remote sensing

MDPI

Article

Enhanced Absence Sampling Technique for Data-Driven Landslide Susceptibility Mapping: A Case Study in Songyang County, China

Zijin Fu [1], Fawu Wang [1,2,*], Jie Dou [3], Kounghoon Nam [1] and Hao Ma [1]

[1] Department of Geotechnical Engineering, College of Civil Engineering, Tongji University, Shanghai 200092, China; fuzijin@tongji.edu.cn (Z.F.); soilnam@tongji.edu.cn (K.N.); mah@tongji.edu.cn (H.M.)
[2] Key Laboratory of Geotechnical and Underground Engineering of the Ministry of Education, Tongji University, Shanghai 200092, China
[3] Badong National Observation and Research Station of Geohazards, China University of Geosciences-Wuhan, Wuhan 430074, China; doujie@cug.edu.cn
* Correspondence: wangfw@tongji.edu.cn

Abstract: Accurate prediction of landslide susceptibility relies on effectively handling absence samples in data-driven models. This study investigates the influence of different absence sampling methods, including buffer control sampling (BCS), controlled target space exteriorization sampling (CTSES), information value (IV), and mini-batch k-medoids (MBKM), on landslide susceptibility mapping in Songyang County, China, using support vector machines and random forest algorithms. Various evaluation metrics are employed to compare the efficacy of these sampling methods for susceptibility zoning. The results demonstrate that CTSES, IV, and MBKM methods exhibit an expansion of the high susceptibility region (maximum susceptibility mean value reaching 0.87) and divergence in the susceptibility index when extreme absence samples are present, with MBKM showing a comparative advantage (lower susceptibility mean value) compared to the IV model. Building on the strengths of different sampling methods, a novel integrative sampling approach that incorporates multiple existing methods is proposed. The integrative sampling can mitigate negative effects caused by extreme absence samples (susceptibility mean value is approximately 0.5 in the same extreme samples and presence-absence ratio) and obtain significantly better prediction results (AUC = 0.92, KC = 0.73, POA = 2.46 in the best model). Additionally, the mean level of susceptibility is heavily influenced by the proportion of absent samples.

Keywords: presence-absence method; data-driven model; landslide susceptibility mapping; absence sampling method; integrative sampling; machine learning

check for updates

Citation: Fu, Z.; Wang, F.; Dou, J.; Nam, K.; Ma, H. Enhanced Absence Sampling Technique for Data-Driven Landslide Susceptibility Mapping: A Case Study in Songyang County, China. *Remote Sens.* **2023**, *15*, 3345. https://doi.org/10.3390/rs15133345

Academic Editors: Stefano Morelli, Veronica Pazzi and Mirko Francioni

Received: 22 May 2023
Revised: 17 June 2023
Accepted: 27 June 2023
Published: 30 June 2023

1. Introduction

Landslides are devastating geological processes that frequently cause massive casualties and economic losses [1,2]. Due to their sudden occurrence, fast movement, and strong impact, landslides pose a serious threat to traffic, buildings, and residents, especially in well-developed mountainous areas. Therefore, accurate landslide forecasting is of paramount importance. In recent years, with the rapid development of 3S technology (remote sensing, geographic information system, and global positioning system), tremendous progress has been made in remote sensing for landslides, covering aspects of landslide detection, monitoring, and hazard assessment and prediction [3–5]. Landslide susceptibility mapping (LSM), which predicts the spatial probability of landslide occurrence under certain environmental conditions in an area, represents important regional landslide research work [6].

In an era of rapid development in computer science and GIS technology, various methods have been proposed for LSM. These methods can be categorized into three groups:

knowledge-based methods, physical methods, and data-driven methods [7]. While the subjectivity of expert knowledge is too strong and the wide-area parameters in physical models possess uncertainty, data-driven methods have more objective results and higher efficiency. Data-driven methods have been widely studied and applied in LSM. The kernel idea of data-based methods is to find a certain relationship by means of data science, relying on a landslide inventory and conditioning factors database. Researchers have developed and applied data-based models in LSM, including bivariate and machine-learning methods [8,9]. The classic and widely used bivariate methods include IV, certainty factor (CF), frequency ratio (FR), and fuzzy logic (FL). There are various types of machine learning models, such as logistic regression (LR), decision tree (DT), random forest (RF), support vector machine (SVM), Bayes, K-Nearest Neighbor (KNN), and artificial neural networks (ANN), as well as more complex models such as auto-encoder [10–15]. Data-driven methods in LSM can be further classified into two groups: presence-absence methods and presence-only methods [16]. The first group needs both landslide presence samples and landslide absence samples. The other group of methods needs only landslide samples, such as bivariate classification.

Sampling strategy, which involves using this acquired data to get a better prediction effect in LSM, is an important issue. The samples used for LSM prediction are usually divided into a training set and a testing set. For most machine learning methods in LSM, this includes presence and absence samples. Extracting and dividing these samples from a map are fundamental steps in LSM. The quality of the samples used for model training significantly impacts the performance of a data-driven model [17]. The presence samples (landslides) in LSM are usually generated from a landslide inventory map produced by field investigations and the interpretation of aerial photos or remote sensing images [18,19]. For the presence sampling locations, there are some commonly used sampling strategies, including the landslide scarp centroid, the centroid of the landslide body, the whole scarp polygon region, and the entire landslide body [20,21]. Different from the present samples, absence samples cannot be obtained directly. Therefore, studying the uncertainty of absence samples is particularly important for data-driven LSM.

Three strategies are commonly employed for absence sampling in current LSM research: geographic generation, feature-based generation, and generation from a prior model [22]. In geographic generation, common methods include random sampling (selection from randomly distributed circular zones) and buffer-controlled sampling (sampling from grid cells outside landslide buffer zones) [23], or selecting samples from areas where landslides are not likely to occur, such as a plain (based on expert judgment) [24,25]. Among these methods, buffer control sampling is the most frequently used [26]. However, this method does not guarantee that the selected area is free from potential landslides or has environmental parameters similar to landslides, either of which reduces the quality of the samples and has a negative influence on the accuracy of the LSM. Some studies aim to reduce the uncertainty associated with randomly selected samples in geographic space to improve the quality of absence samples [27]. In a feature-based generation, the theoretical foundation is the third law of geography, which states that landslides and non-landslides should be separated as much as possible in their feature space. Absence sampling from low-slope areas is an intuitive approach in a feature-based method, but the slope factor is overly magnified because of subjectivity. Some optimized absence sampling methods have been proposed, including target space exteriorization sampling (TSES) and similarity-based methods such as mahalanobis distance, bioclim, domain, and other improved similarity algorithms [22,28–30]. These methods select or create absence samples based on the feature space rather than the geographic space, enhancing the reliability of absence samples. TSES fabricates absence samples by modifying one or more features of a presence sample dataset into random values in the exteriorization feature space; these absence samples, produced in feature space, are not necessarily in geographic space and have some connection with the presence samples. Similarity-based methods select absence samples from the research area through quantitative indicators of similarity to landslide features. Another approach to improving the quality of absence samples is through generation from a prior model, where absence samples are extracted from the low susceptibility zone of the prior model. In LSM, the scenario where labels are available only for landslide samples corresponds to a positive, unlabeled learning problem. The presence-only model is a suitable method for positive-unlabeled learning as a prior model. Presence-only models such as IV and CF are widely applied in hybrid models, such as CF-LR, CF-SVM, CF-RF, LV-LR, IV-SVM, and IV-RF, by combining with a presence-absence model [31–34], but most of these methods use only the output parameters of the presence-only model as the input parameters of the presence-absence model. In fact, the bridge between the two methods can also be built on samples. Research shows that a presence-only method may overestimate the susceptibility of landslides [16,29]. When regional susceptibility is overestimated, it is easier to expose low susceptibility zones from which absence samples can be selected with greater certainty. Some studies have used presence-only as an a priori model for the extraction of missing samples and obtained good prediction results [35,36]. Furthermore, the combination of presence-only methods with unsupervised learning has been explored to further enhance the absence sampling strategy [37]. The two-step approach is also an algorithm for solving the absence sample problem; it trains a prior model with presence samples and unlabeled samples to identify reliable absence samples [38]. Following similar concepts, a multiple-layer perceptron model and a two-step approach based on deep neural networks have been applied in some LSM studies [39–41].

In the field of landslide susceptibility prediction, absence sampling plays a crucial role. However, existing studies often overlook the quality of absence samples in absence sampling methods, leaving the impact of absence sample quality on landslide susceptibility prediction understudied. It is essential to conduct comparative studies using different methods and quality control measures to select appropriate absence sampling methods and ensure the quality of absence samples. Moreover, many LSM studies tend to focus solely on prediction performance scores, neglecting the importance of assessing the reasonableness of landslide susceptibility prediction results. This oversight can lead to disregarding the negative effects of selected absence samples. Therefore, it is necessary to establish a comprehensive and valid evaluation system for absence samples in order to consider their applicability. On this basis, an absence sampling method with strong applicability considering the quality of different methods needs to be proposed.

In this paper, four absence sampling methods are investigated, and an integrative sampling method is proposed based on the fusion of these methods for a data-driven LSM model. In order to control the quality of absence samples, a controlled target space exteriorization sampling (CTSES) method based on improved TSES is proposed. A less commonly used optimization clustering algorithm, mini-batch K-medoids (MBKM), is employed as a prior model. The sampling methods include BCS, CTSES, IV, and MBKM, as well as a novel integrative sampling. Sampling intervals were established for each method, and extensive training was conducted to investigate the effects of absence samples from different sources and of varying quality on LSM. SVM and RF models are applied as predictive models for LSM in Songyang County, Lishui City, Zhejiang Province, China, where five absence sampling methods were tested.

2. Materials

2.1. Study Area

The study area is Songyang County, which is located in the southwest of Zhejiang Province and northwest of Lishui City. The geographical coordinates are 119°10′0″E to 119°42′42″E, 28°14′23″N to 28°36′15″N. The largest diameter from north to south is 40.2 km, with a total area of about 1404 square kilometers. Figure 1 presents the location of the study area.

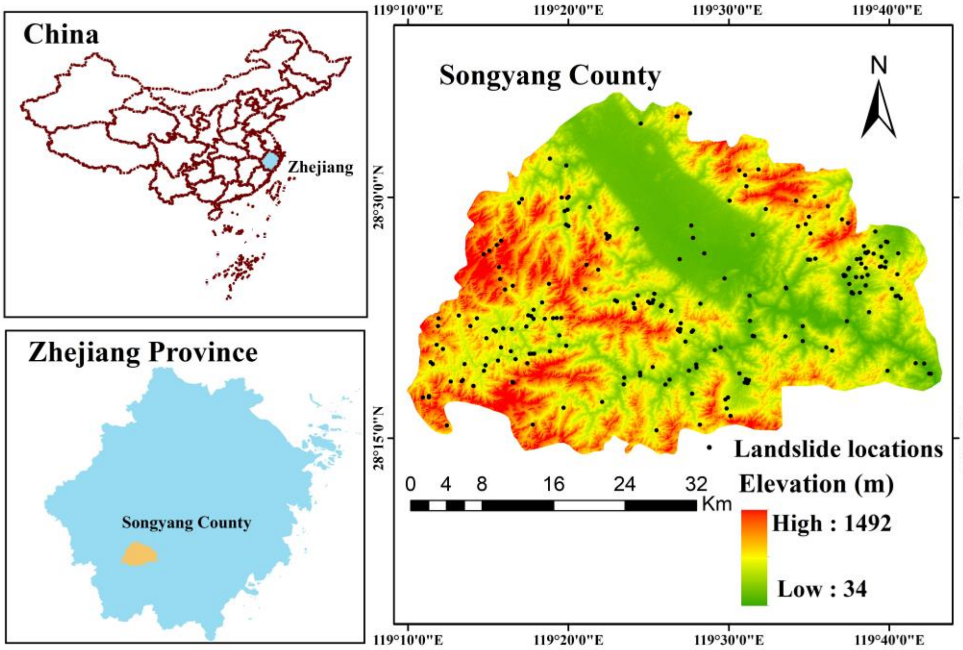


Figure 1. Geographical location of Songyang County.

Songyang County is located in the mountainous region of southern Zhejiang Province. The topography primarily consists of medium and low hills surrounded by mountains, with an open central basin known as "Songgu Basin". The terrain is higher in the northwest and lower in the southeast. The county has obvious layers of landforms, including accumulation landforms, erosion and denudation landforms, and structural erosion landforms. The lithology formation includes rhyolite, tuff, gneiss, granite, sandstone, quartz sandstone, and Quaternary sediments. Faults dominate the geological structure of the Songyang area, with over 30 large and small fault zones identified. Most of these faults align in the NNW and NNE directions, with the NNE fault belonging to the Yuyao-Lishui fault system, which is a deep fault traversing the lithosphere. This fault system formed in a tensioning process in the Cretaceous, with a strike of about 30° and 350 km in Zhejiang Province, and it runs through Songyang County [42,43]. The multistage tectonic activity that occurred in the ancient fault zone has resulted in the formation of a compressive fault zone, which spans 4–30 km in width. The fault zones in Songyang County exhibit intricate tectonic activities and fractured internal rock mass structures, significantly influencing the local geological and geomorphic environment and impacting the stability of the original slope [44]. The seismic peak acceleration in Songyang County is within the range of 0.05 g, indicating a stable crustal zone.

Songyang County is situated in a subtropical monsoon region characterized by a mild climate, abundant rainfall, and distinct seasonal variations, with notable variations in vertical climate. The annual precipitation is 1700 mm, with the period from March to June receiving the highest amount of rainfall, averaging 816.8 mm annually. November experiences the lowest rainfall, 40–50 mm. Additionally, the southeast coastal area is prone to frequent typhoon occurrences, with Songyang County being affected by an average of one to three typhoons each year. From 2015 to 2020, the county experienced the impact of several typhoons, including Sudillo, Moranti, Likima, and Hagupe. Each of these typhoons resulted in increased rainfall and triggered numerous landslides as they passed through the area. The rivers in Songyang County are part of the Oujiang River Basin. Songyin Creek and Xiaogang Creek meander from the northwest and southwest directions, respectively, toward the southeast. The area is characterized by a dense development of tributaries and numerous river valleys. During the rainy season, the water level of the rivers rises sharply.

2.2. Landslide Inventory

A landslide inventory is an important dependent variable in a calculation model and has a great influence on the accuracy of prediction results. In recent years, landslides have continued to occur in the study area. According to a relevant landslide survey report and remote sensing interpretation, 217 landslides were identified in the study area. These landslides occurred between 2000 and 2021, and the years with the most frequent occurrence were 2010, 2012, 2014, and 2016, which were also the years when Songyang County was severely affected by heavy rainfall/typhoons, and other catastrophic effects.

Landslides in Songyang County are divided into soil landslides and rock landslides, with soil landslides being the predominant type (Figure 2). The compositions of landslide masses are generally uniform, composed mainly of silty clay with gravel and a sandy, completely weathered layer in the metamorphic rock and granite distribution area. The thicknesses of landslide masses are generally small, ranging from 1 to 5 m. The volume sizes of landslides are generally below 10,000 m^3; about half of the landslides are less than 1000 m^3, while several reach 100,000 m^3. The smaller ones can measure tens of cubic meters. The soil within the sliding zone is often soft plastic powdery clay, and the sliding bed is often the underlying bedrock or a relatively denser rock-soil layer. From the perspective of a plane, landslides tend to exhibit generally semi-elliptical, tongue-shaped, or wide fan-shaped patterns, with the semi-elliptical shape being the most common. From the perspective of the longitudinal section, landslides primarily appear as linear or stepped features. The transition zone is where the terrain slope changes from steep to gentle, and the upper slope is mostly above 25°. According to an analysis of landslide triggering, fragile geological conditions are the primary causes. Additionally, extreme rainfall and human engineering activities (excavation, slope cutting, plowing, etc.) have created a landslide breeding environment.

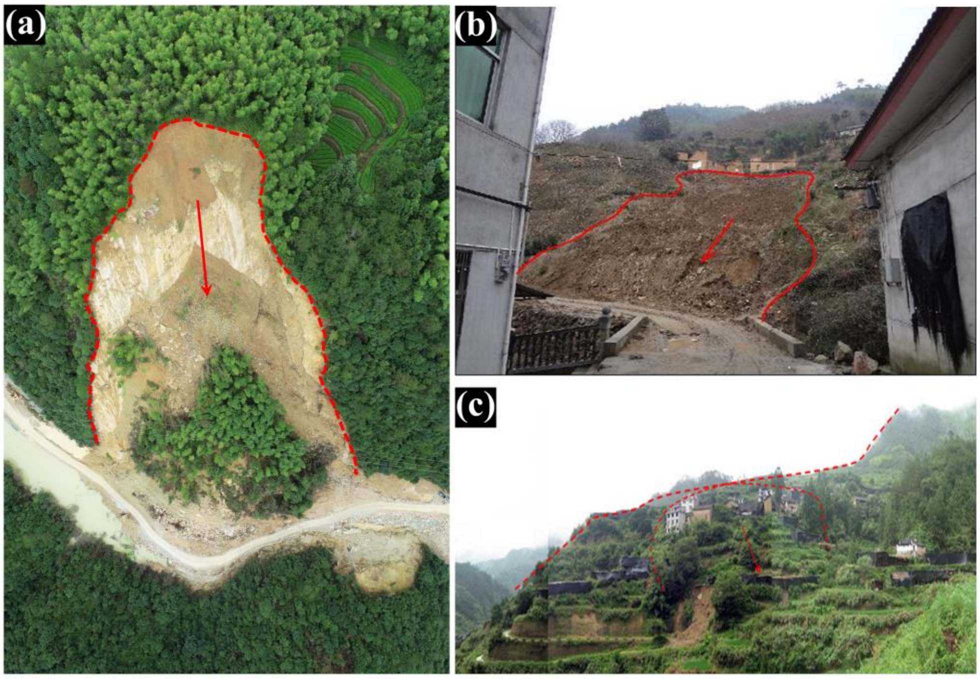


Figure 2. Typical landslides in Songyang County. (**a**) Chengtian landslide; (**b**) Xiangxi town landslide; (**c**) potential landslide in Fanshantui, Shaqiu Village. The red lines indicate the geometric boundary of the landslides and the arrow indicates the direction of the main slide.

