DISCRETIZATION AND FEATURE SELECTION BASED ON GEODETECTOR IN LANDSLIDES SUSCEPTIBILITY MAPPING

Yimo Liu (1)(2), Wanchang Zhang (1)(2)(*), Xiaosong Duo (3), Xiaopeng Zhang (3), Yaning Yi (1)(2)

[1] Key Laboratory of Digital Earth Science, Aerospace Information Research Institute, Chinese Academy of Sciences, No.9 Dengzhuang South Road, Haidian District, Beijing 100094, China
[2] University of Chinese Academy of Sciences, 19 Yuquan Road, Shijingshan District, Beijing, 100049, China
[3] 514 Brigade of North China Geological Exploration Bureau, Chengde, Hebei, 067000, China
Email: zhangwc@radi.ac.cn; liuym@aircas.ac.cn; cds514@126.com
[*] Correspondence: Wanchang Zhang, zhangwc@radi.ac.cn; Tel.: +86-10-8217-8131

KEY WORDS: Landslide causative factors, q-statistic, spatial stratified heterogeneity, multi-layer perceptron

ABSTRACT: Landslide susceptibility mapping is an effective method and key tool for land use planning, management/assessment, and Geo-disaster risk mitigation. The present study is aimed to examine further the integration of terrain data in landslide analyses and to contribute to the recognition of locations where possible Geo-disasters mostly likely take place by discretizing and selecting discrete landslide causative factors with Geo-detector q-statistic model to optimize landslide susceptibility mapping. Firstly, Geo-detector was selected to determine the best subsets of landslide causative factors used for landslide susceptibility zonation in different level of category with different discretization strategies. Subsequently, Landslide susceptibility model were built by means of multi-layer perceptron classifier (MLP) with different number of selected landslide causative factors sorted based on q-statistics obtained from Geo-detector. Finally, area under the receive operating characteristic curve (AUC) was adopted for evaluating model performances as well as for validation of landslide susceptibility mapping. The location chosen for the study was a typical landslide triggered area by the 2018 Mw6.6 Tomakomai Earthquake, Japan, which repeatedly suffered from landslides after heavy rainfall. Results from this study revealed that the ranking orders of q statistics for causative factors determined by Geo-detector under different discretization strategies are basically the same. The highest AUC achieved by the Geo-Detector-MLP using the top ranked eight factors of higher q-statistics was about 0.86, which is about 0.05 higher than that obtained by the traditional MLP, indicating that the selected optimal causative factors based on q statistics of Geo-detector can effectively improve the prediction accuracy of landslide susceptibility mapping.

1. INTRODUCTION

Landslides susceptibility mapping is an important tool to optimize land use planning and policy to reduce the damage of landslide to public property, infrastructure and people's lives (Bui D T, 2016). However, knowledge on the formation of landslides is related to many complex factors such as topography, geology, hydrology, ecology and human activities, how to make rational use of these parameters to generate landslide susceptibility mapping with high reliability is still a challenge in the field of applications.

Landslide susceptibility is determined by qualitative and quantitative analysis of the causative factors obtained in the former disaster area (Jebur M N, 2014). Sufficient causative factors being considered in the model approaches are essential for high-accuracy landslide

susceptibility mapping (Jebur M N, 2014), however, the existence of redundant information may introduce noises that may bring restrictions in accuracy. To access predictive ability of landslide causative factors and exclude the factors with low information content, many studies have been conducted, methods and techniques can be broadly divided into three categories: expert knowledge method, statistical methods, geostatistical methods (Bui D T, 2016; Jebur M N, 2014; Martinez-Alvarez F, 2013; Yang J, 2018) in the literature. The expert knowledge-based approach is too subjective, it needs professional talents who master the prior knowledge of the area, the physical mechanism of landslide and the selection criteria of causative factors. The characteristics of causative factors vary region by region, in this regard, statistical and geostatistical methods own better portability for different areas and data mining ability for various causative factors. Furthermore, statistical methods, such as information gain ratio (Bui D T, 2016; Martinez-Alvarez F, 2013), certainty factor models (Jie D, 2015), only take attribute information into account but ignore spatial pattern characteristic, resulting in low accuracy and big uncertainties (Yang J, 2018). Taking spatial autocorrelation and spatial stratified heterogeneity (SSH) into account, geostatistical method, such as Geo-detector, can achieve better quantitative measurement between independent and dependent variables (Yang J, 2018). The present study is aimed to evaluate the optimization effect of Geo-detector on landslide causative factors. Discretization and feature selection were adopted by using q-statistic in Geo-detector that has already been used in many fields of social and natural sciences but rarely in landslide susceptibility mapping to evaluate predictive performance of the selected causative factors. Then MLP was applied to seven sub-datasets generated from original causative factors with different number of independent variables. Finally, AUC was adopted for evaluating model performances as well as for validation of landslide susceptibility mapping in the selected study sites.

