Scholar Publishing Group
International Journal of Engineering Technology and Construction
https://doi.org/10.38007/IJETC.2023.040104
ISSN 2789-7176 Vol. 4, Issue 1: 48-64

Evaluation of Machine Learning Algorithm for Landslide Sensitive Spatial Model Based on GIS Taking the Area along Sichuan-Tibet Railway as an Example

Feng Zou and Rong Liu*

Northwestern Polytechnical University, Xi'an, China

**corresponding author*

Keywords: GIS Technology, Landslide Sensitivity, Spatial Model, Support Vector Machine (SVM), Logistic Regression Model

Abstract: With the development of GIS-based geospatial information technology, new methods have been provided for landslide disaster research. However, there is currently no suitable way to combine GIS technology with machine algorithms. In order to construct a highly accurate landslide sensitivity spatial model, this paper statistically analyzes the relationship between landslide disasters and various influencing factors in the study area through the GIS spatial analysis function, especially for the actual situation along the Sichuan-Tibet Railway. In this paper, a cross-check method is used to construct a landslide sensitivity evaluation model, and the accuracy of different models is quantitatively evaluated. The results of the fitting accuracy of the logistic regression model and the support vector machine model are: the average accuracy in the modeling stage is 75.722 and 75.65, and the average accuracy in the verification stage is 71.34 and 71.21. At the modeling stage, the SVM model has a fitting accuracy of about 3% higher than that of the logistic regression model; at the verification stage, the fitting accuracy is 0.13% higher than that of the logistic regression model; the AUC results show that the SVM model performs optimally, its AUC value is above 0.9, which achieves a higher accuracy. Compared with the logistic regression model, this value is 0.111 higher in the modeling stage and 0.111 higher in the verification stage.

1. Introduction

Landslides are one of the most widely distributed geological disasters in the world. They not only cause great damage to regional surface cover and ecological environment, but also seriously threaten people's lives and property safety. The large-scale exploitation of coal resources will also trigger a wide range of geological landslides disaster. Based on the analysis of the spatial distribution characteristics of geological hazards, research teams at home and abroad have rarely used numerical modeling and quantitative evaluation of landslide hazards using remote sensing and

geographic information system (GIS) methods.

For the research and analysis of landslide sensitivity, many research teams at home and abroad have spent a lot of time and energy in this area. In [1], the author introduced the K-PSO clustering algorithm and entropy method, and established a landslide sensitivity analysis model. The analysis results based on the K-PSO clustering algorithm show that compared with in situ observations, the sensitivity of the proposed K-PSO method to landslides is an effective water quality analysis for the reservoir area of Xulong Hydropower Station. In [2], the author used drainage density to optimize the catchment area threshold, established a support vector machine sensitivity prediction model based on genetic algorithm, and carried out seismic landslide sensitivity zoning for Baosheng Township. The results show that the accuracy of the seismic landslide sensitivity analysis based on the optimized slope elements reaches 98.72%. In [3-4], the author considered the problem of landslide body formation, and considered the formation of stress-strain state (SSS) and the stability of soil structures (cuts and embankments). The calculation results can see the emergence and evolution of the "plastic" region, or the limit state of the "compressed" and "expanded" regions. In [5], the author used the receiver operating characteristic curve to evaluate the model quality, and also combined sensitivity research and uncertainty assessment to produce a reliable landslide sensitivity model that can be used for regional spatial planning. Intermediate complexity statistical methods are used to assess the relative landslide sensitivity on a regional scale [6]. In [7-8], the authors applied the PFR model to consider the influence of landslide-related factors related to high-resolution images of Google Earth and field observations. The results show that compared with the existing data and previous studies in the same area, the LiDAR-derived DEM using the PFR model improves the landslide sensitivity map with an accuracy performance of 92.59%. In addition, this study shows that all considerations have a relatively positive effect on the landslide sensitivity map in the study, however, the most effective factor for landslide occurrence is 13.7% lithology. In [9], the author puts forward an automatic workflow, from the hourly network-based rain gauge data collection to the generation of spatial difference rainfall prediction, and emphasizes the potential use of citizen scientific data to improve the research on landslide early warning system. In [10-12], the author proposed a new landslide inventory mapping framework based on the spatial characteristics of landslides. This is the first time that high-resolution remote sensing images are used to obtain the landslide spatial information of LIM by integrating multi-scale segmentation of post-event images with MV methods. In [13-14], the author provided a new method for studying the mechanism of earthquake-induced landslide formation, established a geological model and carried out simulation experiments. Model test results show that the effective shear strength of the rock mass can be reduced by 4.4% to 21.6% due to the action of void gas.

Remote sensing is a tool that is important for the production of land use and land cover maps through a process called image classification [15]. In [16], the authors proposed an integration of a geographic information system (GIS) and a gene expression program (GEP) to predict rainfall-induced shallow landslides in Son La province, Vietnam. The predictive power of the model has been verified by the area under curve calculation. In [17], the author applied GIS technology to measure soil material content. Based on the soil nutrient database and GIS and GPS platforms, the authors studied the spatial distribution of soil nutrients in Sanmenxia, Henan. The results showed that the organic matter was lacking in the southwestern area of Sanmenxia, the available phosphorus was moderate, and the available nitrogen was low. In [18-19], the author applied GIS technology to analyze flood disaster index. The author used very high-resolution (VHR) satellite images and multi-criteria analysis (MCA) to make flood hazard maps. He chose the analytic hierarchy process to calculate the weight of each criterion in the flood hazard index (FHI). This work demonstrates the benefits of combining remote sensing data with MCA methods to provide rapid and cost-effective information on hazard assessments. Existing findings about the built

environment are more reliable than those about the food environment. The application of GIS data / methods in obesity research is still limited, and related research faces many challenges [20]. In [21], the author applied GIS technology to the analysis of flood disasters. The author developed a risk-based tool by transferring the Landscape Architecture Technical Information Series (LATIS) to an area with limited data resources, which uses input parameters to estimate the extent of damage. It was found that the map generated showed a spatial change in the cost of damage, which was related to the depth of the flood. In [22], the author describes four shell scripts that perform fast and automatic calculations of morphometric parameters and draw curves that show the changes in the parameters calculated during the entire channel development process. These scripts work on the basis of free and open source software from GRASS GIS and contain the basic characteristics of river channels. In [23], the author analyzed the local field values in the entire GIS geometry, and needed to use the full Maxwell method to simulate the VFT generation process. The authors give example simulations of VFT overvoltage (VFTO) waveforms for two geometries and compare the numerical work required to solve the related magnetic field equations. In [24], the author applied GIS technology to measure soil material content. The authors used remote sensing and GIS data to study land cover changes and CO2 stocks in Indonesia, and processed them with LUMENS software. The study successfully compared land conversions between 1989-2000 and 2000-2013. In [25-26], the authors used pairwise comparisons to assess the consistency ratio between expert opinions and calculate the final weights for each criterion. The article uses a weighted linear combination (WLC) method in a GIS environment to generate a wind turbine adaptability map. It was found that 45% of the study area is very suitable for wind turbines. In [27], the author applied GIS technology to the quantitative assessment of water resources vulnerability. In order to explore a better quantitative assessment method for water resources vulnerability, the authors used GIS and RS technology to evaluate the vulnerability of water resources system in Hengyang Basin.