2.3. Landslide Conditioning Factors

Factors affecting landslides can be divided into two categories: internal disaster-forming conditioning factors and external triggering factors. In this study, a total of 16 parameters were considered as potential conditioning factors for landslides, including elevation, slope, aspect, plan curvature, profile curvature, topographic roughness index (*TRI*), topographic wetness index (*TWI*), sediment transport index (*STI*), normalized difference vegetation index (*NDVI*), lithology, rainfall, distance to faults, distance to river network, distance to road, land use, and soil type (Figure 3). The grouping methods employed subjective division (for factors that are more subjective) and natural breaks (for factors that are more objective). The sources of all indicators are shown in Table 1. The formula for the relevant metrics is shown below:

$$TRI = \frac{A_{mean} - A_{min}}{A_{max} - A_{min}} \tag{1}$$

where the A_{mean}, A_{min}, and A_{max} indicate the mean, minimum, and maximum values of all cells in a 3 × 3 cell, respectively.

$$TWI = \ln\frac{A_s}{\tan\beta} \tag{2}$$

$$STI = \left(\frac{A_s}{22.13}\right)^{0.6} + \left(\frac{\sin\beta}{0.0896}\right)^{1.3} \tag{3}$$

where A_s and β indicate the specific catchment area (m^2/m) and slope gradient, respectively.

$$NDVI = \frac{NIR - Red}{NIR + Red} \tag{4}$$

where NIR is the reflection value in the near infrared band, and *Red* is the reflection value in the red band.

Table 1. Data sources of conditioning factors.

Conditioning Factor	Variable Type	Spatial Resolution (m)	Production Time (year)	Data Source
Altitude	Continuous	30	2009	ASTER GDEM 30M
Slope	Continuous	30	2009	Derived from DEM
Slope aspect	Continuous	30	2009	Derived from DEM
Plan curvature	Continuous	30	2009	Derived from DEM
Profile curvature	Continuous	30	2009	Derived from DEM
TRI	Continuous	30	2009	Derived from DEM
TWI	Continuous	30	2009	Derived from DEM
STI	Continuous	30	2009	Derived from DEM
Lithology	Discrete	30	2019	[41]
Distance to faults	Continuous	30	2019	[41]
Soil type	Discrete	30	2005	https://www.resdc.cn/ (accessed on 1 May 2022)
Annual rainfall	Continuous	30	2000–2021	http://data.cma.cn/ (accessed on 1 May 2022)
Distance to stream	Continuous	30	2021	https://lbs.amap.com/ (accessed on 1 May 2022)
Distance to road	Continuous	30	2021	https://lbs.amap.com/ (accessed on 1 May 2022)
Land use	Discrete	30	2020	https://www.resdc.cn/ (accessed on 1 May 2022)
NDVI	Continuous	30	2021	Landsat8

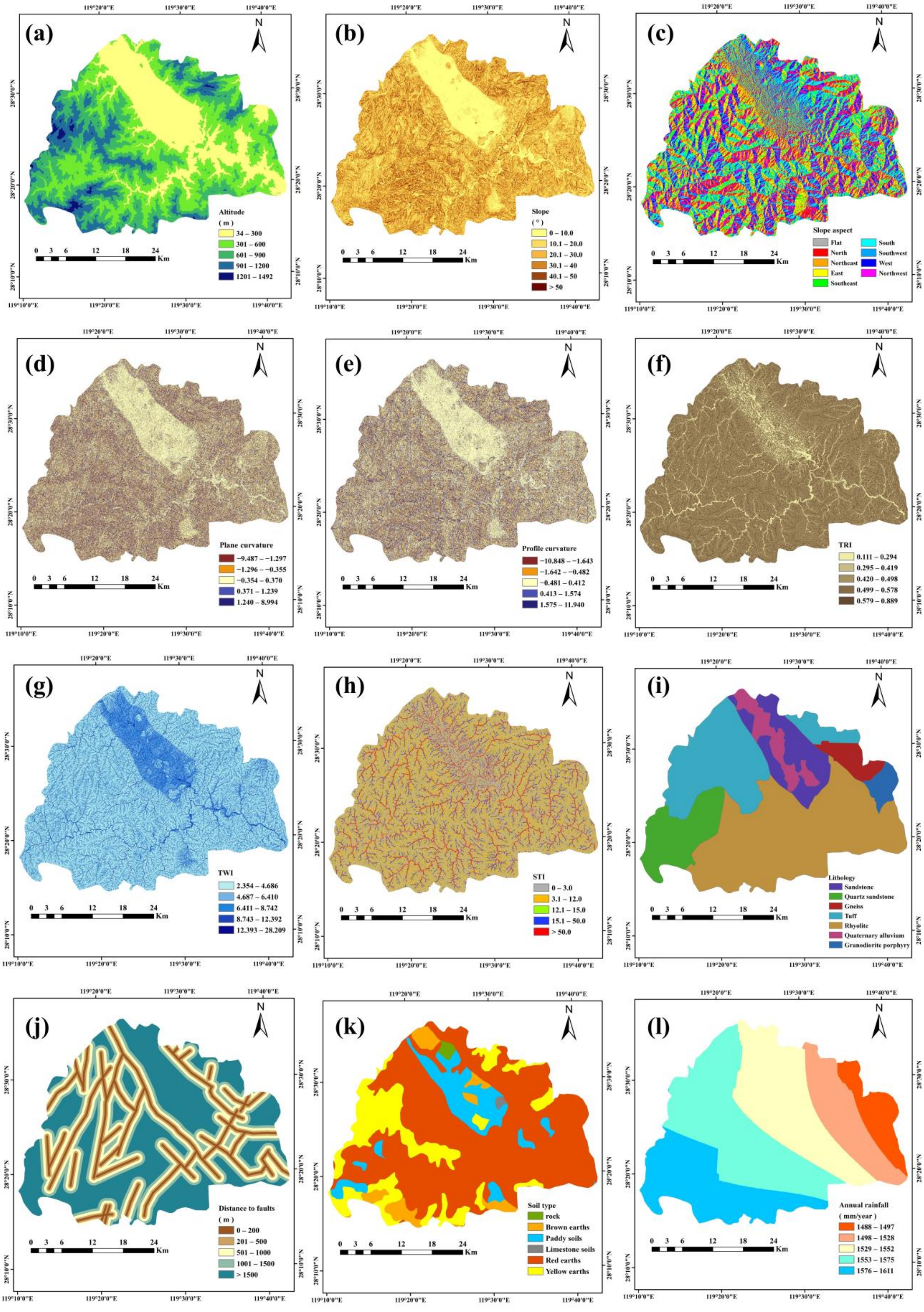


Figure 3. *Cont*.

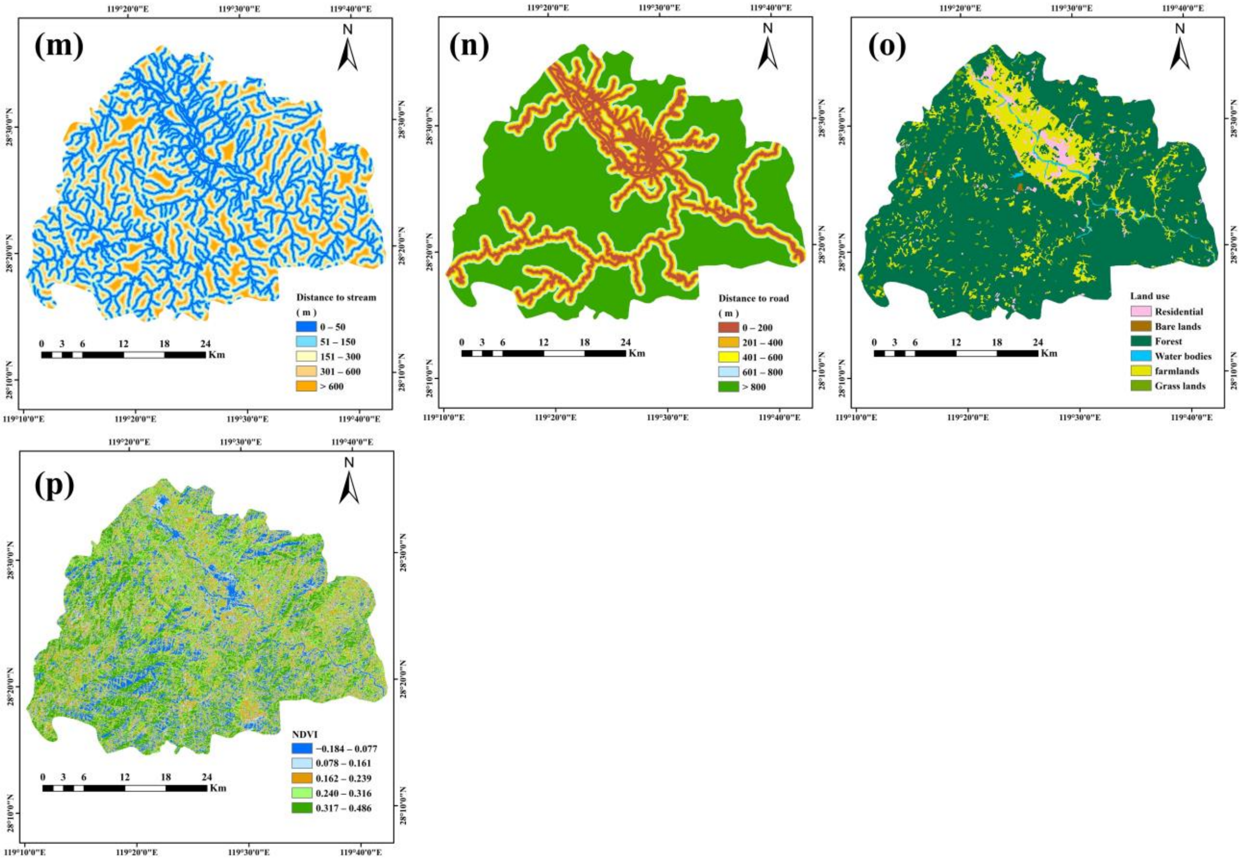


Figure 3. Conditioning factors of LSM. (**a**) Altitude; (**b**) slope; (**c**) slope aspect; (**d**) plan curvature; (**e**) profile curvature; (**f**) *TRI*; (**g**) *TWI*; (**h**) *STI*; (**i**) lithology; (**j**) distance to faults; (**k**) soil type; (**l**) annual rainfall; (**m**) distance to stream; (**n**) distance to the road; (**o**) land use; (**p**) *NDVI*.

As illustrated in Table 1, the production time of each conditioning factor is inconsistent. This is because this study is based on landslides that have occurred in the past 20 years, and numerous real-time data points are not necessary for the long timescale. The ideal data source would be produced at the same time as the landslide occurred. In the case of LSM studies based on abrupt events (earthquakes, extreme weather), the requirement for real-time data is higher. However, it is difficult to achieve consistency in timescales for landslides that have occurred consecutively over 20 years. For less variable data on a long time scale, such as DEM, lithology, faults, soil type, streams, and land use, momentary and reliable data need to be selected to represent the conditions of the region over two decades. For more variable data, such as rainfall, try to choose the average data over the research time as a representative. For the conditioning factors of roads and vegetation, the most recent data are considered representative of the average data of the last two decades by simplification. Based on the above considerations and the limitations of the data sources, more relatively reasonable layers of landslide conditioning factors were collected for this research. Due to the long timescale and the uncertainty of data quality, systematic errors introduced by the data sources are bound to exist but are not further discussed in this study.

Topographical factors have proved to be of great influence in LSM [45]. From the basic DEM, slope, aspect, plan curvature, profile curvature, *TRI*, *TWI*, and *STI* can be calculated using ArcGIS 10.8. Altitude and slope are intuitive factors in a landslide-hazard environment. The altitude values of the study area can be divided into five groups: 34–100 m, 301–600 m, 601–900 m, 901–1200 m, or 1201–1492 m. The slopes are divided into six categories at 10° intervals: 0°–10°, 10.1°–20°, 20.1°–30°, 30.1°–40°, 40.1°–50°, and >50°. The aspect of the study area is divided into the following nine categories: flat, north, northeast, east, southeast, south, southwest, west, and northwest. Profile curvature measures the change in the slope of the surface and affects the acceleration or deceleration of water flow. It can be divided into the following groups by the natural breaks group method: −10.848–1.643, −1.642––0.482, −0.481–0.412, 0.413–1.574, and 1.575–11.940. Plane curvature reflects the change in the lateral slope of the land surface, which affects the convergence or dispersion of surface water flow and, by the natural breaks, is divided into the following five groups: −9.487––1.297, −1.298––0.355, −0.354–0.370, 0.371–1.239, and 1.240–8.994. *TRI* expresses the degree of surface undulation and can be divided into five groups by natural breaks: 0.111–0.294, 0.295–0.419, 0.420–0.498, 0.499–0.578, and 0.579–0.889. *TWI* quantitatively describes the influence and cumulative effect of topography on watershed runoff and predicts the spatial spread of regional soil moisture. By natural breaks, it can be divided into the following five groups: 2.354–4.686, 4.687–6.410, 6.411–8.742, 8.743–12.392, and 12.393–28.209. *STI* can provide important information on the potential for sediment transport in a given river network; due to the uneven distribution of *STI* values in the study area, they are manually divided into the following categories: 0–3.0, 3.1–12.0, 12.1–15.0, 15.1–50.0, and >50.0.

Regional geological factors are important endogenous factors for landslides. The lithology map and fault map are based on a 1:200,000 geological map of China [44]. The lithology includes sandstone, quartz sandstone, gneiss, tuff, rhyolite, quaternary alluvium, and granodiorite porphyry. Areas in the fault map can be divided into five groups by the distance to a fault: 0–200 m, 201–500 m, 501–1000 m, 1001–1500 m, and >1500 m. Soil type was also included as an auxiliary conditioning factor and consists of rock, brown earths, paddy soils, limestone soils, red earths, and yellow earth. Meteorology and hydrology are the most important triggers for landslides. The average annual rainfall from 2000 to 2021 was divided into five groups according to the natural breaks: 1488–1497 mm/year, 1498–1528 mm/year, 1529–1552 mm/year, 1553–1575 mm/year, and 1576–1611 mm/year. The distance to the stream is divided into the following groups: 0–50 m, 51–150 m, 151–300 m, 301–600 m, and >600 m.

Other factors such as traffic, vegetation, and land use are also included. The traffic network can reflect the modification of slopes by human engineering activities. The distance to a road is divided into the following groups: 0–200 m, 201–400 m, 401–600 m, 601–800 m, and >800 m. Land use types involve human modification of the land surface; the classifications detected in the study area are residential, bare lands, forests, water bodies, and farmlands. The vegetation cover is represented by *NDVI* and is divided into the following groups by natural breaks: −0.184–0.077, 0.078–0.161, 0.162–0.239, 0.240–0.316, and 0.317–0.486.

3. Methodology

3.1. Study Route

This study was conducted in three stages (Figure 4). First, layers of the 16 conditioning factors for the study area were prepared, and correlation tests were performed. Then, the absence sample dataset with different qualities was constructed using different methods of absence sampling: BCS, CTSES, IV, MBKM, and integrative sampling. Finally, LSM work was produced by support vector machine and random forest algorithms, and evaluation and comparison of the models produced by each absence sampling method were completed.

