Proc. of the 8th International Symposium on Geotechnical Safety and Risk (ISGSR)
Edited by Jinsong Huang, D.V. Griffiths, Shui-Hua Jiang, Anna Giacomini and Richard Kelly

doi: 10.3850/978-981-18-5182-7_11-017-cd

ISGSR 2022

Introducing a Pattern-Based Approach for Landslide Susceptibility Prediction

Chenxu Su[1], Cong Dai[1], Bijiao Wang[1], Yunhong Lv[1] and Shuai Zhang[*1]

[1] Department of Civil Engineering, Zhejiang University, Hangzhou, 310058, China.
E-mail: suchenxu@zju.edu.cn, 149834307@qq.com, wangbijiao@zju.edu.cn, Lvyunhong@zju.edu.cn, zhangshuaiqj@zju.edu.cn

Abstract: Machine learning (ML) models are extensively used in data-driven landslide susceptibility prediction (LSP). Dataset used in LSP with ML containing positive (landslide) samples and negative (non-landslide) samples, while the spatial biases of non-landslide samples for LSP are frequently ignored. The main objective of this study is to develop a pattern-based approach that properly tackles the spatial biases of non-landslide samples combining two models, i.e. balanced iterative reducing and clustering using hierarchies (BIRCH) and Random Forest (RF). In this study, BIRCH is employed to select four types of non-landslide samples, representing four spatial patterns. In the meanwhile, another set of non-landslide samples is randomly selected to serve as control. RF model is trained to calculate the susceptibility index using these five types of non-landslide samples along with landslide samples derived from landslide inventory, producing five LSP scenarios. Results indicate that the pattern-based approach offers an effective way to find the non-landslide samples and provide sufficient and reliable spatial patterns, and therefore proves itself as a better solution to the LSP.

Keywords: Landslide susceptibility; Machine learning; Risk; Remote sensing.

1 Introduction

A landslide can be characterized as a movement of massive rock and soil denoting slope-forming materials move downward and outward a slope (Zhang et al., 2017). It is obvious that landslides represent a major reason for fatalities and enormous property damage in mountain areas. The 12 May 2008 Wenchuan earthquake triggered approximately 60,000 landslides resulting in significant geological erosion, with more than 20,000 deaths caused by earthquake-induced landslides. Landslide susceptibility is used to illustrate the likelihood of landslide occurrences over space based on environmental predisposing factors (Pham et al., 2018). Reliable, effective and robust landslide susceptibility prediction (LSP) is crucial for mitigating and reducing landslide hazards. Therefore, LSP is considered of predictive importance for landslide-prone areas.

A great number of quantitative appraisal models that can be classified into five categories, namely, inventory-based models, statistical models, machine learning (ML) models, heuristic models and deterministic models are adopted in LSP (Huang et al., 2017). Among those models, the ML models have gradually gained in popularity over recent decades, especially in large-scale mapping. The ML models fall into two primary categories including supervised learning (SL) and unsupervised learning (UL). SL algorithms defined by its use of labeled datasets including support vector machine, logistic regression, artificial neural network and random forest, are widely developed in landslide-prone areas for LSP. UL uses ML algorithms to cluster unlabeled datasets, which offers approaches to find inherent patterns or data clusters without given labels. This study combines the characteristic of the SL and UL algorithms to introduce a pattern-based approach for LSP, which is developed to address the issue of sampling non-landslide samples in LSP. We aim to: (i) find reasonable solutions for sampling non-landside pixels utilizing UL models, (ii) propose a methodological approach that combines SL models and UL models for LSP.

2 Study Area

The study area is located in Yingxiu Town, the epicenter of the Wenchuan earthquake, which is in the transition zone between Sichuan Basin and Qinghai-Tibet Plateau. Province Road 303 (PR303) stretches along the Yuzixi River that is bounded by terrains rugged with steep slopes above 40° in many places presenting significant river incision with "V"-shaped valleys. Widespread co-seismic landslides and post-seismic landslides were triggered on both sides of the PR303. On May 12, 2008, the Wenchuan earthquake induced widespread shallow landslides in the study area, and consequently led to intensively distributed loose deposits. A total of 305 landslides were investigated in detail through field work and satellite images (Figures 1-2). The majority of them are concentrated on the hanging wall of the Yingxiu-Beichuan fault on account of the "hanging wall" effect (Xu et al., 2011). We calculated the covering areas of these landslides delineated with ArcGIS platform, which have an average covering area of 78,763 m^2 and the largest one is in the Dayingou Ravine, located at elevations between 2,300 m and 3,190 m. The geomorphology of the study area is mainly controlled by complex faults, folds and structural fissures. The altitude of the mountains ranges from 1,000 m to 3,000 m showing high relief amplitude

and the slope mainly ranges between 50° and 70°. The bedrock, for the most part, consists of magmatic rock and metamorphic rock. Lithology mainly features diorite, medium fine-grained granite and alluvium. The hard rocks account for the phenomenon that the topography of the study area is quite steep and prone to rockfalls, landslides and debris flows.

