A Landslide Susceptibility Mapping Using CIBD-CURE Algorithm and LEPAM Methods for Baota District, China

Deborah S. Mwakapesa
Jiangxi University of Science and Technology
Lan Xiaoji
Jiangxi University of Science and Technology
I. K. Mensah
Jiangxi University of Science and Technology
Y. A. Nanehkaran
Wang Xiangtai
Jiangxi University of Science and Technology
Ye Li
Jiangxi University of Science and Technology
Chen Liang
Liu Wei
Yimin Mao (mymlyc@163.com)
Jiangxi University of Science and Technology

Research Article

Keywords: Landslide susceptibility mapping, clustering algorithm, CURE, uncertain data, CIBD-CURE, Baota District

Posted Date: April 19th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1508115/v1

Abstract

Landslide susceptibility mapping (LSM) is one of the crucial steps in managing and mitigating landslides. This study targets at developing a new LSM model for Baota District in China, using a new clustering method namely CIBD-CURE, which combines the city block distance (CIBD) and traditional clustering using representatives (CURE) algorithm. It aims at addressing limitations of inability to identify clusters (subclasses) with arbitrary shapes and varying sizes, sensitivity to noise, inability to perform well in large study areas with large dataset and inability to process rainfall (uncertain) data, which affect the results of traditional clustering algorithms in LSM. The CIBD was introduced into the CURE algorithm for processing uncertain data, then CIBD-CURE partitioned the mapping units into arbitrary shaped and sized subclasses with respect to their underlying geology and topography characteristics, and handled noise successfully. Furthermore, LEPAM method was proposed to sort the subclasses into five landslide susceptibility levels. Finally, standard statistical measures were applied to evaluate the model's performance and compared it against CURE, AHC-OLID, HC and KPSO clustering models, along with DTU and NBU classification models. The result analysis showed data the proposed model attained higher performance. This LSM study will be useful for landslides management strategies not only for this study area but also other affected areas around the world.

1. Introduction

Landslide is a brief and promptly happening vandalize natural disaster in mountainous areas caused by immature geology along with other sequential triggering factors which leads to changes of landscape and numerous damages (Cruden, 1991). Recently, the magnitude and frequency of landslide events have increased greatly and global attentiveness has been strained on the landslides studies generally because of the urbanization increasing pressure and its socio-economic effects on habitations (examples shown in Fig. 1) (Akgun, 2012). Thus, identification of landslide-susceptible areas using various strategies such as landslide susceptibility mapping (LSM) has become significantly important for better landslide reduction and management (Das et al., 2012).

Recently, LSM is considered as an integral part of landslide management in various regions, as it maps areas susceptible to landslides on the assumption that there is a probability of landslides to occur in areas with similar geo-environment conditions as well as landslide history (Yimin Mao et al., 2021; Roy et al., 2019). Several machine learning (ML) techniques supported by Geographic Information System (GIS) and Remote Sensing (RS) technologies are extensively applied by many researchers in various parts of the world to develop LSM models. Some of those ML methods are based on classification (supervised learning) and clustering (unsupervised learning) methods. Classification methods work with labeled dataset with multiple conditioning factors and tend to classify this dataset into relevant landslide and non-landslide classes. Such methods includes: classification and regression trees (Li & Chen, 2020), random forest (Hong et al., 2019), SVM, (Lee et al., 2017; Zhao & Zhao, 2021), Boosted Regression (Park & Kim, 2019), Quadratic Discriminant Analysis (Wang et al., 2020), naïve Bayesian classification (Mao et al., 2015), Artificial Neural Networks (Bragagnolo et al., 2020) and Decision tree (Arabameri et al., 2021; Mao et al., 2017). Despite of their wide applications, LSM models based on classification methods depend on large landslide labeled dataset to enhance their performance accuracy, and to acquire such dataset huge efforts are required in surveying sites, a process which becomes a challenge when dealing with large study areas. With this challenge, the application of these methods becomes limited.

Luckily, the clustering methods do not rely on landslide labeled dataset containing mapping units (points) with multiple conditioning factors. They tend to analyze the dataset, discover meaningful similarities and classify the mapping units into subclasses (clusters) based on their similarities. But there are very few LSM models based on clustering methods that have been developed so far, including: Agglomerative hierarchical clustering (Yimin Mao et al., 2021), hierarchical clustering (Pokharel et al., 2021), FCM (Wan, 2013), k-means (Wang et al., 2017), and KPSO (Wan et al., 2015). Unfortunately, most of these methods have limited performance due to their inability to detect subclasses with arbitrary shapes and varying sizes, sensitivity to noise, inability to perform well in large study areas with large datasets as well as lack of ability to successfully process the uncertain (rainfall) data.