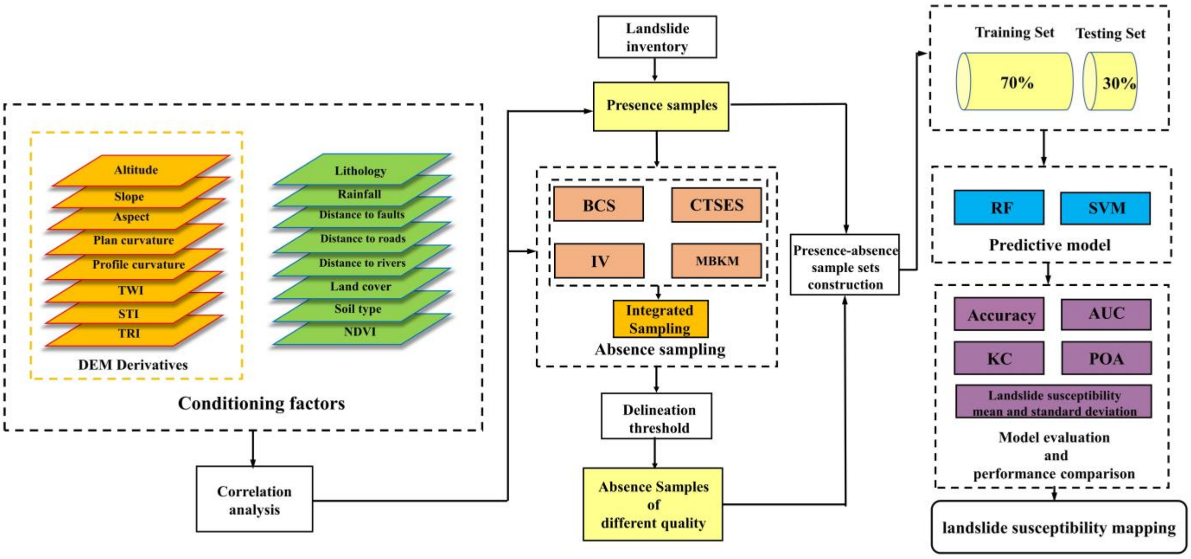


Figure 4. Flowchart of comparison and evaluation process of absence sample sampling methods.

3.2. Correlation Analysis of Conditioning Factors

Here, correlation analysis was performed to determine the statistical correlation among more than two variables. Multiple correlation analysis of conditioning factors in landslide susceptibility mapping is important and represents the screening process for selected indicators. If variables with strong correlations are input into the model, the complexity of the model will increase, and the training process of the model will also be affected. Thus, correlation analysis and screening of conditioning factors are indispensable steps in LSM [46]. In this study, commonly used correlation indicators, including the Pearson correlation coefficient (*PCC*), variance inflation factor (*VIF*), and tolerance (*TOL*), were selected to scan the correlation and multicollinearity between the conditioning factors.

Pearson correlation analysis is used to measure the strength and direction of the linear relationship between two variables. The value of the *PCC* is usually between −1 and 1, where −1 indicates a perfect negative correlation, +1 indicates a perfect positive correlation, and 0 indicates no linear correlation. The closer the absolute value of the PCC is to 1, the stronger the correlation, and the closer the absolute value of the PCC is to 0, the weaker the correlation. When the values are 0–0.2, 0.2–0.4, 0.4–0.6, 0.6–0.8, and 0.8–1, the correlation between the two conditioning factors is very weak, weak, moderate, strong, and very strong, respectively. The formula for calculating the PCC is as follows:

$$\rho_{x,y} = \frac{cov(x,y)}{\sigma_x \sigma_y} \tag{5}$$

where $cov(x,y)$ denotes the covariance of the conditioning factors x and y; σ_x and σ_y denote the standard deviation of x and y.

The *VIF* and *TOL* are indicators to test the multicollinearity of multiple conditioning factors. It is generally considered that when the value of *VIF* is less than 5 and the value of *TOL* > 0.2, the factor has no multicollinearity with other factors. Factors of *VIF* values in the range of 5–10 and *TOL* values in the range of 0.1–0.2 are considered to have weak multicollinearity. Factors with *VIF* values > 10 and *TOL* values < 0.1 are considered to have moderate or higher multicollinearity. The formulas for calculating *VIF* and *TOL* are as follows:

$$VIF_i = \frac{1}{1 - R_i^2} \tag{6}$$

$$TOL = 1 - R_i^2 \tag{7}$$

where R_i^2 is the sample decidability coefficient of the *i*-th conditioning factor.

3.3. Absence Sampling Methods

3.3.1. Buffer Control Sampling (BCS)

BCS is a quick, easy way to create non-landslide samples and is widely used in LSM [47]. The essential basis of BCS is similar to the first law of geography, which states that things close together in geographical space have a stronger connection. Accordingly, in the landslide inventory map, the area adjacent to a landslide is more susceptible to landslides. This is also consistent with the physical mechanics and disaster-forming environment of a landslide. Rock and soil in areas close enough to the landslide are more susceptible to further sliding due to the scraping or accumulation of previous landslides. Moreover, due to the continuity of engineering geological conditions, areas relatively close to landslides are more likely to develop the same geological environment and be exposed to strong external triggers. Therefore, the BCS method could set a buffer zone at a certain distance around the landslide location in the landslide inventory map and complete the sampling work in areas outside the buffer zone within the study area to reduce the strong effect of the spatial proximity of landslide events.

3.3.2. Controlled Target Space Exteriorization Sampling (CTSES)

TSES is proposed to create pseudo-absence data for LSM. The absence of data are produced from feature space instead of geographical space [28]. This method can make up for the defects of sampling in geospatial samples with geographical differences that may be similar in the feature space. The core principle of TSES is to extend landslide features beyond the feature space of all landslides (within the feature space of the study area) to obtain new non-landslide samples. It is reasonable for these absence samples to be created using landslide samples, as doing so preserves the regional character of the study area and enhances variability within landslide samples. Improved target space exteriorization sampling (ITSES) has been proposed to compensate for the non-existence absence samples in geographic space [29]. It maps the absence samples from feature exteriorizing space to geographic space, which gives the absence data more chances to be validated in remote sensing or field work.

Existing TSES methods exteriorize only one feature of the landslide sample during the process. In fact, landslide samples can be simultaneously exteriorized on multiple feature dimensions, and the number of such exteriorizing dimensions determines the degree of difference between absence and landslide samples. Whether this difference could affect the predictive performance of LSM is uncertain. In this study, controlled TSES (CTSES) is proposed to explore absence sampling methods from different exteriorizing dimensions by adding an artificially controlled number *d* of exteriorizing dimensions on the basis of the ITSES (Figures 5 and 6). The code is as follows:

```
Input: landslide samples $S$, number of exteriorizing feature dimension $d$
Output: non-landslide samples $N_d$ with exteriorizing feature dimension $d$
(1) Initialization:
        $N = \varnothing$
(2) For each landslide conditioning factor $A$:
        Calculate the value range $V_{A\,max}$ of feature $A$ of all units on the whole study area;
        calculate the value range $V_A$ of feature $A$ of all landslide samples;
        determine the exteriorized feature space $V'_A = \overline{V}_A \cap V_{A\,max}$.
(3) Traverse every unit $i$ in the study area:
    (a) Set temporary variables a = 0;
    (b) Traverse every landslide conditioning factor A:
            if $A$ of $i$ is in $V'_A$, a = a + 1;
            (c)    If $a = d$, then run (d):
        (d)    $N_d = N_d \cup i$
(4)    Return $N_d$
```

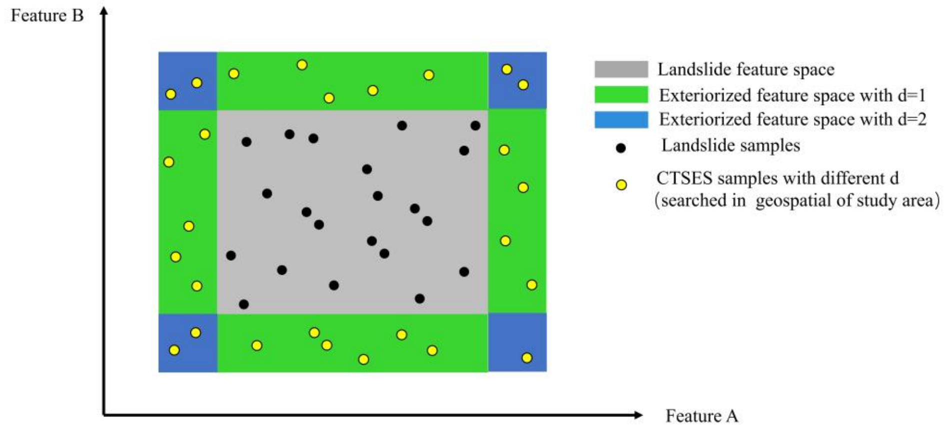
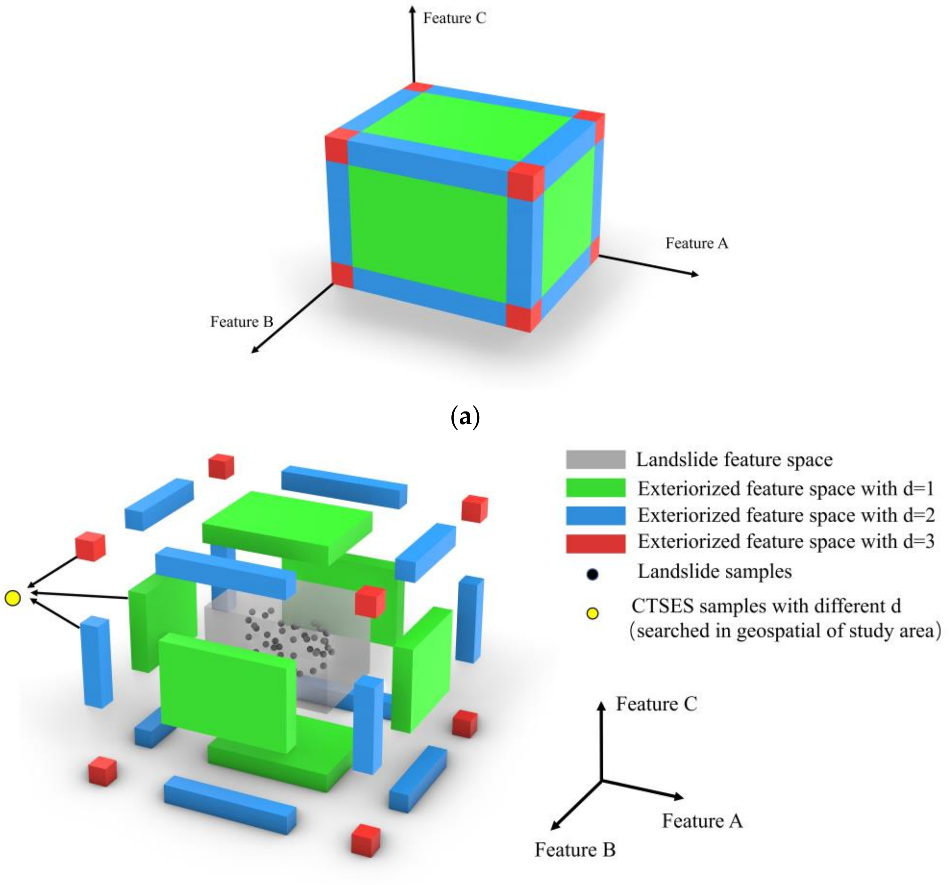

Feature B
Landslide feature space
Exteriorized feature space with d=1
Exteriorized feature space with d=2
Landslide samples
CTSES samples with different d
(searched in geospatial of study area)
Feature A

Figure 5. Schematic of the two-dimensional feature space of the CTSES method.


Feature C
Feature A
Feature B
Landslide feature space
Exteriorized feature space with d=1
Exteriorized feature space with d=2
Exteriorized feature space with d=3
Landslide samples
CTSES samples with different d
(searched in geospatial of study area)
Feature C
Feature A
Feature B

(**a**)

(**b**)

Figure 6. Schematic of the three-dimensional feature space of the CTSES method. (**a**) Three-dimensional feature space of the study area; (**b**) deconstruction map of the three-dimensional feature space in the CTSES method.

3.3.3. Information Value (IV)

As a bivariate analysis method, the IV method is often used in landslide susceptibility analysis. The method is consistent with the principles of several other bivariate analysis methods, including certainty factor (CF) and frequency ratio (FR). Information theory is the basis of the IV method, in which the information gain measures the significance of features to the prediction results. The theory of the IV method assumes that the generation of geohazards depends on the quality and quantity of the information obtained. Probabilistic statistics and comparative mapping are used to derive the magnitude of the contribution of various conditioning factors to landslide generation from historical landslide data. The significance of each conditioning factor is expressed by the IV, and the larger the value, the more likely a geological hazard will occur under this factor.

The theoretical formula for the amount of information is:

$$I(x_i, \mathrm{A}) = \ln \frac{P(x_i \mid A)}{P(x_i)} \tag{8}$$

where $I(x_i, \mathrm{A})$ is the amount of information provided by the single-factor factor x_i to the occurrence of geohazard (A), $P(x_i \mid A)$ is the probability of x_i under geohazard conditions, and $P(x_i)$ is the probability of factor x_i in the study area.

In specific applications, the sampling frequency (area ratio) is often used as:

$$I_{ij} = ln \frac{\frac{N_{ij}}{S_{ij}}}{\frac{N}{S}} \tag{9}$$

where I_{ij} is the amount of information in the j-th classification of the i-th factor, S is the total study area; S_{ij} is the area under the j-th classification of the i-th factor in the study area. N is the total number of geological hazards in the study area, and N_{ij} is the number of the j-th classification of the i-th factor in the study area.

The sum of the I_{ij} of each unit is shown as the landslide sensitivity index, and the formula is:

$$LSI = \sum_{i=1}^{n} I_{ij} \tag{10}$$

3.3.4. Mini-Batch K-Medoids (MBKM)

K-medoids is an improved unsupervised algorithm for clustering based on K-means. Various clustering algorithms are often used for LSM, but the k-medoids method is minimally used in LSM [48,49]. The improvement of the K-medoids method is that it updates the centroid with the most centrally located object in the cluster instead of the center of mass formed by the average value in the K-means method. Therefore, K-medoids are far more adaptable to noise and outliers than K-means.

1. K initial centroids are randomly selected.
2. Assign the remaining points to the cluster represented by the closest medoids.
3. In each class, the sum of distances between each sample point and other points is calculated, and the point with the smallest sum of distances is selected as the new medoid.
4. Repeat the process in steps 2–3 until all medoid points no longer change or the upper limit of iterations is reached.

In this algorithm, the time complexity is much greater than that of K-means due to iterating through all data points, especially with large datasets. When the number of sample points reaches a million, huge computational resources will be required; however, the mini-batch algorithm will solve this problem well.

The MBKM method randomly samples several small sample sets from a large sample set and updates the initial centroids in each small sample set by substituting them into the K-medoids model. The combined training effect on a large number of small samples will be slightly lower than the training effect on the whole sample, but it will save computational resources and time.

3.3.5. Integrative Sampling

At its core, integrative sampling is similar to integrated learning but with added diversity. The specific operation is to mix the absence samples from each sampling method by a certain ratio to form a new absence sample set. The newly formed sample set has more diversity, which results in a more robust and generalized model and a reduced risk of overfitting (similar to the idea of bagging). Under the same sampling interval, the characteristic pattern of the generated absence samples may be singular or one-sided. Overly monotonic sample characteristics can make the model tend to overfit, and extreme samples can make the model perform poorly against new datasets. However, in integrative sampling, the learner receives samples from multiple sampling methods. Due to the different sampling methods for absence samples, the distribution of absence samples in the geographic space and feature space is diverse in order to give comprehensive information to the training model. Therefore, the training model that accepts the integrative sampling set can show good adaptability and generalization ability when faced with a diverse set of prediction data.

Integrated sampling combines four absence sampling methods, namely BCS, CTSES, IV, and MBKM, and the process is shown in Figure 7. The main processing steps are as follows: (1) defining sampling thresholds and sampling for absence samples of each method; (2) pre-training and evaluating combining with landslide samples; and (3) forming the new integrated absence samples by assembling the best-quality sample sets for each method. The best-quality sample sets are discriminated by the comprehensive predictive performance metrics in the pre-training results.

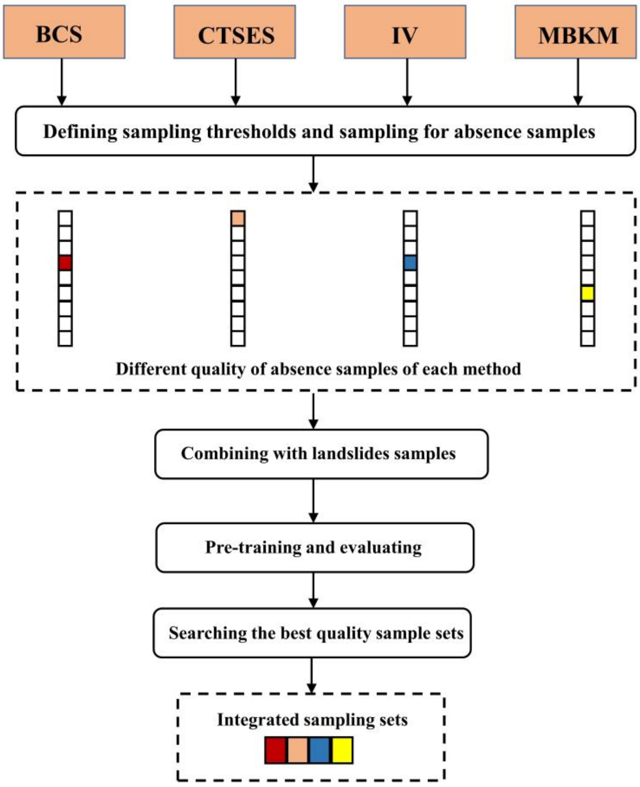

BCS
CTSES
IV
MBKM
Defining sampling thresholds and sampling for absence samples
Different quality of absence samples of each method
Combining with landslides samples
Pre-training and evaluating
Searching the best quality sample sets
Integrated sampling sets

Figure 7. Schematic of the procession of integrated sampling. The different colors represent the best quality sample sets in each absence sampling method.