A large amount of literature on landslide disaster research has sprung up. However, in the study of statistical models of landslide sensitive areas, the following problems have not been reasonably resolved: 1) In terms of the regional scale of mines, there are few studies on the spatial distribution of geological hazards and their dependence on scale; 2) There are not many predictions of geological hazards under the influence of mining. At present, no quantitative evaluation of the impact of mining has been found in the prediction of spatial hazards.

In order to build a spatial model of landslide sensitivity with high accuracy, this paper analyzes the relationship between landslide disaster and various influencing factors in the study area through the spatial analysis function of GIS, especially for the actual areas along the Sichuan Tibet railway, to explore the impact on landslide disaster. In this paper, the sensitivity evaluation model of landslide is constructed by using cross test method, and the accuracy of different models is evaluated quantitatively. The results of the fitting accuracy of the logistic regression model and the support vector machine model are: the average accuracy in the modeling stage is 75.722 and 75.65, and the average accuracy in the verification stage is 71.34 and 71.21. Compared with the logistic regression model, this value is 0.111 higher in the modeling stage and 0.111 higher in the verification stage.

2. Method

2.1. Landslide Sensitivity Evaluation Model for Sichuan-Tibet Railway

(1) Logistic regression method (Logistic)

Similar to linear regression, logistic regression also detects the quantitative relationship between a dependent variable and one or more independent variables through the idea of regression. The dependent variable value is transformed into the logarithm of probability ratio corresponding to its

value state, that is, the probability of dependent variable events is fitted by logistic curve. And the method of solution is changed from the original least square estimation to the maximum likelihood estimation.

Let $X_1, X_2,\ldots, X_p$ represent a sampling value of p independent variables $x_1, x_2,\ldots, x_p$, and Y is a binary variable with a value of 1 or 0. You can use the Logistic function to estimate this sampling Y. A probability of 1 where the relationship between the dependent and independent variables can be expressed as:

$$p = \frac{1}{1+e^{-z}} \tag{1}$$

In the formula, P is the probability of occurrence of dependent variable estimated according to each variable, and its value range is 0 to 1, which can be expressed as an S-shaped curve; Z value is the weighted linear combination of each variable, and its value range is - ∞ to + ∞, which can be expressed by the sum of a series of constant values. The formula is as follows:

$$z = \ln\left(\frac{p}{1-p}\right) = \alpha + \sum_{i=1}^{n} \beta_i x_i \tag{2}$$

In the formula, p / 1-p is the maximum likelihood; α is the intercept of the formula, which is also a constant value; xi is the independent variable, which is the predictor; βi is the coefficient of the independent variable in the polynomial.

(2) Model impact factor

In the logistic regression model, the factor score is used to determine whether the factor needs to enter the model. The calculation method is as follows. Suppose there are n variables, $\alpha_1, \alpha_2,\ldots. \alpha_n$, and m variables that have not entered the model, $b_1, b_2,\ldots b_m$. If bi is not a categorical variable, then the statistical score for bi is : $S_i = \left(L_{n_1}\right)^2 B_{2_2,i}$. In the above, if bi is a categorical variable and there are k categories, then it will be transformed into a k-1 dimensional virtual vector, and these new k-1 variables are denoted as $b_i, \hat{b}_{i+1},\ldots \hat{b}_{i+k-2}$. And the statistical score of the new variable becomes:

$$S_i = \left(L^*_{\hat{\beta}}\right)' B_{22,i} L^*_{\hat{\beta}} \tag{3}$$

In the formula, $\left(L_{\hat{\beta}}\right) = \left(L_{b_1},\ldots L_{\hat{b}_{2+-2}}\right)$, and B22 matrix are (k-1) ×(k-1), expressed as:

$$\begin{gathered} B_{22} = \left(A_{22,i} - A_{21,i} A_{11}^{-1} A_{12,i}\right)^{-1} \\ A_{11} = \alpha' V \alpha, \quad A_{12,i} = \alpha' V b_i, \quad A_{22,i} = b_i' V b_i \end{gathered} \tag{4}$$

In the above formula, α is the design matrix of $\alpha_1, \alpha_2,\ldots \alpha_n$; bi is the design matrix of $b_i, b_{i+1},\ldots b_{i+k-2}$. There are many calculation methods for the importance of model factors. The relative importance of this article is based on the degree to which the uncertainty in predicting the dependent variable can be reduced after using a certain factor. The prediction of the uncertainty of the predicted value is based on the entropy of its distribution. The calculation method of the relative importance of the three models is as follows:

$$H_Y = -\sum_i P(Y=i)\log P(Y=i) \tag{5}$$

In the formula, $H_{Y|x_j}$ represents the entropy value of the conditional distribution probability $f_{y|x_j}$; xj represents the amount of information about whether the conditional distribution probability $f_{y|x_j}$ of Y is greater than the distribution boundary fy.

2.2. Automatic Extraction Model of Landslide Sensitivity based on Machine Learning

(1) Recursive feature elimination method based on random forest (RF-RFE)

In the objective function, the RFE uses DJ(i) and (ωi) 2 to evaluate the criteria that affect the removal of a feature in each iteration. When several features are removed in each iteration, this process becomes a problem of finding the optimal sub solution, which is also necessary to obtain a small feature subset. DJ (i) is the objective function of the OBD algorithm based on the sensitivity analysis feature evaluation. Using DJ (i) as the feature function and expanding by Taylor series, we can get:

$$DJ(i) = (1/2)\frac{\partial^2 J}{\partial \omega_i^2}\left(D\omega_i\right)^2 \tag{6}$$

Among them, the weight change value $D\omega_i = \omega_i$ replaces the corresponding i feature, and the OBD algorithm advocates using DJ (i) as the weight size instead of the weight size based on the pruning standard. For the linear discriminant function, its cost function J is a binomial equation with ωi weights. It is a cost function $J = \sum_{x\in X} \| w \bullet x - y\|^2$ with mean square error classification. Its minimum cost function can be expressed as $J = (1/2)\| w\|^2$ under linear SVM. Therefore, these two expressions provide a theoretical basis for using (ωi) 2 as a criterion for feature criteria.

$$\begin{gathered} e_i^f = A_i - A_i^f, i = 1,\ldots,m, \quad S = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}\left(e_i^f - e^f\right)^2} \\ e^f = \frac{1}{m-1}\sum_{i=1}^{m} e_i^f, \quad f_{imp} = e^f / S \end{gathered} \tag{7}$$

Ai is the OOB accuracy of the i-th tree, efi is the difference between the OOB accuracy of the i-th feature f in the old and new decision trees, S is the variance of the OOB accuracy of the i-th feature f in the new and old decision trees, and ef feature f's influence on the OOB accuracy.

