation and spatial aggregation effect, and the spatial aggregation effect of erosion gullies has a trend of gradually increasing, and the spatial layout of erosion gullies tends to converge in a larger range. The spatial correlation is increasing.

Table 3.1 Gully density in the study area: global Moran's I

Project	**1968**	**2010**	**2020**
Moran's I	0.369	0.365	0.408
Z	5.033	5.034	5.578
P	<0.001	<0.001	<0.001

Table 3.2 Overall fracture degree of erosion ditch in the study area Moran's I

Project	**1968**	**2010**	**2020**
Moran's I	0.358	0.306	0.362
Z	4.863	4.231	4.974
P	<0.001	<0.001	<0.001

3.1.3 Center of gravity migration analysis of erosion trench

In this study, the gravity center-standard deviation ellipse method was used to deeply analyze the spatial dynamic variation process of the erosion trench length in the study area (refer to Figure 3.2). Between 1968 and 2020, the ellipse of standard deviation in the basin shows a pattern of distribution from northeast to southwest. From the perspective of the covered area of the standard deviation ellipse, the area of the ellipse shows a decreasing trend of fluctuation, and the total covered area is reduced by about 1.414km². In terms of the variation of the long and short semi-axis, the length of the short semi-axis increased slightly from 1968 to 2010, by about 28.544m, and then decreased slightly from 2010 to 2020, reflecting that the influence of the erosion trench length on the shape of the standard deviation ellipse was first enhanced and then slightly weakened in the northwest and southeast. The length of the main shaft decreased by 162.534m from 1968 to 2010, and then recovered somewhat from 2010 to 2020, but still did not recover to the initial level, which indicates that the influence of the erosion trench length has also undergone a "weakening - strengthening" change process in the direction of the main axis from northeast to southwest. From the point of view of the gravity center of the development of the length of the erosion trench, its coordinates vary between longitude 125°2 '31 "and 125°2' 37" and latitude 42°44 '16 "and 42°44' 22", and this area is relatively stable. From 1968 to 2010, the center of gravity shifted to the northeast, reaching a distance of 188.690m. Subsequently, between 2010 and 2020, the center of gravity shifted to the southwest again, by about 185.692m. From the perspective of the variation of the azimuth of the standard deviation ellipse, the azimuth of the standard deviation ellipse showed a trend of clockwise rotation from 172.782° to 175.296° during 1968 to 2010, indicating that the influence of the length of erosion trench in the southern region on the shape of the ellipse gradually exceeded that in the northern region during this stage. However, between 2010 and 2020, the azimuth rotated counterclockwise, indicating an increase in the length of the erosion trench in the northern region.

Table 3.3 Erosion channel distribution standard deviation elliptic parameter

Year	Ellipse Area /km²	Areal Coordinates（E，N）	Minor Semi-axis /m	Major Semi-axis /m	Azimuthal Angle /°	Displacement Distance /m
1968	136.826	125° 2′ 36″,42° 44′ 16″	4405.960	9886.112	172.782	——
2010	135.448	125° 2′ 37″,42° 44′ 22″	4434.504	9723.578	175.296	188.690
2020	135.698	125° 2′ 31″,42° 44′ 17″	4420.817	9771.658	174.227	185.692