Hence, from the above analysis, it can be noted that there is still a need to develop, implement and re-evaluate more of improved clustering methods to demonstrate their capacity to yield good susceptibility maps for their implementation in prone areas as they are easy to implement and their performance do not depend on the amount of the available dataset.

Therefore, in the present study, we developed a new clustering algorithm called CIBD-CURE which integrates the CIBD (City Block Distance (de Souza & De Carvalho, 2004)) and CURE (Clustering Using Representatives (Guha et al., 1998)) methods for LSM at Baota District, Shaanxi Province, China. The proposed methodology aims at addressing limitations of inability to identify clusters (subclasses) with arbitrary shapes and varying sizes, sensitivity to noise, inability to perform well in large study areas with large dataset and inability to

process rainfall (uncertain) data, which affect the results of clustering algorithms in LSM. Moreover, during landslide susceptibility mapping, the LEPAM method which includes landslide density and eigenvalues (LE) and Partitioning Around Medoids (PAM, (Rdusseeun & Kaufman, 1987) method is designed and applied to classify the study area into five susceptibility levels (very low, low, moderate, high and very high).

The remainder of the present study is organized as: section 2 includes a brief explanation on the study area; section 3 provides detail information on the research materials and methods used in this study; section 4 covers results and discussions of the study, while the study conclusions are in section 5.

2. Study Area

The study area is Baota district (Fig. 2), a 3,556 km^2 mountainous area and section of the Loess Plateau found in Yan'an city, of Shaanxi Province, China. Its geographic coverage is approximately 109°14'E-110°07' E longitudes and 36°11'N-37°02' N latitudes.

Topographically, there is scarce vegetation and Yanhe River in the northern side, as well as dense vegetation and Fenchuan River in the south. The geomorphology of the area is featured by gorges and curved slopes and the altitude is between 800–1,800 m above sea level. Geologically, the aforesaid area is compound due to existence of sedimentary rocks and the extensive quaternary loess deposits which dominates the area. The annual temperature and annual rainfall is 10°C and 550 mm respectively, and the heavy rainfall varying between 58 and 117 mm extending between June and October (Zhang and Liu 2010). It has also been observed that, rainfall triggers most of landslides in the area (Y.-m. Mao et al., 2021).

In general, we can learn that, the complex nature of the environment in this area, the on-going social and economic activities as well as the growing population, the area is frequently exposed to landslides causing countless losses. Upon dealing with this issue, various authorities apply various measures such as LSM to manage and mitigate landslides and its consequences. We believe that, this LSM study, will be helpful in different ways towards achieving that goal.

3. Research Materials And Methods

In this paper, the methodology is described as follows: (1) Research materials preparation, (2) Description of the proposed research method (3) Landslide susceptibility mapping (4) Models evaluation and comparison. The flow of the study is shown in Fig. 3, and the detailed descriptions are given in the following sections.

3.1 Research Materials

3.1.1 Landslide Database

The information concerning the prevailing landslide distribution as well as the preparation of database containing geospatial distribution information of various factors such as locations, nature, size and varieties of landslides is essential for the determination of likelihood of landslides as well as to conduct LSM. In this research study, the database of 293 landslides (indicated as black spots in the study area map (Fig.. 2) and details of some landslides is presented in Table 1) was acquired using field investigation reports and exploration of the historic data, from the Xi'an Center of Geological Survey. The recorded landslides were used for LSM modelling in this study.

3.1.2 Landslide Conditioning Factors

Landslide occurrences are associated with various conditioning factors. Based on previous research studies (Hu et al., 2019; Yiming Mao et al., 2021; Mao et al., 2017) and data exploration along with field investigation, we selected 7 landslide- conditioning factors (which are here regarded as attributes, described in Table 2) for modeling, which are elevation, slope angle, slope aspect, profile curvature, lithology, vegetation coverage index (NDVI), and rainfall. Elevation is associated with landslide incidences specially in plateau areas (Lee et al., 2018). Slope angle has significant effects on material sliding and flow of water under the influence of gravity thus affects the slope stability (Tran et al., 2021). Profile curvature has effect on the water movement on the surface of Earth resulting to landslide (Nohani et al., 2019). Lithology is the material basis of landslides and an essential in determining the type of rocks/soil exposed to landslides (Zhao & Zhao, 2021). NDVI is an essential ecological factor allied with the soil structure, which is commonly used in landslide susceptibility mapping studies (Yimin Mao et al., 2021; Zhang & Liu, 2010). Also, rainfall was selected since landslides in this area usually occur in the rainy season. Thematic maps (Fig. 4 (a-g) of the aforementioned factors were generated using 25m x 25m cell with the aid of ArcGIS platform.