(2) Committee voting selection method based on random forest (RF-QBC)

There are two main methods for the QBC measurement committee's voting inconsistency. One is based on voting entropy (VE) strategy:

$$X^* = \arg\max_x -\sum_i \frac{V(y_i \mid X)}{C}\log\frac{V(y_i \mid X)}{C} \tag{8}$$

Among them, C is the number of models in the committee, and V (yi|X) is the number of votes labeled by the members of the committee as sample yi. The other is based on the relative entropy (RE) strategy:

$$X^* = \arg\max_x \frac{1}{C}\sum_{c=1}^{C} D\left(P_{\theta^c)} \| P_l\right)$$
$$D\left(P_{\theta^c} \| P_l\right) = \sum_i P_{\theta^c}(y_i \mid X)\log\frac{P_{\theta^c}(y_i \mid X)}{P_{axg}(y_i \mid X)} \quad (9)$$
$$\frac{P_{ay}(y_i \mid X)}{P_{ayg}(y_i \mid X)} P_{avg}(y_i \mid X) = \frac{1}{C}\sum_{c=1}^{c} P_{\theta^c}(y_i \mid X)$$

Among them, θ (c) represents a model member of the committee, and Pavg represents the average of conditional probability of all committee members.

The basic idea of RF-QBC based on random forest based committee voting selection algorithm proposed in this paper can be described as follows: first create a training set to establish a selection and replacement of the original data, and set up the created "bag" as a markable object. Then, use each training set to train a random forest classifier, train the class labels of all points in the candidate set, and iteratively calculate the voting entropy of each sample. The steps are as follows:

1) The stratified random sampling method is used to extract the initial training set S from the original training sample set, and the remaining unsampled samples default to the unlabeled test sample X;

2) The bootstrap randomly selects the training set S, and generates a subset $\{S_k, k=1,2,\ldots,k\}$ of training samples. Each time the unextracted samples form k out-of-bag data (OOB);

3) Generate a single decision tree using each training sample kS as the training set.

4) Repeat steps 2 and 3 until k decision trees are generated;

5) Classify and predict the unlabeled test sample X, calculate the voting entropy H (xl) of the unlabeled test sample, and according to the QBC algorithm voting entropy sampling principle XQBC label the sample with the largest information entropy to the training sample set S:

$$X^{QBC} = -\arg\max_{X_t \in X} H(x_l)$$
$$H(x_l) = \sum_{\alpha=1}^{N_l} p\left(y_l^* = m \mid X_l\right)\log\left[p\left(y_l^* = m \mid X_l\right)\right] \quad (10)$$

H (xl) is an empirical measure of entropy, yl* is the prediction result of unlabeled sample $x_i\left(\forall x_i \in X\right)$, $p\left(y_i^* = m \mid X_i\right)$ is the probability of predicting that unlabeled sample xl belongs to a certain category m, and Nl is the number of categories predicted by the committee to predict the category of unlabeled samples.

2.3. GIS Model based on Support Vector Machine (SVM) and Kernel Function

In general, SVM is designed to solve two types of classification problems, that is, there are both positive samples and negative samples. The goal of the two types of SVMs is to find a hyperplane on an n-dimensional space, to distinguish them at the maximum interval, and to make the separated two types of data points farthest from the classification plane. This hyperplane can be either a plane or a surface. Expressed mathematically:

$$\min \frac{1}{2}\| w\|^2 \quad (11)$$

Where the constraints are: $y_i\left(\left(w \cdot x_i\right)+b\right) \geq 1$

In the above formula, ||w|| is the norm of the hyperplane normal vector, b is a scalar, and $\cdot$ represents a scalar product. The Lagrange multiplier rule can be introduced to find the extreme value, and the auxiliary function is generated as follows:

$$L=\frac{1}{2}\| w\|^2-\sum_{i=1}^{n}\lambda_i\left(y_i\left(\left(w\cdot x_i\right)+b\right)-1\right) \tag{12}$$

In the above formula, λi is a Lagrangian multiplier. Set the partial derivatives of L for w and b equal to 0, and get:

$$w=\sum_{i=1}^{n}\lambda_i y_i x_i,\sum_{i=1}^{n}\lambda_i y_i=0 \tag{13}$$

And subject to: $\lambda_i\geq 0,\ \sum_{i=1}^{n}\lambda_i y_i=0$. The above discussion applies to the case of linear separability. For more general linear inseparable instances, the relaxation factor $\xi_i\ \left(i=1,\ 2,\ ...,\ n\right)$ is introduced to adjust the constraint conditions, as follows:

$$\left(y_i\left(\left(w\cdot x_i\right)+b\right)\geq 1-\xi_i,\quad \xi_i\geq 0\right) \tag{14}$$

Here v∈ (0,1] is a new threshold for error classification. On the other hand, the kernel function K (xi, xj) is introduced into the nonlinear indivisible problem. The selection of kernel functions is very important for SVM model. Although some new kernel functions have been proposed, the four basic kernel functions that are widely recognized are:

Linear kernel function (Linear): $K\left(x_i,x_j\right)=x_i^T x_j$

Polynomial kernel function (Polynomial): $K\left(x_i,x_j\right)=\left(\gamma x_i^T x_j+r\right)^d,\gamma>0$

Radial basis function: $K\left(x_i,x_j\right)=e^{-\gamma\left(x_i-x_j\right)^2}$

Sigmoid function (Sigmoid): $K\left(x_i,x_j\right)=\tanh\left(\gamma x_i^T x_j+r\right)$

Among them, γ, r, and d of the kernel function are parameters, which need to be optimized when building the model to improve the fitting accuracy of the model.

3. Experiment

3.1. Data Source

The Sichuan-Tibet Railway engineering area spans multiple regional geological structural units, and the regional geological conditions are extremely complicated. In particular, the large-scale plate combination zone not only controls the dynamic characteristics of regional geological construction and development and evolution, but also greatly affects the regional engineering geological environment. This article selects the Sichuan-Tibet Railway crossing the Jinsha River combined zone (A), the Lancang River combined zone (B), the Nujiang combined zone (C), the Yarlung Zangbo combined zone (D), and the Sichuan-Yunnan block (I), Beiqiangtang-Qamdo-Simao block (Jiangda-Deqin tectonic magmatic belt (II1); Qamdo-Mangkang basin (II2), Nanchangtang-Zuogong-Baoshan block (III), Gangdise-Nianqing Tanggula land Block (Naqu-Luolong arc pre-basin (IV1); Gangdise-Chayu block magmatic arc (IV2), Baiyu-Linzhi section of the Himalayan orogen (V) as study areas, analysis of the Sichuan-Tibet railway cross-plate combination Engineering geological environment with sections.