3.4. Machine Learning for Landslide Susceptibility Mapping

SVM and RF are among the most classical algorithms for prediction and classification and are widely used and known for their excellent performance in a wide range of industries. In recent years, SVM and RF have been prominent in LSM research due to their excellent performance. In fact, SVM and RF models are capable of not only classification but also a probabilistic prediction that can be transformed into the assessment of susceptibility. Data-driven landslide prediction is a high-dimensional and nonlinear problem that is well suited to be solved with these two algorithms.

3.4.1. Random Forest

Random forest is a supervised machine learning integration algorithm based on bagging. Multiple sample sets and decision trees are constructed through bootstrap sampling (random sampling with put-back), and random features of the sample are selected for splitting on the basis of the decision tree. Due to its randomness, the random forest does not easily fall into overfitting and has good resistance to noise. After each decision tree split is completed (training completed), the set becomes a forest, and each decision tree is voted on to get the final result. Decision trees can achieve the tasks of classification and regression, and for LSM, landslide susceptibility indicators can be obtained by averaging the results of multiple decision trees [50]. Due to its excellent ability to handle high-dimensional and nonlinear data, the random forest method can achieve good prediction results in LSM [51].

3.4.2. Support Vector Machine

A support vector machine is a supervised machine learning algorithm commonly used in landslide susceptibility [52]. An SVM has the ability to solve problems with high-dimensional, small samples, and nonlinear features. The two main ideas of the SVM are the search for the best classification hyperplane and the use of kernel functions [53]. The hyperplane is a subspace of n-1 dimensions in an N-dimensional linear space; it can correctly classify the sample categories. The support vector is the closest training sample point to the hyperplane. The optimal hyperplane has the largest sum of the distances from it to different classes of support vectors. In the two-dimensional feature space, the hyperplane is a straight line that can divide the sample categories, and in the three-dimensional feature space, a hyperplane is a plane that can divide the sample categories. Once no hyperplane in the current feature space can linearly divide the sample categories, the original feature space can be mapped to a higher-dimensional feature space by a kernel function to make the samples linearly divisible in this space. The commonly used kernel functions are the linear kernel, polynomial kernel, radial basis function kernel (RBF-kernel), and sigmoid kernel; the polynomial kernel and the RBF-kernel are most frequently used and have a good predictive effect on landslide susceptibility [54,55].

3.5. Model Evaluation Methods

To assess and compare the predictive performance of models using different absence sample compositions, the models were evaluated from three perspectives. First, the accuracy of the model was examined, encompassing both the accuracy of the training sets and the testing sets. Accuracy is the most intuitive indicator of a model's predictive performance. The accuracy threshold is set to 0.5; it is considered a potential landslide when the prediction result is >0.5. Training accuracy reflects the model's ability to fit onto the training sets; the higher the testing accuracy, the better the model can learn the patterns and laws of the data. Testing accuracy reflects the model's ability to generalize, which represents the ability to perform on unseen data.

The second perspective involves evaluating the comprehensive performance of the model. Area under the curve (AUC), kappa coefficient (KC), and performance overall accuracy (POA) were used to evaluate the combined predictive performance of the model. The receiver operating characteristic curve (ROC) and AUC are commonly used methods to assess the performance of classification models. The AUC represents the area under the ROC curve and ranges from 0.5 to 1.0. A higher AUC value implies a better discrimination performance of the model. A confusion matrix is a table used to evaluate the performance of a classification model by comparing the predicted class labels with the actual class labels. The matrix contains four terms: true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). From the confusion matrix, several metrics can be calculated to evaluate the performance of the classification model, such as precision, recall, accuracy, KC, Matthews correlation coefficient (MCC), and F1-score. KC measures the agreement between the predicted and actual class labels, correcting for the agreement expected by chance. KC values range from −1 to 1, where a value of 1 indicates perfect agreement, a value of 0 indicates agreement by chance, and a value of −1 indicates perfect disagreement. POA, which is the sum of accuracy, MCC, and F1-score, provides a more comprehensive assessment of the model's predictive performance than individual metrics [31,56].

The third is the mean and standard deviation of landslide susceptibility. The mean value reflects the overall overestimation or underestimation of the model's susceptibility, and the standard deviation represents the degree of dispersion of the susceptibility index. Therefore, the mean and standard deviation can be used as auxiliary parameters to assess the reasonableness of the model.

4. Results

4.1. Correlation Analysis

In ArcGIS 10.8, the band collection statistics function was used to calculate the *PCC* between two conditioning factors by inputting a total of 16 conditioning factor layers containing elevation, slope, aspect, profile curvature, plane curvature, *NDVI*, lithology, rainfall, distance to faults, distance to rivers, *TWI*, *TRI*, *STI*, distance to roads, land use, and soil type. The results are entered into Python for heat map visualization (Figure 8). In the heat map, the stronger the correlation, the darker the color, and the weaker the correlation, the lighter the color. The color tends to be brown for positive correlations and purple for negative correlations. Among the 16 conditioning factors, the maximum *PCC* value was −0.59, where the profile curvature and plane curvature were (relatively) most correlated, followed by the *PCC* value of 0.56, showing a relatively positive correlation between elevation and distance to the road. All *PCC* values are less than 0.6, which means that there is no strong correlation among the factors.

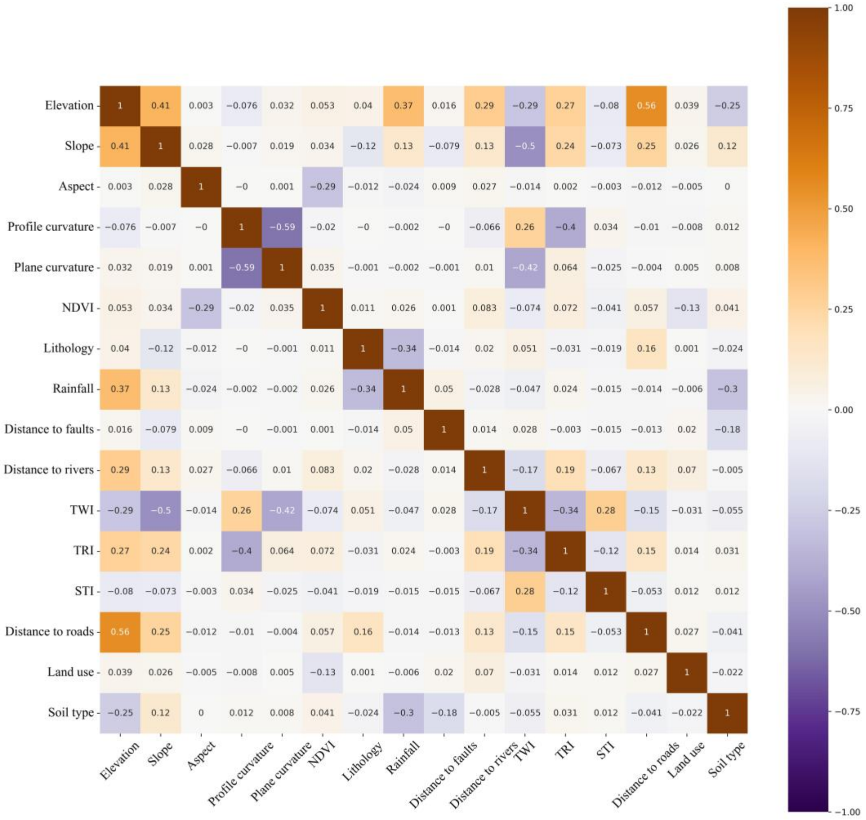


Figure 8. Pearson correlation coefficient heat map of 16 conditioning factors.

In all, 16 layers of conditioning factors were extracted from ArcGIS 10.8 for the whole study area, and the *VIF* and *TOL* values of the 16 conditioning factors were calculated using the statsmodels library in Python (Table 2). Elevation had the highest *VIF* value of 2.6 and the lowest *TOL* value of 0.383, with relatively strong multicollinearity with all other factors. In addition, *VIF* values for all conditioning factors were less than 10 and strictly less than 5, and all *TOL* values were greater than 0.1. Therefore, it can be assumed that there is no strong multicollinearity between these 16 conditioning factors.

Table 2. *VIF* and *TOL* results for 16 conditioning factors.

Factor	*VIF*	*TOL*
Elevation	2.613	0.383
Slope	1.685	0.594
Aspect	1.105	0.905
Profile curvature	2.010	0.498
Plane curvature	2.087	0.479
TRI	1.558	0.642
TWI	2.090	0.478
STI	1.118	0.894
NDVI	1.143	0.875
Lithology	1.223	0.818
Rainfall	1.570	0.637
Distance to faults	1.028	0.972
Distance to rivers	1.172	0.854
Distance to roads	1.613	0.620
Land use	1.032	0.969
Soil type	1.350	0.740

4.2. Results of Absence Sampling

The absence sample set for BCS is constructed by first constructing a buffer for 217 landslide sample points and then creating random points outside the buffer by taking the inverse of the function of the intersection set. By doing so, we can ensure that the random points do not fall inside the buffer or near the landslide. By constructing buffers of different scales, the sampling scales are divided into 300 m, 500 m, 800 m, 1000 m, and 1500 m, and 10 sets of random points are randomly selected under each scale as absence samples. In the BCS method, the relationship between the created absence sample location and the landslide location is shown in Figure 9. The number of absence samples per group was 217, which was used for the final training and testing after extracting 16 conditioning factors. The whole sampling process for BCS was conducted on ArcGIS 10.8.

The program for CTSES was done in Python, and subsequently, all pixels in the study area were entered and traversed in the program (the core code is shown in Section 3.3.2). The number of input pixels was 1,559,752. The search for absence samples of different quality was performed by changing the control parameters. In the sampling results of CTSES, the number of absence samples with d = 1, d = 2, d = 3, and d = 4 were 200,023, 22,765, 2888, and 323, accounting for 12.824%, 1.459%, 0.185%, and 0.021% of the overall pixels, respectively. When the controlled parameter d is increased to 5, the number of absence samples found is 21, and we cannot find any absence samples with d greater than 5. The location of these absence samples mapped to the study area is shown in Figure 10. Absence sample sets were created for training and testing from the CTSES data of d = 1, d = 2, d = 3, and d = 4. In all, 20 sets of 217 absence samples were randomly selected from each interval for final training and testing.

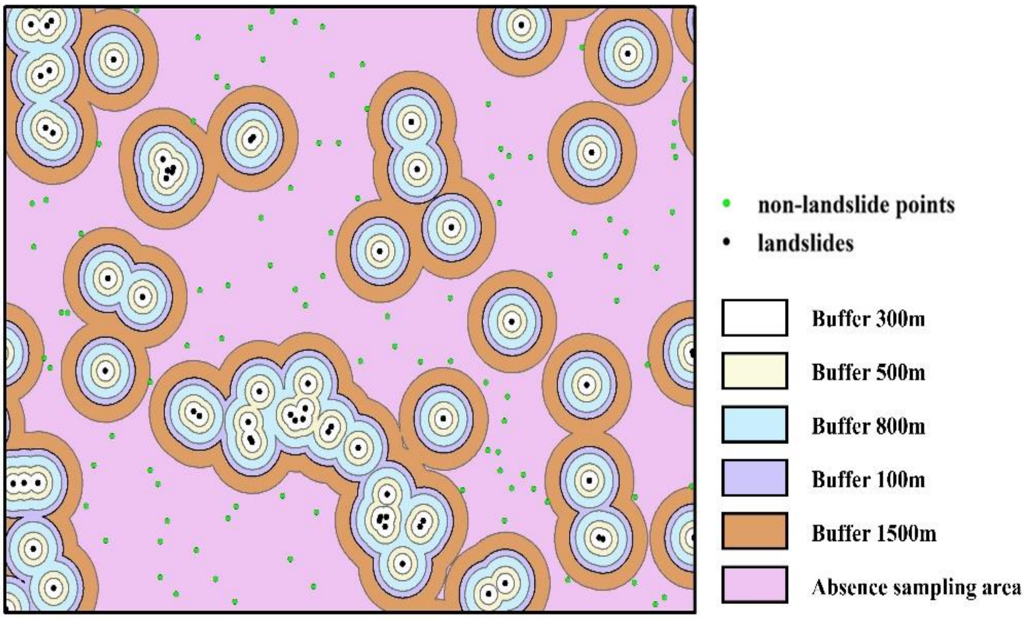


Figure 9. Absence sample location schematic of the BCS method.

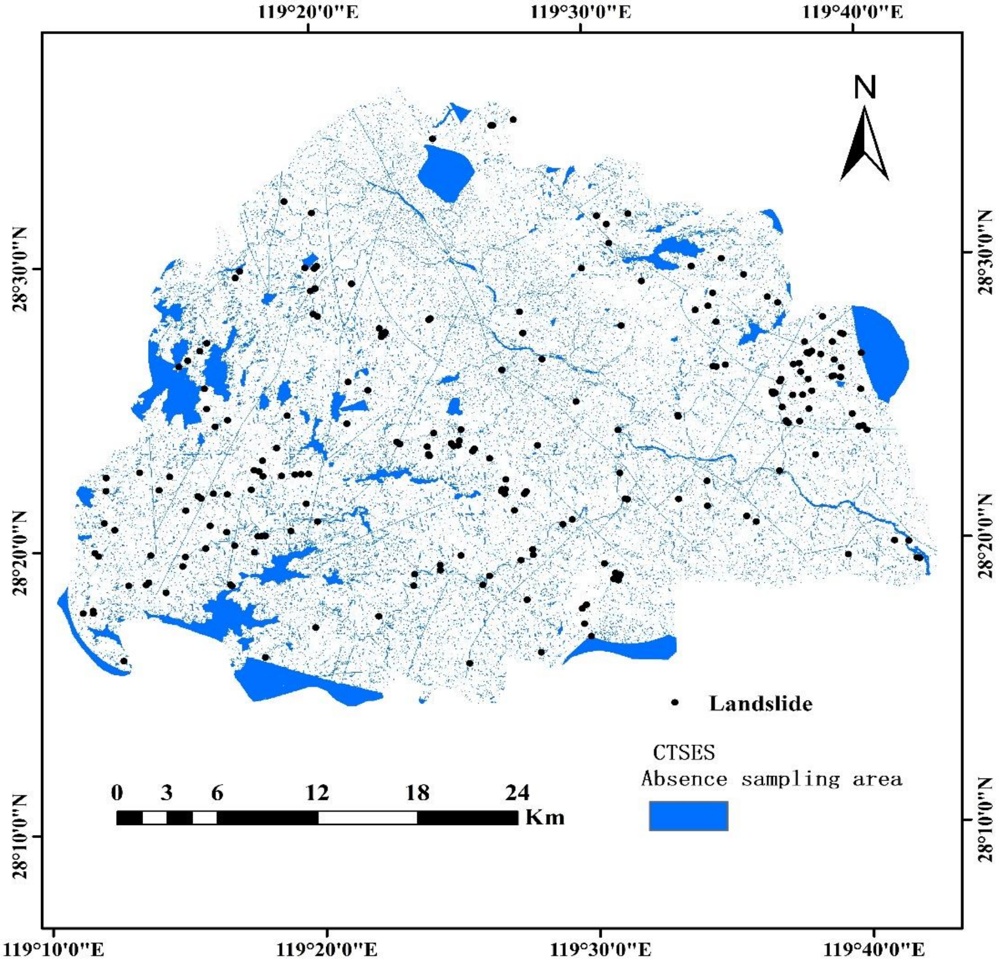


Figure 10. Absence sampling location of CTSES results.