Table 1. Details of some landslides

No.	Location			Length (m)	Width (m)	Thick-ness(m)	Volume ($\times 10^4 m^3$)
	Name	Coordinates					
		Longitude	Latitude				
1	Fengzhuang Nangou Tower	109°25'35"	36°47'46"	200	500	20	200
2	Urban Medical College	109°27'20"	36°33'56"	150	200	3	9
3	Yaoshop ZhaoJiagou	109°41'17"	36°35'54"	220	280	10	17
4	Yuyuan Houjiagou	109°25'03"	36°37'13"	150	35	8	42
5	Dragon Wohu Bay	109°38'34"	36°56'12"	150	200	10	30
6	Baijiaping	109°29'20"	36°34'24"	130	250	7	22.8
7	Liangcun Guojiashi	109°32'12"	36°52'04"	150	200	15	45
8	Wanhua Gaojiagou	109°25'07"	36°32'39"	250	200	14	266.4
..	...	...	...	...	...	...	...

Table 2
Attributes Description

Category	Attribute Name	Attribute type in CIBD-CURE	Classes of discrete attribute
Topography	Elevation	Continuous	None
	Slope Angle	Continuous	None
	Slope Aspect	Discrete	Flat, N, NE, E, SE, S, SW, W, NW
	Profile Curvature	Discrete	<−0.05, − 0.05 to 0.05, > 0.05
Geology	Lithology		1: loess + nearly horizontal paleo-soil,
		Discrete	2: loess + inclined paleo-soil,
			3: loess + paleo-soil layers + bedrock,
			4: loess + paleo-soil layers + the Neogene clay
Underlying surface	NDVI	Continuous	None
Triggering factor	Rainfall	Uncertain	0 60, 60 80, 80 100, 100 110, 110 120, 120mm above

3.2 Research Methods

3.2.1 CURE Algorithm

CURE is a clustering algorithm which performs classification tasks in large-scale datasets (Cai & Liang, 2018; Qian et al., 2002). It adopts works by partitioning the datasets into clusters by using some defined representative points and creating hierarchy amongst clusters in bottom-up approach (Guha et al., 1998; Yimin Mao et al., 2021). Meaning that, the algorithm begins by considering each point as a solo cluster and then iteratively merges two existing and ore similar clusters into one until we obtain desired clusters. This approach facilitates discovery of clusters with arbitrary shapes and varying sizes and make it robust to noise, advantages that cannot be found in most of the clustering algorithms. The algorithm is implemented in two stages as follows:

Stage 1: CURE initialization

1. Select a small random sample of data and cluster it in using bottom-up hierarchical approach

2. Pick a small set of well dispersed points from each cluster to be considered as representative points (RePts), which will later be used in distance minimum (Dmin) cluster merging approach
3. Shift every RePts to a fixed fraction (about 20% or 30%) of the original distance between its current position and its cluster centroid.
Stage 2: CURE completion

1. After completion of the initialization stage, cluster the rest of the points and find the final cluster.
2. Merge two clusters whose pair of RePts are close enough, using Dmin cluster merging approach.
3. After every such merging, select new RePts to represent the new cluster.
4. Repeat the merging step until there are no sufficiently close clusters left.