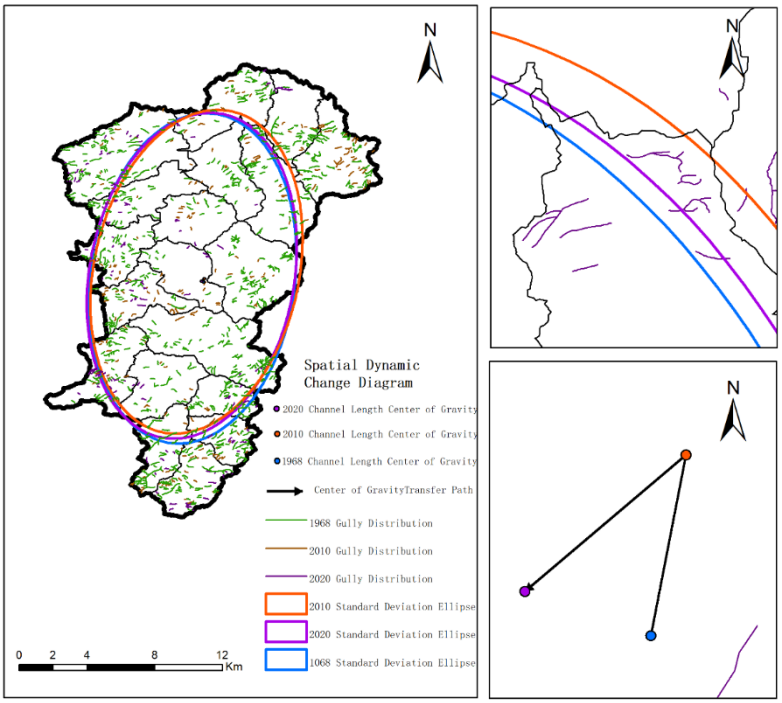


Figure 3.2 Spatial dynamic variation process of erosion trench length

3.2 Multi-collinearity analysis and variable importance analysis

Multicollinearity analysis is one of the commonly used factor selection methods to evaluate the "non-independence" of channel erosion induced factors. The main reason is that there is a strong correlation between variables, which leads to inaccurate test results and unreliable prediction results. Multicollinearity in a dataset is referred to as a linear relationship between two or more gully erosion regulating variables. Tolerance (TOL) and variance inflation factor (VIF) values <0.1 and >10, respectively, indicate good multicollinearity between variables in a dataset. Multicollinearity results (Table 3.4). The results show that all variables remain within the thresholds of tol and vif, which is suitable for GES modeling and evaluation.

Table 3.4 Multicollinearity values of 16 kinds of erosion gullies

Influence Factor	VIF	TOL
NDVI	1.183	0.845
TWI	4.021	0.249
Slope Direction	1.018	0.982
SPI	3.582	0.279
Slope	6.125	0.163
Soil Type	1.228	0.814
Rainfall Erosivity	3.518	0.284
Land-use Type	1.094	0.914
Ground Surface Roughness	2.451	0.408
Distance From Roads	1.412	0.708

Soil Erodibility Factor	1.227	0.815
Distance From Faults	2.472	0.404
DEM	2.607	0.384
Planar Curvature	1.172	0.853
Profile Curvature	1.188	0.842
Distance From Water System	1.145	0.873

16 factors including slope, slope direction, elevation, profile curvature, plane curvature, TWI (topographic wetness index), SPI (stream power index), ground roughness, erosive force, distance from road, distance from water system, distance from fault, land use change, NDVI, soil type and soil erodibility factor were calculated Importance degree and normalization of importance degree. The weights of evaluation factors are shown in Table 3.5.

Table 3.5 Factor weights of random forest evaluation

Influence Factor	**weight**
DEM	0.095
Distance From Roads	0.087
Ground Surface Roughness	0.085
Slope	0.080
SPI	0.078
Distance From Water System	0.077
Rainfall Erosivity	0.076
TWI	0.069
Distance From Faults	0.066
Profile Curvature	0.062
Planar Curvature	0.058
Slope Direction	0.056
NDVI	0.053
Soil Type	0.026
Soil Erodibility Factor	0.023
Land-use Type	0.015

The results show that, among many influencing factors, the weight value of elevation is more prominent, reaching 0.095, close to 0.1, indicating its important position in the model. Distance from the road, ground roughness, slope and SPI also have high weights, and their values reach 0.087, 0.085, 0.080 and 0.078 respectively, all of which

are close to or above 0.075, indicating that the influence of these factors on the results cannot be ignored. The weight values of distance from water system, rainfall erosivity and TWI are 0.077, 0.076 and 0.069 respectively, which decrease slightly but still maintain a high influence. The weight of the three factors, distance from the fault, profile curvature and plane curvature, is concentrated in the range of 0.06 to 0.066, showing a moderate influence. In contrast, the weights of NDVI, slope aspect, soil type and soil erodibility factors were lower, but still exceeded 0.02, indicating that they also contributed to the model results. The weight of land use type is the lowest, 0.015, but it is still included in the evaluation system, which proves its indispensability in the comprehensive evaluation. To sum up, these 16 impact factors are all given weight, regardless of their size, which proves their respective value and significance in the evaluation system. However, for the accuracy of model evaluation, land use type, soil type, soil erodibility factor and NDVI factor were excluded, and the remaining 12 influencing factors were used for modeling.