The in-situ stress testing projects collected by the original in-situ stress data collected in this paper mainly include: Erlangshan Tunnel, Galongsi Tunnel, Sangzhuling Tunnel, Rumei Hydropower Station Dagangshan Hydropower Station, Lianghekou Hydropower Station, etc. See

Table 1.

According to Table 2, in the study area, the Jinshajiang fault is shown as a right-handed squeeze motion, the Lancangjiang fault is shown as a right-handed strike-slip motion, the Nujiang fault is shown as a right-handed strike-slip motion, and the Yarlung Zangbo River fault is characterized as a right-handed squeeze motion .

Table 1. Statistics of in-situ stress measured along the Sichuan-Tibet Railway and adjacent areas

Area code	Buried depth	Stress 1	Stress 2
Area 1	215	12.55	5.66
Area 2	27	13.14	8.04
Area 3	286	19.45	11.06
Area 4	255	19.88	4.23
Area 5	268	16.43	8.94
Area 6	195	15.22	10.57
Area 7	166	20.16	5.07
Area 8	191	29.92	9.77
Area 9	160	16.44	8.39
Area 10	194	17.08	1.94
Area 11	210	19.34	6.99
Area 12	15	9.99	9.20
Area 13	18	16.44	1.50
Area 14	401	13.09	8.96
Area 15	99	16.45	2.66

Table 2. Movement characteristics of major faults in the study area

Fault name	Position	Parcel movement rate		Fault shift rate		Fault properties
		Towards	Tendency	Towards	Tendency	
A	Beidongpan	49.12	18.34	3.61	3.05	Right hand squeeze
	Nanxipan	44.26	20.13			
B	Beidongpan	40.19	36.48	2.83	1.77	Right hand squeeze
	Nanxipan	39.99	38.81			
C	Beidongpan	29.64	42.15	2.43	0.21	Right hand slip
	Nanxipan	26.49	43.55			
D	Beidongpan	30.11	40.16	1.97	2.84	Right hand squeeze
	Nanxipan	26.43	40.22			

3.2. Evaluation Factor Data

The landslide sensitivity evaluation factor data sources used in this paper include DEM data (ASTER-GDEM) with a resolution of 30m, remote sensing image data, and other statistical data. DEM data and remote sensing images are downloaded from the Internet. Other statistical data include formation rocks. The sex maps (see Table 3) were scanned and digitized by paper maps.

Based on the remote sensing image, with the support of ENVI remote sensing software, the land use distribution map and NDVI distribution map were obtained through interpretation; the elevation, slope, aspect, and curvature factors were extracted based on DEM; Factors such as distance from road, distance from river, distance from fault, etc. were extracted on the basis of each thematic map.

The software used in this article mainly includes ARCGIS map mapping and spatial analysis software, ENVI remote sensing image processing software, IBM SPSS Statistic 19 statistical analysis software, and IBM SPSS Modeler 18 data mining software trial version.

Table 3. Evaluation basic data table

Data item	Data source	Scale	Data sources
Height	ASTER-GDEM	30m	Download online
Slope			
Aspect			
Plane curvature			
Section curvature			
Land use type NDVI	Remote sensing image (TM)	30m	Download online
Way	Paper map	1:50 thousand	Land and Resources Department
Formation lithology			
River network			
Fault			
Mining disturbance			
Coal field boundary			

3.3. Space Coordinate System and Its Parameters

1) Coordinate system and map projection

For the study area, the mapping scale is also large due to the high resolution of remote sensing images. In such a large scale, if the two different geodetic coordinate systems are used, the same geographic feature will have significant displacement and deformation under the two geodetic coordinate systems. Therefore, it is necessary to uniformly specify which geodetic coordinate system is used. This study used the 1980 Xi'an coordinate system, and the map projection was a Gauss-Kruger projection (3 degree band or 6 degree band).

2) Scale

For research areas, the scale of the results is generally 1: 2000, 1: 5000, 1: 10000, a few are 1: 500, 1: 25000, 1: 50000, 1: 250000, etc. Among them: 1: 500, 1: 1000, 1: 2000, 1: 5000, 1: 10000 use Gaussian 3 degree band projection, 1: 25000, 1: 50000 use Gaussian 6 degree band projection (see Table 4 for details).

Table 4. Scale code table

Number	Scale	Scale code	Number	Scale	Scale code
1	1:1000000	A	11	1:500	K
2	1:500000	B	12	1:200000	L
3	1:250000	C	13	1:1500000	M
4	1:100000	D	14	1:2500000	N
5	1:50000	E	15	1:4000000	O
6	1:25000	F	16	1:5000000	P
7	1:10000	G	17	1:6000000	Q
8	1:5000	H	18	1:8000000	R
9	1:2000	I	19	1:10000000	S
10	1:1000	J	20	No scale or nothing to do with it	T

3.4. Evaluation Criteria of Sensitivity Model

(1) Accuracy evaluation parameters

The most direct way to evaluate the effect of the model is to look at the fitting accuracy of the model. In addition to the fitting accuracy, this paper further calculates four parameters on the basis of the confusion matrix, namely sensitivity, specificity, positive predictive value and negative predictive value. The formula expression is as follows :

$$Accuray = \frac{TP+TN}{TP+TN+FP+FN},\ Sensitivity\ = \frac{TP}{TP+FN},\ Specificity\ = \frac{TN}{TN+FP} \quad (15)$$

$$Positive\ predictive\ value = \frac{TP}{TP+FP},\ Negative\ predictive\ value = \frac{TN}{TN+FN} \quad (16)$$

In the formula, TP and TN represent the positive and negative values respectively, that is, the number of disaster points and non disaster points correctly assigned by the model; while FP and FN represent the positive and negative values respectively, that is, the number of disaster points and non disaster points wrongly assigned by the model.

(2) ROC curve and AUC value

Usually, the area under the curve (Area Under Curve, AUC) is used as a standard to measure the prediction accuracy of the model. The range of AUC value is 0.5~1, and the value is 0.5. It is completely valueless prediction. When the value reaches 1, it is a completely ideal prediction. The ROC is calculated as follows:

$$Sensitivity = \frac{TP}{TP+FN}, 1-Specificity = \frac{FP}{FP+TN} \quad (17)$$

(3) Landslide point density

The landslide point density is the basic indicator for testing the results of sensitive zoning. The landslide point density (LDD) is calculated as follows:

$$LDD_i = 100 \times L_i / Npix(S_i) \quad (18)$$

Among them, $L_i (i \in \{1,2,3,4,5\})$ is the number of historical landslide points in the i-th sensitive area, and this value can be obtained from the landslide sensitivity zoning map and historical landslide point spatial distribution map; $Npix(S_i)$ is the area size of the i-th sensitive area.