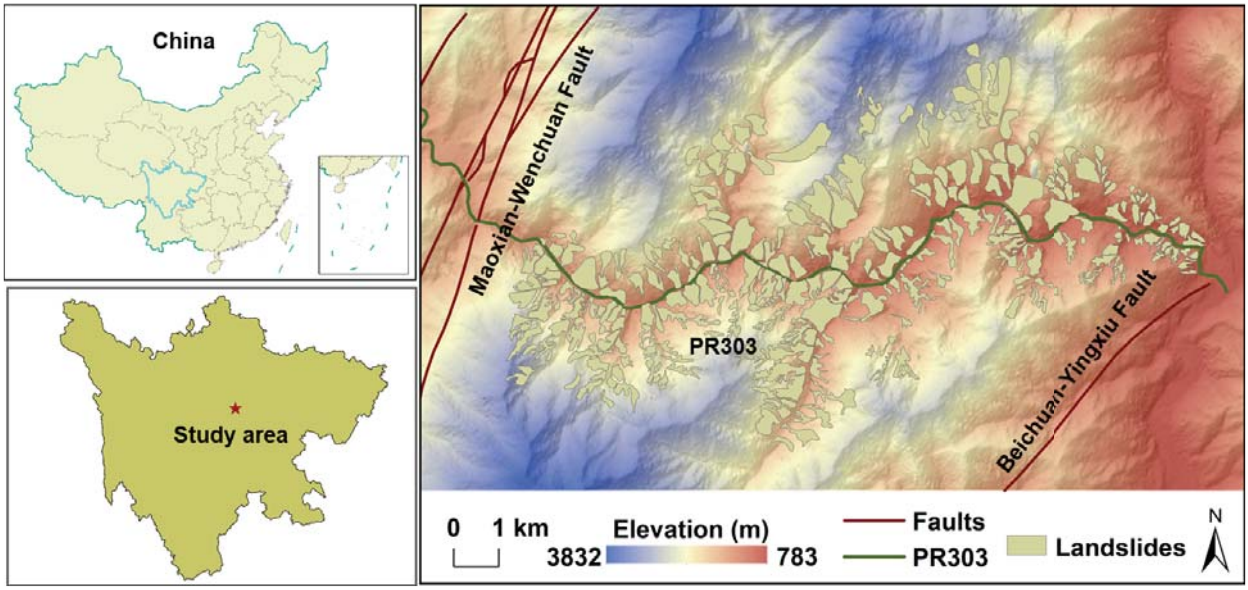

Figure 1. Location of the study area and elevation map with landslide inventory.

3 Methods

We first prepared the landslide inventory through field investigation and remote sensing images. Six conditioning factors for LSP were extracted and discretized using natural breaks method. Then the data was used to power model construction. The UL algorithm, Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH), was introduced to sample the non-landslide pixels. Moreover, a set of randomly selected non-landslide samples are used here to serve as control. Random Forest (RF) models feed on these samples were constructed and the predictive accuracy is compared and discussed according to the ROC curves.

3.1 Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH)

Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) is a clustering algorithm, a highly effective method which clusters datasets, particularly large datasets by first producing a brief summary of the datasets that is called the clustering feature (CF) tree and can represent sufficient patterns (Charest & Plante, 2014). CF is defined as the triple:

$$CF = (N, \boldsymbol{LS}, SS) \tag{1}$$

where N denotes the number of data points, $\boldsymbol{LS}$ is the liner sum and SS is the square sum of the data points. Then the summary statistic, i.e. CF tree, is then clustered instead of clustering the original dataset. For a given dataset, the BIRCH can make clustering decisions without considering all data points and existing clusters. It incrementally and dynamically clusters multi-dimensional metric data points in an attempt to derive the high quality clustering. And clusters number K is not necessarily for BIRCH as it can be determined according to CF.