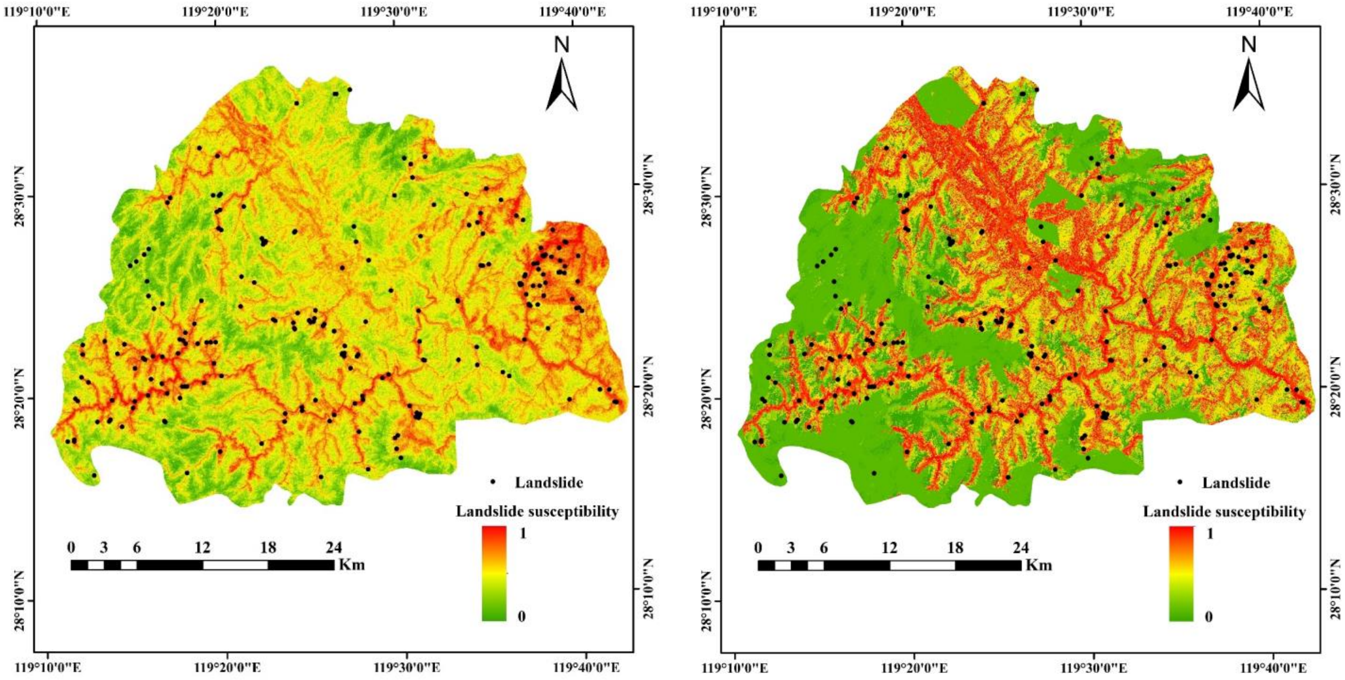
The total number of landslides in the study area was 217, and the total number of pixels was 1,559,752, which was used to replace N and S in the IV model. The statistics of the number of landslides N_{ij} and the corresponding area S_{ij} within the interval of each conditioning factor are shown in Table 3. The IV values of the grading of each factor were calculated. After the initial susceptibility mapping of the IV model was conducted, the susceptibility was divided into 10 intervals based on the natural break method, as shown in Figure 11a. The absence sampling threshold was set according to the natural break, and a total of 10 sampling intervals were divided: <0.1, <0.2, <0.3, <0.4, <0.5, <0.6, <0.7, <0.8, <0.9, and <1.0.

Table 3. IV calculation table of IV model.

Factor	Class	No. of Landslides	No. of Pixels in Domain	IV
Altitude (m)	34–400	65	400,676	0.154
	301–600	105	529,308	0.355
	601–900	40	424,249	−0.389
	901–1200	7	191,726	−1.338
	1201–1492	0	13,793	0.000
Slope (°)	0–10	35	285,800	−0.128
	10.1–20	75	353,914	0.421
	20.1–30	60	484,962	−0.117
	30.1–40	37	339,129	−0.243
	40.1–50	10	84,504	−0.162
	>50	0	5759	0.000
Aspect (°)	0	0	5939	0.000
	0–22.5	11	89,012	−0.118
	22.6–67.5	34	202,755	0.187
	67.6–112.5	32	216,747	0.059
	112.6–157.5	35	202,523	0.217
	157.6–202.5	30	178,953	0.186
	202.6–247.5	24	182,953	−0.059
	247.6–292.5	26	194,308	−0.039
	292.6–337.5	16	192,853	−0.517
	337.6–360	9	88,025	−0.308
Plane curvature	(−9.487–−1.297)	7	90,832	−0.591
	(−1.296–−0.355)	57	358,239	0.134
	(−0.354–0.370)	89	646,319	−0.010
	(−0.371–1.239)	56	354,515	0.127
	(1.240–8.994)	8	109,847	−0.647
Profile curvature	(−10.848–−1.643)	4	78,385	−1.003
	(−1.642–−0.482)	36	318,609	−0.208
	(−0.481–0.412)	91	696,850	−0.063
	(0.413–1.574)	70	376,172	0.291
	(1.575–11.940)	16	89,736	0.248
TWI	(2.354–4.686)	85	656,862	−0.072
	(4.687–6.410)	69	524,863	−0.057
	(6.411–8.742)	40	250,590	0.137
	(8.743–12.392)	17	92,287	0.281
	(12.393–28.209)	6	29,466	0.381
TRI	(0.111–0.294)	23	85,999	0.654
	(0.295–0.419)	45	246,752	0.271
	(0.420–0.498)	80	535,396	0.071
	(0.499–0.578)	52	531,083	−0.351
	(0.579–0.889)	17	208,097	−0.532

Table 3. *Cont.*

Factor	Class	No. of Landslides	No. of Pixels in Domain	IV
STI	(0–3.0)	84	715,531	−0.170
	(3.1–12.0)	64	586,030	−0.242
	(12.1–15.0)	11	50,604	0.446
	(15.1 -50.0)	31	140,393	0.462
	(>50.0)	27	75,298	0.947
NDVI	(−0.184–0.077)	32	214,619	0.069
	(0.078–0.161)	54	270,357	0.362
	(0.162–0.239)	66	361,211	0.273
	(0.240–0.316)	49	403,597	−0.136
	(0.317–0.486)	16	309,958	−0.991
Rainfall (mm/year)	1488–1497	33	118,347	0.695
	1498–1528	33	211,130	0.116
	1529–1552	28	398,437	−0.683
	1553–1575	70	494,390	0.018
	1576–1611	53	337,453	0.121
Distance to rivers (m)	0–50	39	169,598	0.503
	51–150	79	327,307	0.551
	151–300	52	338,629	0.099
	301–600	39	498,642	−0.576
	>600	8	175,619	−1.116
Distance to roads (m)	0–200	66	222,151	0.759
	201–400	21	147,406	0.024
	401–600	8	119,129	−0.728
	601–800	13	102,609	−0.094
	>800	109	968,500	−0.212
Distance to faults (m)	0–200	17	146,962	−0.185
	201–500	21	213,815	−0.348
	501–1000	61	318,946	0.318
	1001–1500	37	256,177	0.037
	>1500	81	623,895	−0.069
Land use	Residential	3	42,047	−0.668
	Bare land	0	2209	0.000
	Forest	155	1,264,284	−0.126
	Water body	1	12,156	−0.525
	Farmland	56	216,100	0.622
	Grassland	2	21,548	−0.405
Soil type	Rock	0	9137	0.000
	Brown earth	3	78,111	−1.287
	Paddy soil	20	170,108	−0.168
	Limestone soil	1	4313	0.511
	Red earth	179	1,011,069	0.241
	Yellow earth	14	288,375	−1.053
Lithology	Sandstone	10	153,095	−0.756
	Quartz sandstone	47	179,684	0.631
	Gneiss	18	56,402	0.830
	Tuff	38	373,449	−0.313
	Rhyolite	86	679,184	−0.094
	Quaternary alluvium	3	78,859	−1.297
	Granodiorite Porphyry	15	38,986	1.017



(**a**) (**b**)

Figure 11. Results of the prior model. (**a**) IV model; (**b**) MBKM model. Absence samples were created by random sampling within 10 sampling intervals delineated by landslide susceptibility.

The 1,559,752 pixels with information values were normalized and processed into the MBKM algorithm for clustering calculations. The calculation process was based on Python's msmbuilder library [57]. The specific parameters were set as follows: n_clusters is 10, max_iter is 1000, batch_size is 20,000, and max_no_improvement is 50. After the clustering calculation, 10 clusters and 10 susceptibility rank intervals were obtained by sorting the size according to the information content value of each cluster center. The preliminary mapping of landslide susceptibility based on MBKM is shown in Figure 11b. The absence sample sampling threshold was set according to the natural break, and a total of 10 sampling intervals were divided as in the IV model: <0.1 through <1.0.

Absence samples extracted from low susceptibility intervals exhibited a lower probability of landslide occurrence in the IV and MBKM models. When the sampling threshold was close to 1.0, the absence samples were almost randomly selected over the entire area. Twenty sets of samples were created from each interval for training and prediction in the prior models of IV and MBKM.

The idea behind integrative sampling is to feed the model with as many samples with different characteristics as possible. Since the four sampling methods of BCS, CTSES, IV, and MBKM have different sampling intervals, the default interval with the best prediction performance metrics was the target interval for integrative sampling. According to the results (reported in Section 4.4), the 1000 m interval in BCS, the d = 4 intervals in CTSES, and the 0–0.1 interval in IV and MBKM had the best metrics (accuracy, AUC, KC, POA) such that the samples in them could input more correct absence sample information to the model. The selected absence sample sampling groups were BCS-1000m, CTSES-d = 4, IV-0.1, and MBKM-0.1 for integrative sampling. The final input from the integrative samples was obtained using random sampling from the intervals of different absence sampling methods. The ratio of mixing was 1:1:1:1 to evaluate the mixing effect simply. The number of absence samples in the input of prediction models was set to 217 and 506 to achieve presence-to-absence sample ratios of 1:1 and 3:7, respectively. The purpose was to test the effect of the training ratio on susceptibility prediction. This setting can explore whether increasing the proportion of absence samples reduces the overestimation of susceptibility levels by comparing the results of different sample ratios of 1:1 and 3:7.

4.3. Landslide Susceptibility Mapping

4.3.1. LSM Results for Four Sampling Strategies with Different Sampling Intervals

The absence samples obtained in each of the above sampling methods were fed into the SVM and RF prediction models for training and prediction. The ratio of landslide samples to absence samples was set to 1:1. Optimal hyperparameters for SVM and RF prediction models were found by grid search based on 10-fold cross-validation. In the SVM training process, the RBF-kernel was used; the grid search range for hyperparameters C and gamma was set to 0–20 and 0–1.2. In the training process of RF, the grid search range for the hyperparameters n_estimators, max_depth, and min_samples_split were set to 10–100, 10–50, and 10–40, respectively. There were 10/20 sets of samples under each sampling threshold for each method, and the set with the highest AUC was selected for LSM visualization.

Figure 12 shows the results of the SVM-based models for different sampling intervals under four absence sampling strategies. In the BCS sampling method, the high landslide susceptibility areas were concentrated in the eastern part of Songyang County and near the east-west spreading gullies. Districts in the central flatlands exhibit a lower susceptibility to landslides. As the size of the landslide buffer increases, the distribution pattern of landslide susceptibility size little changes until it increases to 1000 m and 1500 m when the area of the high susceptibility zone in the east increases significantly. Under the CTSES sampling method, the high susceptibility areas at $d = 1$ are distributed in the eastern and central districts around the county seat and on the slopes on both sides of the gully. With the increase of parameter d, expansion of the high landslide susceptibility area and dispersion of the susceptibility index (deepening of LSM color) can be observed. When $d = 4$, most of the area is classified as having high susceptibility except for a small number of gullies, mountain fronts, and flatlands in the central county. The same expansion and dispersion of susceptibility are observed in the lower interval sampling of the IV prior model. When the sampling threshold of IV is 0.1 or 0.2, the high susceptibility area covers the majority of the study area. These two phenomena decrease as the IV sampling threshold increases and eventually stabilize above a sampling threshold of 0.5; they are close to the BCS predictions. The prediction model law based on the MBKM method of sampling is similar to that of the IV model, but the polarization and expansion phenomena are greatly reduced compared with the IV model, and most high landslide susceptibility areas in the low sampling interval are distributed on both sides of the gullies, which seems more reasonable. In addition, more areas of high landslide susceptibility occur in the central flatland county in the low sampling interval (threshold < 0.6) of the IV and MBKM sampling models and in the $d = 3$ and 4 intervals of the CTSES, which are mitigated in the other sampling intervals.

The RF-based model results for different sampling intervals under four absence sampling strategies are shown in Figure 13. The distribution pattern of susceptibility in the RF model is similar to that of the SVM, but there are significant differences. In the BCS method, when the buffer zone is 500 m, 800 m and 1000 m, the high susceptibility zone is more concentrated in the gully without excessive expansion or confusion. In the three sampling methods of CTSES, IV, and MBKM, expansion of high landslide susceptibility areas and dispersion of landslide susceptibility indices can be found with changes in sampling interval, which is consistent with the SVM prediction model. However, it is important to note that the expansion and polarization of the RF model are milder than those of the SVM model. Regardless, the expansion of high landslide susceptibility in the intermediate flatland counties can still be seen in the low sampling interval of the IV and MBKM sampling models and in the $d = 4$ sampling interval of CTSES.

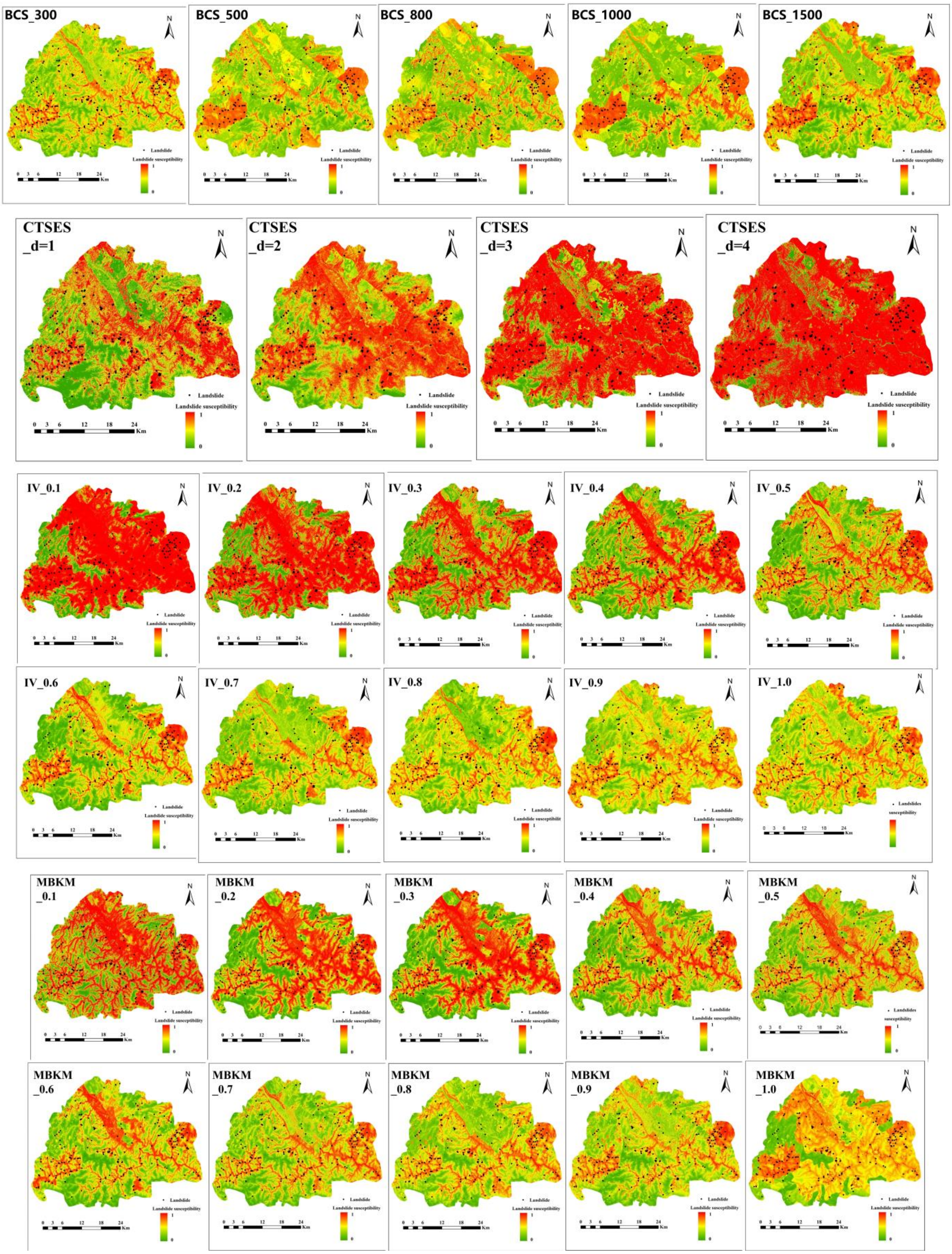


Figure 12. SVM-based LSM results of four absence sampling methods.