4. Results and Discussions

4.1. Analysis of Optimization Results of Sensitivity Space Model

In the landslide sensitivity evaluation along the Sichuan-Tibet Railway, the parameters that entered the logistic regression model were six factors: elevation, slope, NDVI, distance from the road, lithology, and land use type. Their significance levels were above 95%. In addition, the entry of factors between different cross-tests is also different. The five factors of elevation, slope, NDVI, distance from road, and lithology are in 5 cross-checks (k = 1, k = 2, k = 3, k = 4 and k = 5), they entered the logistic regression model, which reflected that the model entry factors under different cross-tests were mostly the same.

Through cross-validation of the two models, the confusion matrices of the three models at the 30m scale were obtained, and the Moses fitting accuracy was calculated based on this. The results are shown in Figure 1 and the results of each evaluation parameter, as shown in Figure 2.

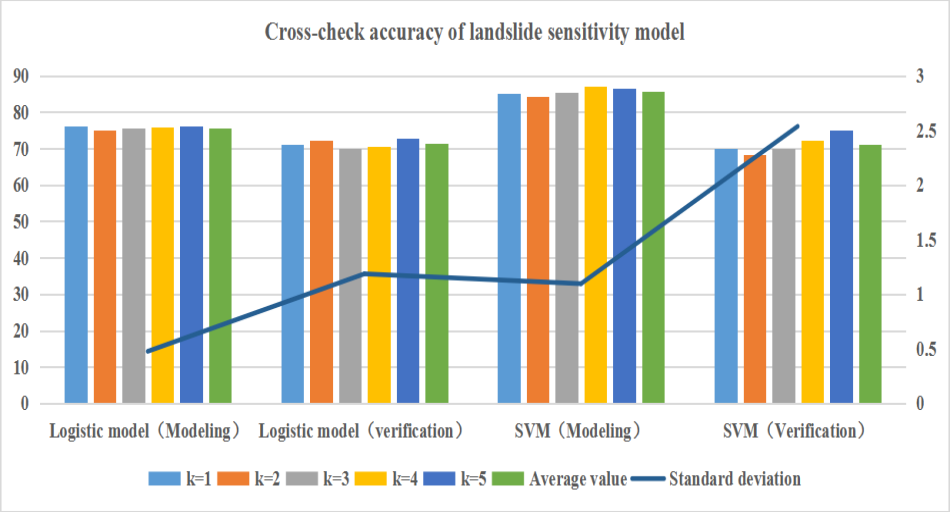


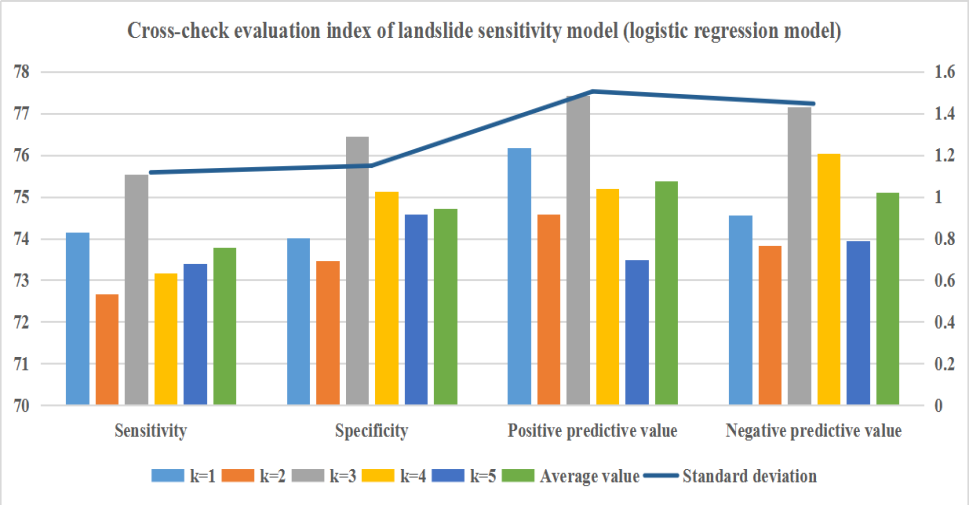
Figure 1. Cross-check accuracy of landslide sensitivity models in areas along the Sichuan-Tibet Railway

From the results in Figure 1, it can be seen that the average accuracy of the logistic regression model and the SVM cross-check is very small. The average accuracy of the modeling stage is 75.722 and 75.65, and the average accuracy of the verification stage is 71.34 and 71.21. Among the two models, the highest fitting accuracy is the support vector machine, and the average accuracy during the modeling and verification phases is 75.722 and 71.34, respectively. However, comparing the modeling stage and the verification stage, it can be found that although the SVM model has a higher fitting accuracy than the other model, which is about 3% higher, during the verification stage, the fitting accuracy is not much higher than the other two models, which is only 0.13% higher. From this perspective, the accuracy levels of the two models are comparable.

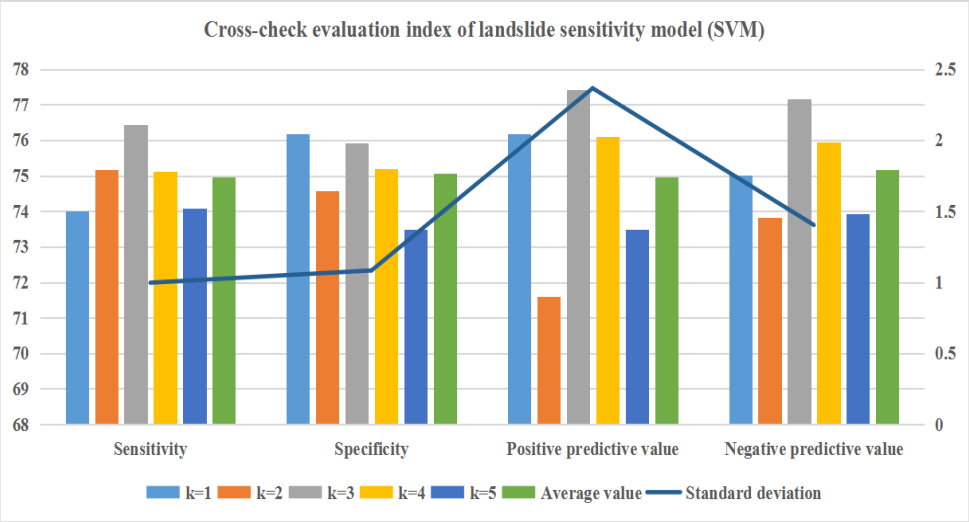
From the standard deviation values in Figure 1, it can be found that in the modeling stage, the stability of both models is better, and the standard deviation values are small, but in the model verification stage, the model stability is worse than the modeling. In summary, in the landslide sensitivity evaluation analysis along the Sichuan-Tibet Railway, the support vector machine is the best of the two models. To further evaluate the model's ability to simulate areas along the Sichuan-Tibet Railway, four evaluation indicators of sensitivity, specificity, positive predictive value, and negative predictive value were calculated based on the confusion matrix. The results are shown in Figure 2.