3.2 Random Forest (RF)

Random forest (RF), a well-known ML model utilizes the methodology of classification and regression tree (CART) algorithm, which takes advantage of the bootstrap aggregating algorithm which randomly extracts samples from original datasets to produce new discrepant training datasets with pattern spaces constructed by the algorithm. Independent training samples are used to establish decision trees that are assembled to form a RF model (Wang et al., 2021). Metrics including Gini impurity, information gain, and mean square error (MSE), can be used to assess the performance of the split and the final results of the RF should be determined based on the voting or averaging of these CART. Moreover, the remaining dataset called out of bag (OOB) can be used to improve the performance of the RF model to avoid overfitting.

4 Materials

4.1 Dataset preparation

According to previous researches and the characteristics of the study area, we selected six landslide-related conditioning factors, including elevation, slope angle, aspect, lithology, topographic wetness index (TWI), normalized difference vegetation index (NDVI). The data sources used to extract landslide-related environmental factors mainly include the following: (1) Digital Elevation Model (DEM); (2) Landsat TM 8 Remote sensing images and (3) Lithology distribution map. The ALOS-PALSAR DEM with a spatial resolution of 12.5×12.5 m was used to extract topographic factors including slope, aspect and topographic wetness index (TWI). Landsat TM 8 satellite images with 30×30 m spatial resolution were used to extract normalized difference vegetation index (NDVI). The lithology maps at a scale of 1:100,000 were collected from the local Land and Resources Bureau. All the landslide-related conditioning factors were converted into raster format with a spatial resolution of 12.5×12.5 m.

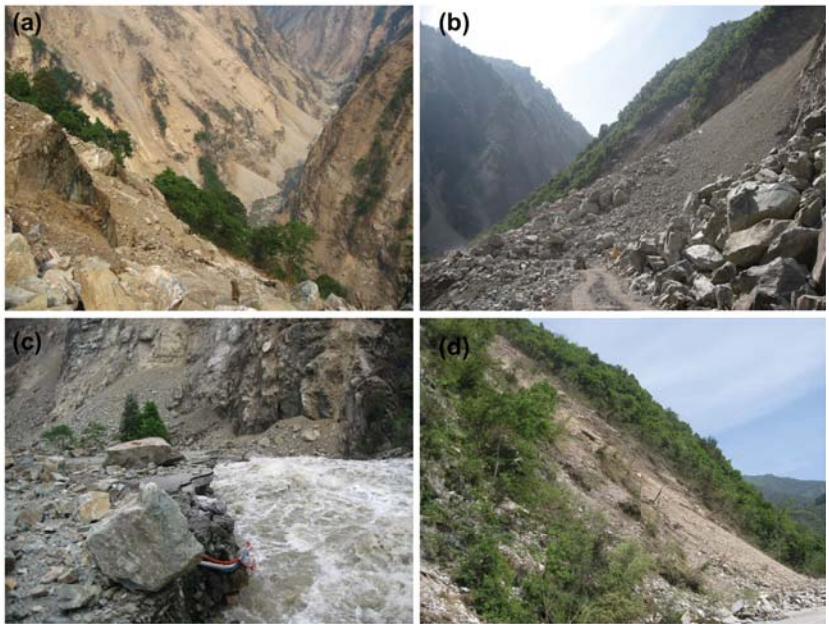

Figure 2. Damage of the earthquake to the study area: (a) overall view; (b) and (c) damaged road; (d) shallow landslide.

4.2 landslide-related conditioning factors

The elevation is usually linked to anthropic engineering activities, and linked to vegetation characteristics. The elevation map is the primary data source of topographic factors and is the most significant factor concerning landslide occurrences. It was classified into eight categories using natural breaks method, i.e., (783–1,145, 1,145–1,479, 1,479–1,793, 1,793–2,103, 2,103–2,423, 2,423–2,777, 2,777–3,174 and 3,174–3,832 m). Slope influences the stability of a slope regime as the gravitational forces increase with an increasing slope angle, and therefore making it prone to landslide. Owing to the mountainous characteristic of the study area, it was significant for determining the landslide-prone areas. The slope map of the study area was categorized into eight subclasses (0–15, 15–25, 25–32, 32–39, 39–44, 44–51, 51–60 and 60–88°). It is well acknowledged that earth surface characterized by a lower elevation and a smoother slope, appears to have lower landslide frequency compared with high elevation areas with steep slopes (Hong et al., 2020).