3.3 Model construction and performance evaluation

In this study, five evaluation indicators were used to validate two models (Random Forest and Logistic Regression) in order to assess their predictive power. In the training stage, the AUC value of the training set is used as the basis to evaluate the training effect of the model. In the model verification stage, the predictive ability of the model is evaluated by testing the AUC value of the data set. Through this comprehensive evaluation method, the performance of the model can be analyzed and understood comprehensively and deeply.

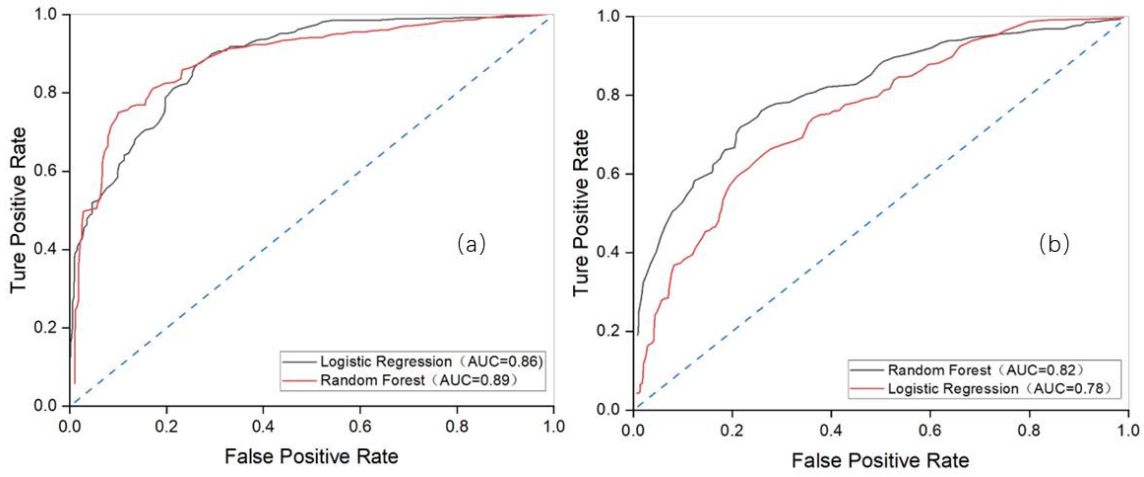


Figure 3.3 ROC and AUC values of Logistic regression and random forest model （a）training set and （b）validation set

As shown in Figure 3.3, the AUC value of the random forest regression model is 0.86 and that of the Logistic regression model is 0.89 on the training data set. On the test dataset, the AUC value of the random forest regression model is 0.78, and the AUC value of the Logistic regression model is 0.82. After comparison, it can be observed that the random forest regression model performs better than the Logistic regression model in AUC values on training and test data sets, showing a better fitting effect. In addition, in order to evaluate the quality of the training data set, five indexes calculated from the confusion matrix were used for evaluation. Detailed results are shown in Table 3.6.

Table 3.6 Accuracy index evaluation

Model	Evaluation Index				
	A	P	R	F1-score	AUC
Random Forest	0.85	0.83	0.88	0.85	0.82
Logistic Regression	0.79	0.78	0.82	0.80	0.78

Table 3.7 Confusion matrix of random forest and Logistic test set

Pedicted Value	True Value		
	Random Forest /Logistic Regression		
	1	0	Class error
1	327/304	46/69	0.12/0.18
0	65/87	308/286	0.17/0.23

Compared with the Logistic regression model, random forest regression showed more prominent prediction performance. Therefore, the random forest model was used to map and analyze the sensitivity of erosion ditch in the northeast black soil region.