The results in Figure 2 show that during the evaluation of the landslide sensitivity evaluation model along the Sichuan-Tibet Railway, the standard deviations of the four validation indicators, such as the sensitivity of the logistic regression model, are small during the modeling stage, and the indicators are much more stable. The standard deviation of the 5-fold cross-validation is basically below 2, so it can be seen that the logistic regression model should be stable in terms of models. The results in Figure 2 show that in the evaluation of the landslide sensitivity model along the Sichuan-Tibet Railway, the stability of the four indicators in the verification phase compared to the modeling phase has decreased. The standard deviation of the verification phase is also above 3. The comparison between different models found that for four indicators and five cross-standard deviations, the sensitivity is expressed as support vector machine> logistic regression model; the specificity is expressed as support vector machine> logistic regression model; the positive predictive value is expressed as support vector machine> logistic regression model; the negative predictive value is expressed as support vector machine> logistic regression model, indicating that the stability of the support vector machine on the four indicators has its own advantages.

From the average value of the four indicators, the sensitivity shows as support vector machine > logistic regression model; the specificity shows as support vector machine > logistic regression model; the positive predictive value shows as support vector machine > logistic regression model; the negative predictive value shows as support vector machine > logistic regression model, which shows that the support vector machine model is the best in the evaluation of areas along the Sichuan Tibet railway.



(a) Cross-check of logistic regression model



(b) SVM cross-check

Figure 2. Cross-examination evaluation index of landslide sensitivity model along the Sichuan-Tibet Railway

4.2. Accuracy Analysis of Sensitivity Spatial Model Fitting

In this paper, the ROC curve of the two models is drawn by using the specific value of landslide disaster / non landslide disaster at various points in the modeling and verification stage and the predicted value of the model, and the corresponding AUC value is calculated based on the obtained curve. The results are shown in Figure 3.

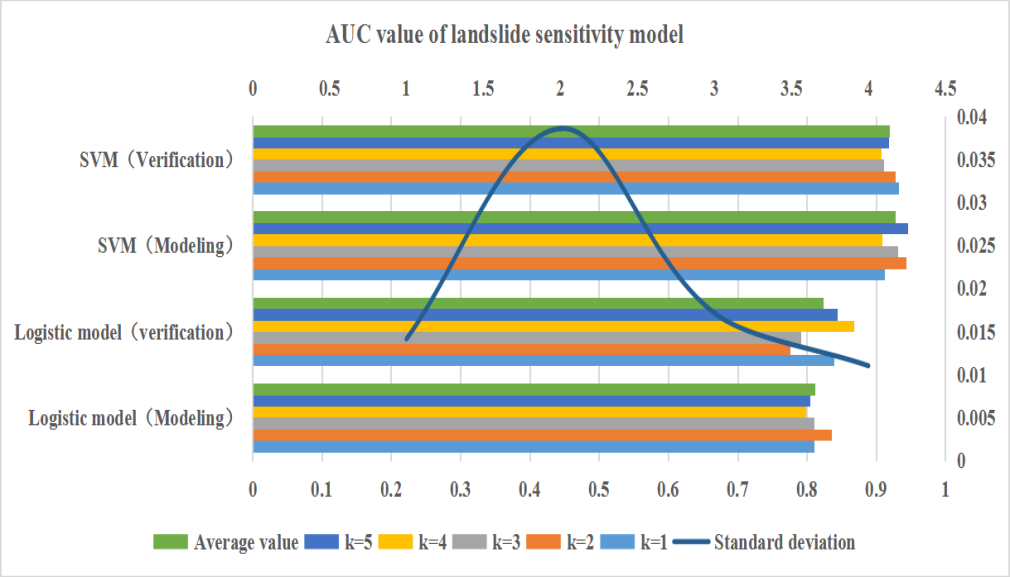


Figure 3. AUC value of landslide sensitivity model along Sichuan Tibet Railway

From the results in Figure 3, it can be seen that the AUC value of the two models is higher than 0.5 and basically above 0.8 regardless of the modeling stage or the verification stage, which proves that the models are valid and can be used to evaluate the landslide sensitivity along the Sichuan-Tibet Railway. Comparing the two specific models, the support vector machine model performs optimally. Its AUC value is above 0.9, which achieves a higher accuracy. Compared with the logistic regression model, this value is 0.111 higher than the logistic regression model (logistic regression model); During the verification phase, it was 0.111 (logistic regression model) higher. From the perspective of the ROC curves and AUC values of the two models in the modeling and verification phases, it also shows that the support vector machine model is the optimal model.

In order to further compare the models, the average value of the relative importance of different cross-validation is calculated, and the distribution map of the factor importance model is drawn.

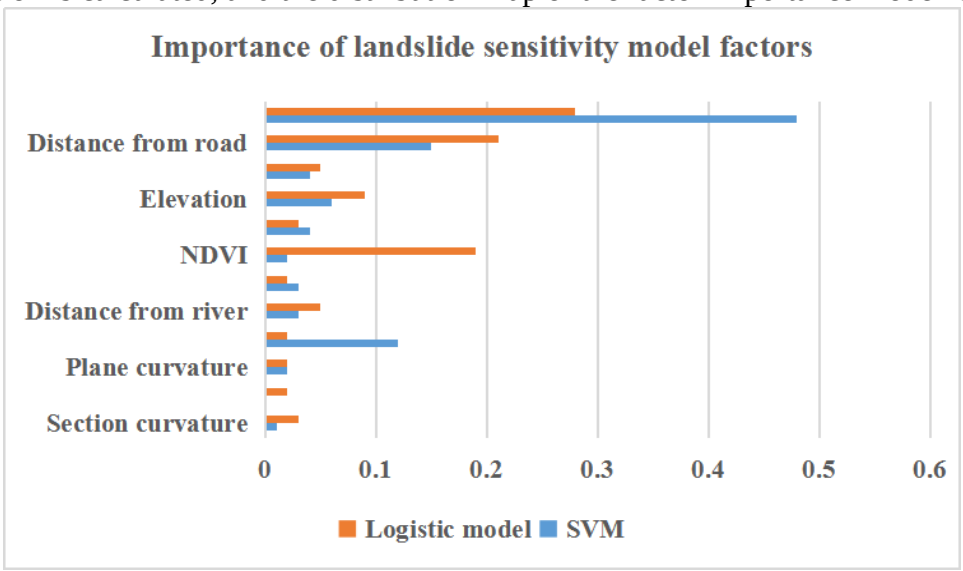


Figure 4. Importance of landslide sensitivity model factors in areas along the Sichuan-Tibet Railway