The aspect, to a certain extent, controls the response of the slope regime to climate conditions such as precipitation and wind direction. The aspect is defined as the azimuth angle of projection of slope normal on horizontal plane, and was classified into nine categories including flat, north, northeast, east, southeast, south, southwest, west and northwest. The lithology is one of the most important factors showing a direct effect on the occurrences of landslides as the lithological largely controls the physical and mechanical properties of bedrock and deposits. The strata in the study area can be classified into 6 groups, i.e., schist, phyllite, alluvium, granite, limestone and diorite.

The topographic wetness index (TWI) can quantitatively represent the topography and condition of soil moisture in a watershed, which directly influences the slope stability. The TWI was classified into eight classes (0–3.4, 3.4–4.3, 4.3–5.1, 5.1–6.1, 6.1–7.3, 7.3–8.8, 8.8–10.8 and 10.8–18.9). The normalized difference vegetation index (NDVI) is a widely used factor to quantitatively calculate the relationship vegetation on landslides. The NDVI was divided into eight subclasses (-0.33–0.03, 0.03–0.14, 0.14–0.25, 0.25–0.37, 0.37–0.51, 0.51–0.64, 0.64–0.76 and 0.76–0.91).

5 Results

5.1 LSP using the single RF model

We first used randomly selected non-landslide samples to perform LSP. A total of 1,265,416 grid cells with six landslide-related conditioning factors forms a 1,265,416*6 matrix used as the dataset for modeling. 146,386 recorded landslide grid cells and the same number of randomly selected non-landslide samples were divided into 7:3 ratios as two parts (i.e., training set and testing set). Moreover, the labels of recorded landslide grid cells were set to 1 and the labels of non-landslide samples were set to 0.

For the implementation of the RF model, the number of trees in RF and the number of variables considered for each tree are the most important parameters. In this study, the number of trees was set to 100 and the variables considered for each tree was set to 3 through grid research, and other parameters were left at the default setting. The landslide susceptibility map using the trained RF model is shown in Figure 3. The susceptibility was divided into five levels: very high (22.98%), high (21.81%), moderate (22.59%), low (20.65%) and very low (11.97%). The landslide susceptibility index predicted by the single RF model, on the whole, can be approximately regarded as the uniform distribution. It can be seen from Figure 3 that the very high and high susceptibility zones are chiefly distributed along the valley slopes well in line with the landslide inventory.

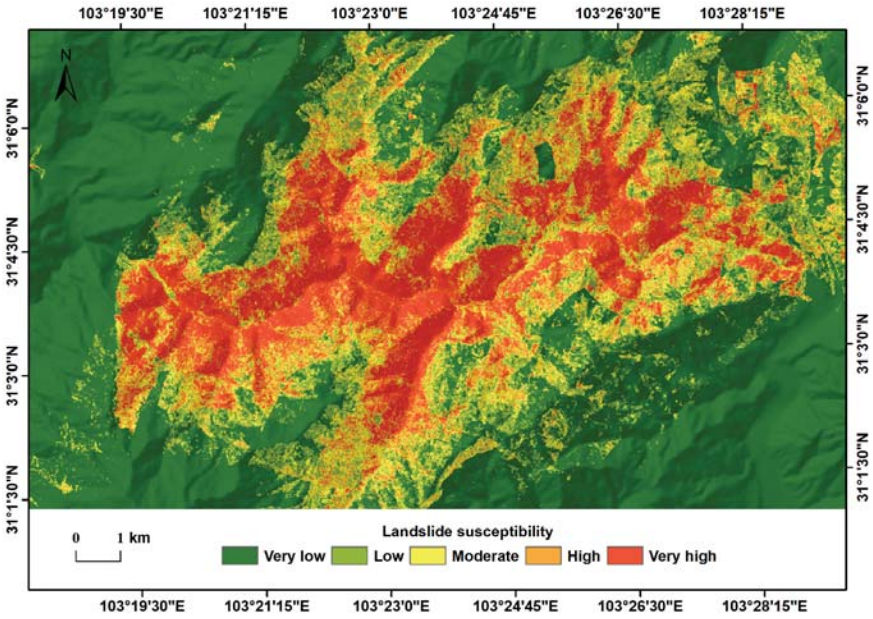

Figure 3. LSP using single RF model.