2. METHODOLOGY AND PRINCIPLES

2.1 Geo-Detector

The quality and the quantity of the data used as causative factors are vital factors which determine the reliability of landslide model (Bui D T, 2016; Jebur M N, 2014; Yang J, 2018). Evaluation and optimization of the predictive capability of causative factors are essential to improve the accuracy of model.

Geo-detector is a tool capable of detecting and utilizing SSH of geographical phenomenon, determining and evaluating the degree and rank order of its causative factors on contribution of those SSH. Such approach has been applied in many fields of natural and social sciences. It is worthwhile to mention that an assumption behand Geo-detector is: the greater the influence of an independent causative factor on a dependent consequence, the more similar their spatial distribution will be (Wang J, 2017). The q-statistic in Geo-detector that used to measure spatial similarity is usually calculated as:

$$q = 1 - \frac{\sum_1^L N_i \sigma_i^2}{N\sigma^2} \quad (1)$$

where i=1, 2, …, L, represents the stratum i of a causative factor. N_i and N represent the unit numbers of stratum i and the whole region of interest, respectively. σ_i^2 and σ^2 represent the variance of stratum i and the whole area, respectively. And the second term in the right side of equation above denotes the ratio of Within Sum of Squares (WSS) over the Total Sum of Squares (TSS). According to different basis of stratification, there are two meanings of q-statistics: SSH detection, and factor detection. In SSH detection, stratification is based on the causative factor itself, and the more obvious the SSH is, the larger q-statistic will be; in factor detection,

stratification of dependent consequence Y is based on independent causative factor X, so larger q-statistic represents stronger explanatory power of independent causative factor X on dependent consequence Y (Wang J, 2017).

2.2 Discretization of Continuous Causative Factors

The numerical independent causative factors were discretized into the category data as inputs to Geo-Detector. Common discretization strategies include but are not limited to uniform-interval-method, quantile-method and k-means- method, correspondingly.
50 landslide samples and 50 non-landslide samples were extracted as an example of the discretization effect of different methods in present study, as shown in Figure 1. The characteristic of numerical relationship of the samples selected was fully retained with uniform-interval-method, while the distribution characteristics of the samples selected were found better expressed when using k-means as discretization strategy, and quantile-method guaranteed the approximate number of samples within each bin.