From the analysis of the relative importance of different model factors, we can see that in the

evaluation of landslide sensitivity along the Sichuan-Tibet Railway, some factors are important to all models, including the two factors of lithology and distance from the road. Both models have higher scores, and their relative importance in the two models are: The relative importance of lithology is 0.2915 (logistic regression model) and 0.4288 (support vector machine); the relative importance of elevation is 0.2519 (logistic regression model) and 0.1022 (support vector machine); the relative importance of distance from road is 0.2108 (Logistic Regression Model) and 0.1771 (Support Vector Machine); the relative importance of the slope is 0.1366 (Logistic Regression Model) and 0.0915 (Support Vector Machine).

4.3. Analysis of Evaluation Results of Sensitivity Space Model

The SVM model and RVM model obtained from the training are used to predict the sample data in the study area, and the landslide sensitivity index distribution in the study area is obtained. Based on the natural breakpoint method, the landslide sensitivity is divided into: VHS (HS), Moderate Sensitivity (MS), Low Sensitivity (LS), and Very Low Sensitivity (VLS). We use the landslide data to generate a landslide interval distribution map as shown in Figure 5.

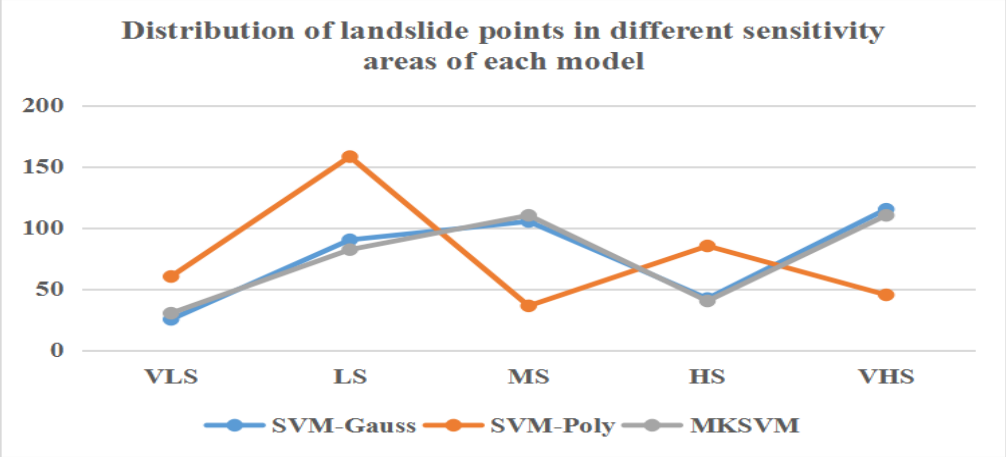


Figure 5. Distribution of landslide points in different sensitive areas of each model

It can be seen from Figure 5 that the number of landslides in the two models of SVM_Gauss and MKSVM is very low in the HS region. In addition, the SVM_Poly model concentrates a large number of landslide points in the LS region. In the distribution of landslide points of the three SVM models, the landslide distribution trend of the MKSVM model gradually increased from VLS to VHS, which is closest to the actual landslide distribution rule. Although the distribution trends of SVM_Poly and SVM_Gauss are not as good as MKSVM, it is also difficult to indicate where is the gap, so the landslide point density needs to be used to further analyze the superiority of the RVM model. The landslide point density calculated in this paper is shown in Table 5.

Table 5. Calculation results of landslide point density in each sensitive area

Sensitive area	Landslide point density		
	SVM-Gauss	SVM-Poly	MKSVM
VLS	0.51	0.67	0.47
LS	0.99	1.64	0.87
MS	1.78	2.05	1.80
HS	2.05	3.46	2.03
VHS	4.55	6.18	4.68

According to the analysis of three SVM models in Table 5, the density value of SVM - poly in VHS is the highest (6.18), but the value of VLS (0.67) is not the lowest. The value of MKSVM in VLS is the lowest (0.47), but it is not the highest in VHS, so it is difficult to describe the best model. Here, the sum of the two high sensitivity regions (VHS and HS) and the two low sensitivity regions (VLS and LS) is calculated respectively to get table 6.

Table 6. Statistical results of the sum of density of landslide points

Sensitive area	Landslide point density (Pcs / 100 km2)		
	SVM-Gauss	SVM-Poly	MKSVM
Sum of VLS and LS	1.65	2.11	1.58
Sum of HS and VHS	5.67	9.01	5.55

It can be seen from table 6 that the effect of SVM model is generally good, and the prediction effect of MKSVM is closest to the real landslide distribution law. The sum of the two low sensitivity is only 1.58 /100 km2, and the sum of the two high sensitivity is 9.01 /100 km2.

5. Conclusion

Based on the spatial analysis function of GIS, this paper analyzes the relationship between the landslide disaster and the influencing factors in the study area. The basic factors to be selected for evaluation include topography, geology, hydrology, surface cover and human activities. In particular, in view of the actual situation of the areas along the Sichuan Tibet railway, the mining disturbance factor is introduced to explore the impact on landslide disaster.

In this paper, a cross-check method is used to construct a landslide sensitivity evaluation model (logical regression model and support vector machine model), and the model is optimized, and the accuracy of different models is quantitatively evaluated. The results of the fitting accuracy of the logistic regression model and the support vector machine model are: the average accuracy in the modeling stage is 75.722 and 75.65, and the average accuracy in the verification stage is 71.34 and 71.21. Through the above comparison, the support vector machine model is the optimal model for landslide sensitivity evaluation along the Sichuan-Tibet Railway.

At the modeling stage, the SVM model has a fitting accuracy of about 3% higher than that of the logistic regression model; at the verification stage, the fitting accuracy is 0.13% higher than that of the logistic regression model; the AUC results show that the SVM model performs optimally, its AUC value is above 0.9, which achieves a higher accuracy. Compared with the logistic regression model, this value is 0.111 higher in the modeling stage and 0.111 higher in the verification stage.

Funding

This article is not supported by any foundation.

Data Availability

Data sharing is not applicable to this article as no new data were created or analysed in this study.

Conflict of Interest

The author states that this article has no conflict of interest.