5.2 LSP using BIRCH-RF model.

The UL model, Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) model was used to process unclassified datasets into classes represented by their susceptibility patterns, from which the non-landslide samples were selected. The 1,265,416*6 matrix was used as the input dataset while the labels were no longer needed. All grid cells were categorized into five groups that represent five susceptibility classes. The theory of BIRCH is introduced in Section 3. Clusters number K was set to five due to the five susceptibility classes considered. Without prior knowledge of landslide inventory, a landslide susceptibility map can be automatically produced utilizing BIRCH as shown in Figure 4. It can be seen that the spatial patterns of the landslide susceptibility are well distinguished, therefore, effective clustering is achieved through BIRCH model. Moreover, the very high and high susceptibility areas are well recognized that the majority of the recorded landslide pixels fall into the very high and high susceptible areas while only few recorded landslide samples fall into the very low and low areas.

Table 1. Four sampling selections of non-landslide grid cells.

Sampling selection	Description of selections[a]
Selection A	Sampling from the very low, low, moderate and high susceptible regions.
Selection B	Sampling from the very low, low and moderate susceptible regions.
Selection C	Sampling from the very low and low susceptible regions.
Selection D	Sampling from the very low susceptible regions.

[a] The A, B, C and D sampling selections are different methods to sample non-landslide grid cells using landslide susceptibility maps generated by BIRCH model.

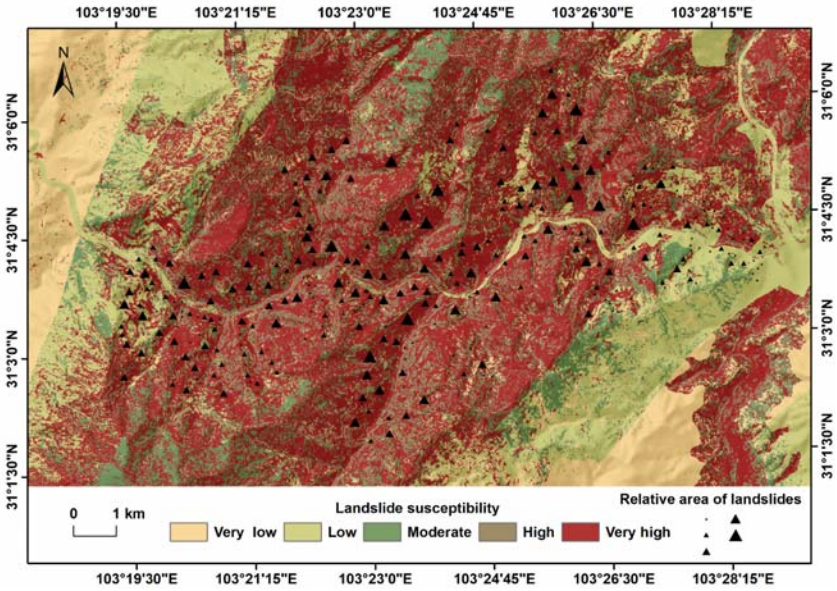

Figure 4. LSP using BIRCH model.

The next step was to select non-landslide samples from the landslide susceptibility map generated by the BIRCH model. As shown in Table 1, four selections of sampling non-landslide pixels were proposed. Selected non-landslide samples as well as recorded landslide samples were used to establish RF model. The landslide susceptibility maps using four different non-landslide selections are shown in Figure 5. To illustrate this, the four selections of non-landslide samples are displayed in Figure 5, which show intuitively the process of the sampling. Pixels labeled with digit 1 denote selected non-landslide samples while others labeled with digit 0 are not selected. As we can see, from selection A to D, the sampling points gradually gather suggesting decreasing patterns for non-landslide samples. Selected non-landslide samples together with recorded landslide samples are used to establish RF model and the landslide susceptibility maps using four different non-landslide selections are shown in Figure 6. Through the comparison of Figure 5 and Figure 6, it can be seen that the landslide susceptibility maps are well in line with the map of non-landside selections.