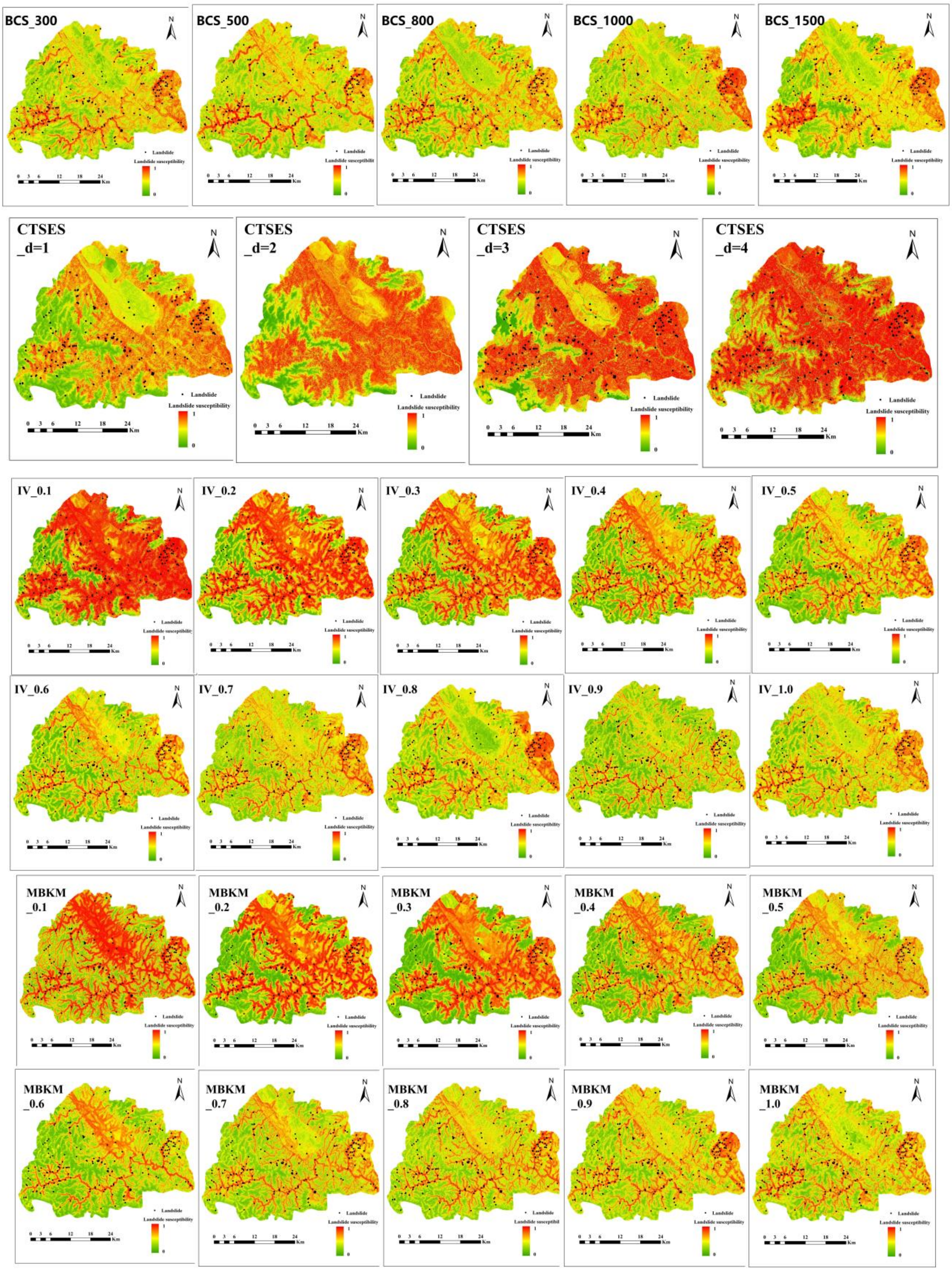


Figure 13. RF-based LSM results of four absence sampling methods.

4.3.2. LSM Results of the Integrative Sampling Method

The prediction plots of landslide susceptibility in SVM and RF prediction models with integrative sampling methods are shown in Figure 14; 1:1 and 3:7 presence-absence ratio sample sets are included respectively. In the training process of the SVM and RF, the hyperparameters are searched in the same range as in Section 4.3.1 for the grid search based on 10-fold cross-validation. The high landslide susceptibility areas are basically distributed on and on both sides of the gully, in the eastern part of Songyang County, and around the central prefecture. Widespread mountain peaks and flat areas in the central prefecture are designated as safety zones. The results of the integrative sampling did not reveal over-expansion or diversion of the high landslide susceptibility area, which is more reasonable than the results obtained from the previous CTSES, IV, and MBKM sampling. Comparing the results of presence-absence sample ratios of 1:1 and 3:7, the former has a larger area of high landslide susceptibility. The high susceptibility area predicted by SVM is slightly higher than in RF.

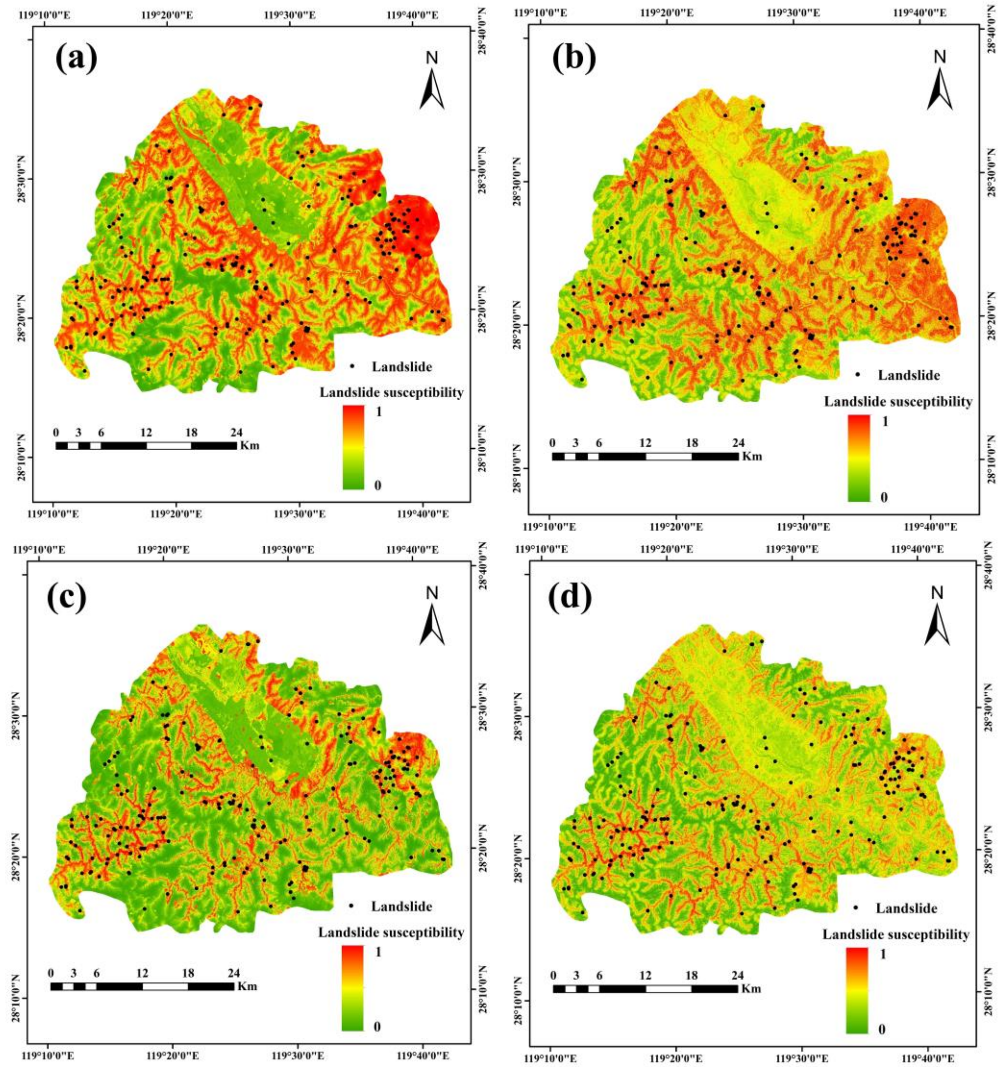


Figure 14. LSM of integrative sampling with different ratios. (**a**) SVM_IS_1:1; (**b**) RF_IS_1:1; (**c**) SVM_IS_3:7; (**d**) RF_IS_3:7.

4.4. Evaluation of Different Absence Sampling Methods

4.4.1. Model Accuracy of Four Absence Sampling Methods with Respective Sample Intervals

The training and testing accuracy of the models for different sampling intervals under each absence sampling method is presented in the form of box-line plots in Figures 15 and 16.

In the SVM and RF prediction models, a consistent pattern was found in the variation of accuracy with different sampling intervals. In the BCS method, the accuracy of the training and testing under SVM prediction increased slightly overall when the buffer rose from 300 to 1000 m and decreased slightly by 1500 m. In the RF model, training accuracy was relatively stable, mostly maintained above 0.9, and the amount and trend of testing accuracy were similar to those of the SVM prediction model. In CTSES, as *d* increased from 1 to 4, the accuracy of the training and testing sets under the SVM and RF prediction models tended to increase. The testing accuracy of SVM and RF reached 0.95 (nearly 1), as *d* = 4 and 3, respectively. In the IV and MBKM models, the accuracy of SVM decreased with increasing sampling threshold, and the testing accuracy of RF also decreased, but the training accuracy was maintained at a high level (>0.9). In addition, the testing accuracy of IV was higher than that of MBKM at the sampling thresholds of 0.1, 0.2, 0.3, 0.4, and 0.5; when the threshold was lowest, the testing accuracy was generally greater than 0.9 or close to 1.

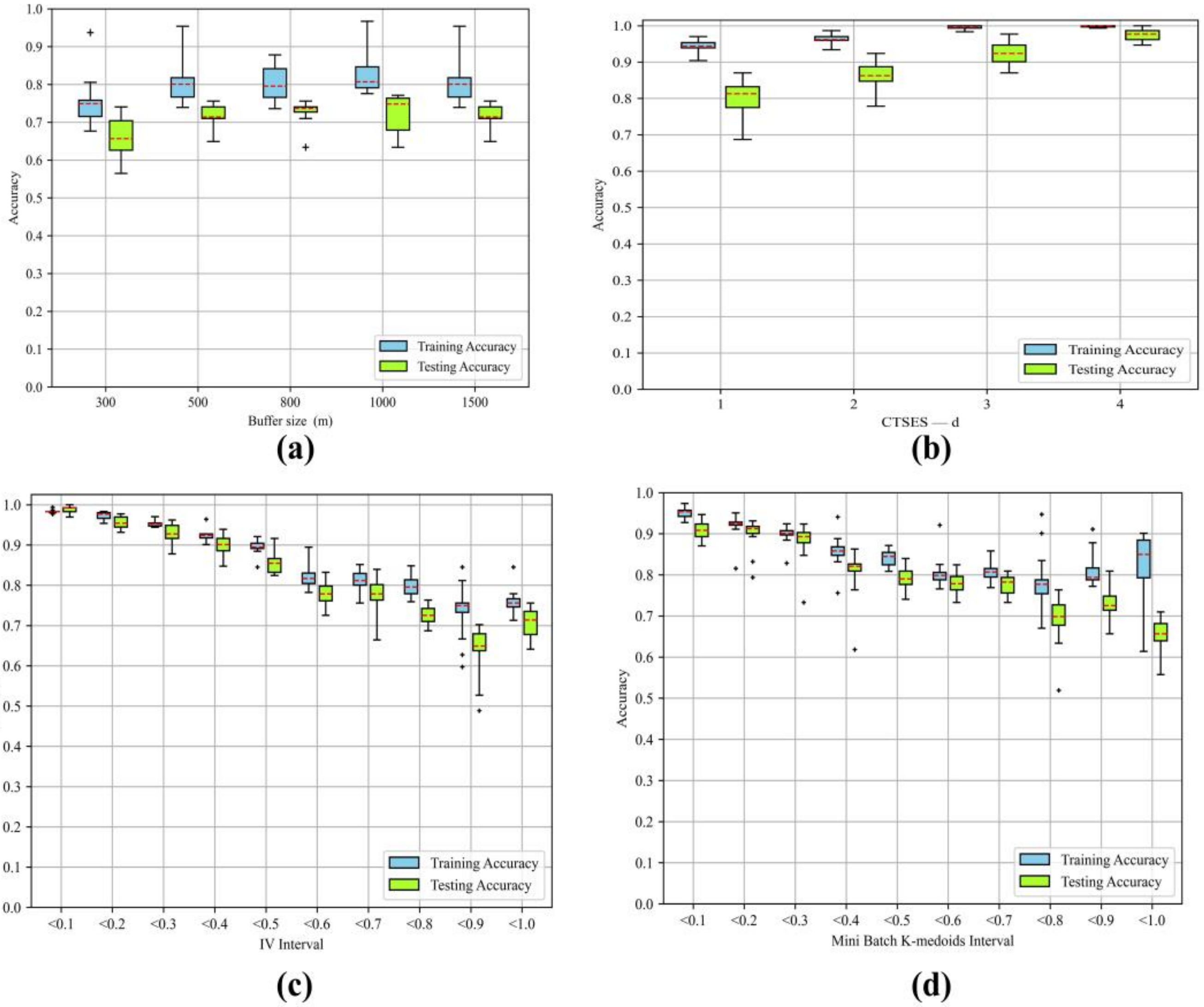


Figure 15. Accuracy results of SVM-based absence sampling. (**a**) BCS; (**b**) CTSES; (**c**) IV; (**d**) MBKM. The "+" represents a small number of abnormal values that are outside the normal range.

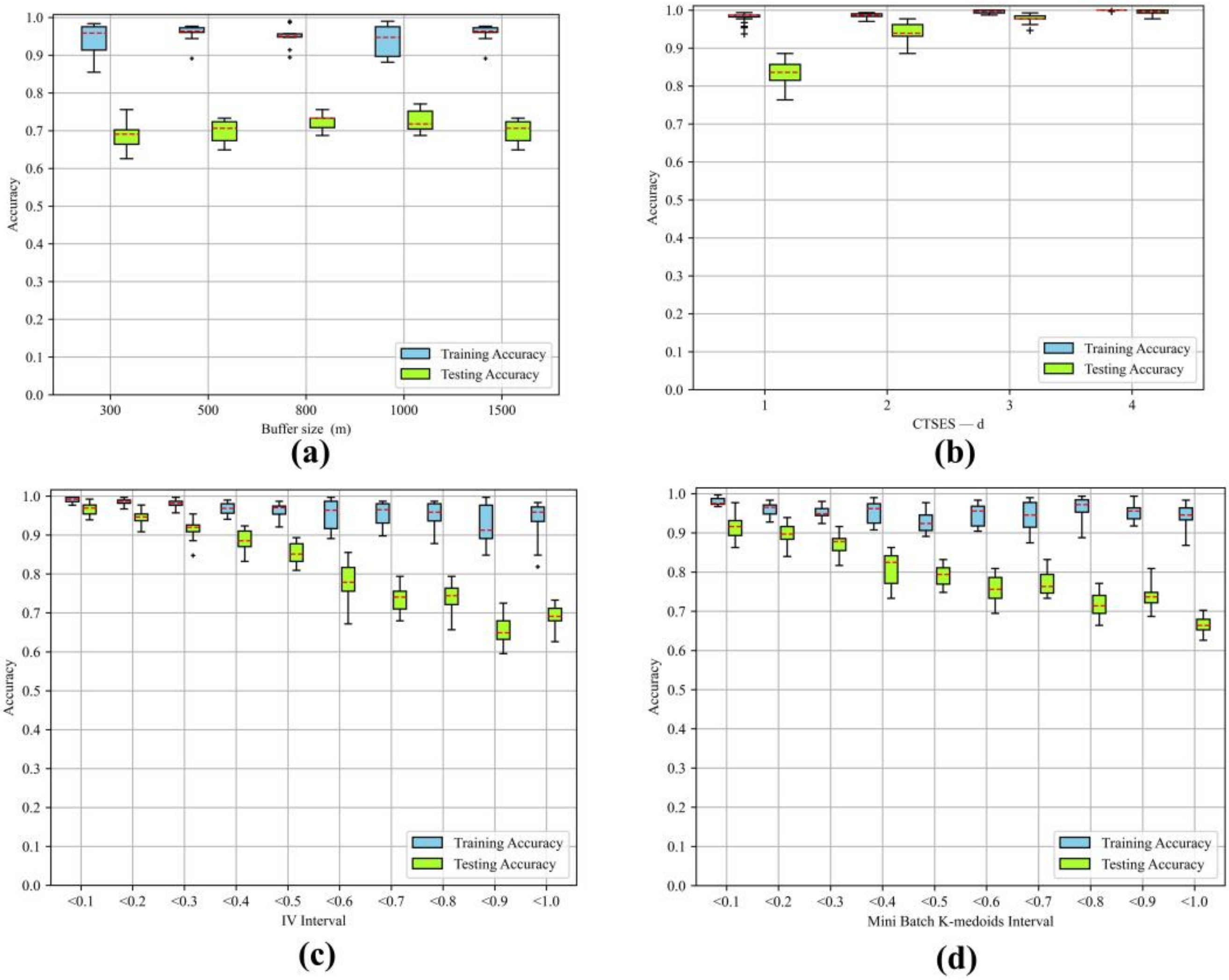


Figure 16. Accuracy results of RF-based absence sampling. (**a**) BCS; (**b**) CTSES; (**c**) IV; (**d**) MBKM. The "+" represents a small number of abnormal values that are outside the normal range.

4.4.2. Model Comprehensive Predictive Performance of Four Absence Sampling Methods with Respective Sample Intervals

Three metrics, namely AUC, KC, and POA, were used to weigh the comprehensive predictive performance of each model consisting of each group of samples. They are presented as box plots in Figures 17 and 18. The three metrics have a consistent pattern. In BCS, they increased slightly when the buffer size changed from 300 to 1000 m and decreased slightly to 1500 m. In BCS, the AUC was around 0.8, the KC did not exceed 0.6, and the POA reached a maximum of approximately 2.0. In CTSES, the three metrics gradually increased as *d* varied from 1 to 4. In IV and MBKM, the three metrics gradually decreased as the sampling threshold increased. In addition, the AUC approached 1 and even the POA approached 3 at low sampling intervals in IV and MBKM or a higher *d* in CTSES. In both the RF and SVM prediction models, the above-mentioned pattern was present.