Note

Points in this algorithm are referred to as mapping units (each containing the conditioning factors) in LSM modelling. Also, in step 2, the minimum distance from the dispersed points inside and outside the selected sample is computed using Euclidean distance, whereby, the point with minimum distance to the dispersed point inside the sample is taken and merged into the sample

Figure 5 Clustering of points in the dataset by CURE algorithm (a) A random sample of data (b) 3 clusters with representative points (red points) (c) Merge clusters with closest representative points (d) Shrink the representative points

3.2.2 Uncertain Data and CIBD-CURE algorithm

Uncertain data is the type of data which is in some range, and its specific value is not well-known (Ren et al., 2009). For example in real life scenarios, person's blood pressure may be recorded by (120, 129) values, daily rainfall (14, 28mm). In practice, this data is presented with its lower and upper bounds such as $u_a^b = (i_a^b, j_a^b)$where $i_a^b$is the lower bound while j_a^b denotes the upper bound. Alternatively, the data can be denoted using its midpoint (m) and half its length or radius (r) as, $u_a^b = (m_a^b, r_a^b)$ where $m_a^b = (i + j)/2$and $r_a^b = (j - i)/2$.

In landslide susceptibility modelling, factors (example slope angle, profile curvature and rainfall) with various data types such as continuous, discrete and uncertain types are considered. Most of the existing landslide clustering methods do not take the uncertain data such as rainfall into consideration and tend to process it like a discrete type, while for areas where landslides are mostly induced by rainfall, it is important to note such consideration as it has great impact on the modelling results and since safety is concerned.

CURE algorithm depend on the Euclidean distance function to compute the minimum distance among the points while forming clusters. However, this function can work well with continuous and discrete data types but fails to process the uncertain data. Hence, limits the application of CURE algorithm especially for landslide susceptibility modelling. To improve the performance of CURE algorithm, we propose the use of city-block distance (CIBD) which can successfully handle the uncertain data.

Given a dataset containing points, and$u_a^b = (i_a^b, j_a^b)$ and$v_a^b = (\alpha_a^b, \sigma_d^b)$ be two random uncertain points in the datasets. The CIBD $d_c(u_a^b, v_d^b)$ between these points is expressed in Eq. (1).

$$d_c(u_a, v_d) = \sum_{b=1}^{N} \lambda^b [|i_a^b - \alpha_d^b| + |j_a^b - \sigma_d^b|]$$

1

Whereas $\lambda^b$is a weight vector forpoints.

Also, since the uncertain data can be represented using radius and midpoints, the CIBD $d_c(u_a^b, v_d^b)$can be written as follows:

$$d_c(u_a, v_d) = \sum_{b=1}^{N} \lambda^b [|m(u_a^b) - m(v_d^b)| + |r(u_a^b) - r(v_d^b)|]$$

2

Using this CIBD function in Eq. (2) instead of the Euclidean distance step 2 to calculate the minimum distance among the points leads to the development of a new algorithm titled CIBD-CURE algorithm which is an improvement of the traditional CURE algorithm for modeling landslide susceptibility.

3.2.3 LEPAM Method for Landslide Susceptibility Classification

This methods includes the landslide density and eigenvalues (LE) strategy and Partitioning around medoids (PAM) clustering algorithm for landslide susceptibility classification.

LE Strategy

To specify landslide susceptibility levels in the area, we apply a strategy based on landslide density and eigenvalues. Landslide density (LD, (Mao et al., 2017)) is computed using the number of landslide (N) per square kilometer (km^2) of a mapping unit in a subclass, and is applied to indicate the susceptibility level of that subclass. When N = 0 in a subclass, means the LD is also equal to zero, then, the eigenvalues (which describe more the characteristics of the area) based on geology expertise will be applied to specify the susceptibility level.

PAM Clustering Algorithm

PAM is a partitioning clustering algorithm which partition data to some clusters based on selected points called medoids which represent the number of clusters to be obtained. In this present study, PAM is applied to partition the subclasses obtained from the CIBD-CURE algorithm, into five landslide susceptibility levels (hence,$m = 5$) based on LE strategy described above. The algorithm follows the steps below:

1. Fix the value of to 5, to represent the susceptibility levels to be obtained
2. From the input data randomly choose 5 subclasses as medoids for each susceptibility levels
3. Each subclass gets assigned to the susceptibility level to which its nearest medoid belongs.
4. For each subclass of susceptibility level, its distance from all other subclasses is computed and added. The subclass of i^{th} susceptibility level for which the computed sum of distances from other subclasses is minimal is assigned as the medoid for that susceptibility level.
5. Steps (3) and (4) are repeated until the medoids stop changing.