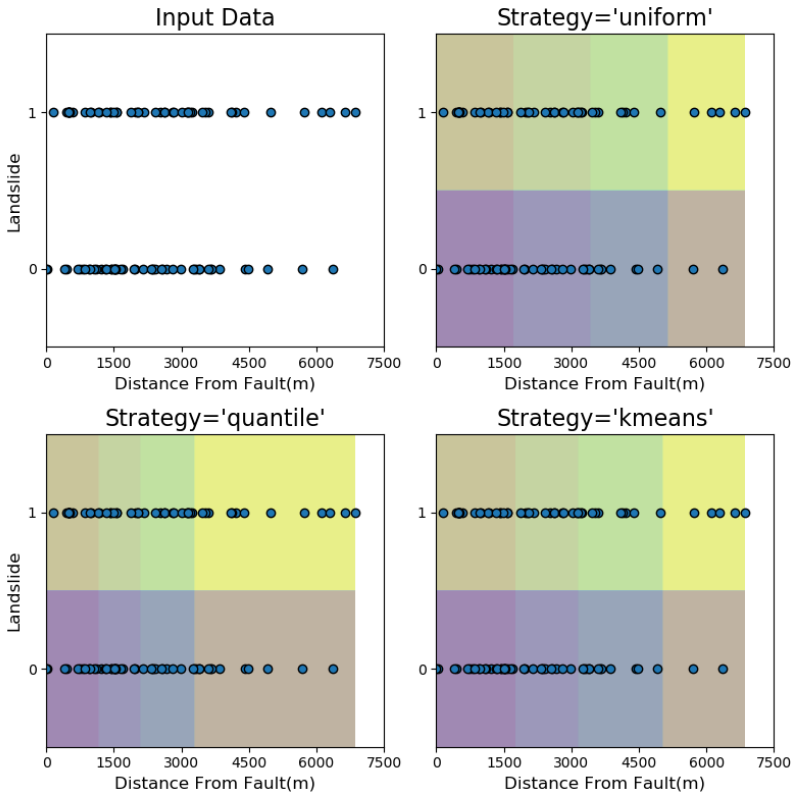

Figure 1. Discretization Results Obtained with Different Strategies

2.3 Multi-layer Perceptron Classifier

MLP is a supervised learning algorithm which consists of an input layer, an output layer, and one or more hidden layers that are non-linear between input layer and output layer (Kavzoglu T, 2003). Connection weight matrixes between every two adjacent layers were used to calculate the values of neurons in the current layer based on those of previous layer and modified ground on the difference between output consequences and the ideal results. Using back-propagation algorithm, MLP learnt from a non-linear function by training on a given set of landslide causative factors and landslide binary labels for classification.

3. EXPERIMENT

In this study, we conducted susceptibility mapping for a selected experimental site using Geo-Detector-MLP model for partial typical landslides triggered area by the 2018 Mw6.6 Tomakomai Earthquake, Japan, which repeatedly suffered from landslides after heavy rainfall (Shao X, 2019).

3.1 Spatial Dataset

The detailed and reliable dataset in GIS-based format such as location and scale of the earthquake-triggered landslides in the study area are derived from https://www.jma.go.jp/. The landslide index in the landslide susceptibility mapping was separated into two categories: value of 0 stands for the area of non-landslide, while value of 1 for area of landslide.
According to the prior knowledge on landslide mechanism and other relevant information, the appropriate initial set of causative factors was prepared by using the data from conventional field survey and remote sensing (Bui D T, 2016; Jebur M N, 2014; Yi Y, 2019). The causative factors selected in this study could be divided into four categories: topographic (elevation, slope, aspect, plan curvature and profile curvature), geological (lithology, distance to faults, and seismic peak ground acceleration (PGA)), hydrology (topographic wetness index (TWI)), and ecology (distance to roads, distance to river and landcover). Details of resolution, data formats and sources of dependent and independent variables were listed in Table 1.

Table 1. Spatial Dataset Introduction

Factor Type	**Name**	**Data type**	**Source**
	Landslide	Polygon	https://www.jma.go.jp/
Topographic	Elevation	Raster	http://earthexplorer.usgs.gov/
	Slope	Raster	Extracted from DEM,30m
	Aspect	Raster	Extracted from DEM,30m
	Plan curvature	Raster	Extracted from DEM,30m
	Profile curvature	Raster	Extracted from DEM,30m
Geological	Lithology	Polygon	https://www.gsj.jp/
	Faults	Line	https://www.gsj.jp/
	PGA	Polygon	http://earthquake.usgs.gov/
Hydrology	TWI	Raster	Extracted from DEM,30m
Ecology	Road	Line	https://www.gsj.jp/
	River	Line	https://www.gsj.jp/
	Landcover	Raster	https://data.ess.tsinghua.edu.cn/

3.2 Discretization and Geo-Detector

The parameter combinations for three discretization strategies including “uniform”, “quantile” and

"kmeans", and different values of the bin ranging from 5 to 15, were used for discretization of numerical variables. 2000 samples were randomly selected at ratio of 1:1 for landslide and non-landslide samples. After that, q-statistics of each classification results were evaluated by using Geo-Detector on the sample dataset obtained from previous processes and was visualized as shown in Figure 2 for illustration of category variables, i.e. landcover, lithology and PGA with fixed q-statistics. The results suggested that the order of q-statistics of independent causative factors, except for the lithology, always took the top ranking but the aspect always took the bottom ranking, varied with the value of bin in all three discretization methods. Results from this study revealed that the ranking orders of q-statistics for causative factors determined by Geo-detector under different discretization strategies are basically the same.