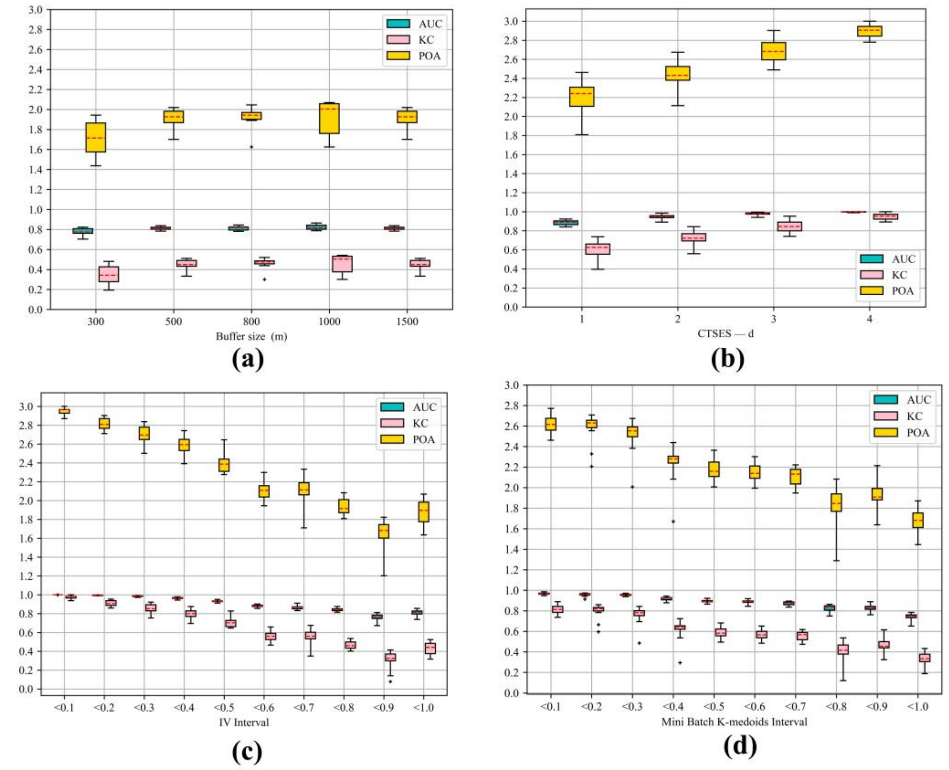


Figure 17. Prediction performance results of SVM-based absence sampling. (**a**) BCS; (**b**) CTSES; (**c**) IV; (**d**) MBKM.

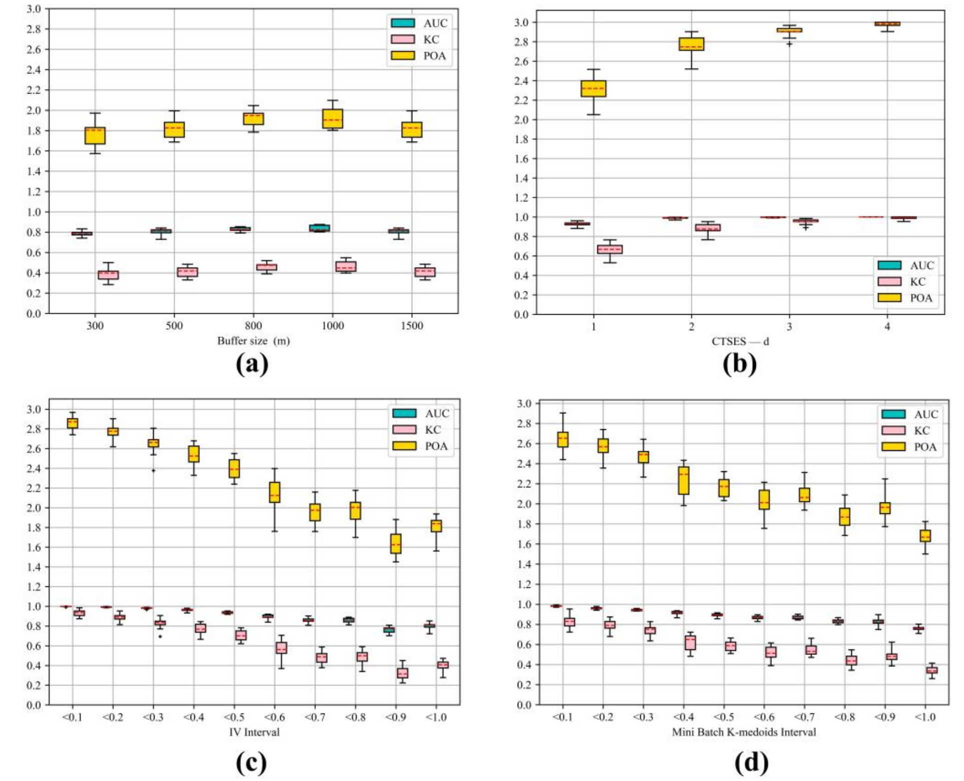


Figure 18. Prediction performance results of RF-based absence sampling. (**a**) BCS; (**b**) CTSES; (**c**) IV; (**d**) MBKM.

4.4.3. Model Susceptibility Distribution of Four Absence Sampling Methods with Respective Sample Intervals

Figure 19 shows the means and standard deviations of susceptibility for four absence sampling strategies with different intervals. In BCS, the change in mean and standard deviation was negligible. In the remaining three models, as *d* increased or the sampling threshold decreased, there was a significant increase in the mean and a weaker upward trend in the standard deviation. In CTSES, the change in the mean value was particularly pronounced, rising from 0.4 to 0.87. In IV and MBKM, the maximum mean values were approximately 0.65 and 0.8, respectively, when the threshold was 0.1. In terms of standard deviation, BCS had the lowest with a maximum value of about 0.2, while the other methods had a maximum value of 0.3 or above. In the comparison of the prediction models, the mean distributions of SVM and RF were about the same in BCS and IV, slightly higher for RF than SVM in CTSES, and higher for SVM than RF at low thresholds in MBKM. In addition, almost all standard deviations of the SVM model were larger than those of the RF model.

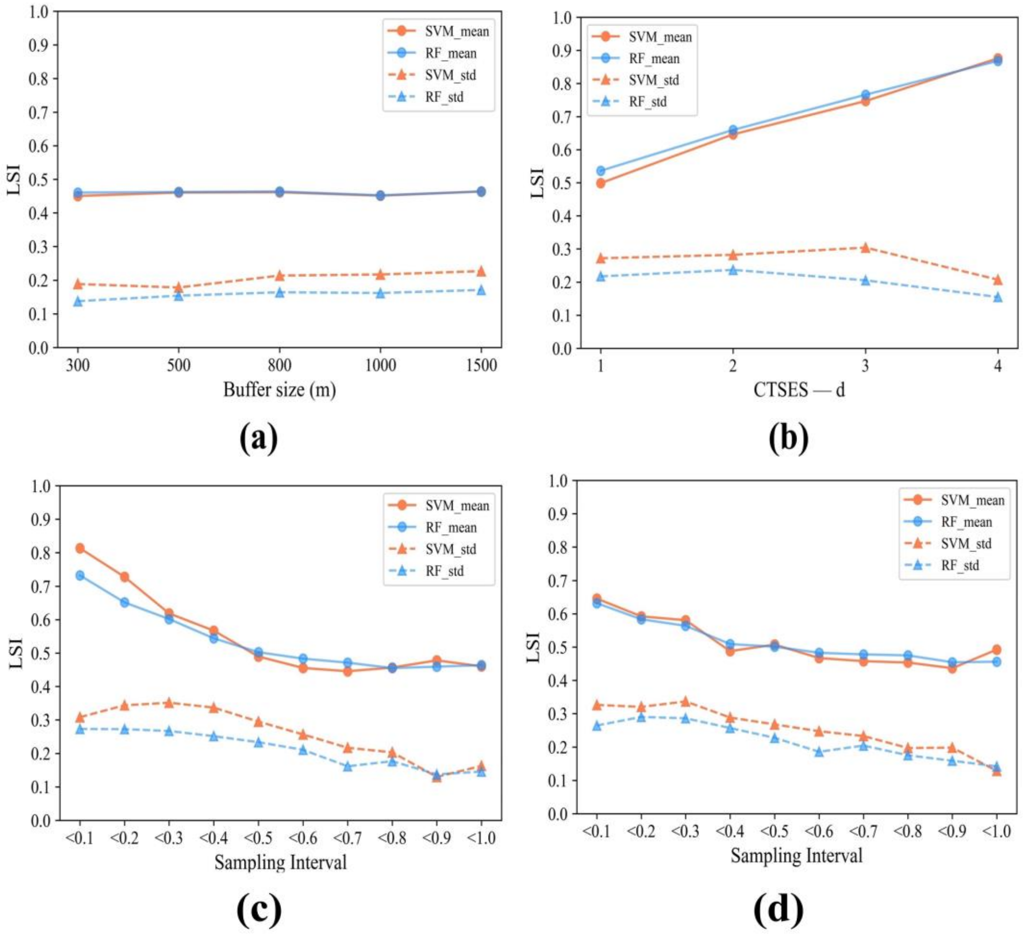


Figure 19. Means and standard deviations of four absence sampling methods with different intervals. (**a**) BCS; (**b**) CTSES; (**c**) IV; (**d**) MBKM.

4.4.4. Evaluation of the Integrative Sampling Model

The prediction model evaluation metrics for the integrative sampling approach are shown in Table 4. The training set accuracies were all greater than 0.9, while the training set accuracy of RF was maintained at a high level (>0.95). The testing accuracy and comprehensive predictive performance metrics, including the AUC, KC, and POA of SVM, were less than those of RF. When the ratio was 1:1, the various metrics fell within the region of the BCS and the other three models. The combined predictive performance of the integrative sampling model was stronger than that of BCS but weaker than the performances of CTSES, IV, and MBKM under the same set of absence sample intervals. When the ratio changed from 1:1 to 3:7, each parameter decreased slightly, resulting in a slight decrease in prediction performance. In terms of susceptibility distribution, the mean value of integrative sampling was slightly higher than that of BCS but significantly lower than those of CTSES, IV, and MBKM. When the ratio changed from 1:1 to 3:7, further significant decreases in the mean value occurred (from 0.52 and 0.56 to 0.36 and 0.39).

Table 4. Metrics of the integrative sampling model.

Predictive Model	Presence: Absence	Training Accuracy	Testing Accuracy	AUC	KC	POA	Susceptibility Mean	Susceptibility SD
SVM	1:1	0.90	0.77	0.89	0.55	2.13	0.52	0.29
	3:7	0.93	0.81	0.87	0.56	2.06	0.36	0.28
RF	1:1	0.96	0.86	0.92	0.73	2.46	0.56	0.23
	3:7	0.97	0.83	0.91	0.60	2.15	0.39	0.21

5. Discussion

5.1. Effects of Absence Sampling Strategies and Sample Quality on LSM

Each of the four absence sample sampling models has its own control parameters. Figures 12 and 13 show the sensitivity of LSM to these parameters, and Figures 15–19 show the sensitivity of the model evaluation metrics to the parameters.

For BCS, various evaluation metrics show that buffer size has the best performance effect at 1000 m. The model's performance rises first when the buffer increases and then decreases when it increases to a certain level (1500 m). This scenario suggests that there may be a buffer size that is most suitable for generating negative samples. However, compared to the feature screening methods in CTSES, IV, and MBKM, the LSM prediction patterns and prediction performance metrics, as well as the susceptibility distribution of BCS, show little variation. This indicates that the improved geospatial-based absence sampling method has a certain upper limit and is less sensitive.

CTSES, IV, and MBKM were found to have significant sensitivity in LSM as well as various metrics to the control parameters, the main reason for which was absence sampling methods based on feature screening patterns. The control parameter d in CTSES and the sampling thresholds in IV and MBKM can be good representations of the negative sample quality. The larger the d or the lower the sampling threshold, the greater the difference between the features of the absence samples and the landslide samples, and the safer the region of the absence samples. These control parameters control the reliability or certainty of the absence of samples in the input model.

When the reliability and certainty of the absence samples are enhanced, an expansion of the high landslide susceptibility area and a divergence of the susceptibility index can be observed in the LSM results of these models and in the mean susceptibility level. Similar results can also be seen in other studies that have classified absence sample reliability [22]. These phenomena are particularly pronounced in IV and CTSES and slightly less pronounced in MBKM. At the same time, the model's predictive power indicators (accuracy, AUC, KC, and POA) are significantly improved or even close to full scores. These phenomena are due to the fact that when the non-landslide samples are too absent, the prediction model learns only the extreme features of the absent samples and the features of the landslides in the learning process. For those intermediate points between the extreme absence sample and the deterministic landslide sample, the model does not have enough information to assess them and overestimates them to a higher level of susceptibility. Since landslide samples and extreme absence samples are used for validation scores, the model's susceptibility overestimation will not lead to misjudgment of these samples; therefore, very high prediction scores can be obtained. The overestimation of the overall regional susceptibility by the model due to extreme samples is obviously unreasonable and difficult to apply. This can be mitigated by reducing the *d* in CTSES or increasing the sampling threshold in IV and MBKM, but it will cause some degradation in prediction performance.

5.2. Advantages of Integrative Sampling

The integrative sampling absence sample set is combined with highly reliable absence samples, including CTSES-d = 4, IV-0.1, MBKM-0.1, and BCS-1000. Looking at the LSM prediction results for the integrative sampling method, it is surprising to see that the prediction pattern appears very reasonable. The flatland county in the central part of Songyang County is no longer misclassified as a high susceptibility area, and the excessive extension of the high susceptibility area in the gully to the banks has been significantly suppressed. The high landslide susceptibility areas are basically distributed in the gully and both sides of the gully, the eastern part of Songyang County, and around the central prefecture. The same samples (CTSES-d = 4, IV-0.1 and MBKM-0.1) would still be susceptible to overestimation and differentiation if returned to the original single absence sample sampling model, as Figures 12–14 show, but the integrated sampling method overcomes this. In addition, the normal mean level at a presence-absence ratio of 1:1 also indicates that the overall overestimation and divergence of regional susceptibility are greatly mitigated in integrative sampling (0.52 in SVM and 0.56 in RF compared with a max mean value of 0.87 in CTSES). Although the predictive assessment metrics (accuracy, AUC, KC, and POA) are lower than in the CTSES, IV, and MBKM methods, they are still much higher than in the BCS method. As shown in Table 4, the AUC of the integrative sampling method is around 0.9 for both SVM and RF models, with presence-absence ratios of 1:1 and 3:7, and the other metrics are within the excellent range (KC > 0.55, POA > 2.06).

Integrative sampling has the above advantages, mainly because the diversity of absence samples allows the prediction model to learn more non-landslide features. When predicting objects that are ambiguous, the model is able to synthesize the susceptibility value of the object based on the sample characteristics of each source. In the BCS individual model, the selection of absence samples is highly random and has the potential to give the model the wrong samples. In the CTSES, IV, and MBKM single models, the combination of features with high reliability absence samples may be highly similar, with insufficient information for the model to predict medium susceptibility objects. Integrative sampling combines various types of absence samples to obtain more information about the combination of absence sample features of CTSES, IV, and MBKM and also shares the error introduction rate of BCS. Thus, integrative sampling would give the predictive model a broader predictive capability rather than focusing on very low or very high susceptibility zones.

In addition, when the presence-absence ratio was changed from 1:1 to 3:7 in integrative sampling, a reduction in the high susceptibility zone and a decrease in the mean value of susceptibility were observed. Metrics of model predictive performance are reduced but within reasonable limits. Therefore, it is inferred that the overestimation level of regional susceptibility can be reduced by increasing the proportion of absence samples in the training process, as in another study [29]. Such an approach can be used in situations where the overall regional susceptibility is too high or where very high susceptibility areas need to be located quickly.

6. Conclusions

Four absence sampling methods were employed in LSM, namely BCS, CTSES, IV, and MBKM, and a new method of integrative sampling was proposed. In these methods, the TSES method was improved to obtain the CTSES method, and the less commonly used optimized clustering algorithm MBKM was introduced and showed a more reasonable prediction pattern than the IV model with the same set of samples. The absence of samples generated by these four methods were mixed for the test of the integrative sampling method. The above methods were applied to LSM in Songyang County, Lishui City, Zhejiang Province, China, and the effects of various absence sampling methods and the quality of absence samples in LSM were discussed. Multiple sets of sampling intervals were set for each method, and extensive training computations were conducted in each interval. The results indicate that, as the certainty and reliability of the sample increase, expansion of the high susceptibility region and divergence of the susceptibility index occur in the CTSES, IV, and MBKM methods. This phenomenon can be significantly reduced, and very satisfactory prediction results can be achieved by employing the proposed integrative sampling method or a higher proportion of absence samples. Integrative sampling has a stronger prediction performance than the general independent absence sample sampling method and can combine the sample information of each sampling method for a more reasonable classification of high, low, and medium susceptibility areas. The integrative sampling method presents a valuable approach for constructing absence samples in LSM based on data-driven models.