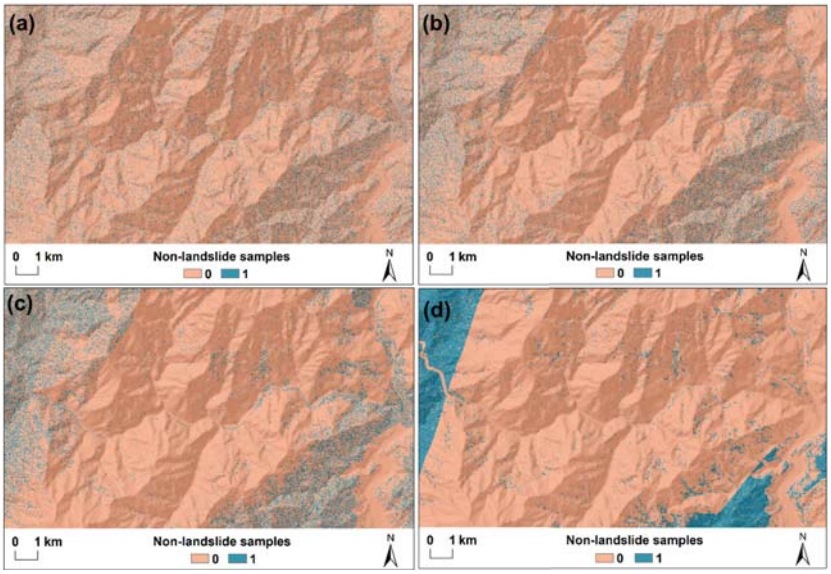

Figure 5. LSP using BIRCH model.

5.3 Analysis of the pattern-based approach

The performance of the BIRCH-RF model was evaluated with the ROC curve and the area under curve (AUC) (Figure 7). It can be observed that the BIRCH model considering the non-landslide selections exhibit better performance than the single RF model where the non-landslide samples were randomly selected. Therefore, applying UL models to perform sampling of non-landslide pixels is effective to improve predictive accuracy.

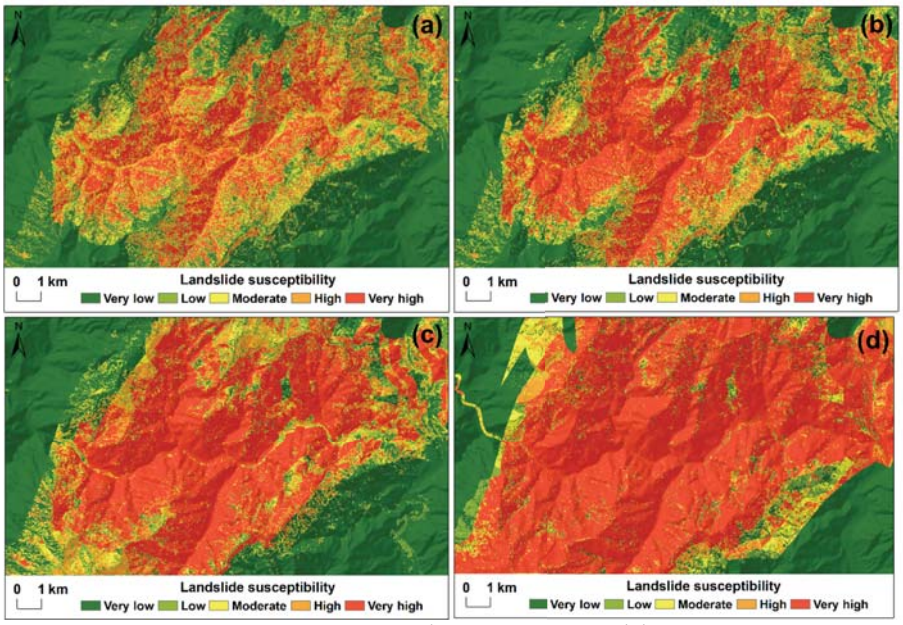

Figure 6. LSP using BIRCH-RF model.

As we can see from Figures 6&7, though expanding regions of very high and high susceptible zones can improve the predictive capacity, it may overestimate the susceptibility. This study used a sensitivity index, accuracy improvement ratio (AIR), to quantify the overestimation of different sampling selections. The AIR is defined as:

$$\mathrm{AIR} = \frac{\Delta \mathrm{AUC}}{\Delta \mathrm{P}} \tag{2}$$

Where ΔAUC is the difference of AUC value between the BIRCH-RF models with different selections and single RF model (Figure 7). And the solution of ΔP is similar, which means the difference of the percentages of susceptibility above 0.5 between the models. The AIR evaluate the cost of susceptibility overestimation when use the pattern-based approach (Table 2).

Table 2. Accuracy improvement ratio (AIR) of the BIRCH-RF models.