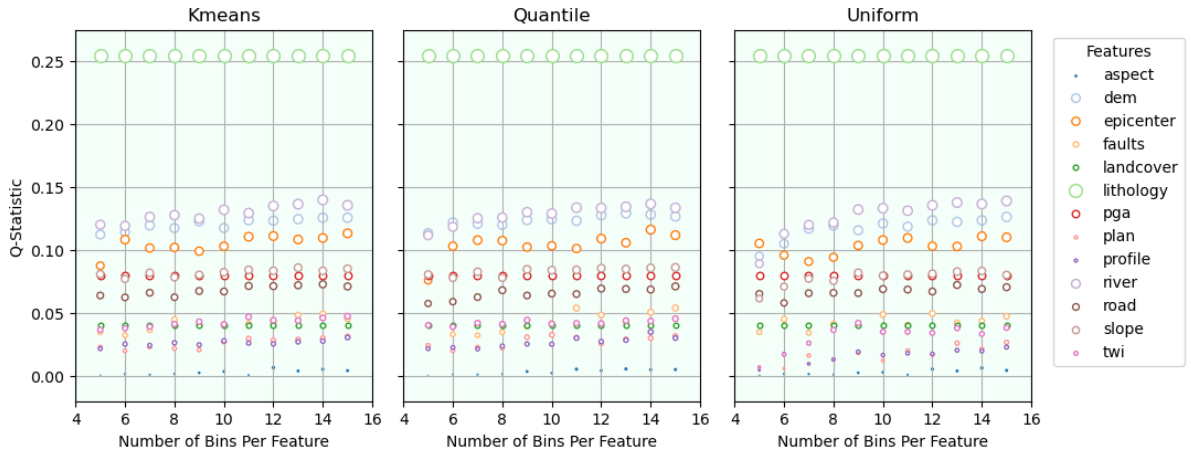

Figure 2. Q-statistics of Causative Factors on Landslide Distribution with Different Discretization Parameters

3.3 Feature Selection and MLP Model

The discretization methods with highest q-statistic of each feature were used to generate the dataset for training and evaluation of landslide susceptibility models, and their corresponding q-statistics were shown in Figure 3, based on the feature selection being implemented.

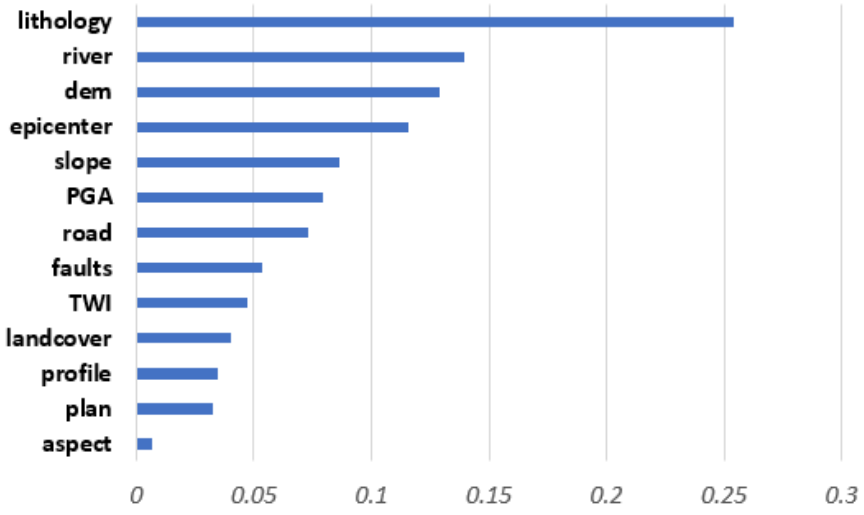

Figure 3. Q-statistics of Causative factors on landslide distribution with optimal discretization parameters

The number of causative factors affects the precision of the generated landslide susceptibility mapping. Seven datasets containing different number of features, ranging from 7 to 13, were generated based on the ranking of the q-statistics corresponding to each feature. Subsequently, seven landslide models were built using MLP. The performances of the obtained models were evaluated by 10-fold cross-validation by using AUC of ROC as shown in Figure 4. The result revealed that the model obtained on the dataset generated from eight selected causative factors, including lithology, the distance to river, dem, distance to epicenter, slope, PGA, distance to road, and distance to faults, was optimal with the highest AUC of 0.86. The lowest AUC was 0.81 with only one feature, aspect, being discarded. Between 7 to 12, as the number of features utilized were increased, the AUC of the generated landslide susceptibility model decreased gradually. The result implied that selection of proper features is very essential and this could affect the precision of the landslide susceptibility model in certain levels.