References

[1] Ruan, Y. K. , Zhan, J. W. , Chen, J. P. , & Li, Y. Y. . (2017). Landslide sensitivity based on k-pso clustering algorithm and entropy method. Dongbei Daxue Xuebao/Journal of Northeastern University, 38(4), 571-575.

[2] D. Qiu, & R. Niu. (2017). Susceptibility analysis of earthquake-induced landslides based on slope units. Journal of Natural Disasters, 26(2), 144-151.

[3] N Gorshkov, S Zhdanova, & M Dobromyslov. (2018). Formation of landslide bodies at numerical calculations of making soil constructions (cut and embankment). IOP Conference Series Materials Science and Engineering, 463(4), 042068.

[4] Songtang He, Daojie Wang, Yingchao Fang, & Huijuan Lan. (2017). Guidelines for integrating ecological and biological engineering technologies for control of severe erosion in mountainous areas – a case study of the xiaojiang river basin, china. International Soil & Water Conservation Research, 5(4), 335-344.

[5] Jewgenij Torizin, Michael Fuchs, Adnan Alam Awan, Ijaz Ahmad, & Ahsan Jamal Khan. (2017). Statistical landslide susceptibility assessment of the mansehra and torghar districts, khyber pakhtunkhwa province, pakistan. Natural Hazards, 89(4), 757-784.

[6] Chao, G. , Bang, L. , Yong-Cai, G. , & Zheng-Wei, Z. . (2017). Five mn force sensor based on fiber bragg grating. Optics & Precision Engineering, 25(4), 857-866.

[7] Saied Pirasteh, & Jonathan Li. (2017). Probabilistic frequency ratio (pfr) model for quality improvement of landslide susceptibility mapping from lidar-derived dems. Geoenvironmental Disasters, 4(1), 19.

[8] Binh Thai Pham, Khabat Khosravi, & Indra Prakash. (2017). Application and comparison of decision tree-based machine learning methods in landside susceptibility assessment at pauri garhwal area, uttarakhand, india. Environmental Processes, 4(3), 711-730.

[9] Ekrem Canli, Bernd Loigge, & Thomas Glade. (2018). Spatially distributed rainfall information and its potential for regional landslide early warning systems. Natural Hazards, 91(Suppl. 1), 103-127.

[10] Zhi Yong Lv, Wenzhong Shi, Xiaokang Zhang, & Jon Atli Benediktsson. (2018). Landslide inventory mapping from bitemporal high-resolution remote sensing images using change detection and multiscale segmentation. IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing, 11(5), 1520-1532.

[11] Michele Calvello, Dario Peduto, & Livia Arena. (2016). Combined use of statistical and dinsar data analyses to define the state of activity of slow-moving landslides. Landslides, 14(2), 1-17.

[12] Lou, Y. , Clark, D. , Marks, P. , Muellerschoen, R. J. , & Wang, C. C. . (2016). Onboard radar processor development for rapid response to natural hazards. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 9(6), 2770-2776.

[13] Zhou Zhou, Xiao-qun Wang, Yu-feng Wei, Jun-hui Shen, & Man Shen. (2019). Simulation study of the void space gas effect on slope instability triggered by an earthquake. Journal of Mountain Science, 16(6), 1300-1317.

[14] L. Wang, Y. Shen, & C.-N. Bai. (2017). Three-dimensional analysis of force and deformation characteristics of oblique crossing of railway tunnel with landslide. Journal of Railway Engineering Society, 34(1), 16-22.

[15] Rwanga, S. S. , & Ndambuki, J. M. . (2017). Accuracy assessment of land use/land cover classification using remote sensing and gis. International Journal of Geosciences, 8(4), 611-622.

[16] Nhat-Duc Hoang, & Dieu Tien Bui. (2018). Spatial prediction of rainfall-induced shallow landslides using gene expression programming integrated with gis: a case study in vietnam. Natural Hazards, 92(3), 1871–1887.

[17] Chen Shengli, Qiao Hongbo, Wang Hongqi, Jiang Jinwei, Ma Jisheng, & Li Qingchang. (2017). Soil nutrient management of tobacco field based on gis and gps. Tobacco Science & Technology,50(3), 23-30.

[18] Francesca Franci, Gabriele Bitelli, Emanuele Mandanici, Diofantos Hadjimitsis, & Athos Agapiou. (2016). Satellite remote sensing and gis-based multi-criteria analysis for flood hazard mapping. Natural Hazards,83(1), 31-51.

[19] Gilbert, J. T. , Macfarlane, W. W. , & Wheaton, J. M. . (2016). The valley bottom extraction tool (v-bet): a gis tool for delineating valley bottoms across entire drainage networks. Computers & Geosciences, 97(December), 1-14.

[20] Hanne Glas, M. Jonckheere, A. Mandal, S. James-Williamson, & Greet Deruyter. (2017). A gis-based tool for flood damage assessment and delineation of a methodology for future risk assessment: case study for annotto bay, jamaica. Natural Hazards, 88(8), 1867-1891.

[21] Aldo Clerici, & Susanna Perego. (2016). A set of grass gis-based shell scripts for the calculation and graphical display of the main morphometric parameters of a river channel. International Journal of Geosciences, 7(7), 135-143.

[22] Szewczyk, M. , Kutorasinski, K. , Wronski, M. , & Florkowski, M. . (2017). Full-maxwell simulations of very fast transients in gis: case study to compare 3d and 2d-axisymmetric models of 1100 kv test set-up. IEEE Transactions on Power Delivery, 32(2), 733-739.

[23] D. Thinh Nguyen, Iskhaq Iskandar, & Son Ho. (2016). Land cover change and the co2 stock in the palembang city, indonesia: a study using remote sensing, gis technique and lumens. Egyptian Journal of Remote Sensing & Space Science, 19(2), 313-321.

[24] Abdel Rahman Al-Shabeeb, Rida Al-Adamat, & Atef Mashagbah. (2016). Ahp with gis for a preliminary site selection of wind turbines in the north west of jordan. International Journal of Geosciences, 07(10), 1208-1221.

[25] Raju Thapa, Srimanta Gupta, D. V. Reddy, & Harjeet Kaur. (2017). An evaluation of irrigation water suitability in the dwarka river basin through the use of gis-based modelling. Environmental Earth Sciences, 76(14), 471.

[26] Zou, J. , Zhang, W. , & Yang, Y. . (2016). Evaluation of water resources system vulnerability in southern hilly rural region based on the gis/rs take hengyang basin as an example. Scientia Geographica Sinica, 34(8), 1010-1017.

[27] James, P. , Jankowska, M. , Marx, C. , Hart, J. E. , Berrigan, D. , & Kerr, J. , et al. (2016). "spatial energetics": integrating data from gps, accelerometry, and gis to address obesity and inactivity. American Journal of Preventive Medicine, 51(5), 792-800.