Model	Non-landslide sampling	ΔP (%)	ΔAUC	AIR
BIRCH-RF	Selection A	1.6	0.0374	2.33
	Selection B	7.5	0.0778	1.04
	Selection C	16.3	0.0895	0.55
	Selection D	41.2	0.0933	0.23

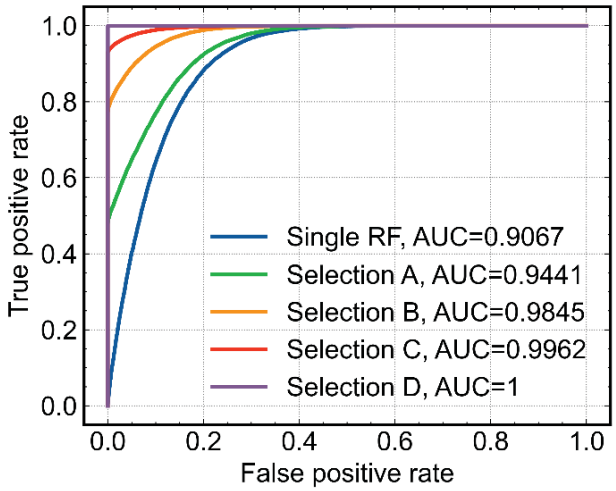

Figure 7. AUC accuracy of the BIRCH-RF model.

The selection A of the BIRCH-RF model exhibited the highest AIR value of 2.33 far outweighing other selections, indicating that the BIRCH-RF model with selection A does improve the predictive accuracy but barely overestimate the susceptibility. It can be concluded that the non-landslide patterns of the UL models were improperly used when utilizing the selections B, C and D, and selection A of the BIRCH-RF model with the AIR value of 2.33 is the best selection among others.

6 Conclusions

This study proposed a pattern-based approach to construct a UL-SL model in which non-landslide samples can be reasonably selected. The working mechanism of the pattern-based model is to find the inherent non-landslide spatial patterns in datasets using UL algorithm, and to extract the patterns by sampling non-landslide pixels, which brings performance advantages for LSP. The BIRCH-RF model was used to perform the LSP, while the RF model served as control. Four sampling selections of non-landslide pixels were compared in detail. The comparative research indicated that the proposed approach possesses superior predictive capacity to the single RF model, offering an effective way to find the non-landslide patterns. Nonetheless, the pattern-based models can lead to the overestimation of susceptibility once the improper use of the extracted non-landslide samples is involved. We therefore proposed the sensitivity parameter AIR to introduce an explanation for the performance of the approach. The AIR was defined to evaluate the overestimation of the models utilizing those four sampling selections. The result shows that the selection A of the BIRCH-RF model with the AIR value of 2.33 is regarded as the best solution, whereas the non-landslide patterns are improperly used when taking other selections, which have higher cost of the overestimation.

Acknowledgments

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Charest, L., & Plante, J. F. (2014). Using balanced iterative reducing and clustering hierarchies to compute approximate rank statistics on massive datasets. *Journal of Statistical Computation and Simulation*, 84(10), 2214–2232.

Hong, H., Tsangaratos, P., Ilia, I., Loupasakis, C., & Wang, Y. (2020). Introducing a novel multi-layer perceptron network based on stochastic gradient descent optimized by a meta-heuristic algorithm for landslide susceptibility mapping. *In Science of the Total Environment* (Vol. 742).

Huang, F., Yin, K., Huang, J., Gui, L., & Wang, P. (2017). Landslide susceptibility mapping based on self-organizing-map network and extreme learning machine. *Engineering Geology*, 223, 11–22.

Pham, B. T., Prakash, I., & Tien Bui, D. (2018). Spatial prediction of landslides using a hybrid machine learning approach based on Random Subspace and Classification and Regression Trees. *Geomorphology*, 303, 256–270.

Wang, Y., Wen, H., Sun, D., & Li, Y. (2021). Quantitative Assessment of Landslide Risk Based on Susceptibility Mapping Using Random Forest and GeoDetector. *Remote Sensing*, 13(13), 2625.

Xu, Q., Zhang, S., & Li, W. (2011). Spatial distribution of large-scale landslides induced by the 5.12 Wenchuan Earthquake. *Journal of Mountain Science*, 8(2), 246–260.

Zhang, S., Xu, Q., & Zhang, Q. (2017). Failure characteristics of gently inclined shallow landslides in Nanjiang, southwest of China. *Engineering Geology*, 217, 1–11.