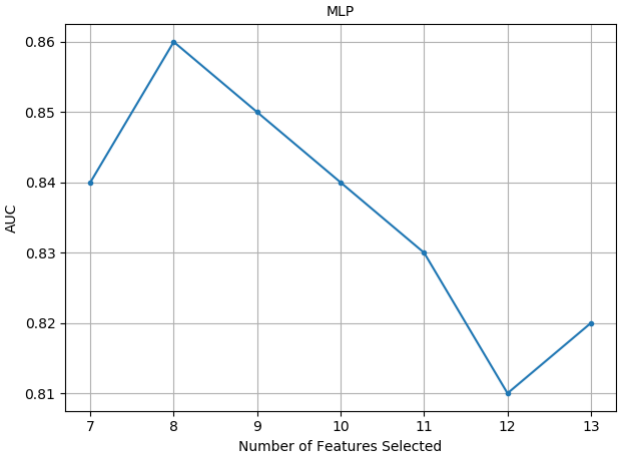

Figure 4. AUC of MLP Models Trained on Datasets with Different Number of Causative Factors

4. CONCLUSION

In this study, we implemented and evaluated the Geo-Detector-MLP model on landslide susceptibility mapping based on the triggered landslides inventory of 2018 Mw6.6 Earthquake Tomakomai, Japan. Discretization of landslide causative factors whose datatype is continuous were evaluated and optimized by Geo-Detector. Feature selection was implemented according to the q-statistics obtained by Geo-Detector. Classification models were built by using MLP, and were trained and evaluated on the features selected. All the experimental results indicated that discretization and selection of the optimal causative factors based on q-statistics of Geo-detector can effectively improve the prediction accuracy of landslide susceptibility mapping. The approach proposed in this study can provide a reliable data preprocessing method for determining the optimal landslide causative factors.

5. REFERENCES

Bui D T, Tran A T, Klemp H, et al, 2016. Spatial prediction models for shallow landslide hazards: a comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree. Landslides, 13(2), pp. 361-378.
Jebur M N, Pradhan B, Tehrany M S, 2014. Optimization of landslide conditioning factors using very high-resolution airborne laser scanning (LiDAR) data at catchment scale. Remote Sensing

of Environment, 152, pp. 150-165.
Jie D, Dieu T B, Ali P Y, et al, 2015. Optimization of Causative Factors for Landslide Susceptibility Evaluation Using Remote Sensing and GIS Data in Parts of Niigata, Japan. PLoS ONE, 10(7), pp. e0133262.
Kavzoglu T, Mather P M, 2003. The use of backpropagating artificial neural networks in land cover classification. International Journal of Remote Sensing, 24(23), pp. 4907-4938.
Martinez-Alvarez F, Reyes J, Morales-Esteban A, et al, 2013. Determining the best set of seismicity indicators to predict earthquakes. Two case studies: Chile and the Iberian Peninsula. Knowledge-Based Systems. 50(sep.), pp.198-210.
Shao X, Ma S, Xu C, et al, 2019. Planet Image-Based Inventorying and Machine Learning-Based Susceptibility Mapping for the Landslides Triggered by the 2018 Mw6.6 Tomakomai, Japan Earthquake. Remote Sensing, 11(8).
Wang J, Xu C, 2017. Geodetector: Principle and prospective. Acta Geographica Sinica.
Yang J, Song C, Yang Y, et al, 2018. New method for landslide susceptibility mapping supported by spatial logistic regression and GeoDetector: A case study of Duwen Highway Basin, Sichuan Province, China. Geomorphology, 324. 62-71.
Yi Y, Zhang Z, Zhang W, et al, 2019. GIS-based earthquake-triggered-landslide susceptibility mapping with an integrated weighted index model in Jiuzhaigou region of Sichuan Province, China. Natural Hazards and Earth System ences, 19(9), pp. 1973-1988.