3.2.4 Model Evaluation and Comparison

Evaluation Metrics

Developed LSM models need to be validated to check their prediction capability (Pham et al., 2020) and up to date there are no universal metrics to perform this task. In this study, we apply some statistical metrics namely accuracy (Ac, for the correctly predicted landslide and non-landslide samples), sensitivity (St), specificity (Sp), kappa (ka), and *AUC* (area under the receiver operating curve, plotted using St (y-axis) against$1 - Sp$ (x-axis). These metrics are computed on account of four prediction indices: true positive (tp), true negative (tn), false positive (fp) and false negative (fn) (Dou et al., 2020; Pham et al., 2020). $tp$and$fp$are the landslide samples that have been correctly predicted as landslide and non-landslide samples respectively, while, $tn$and $fn$are the landslide samples that have been incorrectly predicted into landslide and non-landslide classes respectively. The metrics are expressed in equations below:

$$Ac = \frac{tp + tn}{tp + tn + fp + fn}$$

3

$$St = \frac{tp}{tp + fn}$$

4

$$Sp = \frac{tn}{tn + fp}$$

5

$$ka = \frac{P_a - P_{\exp}}{1 - P_{\exp}}$$

6

Whereby, $P_a = (tp + tn)/(tp + tn + fp + fn)$and

$$P_{\exp} = (((tp + fn)(tp + fp) + (tn + fp)(tn + fn))/(\sqrt{(tp + tn + fp + fn)})$$

Ac and kaclose to 1 indicates that the model is reliable, while close to 0 means the model is not reliable (Landis & Koch, 1977). Also, when *AUC* is almost 1 implies that the model is perfect while when AUC = 0.5 means the model is inaccurate (Youssef et al., 2016)

Comparison Methods

To assess the CIBD-CURE algorithm performance for landslide susceptibility modelling, we compared its results with four other clustering algorithms CURE, AHC-OLID, HC and KPSO as benchmark clustering methods. The comparison based on the evaluation metrics mentioned above. The objective of comparison is to evaluate and show the possible differences between the proposed method and compared methods in LSM. Furthermore, using the same dataset, DTU (uncertain decision tree (Mao et al., 2017)) and NBU (uncertain naïve Bayesian (Mao et al., 2015)) LSM classification methods were applied for comparison with the proposed clustering method on the basis of their performance accuracy.

4. Results

4.1 Clustering Results

Based on the procedures mentioned in Section 3.2.1 and 3.2.1, the proposed CIBD-CURE algorithm divided the mapping units into various dispersed subclasses. The attributes of every mapping units were first standardized (normalized) and then used as inputs the CIBD-CURE algorithm. The mapping units in the study area were clustered into 483 distinguished subclasses with varying shapes and sizes.

4.2 Landslide Susceptibility Mapping

Although the study area has been partitioned to subclasses, but the model cannot yet indicate which subclasses are susceptible to landslides and to what extent. So, for that purpose, in this study, using the gotten subclasses, we developed LEPAM method for classifying and constructing the landslide susceptibility map in ArcGIS platform. Firstly, landslide density for every subclass was calculated. Then PAM clustering method was used to divide those subclasses into five susceptibility levels (very low, low, moderate, high and very high). To assign the subclasses to their respective levels, landslide density and eigenvalues were applied using the principle that high landslide density implies high level of susceptibility and low landslide density implies low susceptibility level. Meanwhile, for landslide density equal to zero, eigenvalues and experts geology knowledge were applied to determine the susceptibility level. Table 3 presents some subclasses alongside with their eigenvalues, landslide densities as well as susceptibility levels.

Table 3
Eigenvalues and landslide density of subclasses

Sub-class No.	Eigenvalues							Landslide Density			Susceptibility Level
	Elevation	Slope Angle	Profile Curvature	Slope Aspect	Lithology	NDVI	Rainfall	Area (km^2)	Landslides	LD ($/km^2$)	
1	30.21	26.89	0.028	S	II	0.66	24–233	9.54	1	0.1	Low
2	25.35	20.19	0.033	SE	I	0.54	20–192	6.53	5	0.77	High
…	…	…	…	…	…	…	…	…	…	…	…
235	21.97	41.23	0.59	NE	III	0.59	28–187	14.25	0	0	Determined by expert
…	…	…	…	…	…	…	…	…	…	…	…
410	19.89	33.19	0.47	N	II	0.49	33–267	25.06	16	0.64	Moderate
…	…	…	…	…	…	…	…	…	…	…	…

Table 4
CIBD-CURE analysis of landslide susceptibility classification

Landslide density	% of Subclasses	Landslide susceptibility levels
0.90–1.70	17	Very High
0.70–0.90	19	High
0.14–0.70	34	Moderate
0.04–0.14	16	Low
0–0.04	14	Very Low

Table 4 presents the CIBD-CURE analysis of landslide susceptibility classification. From the table, it is clear that the most subclasses (34%) are susceptible to landslides at moderate level, followed by high level (19%) and very high level (17%). Few subclasses (14%) fell into the very low susceptibility level. The constructed susceptibility map is shown in Fig. 6.