3.4 Sensitivity mapping of erosion ditch

The erosion channel sensitivity mapping in the study area is shown in Figure 5.5, and the classification of erosion channel sensitivity mapping is shown in Table 5.5. In this paper, the probability of erosion gully occurrence is between 0 and 1 when drawing the sensitivity evaluation map of erosion gully in the study area. The prediction probability of each grid obtained from the random forest model is divided into four levels: 0-0.25, 0.25-0.50, 0.50-0.75, 0.75-1.00, and the sensitivity level corresponds to low, medium, high, and extremely high. Based on the random forest, the erosion channel sensitivity evaluation map of the study area is drawn, as shown in Figure 3.8.

Table 3.8 Sensitivity mapping classification of erosion gully

Classification	Grid number	Area/km^2	Proportion
Low	6067332	60.20	26.93
Medium	5547846	55.05	24.63
High	5300344	52.59	23.53
Extremely High	5610660	55.67	24.91

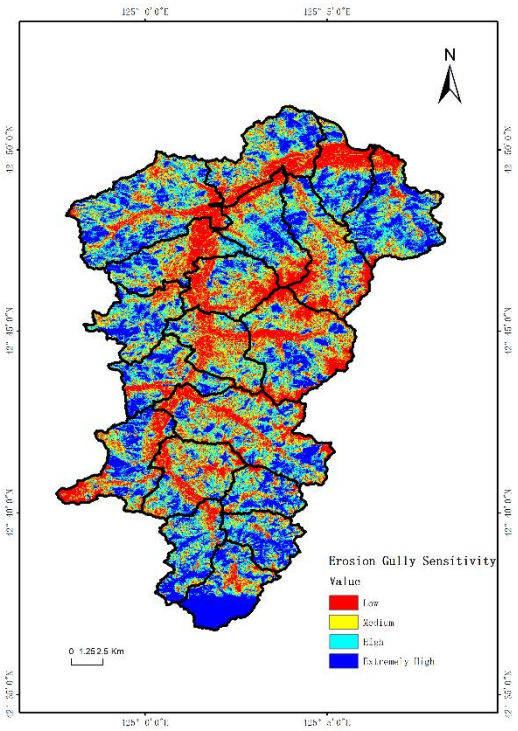


Figure 3.5 Sensitivity distribution of erosion gully

According to ArcGIS regional statistics, in the sensitivity distribution map of random forest, the area of extremely sensitive area was 52.58 km^2, accounting for 24.91%. The area of highly sensitive area was 47.38km^2, accounting for 23.53%. The area of the sensitive area was 51.93 km^2, accounting for 24.63%. The area of low sensitive area was 53.63km^2, accounting for 26.93%. The area of extremely sensitive areas and highly sensitive areas distributed in the surrounding small watershed accounted for a relatively high proportion of the total area of small watershed, and the area proportion of extremely sensitive areas was generally consistent with the development trend of erosion ditch cleavage. The small watershed located in the middle of the basin has lower sensitivity to erosion channel. The extremely sensitive areas and the highly sensitive areas are mainly distributed in the medium-high altitude distribution areas. The future development of erosion gullies can be predicted and effective measures can be reasonably laid according to the sensitivity distribution map.

Table 3.9 Erosion area and sensitivity distribution characteristics of each small watershed

Watershed Number	Watershed Area km^2	Cut Degree of Erosion Ditch	Erosion Gully area (ha)	Low Sensitive Area/km^2	Medium Sensitive Area/ km^2	High Sensitive Area/ km^2	Extremely High Sensitive Area/ km^2
1	15.56	1.77	27.53	3.85	3.66	3.53	4.52
2	26.93	2.33	60.13	8.09	6.89	4.14	7.47