Author Contributions: Z.F. drafted the manuscript and was responsible for the research design, experiment, and analysis; F.W. provided the supervision, funding acquisition, and overall recommendation for the study area. F.W., J.D., K.N. and H.M. contributed to reviewing and editing the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the National Natural Science Foundation of China (grant nos. 42230715).

Data Availability Statement: The data that support the findings of this study are available from the first author upon reasonable request.

Acknowledgments: Thanks to the Geological hazard risk census report of Songyang County, Zhejiang Province for the basic landslide field information for this research.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wang, F.; Sassa, K. Landslide Simulation by a Geotechnical Model Combined with a Model for Apparent Friction Change. *Phys. Chem. Earth Parts ABC* **2010**, *35*, 149–161. [CrossRef]
2. Adnan, M.S.G.; Rahman, M.S.; Ahmed, N.; Ahmed, B.; Rabbi, M.F.; Rahman, R.M. Improving Spatial Agreement in Machine Learning-Based Landslide Susceptibility Mapping. *Remote Sens.* **2020**, *12*, 3347. [CrossRef]
3. Scaioni, M.; Longoni, L.; Melillo, V.; Papini, M. Remote Sensing for Landslide Investigations: An Overview of Recent Achievements and Perspectives. *Remote Sens.* **2014**, *6*, 9600–9652. [CrossRef]
4. Zhao, C.; Lu, Z. Remote Sensing of Landslides—A Review. *Remote Sens.* **2018**, *10*, 279. [CrossRef]
5. Mohan, A.; Singh, A.K.; Kumar, B.; Dwivedi, R. Review on Remote Sensing Methods for Landslide Detection Using Machine and Deep Learning. *Trans. Emerg. Telecommun. Technol.* **2021**, *32*, e3998. [CrossRef]
6. Guzzetti, F.; Reichenbach, P.; Cardinali, M.; Galli, M.; Ardizzone, F. Probabilistic Landslide Hazard Assessment at the Basin Scale. *Geomorphology* **2005**, *72*, 272–299. [CrossRef]
7. Corominas, J.; van Westen, C.; Frattini, P.; Cascini, L.; Malet, J.-P.; Fotopoulou, S.; Catani, F.; Van Den Eeckhaut, M.; Mavrouli, O.; Agliardi, F.; et al. Recommendations for the Quantitative Analysis of Landslide Risk. *Bull. Eng. Geol. Environ.* **2014**, *73*, 209–263. [CrossRef]
8. Li, L.; Lan, H.; Guo, C.; Zhang, Y.; Li, Q.; Wu, Y. A Modified Frequency Ratio Method for Landslide Susceptibility Assessment. *Landslides* **2017**, *14*, 727–741. [CrossRef]
9. Dou, J.; Xiang, Z.; Qiang, X.; Zheng, P.; Wang, X.; Su, A.; Liu, J.; Luo, W. Application and Development Trend of Machine Learning in Landslide Intelligent Disaster Prevention and Mitigation. *Earth Sci.* **2022**.
10. Merghadi, A.; Yunus, A.P.; Dou, J.; Whiteley, J.; ThaiPham, B.; Bui, D.T.; Avtar, R.; Abderrahmane, B. Machine Learning Methods for Landslide Susceptibility Studies: A Comparative Overview of Algorithm Performance. *Earth-Sci. Rev.* **2020**, *207*, 103225. [CrossRef]
11. Nam, K.; Wang, F. An Extreme Rainfall-Induced Landslide Susceptibility Assessment Using Autoencoder Combined with Random Forest in Shimane Prefecture, Japan. *Geoenvironmental Disasters* **2020**, *7*, 6. [CrossRef]
12. Nam, K.; Wang, F. The Performance of Using an Autoencoder for Prediction and Susceptibility Assessment of Landslides: A Case Study on Landslides Triggered by the 2018 Hokkaido Eastern Iburi Earthquake in Japan. *Geoenvironmental Disasters* **2019**, *6*, 19. [CrossRef]
13. Huang, F.; Zhang, J.; Zhou, C.; Wang, Y.; Huang, J.; Zhu, L. A Deep Learning Algorithm Using a Fully Connected Sparse Autoencoder Neural Network for Landslide Susceptibility Prediction. *Landslides* **2020**, *17*, 217–229. [CrossRef]
14. Huang, F.; Cao, Z.; Guo, J.; Jiang, S.-H.; Li, S.; Guo, Z. Comparisons of Heuristic, General Statistical and Machine Learning Models for Landslide Susceptibility Prediction and Mapping. *Catena* **2020**, *191*, 104580. [CrossRef]

15. Huang, F.; Xiong, H.; Yao, C.; Catani, F.; Zhou, C.; Huang, J. Uncertainties of Landslide Susceptibility Prediction Considering Different Landslide Types. *J. Rock Mech. Geotech. Eng.* **2023**. [CrossRef]
16. Zhu, A.-X.; Miao, Y.; Yang, L.; Bai, S.; Liu, J.; Hong, H. Comparison of the Presence-Only Method and Presence-Absence Method in Landslide Susceptibility Mapping. *Catena* **2018**, *171*, 222–233. [CrossRef]
17. Atkinson, P.M.; Massari, R. Generalised linear modelling of susceptibility to landsliding in the central apennines, Italy. *Comput. Geosci.* **1998**, *24*, 373–385. [CrossRef]
18. Hearn, G.J.; Hart, A.B. Landslide Susceptibility Mapping: A Practitioner's View. *Bull. Eng. Geol. Environ.* **2019**, *78*, 5811–5826. [CrossRef]
19. Carrara, A.; Cardinali, M.; Guzzetti, F.; Reichenbach, P. Gis Technology in Mapping Landslide Hazard. In *Geographical Information Systems in Assessing Natural Hazards*; Carrara, A., Guzzetti, F., Eds.; Advances in Natural and Technological Hazards Research; Springer: Dordrecht, The Netherlands, 1995; pp. 135–175. ISBN 978-94-015-8404-3.
20. Dou, J.; Yunus, A.P.; Merghadi, A.; Shirzadi, A.; Nguyen, H.; Hussain, Y.; Avtar, R.; Chen, Y.; Pham, B.T.; Yamagishi, H. Different Sampling Strategies for Predicting Landslide Susceptibilities Are Deemed Less Consequential with Deep Learning. *Sci. Total Environ.* **2020**, *720*, 137320. [CrossRef]
21. Pourghasemi, H.R.; Kornejady, A.; Kerle, N.; Shabani, F. Investigating the Effects of Different Landslide Positioning Techniques, Landslide Partitioning Approaches, and Presence-Absence Balances on Landslide Susceptibility Mapping. *Catena* **2020**, *187*, 104364. [CrossRef]
22. Zhu, A.-X.; Miao, Y.; Liu, J.; Bai, S.; Zeng, C.; Ma, T.; Hong, H. A Similarity-Based Approach to Sampling Absence Data for Landslide Susceptibility Mapping Using Data-Driven Methods. *Catena* **2019**, *183*, 104188. [CrossRef]
23. Sameen, M.I.; Pradhan, B.; Bui, D.T.; Alamri, A.M. Systematic Sample Subdividing Strategy for Training Landslide Susceptibility Models. *Catena* **2020**, *187*, 104358. [CrossRef]
24. Lucchese, L.V.; de Oliveira, G.G.; Pedrollo, O.C. Investigation of the Influence of Nonoccurrence Sampling on Landslide Susceptibility Assessment Using Artificial Neural Networks. *Catena* **2021**, *198*, 105067. [CrossRef]
25. Pradhan, B.; Lee, S.; Buchroithner, M.F. A GIS-Based Back-Propagation Neural Network Model and Its Cross-Application and Validation for Landslide Susceptibility Analyses. *Comput. Environ. Urban Syst.* **2010**, *34*, 216–235. [CrossRef]
26. Wang, H.; Zhang, L.; Luo, H.; He, J.; Cheung, R.W.M. AI-Powered Landslide Susceptibility Assessment in Hong Kong. *Eng. Geol.* **2021**, *288*, 106103. [CrossRef]
27. Chang, Z.; Huang, J.; Huang, F.; Bhuyan, K.; Meena, S.R.; Catani, F. Uncertainty Analysis of Non-Landslide Sample Selection in Landslide Susceptibility Prediction Using Slope Unit-Based Machine Learning Models. *Gondwana Res.* **2023**, *117*, 307–320. [CrossRef]
28. Xiao, C.; Tian, Y.; Shi, W.; Guo, Q.; Wu, L. A New Method of Pseudo Absence Data Generation in Landslide Susceptibility Mapping with a Case Study of Shenzhen. *Sci. China Technol. Sci.* **2010**, *53*, 75–84. [CrossRef]
29. Hong, H.; Miao, Y.; Liu, J.; Zhu, A.-X. Exploring the Effects of the Design and Quantity of Absence Data on the Performance of Random Forest-Based Landslide Susceptibility Mapping. *Catena* **2019**, *176*, 45–64. [CrossRef]
30. Rabby, Y.W.; Li, Y.; Hilafu, H. An Objective Absence Data Sampling Method for Landslide Susceptibility Mapping. *Sci. Rep.* **2023**, *13*, 1740. [CrossRef]
31. Yuan, X.; Liu, C.; Nie, R.; Yang, Z.; Li, W.; Dai, X.; Cheng, J.; Zhang, J.; Ma, L.; Fu, X.; et al. A Comparative Analysis of Certainty Factor-Based Machine Learning Methods for Collapse and Landslide Susceptibility Mapping in Wenchuan County, China. *Remote Sens.* **2022**, *14*, 3259. [CrossRef]
32. Zhao, B.; Ge, Y.; Chen, H. Landslide Susceptibility Assessment for a Transmission Line in Gansu Province, China by Using a Hybrid Approach of Fractal Theory, Information Value, and Random Forest Models. *Environ. Earth Sci.* **2021**, *80*, 441. [CrossRef]
33. Xu, C.; Zhang, W.; Yi, Y.; Xu, Q. Landslide Susceptibility Mapping Using Logistic Regression Model Based on Information Value for the Region Along China-Thailand Railway from Saraburi to Sikhio, Thailand. In Proceedings of the IGARSS 2019—2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 31 July 2019; pp. 9650–9653.
34. Zhao, Z.; Liu, Z.Y.; Xu, C. Slope Unit-Based Landslide Susceptibility Mapping Using Certainty Factor, Support Vector Machine, Random Forest, CF-SVM and CF-RF Models. *Front. Earth Sci.* **2021**, *9*, 589630. [CrossRef]
35. Ji, J.; Zhou, Y.; Cheng, Q.; Jiang, S.; Liu, S. Landslide Susceptibility Mapping Based on Deep Learning Algorithms Using Information Value Analysis Optimization. *Land* **2023**, *12*, 1125. [CrossRef]
36. Li, Y.; Deng, X.; Ji, P.; Yang, Y.; Jiang, W.; Zhao, Z. Evaluation of Landslide Susceptibility Based on CF-SVM in Nujiang Prefecture. *Int. J. Environ. Res. Public. Health* **2022**, *19*, 14248. [CrossRef]
37. Huang, F.; Yin, K.; Huang, J.; Gui, L.; Wang, P. Landslide Susceptibility Mapping Based on Self-Organizing-Map Network and Extreme Learning Machine. *Eng. Geol.* **2017**, *223*, 11–22. [CrossRef]
38. Kaboutari, A.; Bagherzadeh, J.; Kheradmand, F. An Evaluation of Two-Step Techniques for Positive-Unlabeled Learning in Text Classification. *Int. J. Comput. Appl. Technol. Res.* **2014**, *3*, 592–594. [CrossRef]
39. Huang, F.; Cao, Z.; Jiang, S.-H.; Zhou, C.; Huang, J.; Guo, Z. Landslide Susceptibility Prediction Based on a Semi-Supervised Multiple-Layer Perceptron Model. *Landslides* **2020**, *17*, 2919–2930. [CrossRef]
40. Yao, J.; Qin, S.; Qiao, S.; Liu, X.; Zhang, L.; Chen, J. Application of a Two-Step Sampling Strategy Based on Deep Neural Network for Landslide Susceptibility Mapping. *Bull. Eng. Geol. Environ.* **2022**, *81*, 148. [CrossRef]

41. Chang, Z.; Du, Z.; Zhang, F.; Huang, F.; Chen, J.; Li, W.; Guo, Z. Landslide Susceptibility Prediction Based on Remote Sensing Images and GIS: Comparisons of Supervised and Unsupervised Machine Learning Models. *Remote Sens.* **2020**, *12*, 502. [CrossRef]
42. Zhu, L.P.; Chen, R.; Zeng, J.W.; Liao, S.B.; Yang, Z.L. Main structural characteristics of Yanshanian in Shengzhou area of Yuyao-Lishui fault zone (in Chinese). *Chin. Geol. Surv.* **2018**, *5*, 49–57.
43. Chen, L.F. Study on the Activity of NE Trending Faults along the Coast of Zhejiang Province (in Chinese). Master's Thesis, Zhejiang University, Hangzhou, China, 2010.
44. Wang, F.; Chen, Y.; Peng, X.; Zhu, G.; Yan, K.; Ye, Z. The Fault-Controlled Chengtian Landslide Triggered by Rainfall on 20 May 2021 in Songyang County, Zhejiang Province, China. *Landslides* **2022**, *19*, 1751–1765. [CrossRef]
45. Fabbri, A.G.; Chung, C.-J.F.; Cendrero, A.; Remondo, J. Is Prediction of Future Landslides Possible with a GIS? *Nat. Hazards* **2003**, *30*, 487–503. [CrossRef]
46. Yi, Y.; Zhang, Z.; Zhang, W.; Jia, H.; Zhang, J. Landslide Susceptibility Mapping Using Multiscale Sampling Strategy and Convolutional Neural Network: A Case Study in Jiuzhaigou Region. *Catena* **2020**, *195*, 104851. [CrossRef]
47. Xi, C.; Han, M.; Hu, X.; Liu, B.; He, K.; Luo, G.; Cao, X. Effectiveness of Newmark-Based Sampling Strategy for Coseismic Landslide Susceptibility Mapping Using Deep Learning, Support Vector Machine, and Logistic Regression. *Bull. Eng. Geol. Environ.* **2022**, *81*, 174. [CrossRef]
48. Hu, J.; Xu, K.; Wang, G.; Liu, Y.; Khan, M.A.; Mao, Y.; Zhang, M. A Novel Landslide Susceptibility Mapping Portrayed by OA-HD and K-Medoids Clustering Algorithms. *Bull. Eng. Geol. Environ.* **2021**, *80*, 765–779. [CrossRef]
49. Pokharel, B.; Althuwaynee, O.F.; Aydda, A.; Kim, S.-W.; Lim, S.; Park, H.-J. Spatial Clustering and Modelling for Landslide Susceptibility Mapping in the North of the Kathmandu Valley, Nepal. *Landslides* **2021**, *18*, 1403–1419. [CrossRef]
50. Kim, J.-C.; Lee, S.; Jung, H.-S.; Lee, S. Landslide Susceptibility Mapping Using Random Forest and Boosted Tree Models in Pyeong-Chang, Korea. *Geocarto Int.* **2018**, *33*, 1000–1015. [CrossRef]
51. Dou, J.; Yunus, A.P.; Tien Bui, D.; Merghadi, A.; Sahana, M.; Zhu, Z.; Chen, C.-W.; Khosravi, K.; Yang, Y.; Pham, B.T. Assessment of Advanced Random Forest and Decision Tree Algorithms for Modeling Rainfall-Induced Landslide Susceptibility in the Izu-Oshima Volcanic Island, Japan. *Sci. Total Environ.* **2019**, *662*, 332–346. [CrossRef]
52. Dou, J.; Yunus, A.P.; Bui, D.T.; Merghadi, A.; Sahana, M.; Zhu, Z.; Chen, C.-W.; Han, Z.; Pham, B.T. Improved Landslide Assessment Using Support Vector Machine with Bagging, Boosting, and Stacking Ensemble Machine Learning Framework in a Mountainous Watershed, Japan. *Landslides* **2020**, *17*, 641–658. [CrossRef]
53. Yao, X.; Tham, L.G.; Dai, F.C. Landslide Susceptibility Mapping Based on Support Vector Machine: A Case Study on Natural Slopes of Hong Kong, China. *Geomorphology* **2008**, *101*, 572–582. [CrossRef]
54. Hong, H.; Pradhan, B.; Jebur, M.N.; Bui, D.T.; Xu, C.; Akgun, A. Spatial Prediction of Landslide Hazard at the Luxi Area (China) Using Support Vector Machines. *Environ. Earth Sci.* **2015**, *75*, 40. [CrossRef]
55. Marjanović, M.; Kovačević, M.; Bajat, B.; Voženílek, V. Landslide Susceptibility Assessment Using SVM Machine Learning Algorithm. *Eng. Geol.* **2011**, *123*, 225–234. [CrossRef]
56. Wang, H.; Zhang, L.; Yin, K.; Luo, H.; Li, J. Landslide Identification Using Machine Learning. *Geosci. Front.* **2021**, *12*, 351–364. [CrossRef]
57. Beauchamp, K.A.; Bowman, G.R.; Lane, T.J.; Maibaum, L.; Haque, I.S.; Pande, V.S. MSMBuilder2: Modeling Conformational Dynamics on the Picosecond to Millisecond Scale. *J. Chem. Theory Comput.* **2011**, *7*, 3412–3419. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.