3	11.47	2.03	23.26	4.24	3.29	1.49	2.44
4	5.34	1.96	10.48	2.27	1.26	1.23	0.57
5	9.56	0.67	6.45	2.63	3.06	1.35	2.52
6	6.92	0.83	5.77	1.47	1.67	1.43	2.35
7	10.09	2.06	20.81	2.85	2.48	2.12	2.71
8	7.34	1.65	12.12	1.98	1.39	1.87	2.10
9	11.87	1.61	19.09	2.73	4.47	3.05	1.56
10	10.53	1.65	17.32	1.94	1.88	3.09	3.68
11	9.25	1.15	10.61	1.47	1.94	2.50	3.34
12	8.48	1.5	12.75	2.11	3.07	1.55	1.75
13	9.17	2.91	26.71	3.21	1.59	2.51	1.86
14	14.63	0.97	14.19	5.37	5.02	2.69	1.55
15	15.75	2.31	36.41	2.98	4.10	5.39	3.29
16	10.1	2.16	21.79	2.22	1.99	3.03	2.86
17	6.31	1.52	9.62	1.53	1.72	1.85	1.21
18	10.01	3.26	32.73	1.27	1.15	2.68	4.93
19	6.47	1.47	9.53	1.41	1.29	1.89	1.88

4 CONCLUSION

In this study, high-resolution remote sensing images from different periods were used to analyze the spatial pattern of the erosion trench, revealing the number, area, length and other morphological characteristics and their spatial distribution, which provided an important basis for evaluating the regional erosion state. On top of this, in order to comprehensively explore the influence mechanism of natural and man-made multiple factors on the development of erosion gully, comprehensively study the influence of natural and man-made multiple factors on the development of erosion gully, including terrain, precipitation, soil, human activities and other factors, a comprehensive database of influencing factors is built, and through linear regression and correlation analysis, The relationship between each factor and the development of erosion gully was deeply understood. Finally, in order to predict the occurrence risk of erosion gully, the concept of sensitivity was introduced to evaluate the accuracy of the sensitivity prediction of erosion gully by two advanced machine learning algorithms, Logistic regression model and random forest model, and the model with higher accuracy was selected to predict the sensitivity of erosion gully in the black soil area of Northeast China. To provide directions for the protection and sustainable utilization of land resources in the northeast black soil region, the conclusions of this study are as follows:

Through comparative analysis of the basic data of erosion trenches in different years, this study found that the number, total length, average length, total area and average area of erosion trenches in the Northeast black soil region showed a significant growth trend from 1968 to 2020. It is particularly noteworthy that the average area of erosion gullies increased by 34.00% to 114.28%, which strongly indicates the severity and increasing trend of soil erosion in this area. At the same time, the significant increase of gully density and fragmentation degree further confirms the intensification of soil erosion in this area. From the perspective of spatial distribution, this study

reveals that the distribution of erosion gullies has significant spatial positive correlation and aggregation effect, and this aggregation gradually strengthens with the advance of time. Through the in-depth analysis of the migration of the gravity center of the erosion trench, it is also found that the erosion process presents a dynamic change, which is manifested in that the gravity center of the growth of the length of the erosion trench first shifts to the northeast, and then turns to the southwest. In addition, using the method of hot and cold spot analysis, this study successfully identified the accumulation area of erosion gully activity, which provides a strong scientific support for the implementation of accurate soil erosion control measures.

Through comparative analysis of the basic data of erosion trenches in different years, this study found that the number, total length, average length, total area and average area of erosion trenches in the Northeast black soil region showed a significant growth trend from 1968 to 2020. It is particularly noteworthy that the average area of erosion gullies increased by 34.00% to 114.28%, which strongly indicates the severity and increasing trend of soil erosion in this area. At the same time, the significant increase of gully density and fragmentation degree further confirms the intensification of soil erosion in this area. From the perspective of spatial distribution, this study reveals that the distribution of erosion gullies has significant spatial positive correlation and aggregation effect, and this aggregation gradually strengthens with the advance of time. Through the in-depth analysis of the migration of the gravity center of the erosion trench, it is also found that the erosion process presents a dynamic change, which is manifested in that the gravity center of the growth of the length of the erosion trench first shifts to the northeast, and then turns to the southwest. In addition, using the method of hot and cold spot analysis, this study successfully identified the accumulation area of erosion gully activity, which provides a strong scientific support for the implementation of accurate soil erosion control measures.

REFERENCES

[1] Wang R, Wang N, Fan Y, et al. Quantitative Attribution Analysis of the Spatial Differentiation of Gully Erosion in the Black Soil Region of Northeast China[J]. Geofluids, 2022, 2022.

[2] García-Ruiz J M, Nadal-Romero E, Lana-Renault N, et al. Erosion in Mediterranean landscapes: Changes and future challenges[J]. Geomorphology, 2013, 198: 20-36.

[3] WANG W, DENG R, ZHANG S. Preliminary research on risk evaluation of gully erosion in typical black soil area of Northeast China[J]. Journal of Natural Resources, 2014, 29(12): 2058-2067.

[4] Ying Z ,Bin Z ,Wei Q , et al.Spatial differentiation of gully clusters based on the regional scale: an example from northeastern China.[J].PeerJ,2020,8e9907-e9907.

[5] Zhou Y ,Zhang B ,Qin W , et al.Primary environmental factors controlling gully distribution at the local and regional scale: An example from Northeastern China[J].International Soil and Water Conservation Research,2020,(prepublish):

[6] Frankl A ,Zwertvaegher A ,Poesen J , et al.Transferring Google Earth observations to GIS-software: example from gully erosion study[J].International Journal of Digital Earth,2013,6(2):196-201.

[7] Karydas C ,Panagos P .Towards an Assessment of the Ephemeral Gully Erosion Potential in Greece Using Google Earth[J].Water,2020,12(2):603-603.

[8] Real C S L ,Crestana S ,Ferreira M R R , et al.Proposition for a new classification of gully erosion using multifractal and lacunarity analysis: A complex of gullies in the Palmital stream watershed, Minas Gerais (Brazil)[J].Catena,2020,186104377-104377.

[9] Deng Q, Miao F, Zhang B, et al. Planar morphology and controlling factors of the gullies in the Yuanmou Dry-hot Valley based on field investigation[J]. Journal of Arid Land, 2015, 7: 778-793.

[10]Vanmaercke M ,Poesen J ,Mele V B , et al.How fast do gully headcuts retreat?[J].Earth-Science Reviews,2016,154336-355.

[11] Luffman E I ,Nandi A ,Spiegel T .Gully morphology, hillslope erosion, and precipitation characteristics in the Appalachian Valley and Ridge province, southeastern USA[J].Catena,2015,133221-232.

[12] Wischmeier W H, Smith D D, Uhland R E. Evaluation of factors in the soil-loss equation[J]. 1958.

[13] Vanmaercke M, Chen Y, Haregeweyn N, et al. Predicting gully densities at sub-continental scales: a case study for the Horn of Africa[J]. Earth Surface Processes and Landforms, 2020, 45(15): 3763-3779.

[14] Zhu Y, Cai Q. Rill erosion processes and its factors in different soils[J]. Gully Erosion Under Global Change. Sichuan Science and Technology Press, Chengdu, China, 2004: 96-108.

[15] Li G, Klik A, Wu F, et al. Gully erosion features and its causes of formation on the (Yuan) land in the Loess Plateau, China[J]. Gully Erosion Under Global Change. Sichuan Science and Technology Press, Chengdu, China, 2004: 131-142.

[16] Thomas J T, Iverson N R, Burkart M R, et al. Long-term growth of a valley-bottom gully, western Iowa[J]. Earth Surface Processes and Landforms: The Journal of the British Geomorphological Research Group, 2004, 29(8): 995-1009.

[17] Christian C ,Valerio A ,Silvia A , et al.A GIS-based approach for gully erosion susceptibility modelling: a test in Sicily, Italy[J].Environmental Earth Sciences,2013,70(3):1179-1195.

[18] Chowdhuri I, Pal S C, Saha A, et al. Evaluation of different DEMs for gully erosion susceptibility mapping using in-situ field measurement and validation[J]. Ecological Informatics, 2021, 65: 101425.

[19] Zabihi M, Mirchooli F, Motevalli A, et al. Spatial modelling of gully erosion in Mazandaran Province, northern Iran [J]. 2018, 161: 1-13.

[20 Debanshi S, Pal S. Assessing gully erosion susceptibility in Mayurakshi river basin of eastern India[J]. Environment, development and sustainability, 2020, 22: 883-914.

[21] Eustace H A ,Pringle J M ,Denham J R .A risk map for gully locations in central Queensland, Australia[J].European Journal of Soil Science,2011,62(3):431-441.

[22] Band S S, Janizadeh S, Chandra Pal S, et al. Novel ensemble approach of deep learning neural network (DLNN) model and particle swarm optimization (PSO) algorithm for prediction of gully erosion susceptibility[J]. Sensors, 2020, 20(19): 5609.

[23] Pourghasemi R H ,Yousefi S ,Kornejady A , et al.Performance assessment of individual and ensemble data-mining techniques for gully erosion modeling[J].Science of the Total Environment,2017,609764-775.

[24] Yang Z ,Ping J ,Liye C , et al.Study on the spatial variation of China's territorial ecological space based on the standard deviation ellipse[J].Frontiers in Environmental Science,2022,10

[25]Rahmati, O.; Tahmasebipour, N.; Haghizadeh, A.; Pourghasemi, H.R.; Feizizadeh, B. Evaluating the influence of geoenvironmental factors on gully erosion in a semi-arid region of Iran: An integrated framework. Sci. Total Environ. 2017, 579, 913–927.

[26] Arabameri, A.; Pradhan, B.; Pourghasemi, H. Rahmati, O.; Tahmasebipour, N.; Haghizadeh, A.; Pourghasemi, H.R.; Feizizadeh, B. Evaluating the influence of geoenvironmental factors on gully erosion in a semi-arid region of Iran: An integrated framework. Sci. Total Environ. 2017, 579, 913–927.

[27] Arabameri, A.; Pradhan, B.; Pourghasemi, H.R.; Rezaei, K.; Kerle, N. Spatial modelling of gully erosion using GIS and

R programing: A comparison among three data mining algorithms. Appl. Sci. 2018, 8, 1369.

[28]. Garosi, Y.; Sheklabadi, M.; Pourghasemi, H.R.; Besalatpour, A.A.; Conoscenti, C.; Van Oost, K. Comparison of differences in resolution and sources of controlling factors for gully erosion susceptibility mapping. Geoderma 2018, 330, 65–78.

[29].Amiri, M.; Pourghasemi, H.R.; Ghanbarian, G.A.; Afzali, S.F. Assessment of the importance of gully erosion effective factors using Boruta algorithm and its spatial modeling and mapping using three machine learning algorithms. Geoderma 2019, 340, 55–69.

[30]. Chowdhuri, I.; Pal, S.C.; Arabameri, A.; Saha, A.; Chakrabortty, R.; Blaschke, T.; Pradhan, B.; Band, S.S. Implementation of Artificial Intelligence Based Ensemble Models for Gully Erosion Susceptibility Assessment. Remote Sens. 2020, 12, 3620.

[31]Fawcett, T. An introduction to ROC analysis. Pattern Recognit. Lett. 2006, 27, 861–874.

[32]Williams, G. Data Mining with Rattle and R: The Art ofExcavating Data for Knowledge Discovery; Springer: New York, NY, USA, 2011; p. 382.

ABOUT THE AUTHOR

Hongfeng Yu

School of Soil and Water Conservation ,Beijing Forestry University, Beijing 100083, Beijing , China.

E-mail: Yhf727498@bjfu.edu.cn

Mingchang Shi

School of Soil and Water Conservation ,Beijing Forestry University, Beijing 100083, Beijing , China.

E-mail: shimc@bjfu.edu.cn