4.3 Validation and Comparison

4.3.1 Comparison among the clustering models

The evaluation results of proposed model and comparison results are presented in Table 5. From the table, it can be observed that performance of the CIBD-CURE algorithm is the best (St= 0.8874, Sp = 0.8920, Ac = 0.8893, and ka = 0.7744) as compared to the other models: CURE (St= 0.8634, Sp = 0.8685, Ac = 0.8676, and ka = 0.7307), AHC-OLID (St= 0.8532, Sp = 0.8451, Ac = 0.8636, and ka = 0.7219), HC (St= 0.802, Sp = 0.7606, Ac = 0.8353, and ka = 0.7737) and KPSO (St= 0.6724, Sp = 0.6479, Ac = 0.6621, and ka = 0.3161). These results showed strong clustering capability of CIBD-CURE compared to the other clustering models. In addition, the proposed CIBD-CURE algorithm showed the highest AUC of 0.857 in the ROC shown in Fig. 7.

Table 5
Evaluation and comparison results

Models	*tp*	*tn*	*fp*	*fn*	*St*	*Sp*	*Ac*	*Ka*
CIBD-CURE	260	190	23	33	0.8874	0.892	0.8893	0.7744
CURE	253	186	27	40	0.8634	0.8685	0.8676	0.7307
AHC-OLID	250	180	33	43	0.8532	0.8451	0.8636	0.7219
HC	235	162	51	58	0.802	0.7606	0.8353	0.6637
KPSO	197	138	75	96	0.6724	0.6479	0.6621	0.3161

4.3.2 Comparison among clustering and classification methods

To examine the performance accuracy of CIBD-CURE clustering model, we also compared it with DTU and NBU classification models. To construct and evaluate the classification models, the dataset was divided into 30% for training and 70% for validation, whereby, the process was carried out iteratively by adding 10% of the data from the validation set to the training set until there was 70% of the data in the training set. From the comparison results (Fig. 8), the CIBD-CURE model showed nearly constant performance accuracy during the experiment while the accuracies of the DTU and NBU model were low at the beginning and kept on increasing as additional data was inputted to the training set.

5. Discussion

Nowadays, machine learning methods such as clustering approaches are utilized for construction of landslide susceptibility maps in various regions. In this study, a new clustering method named CIBD-CURE was developed for LSM and construction of landslide susceptibility map for Baota District, which is amongst the landslide-vulnerable areas in China. Seven landslide conditioning factors were employed. The CIBD-CURE model was able to divide the study area into subclasses of varying shapes and sizes, whereby, each subclass had distinguished geology and topography characteristics. These results indicate that the model has good and effective clustering capability. The obtained results were further applied LEPAM in classifying and constructing the susceptibility map, whereby, it was revealed that about 16% and 21%% of the study area were observed in the high and very high susceptibility levels.

For comparison with other clustering models, the proposed model was compared with the traditional CURE, AHC-OLID, HC and KPSO clustering algorithms using standard statistical metrics: accuracy, sensitivity, specificity, kappa and AUC. The result analysis indicated that, the CIBD-CURE model was able to obtain better performance than the compared models in terms of statistics metrics. This is because of the enhanced features found in the proposed algorithm, including: improved capability to process the uncertain (rainfall) data using CIBD function, ability to detect subclasses of varying shapes and sizes from the dataset, it is robust to noise and its ability to work well with large dataset. Moreover, the proposed model achieved higher AUC than the other models, implying that it is very accurate and effective for assessing landslide susceptibility in this area.

Furthermore, the proposed clustering model was also compared with the DTU and NBU classification models based on their performance accuracies. From the comparison results, as a clustering method, the performance of CIBD-CURE did not require training data and showed nearly constant accuracy throughout the process indicating guaranteed performance even when the available dataset is large. On the contrary, the classification models depend on the training data and their accuracies tends to vary with respect to the variation of the training data, that is, with small training dataset the accuracy was low and increased when additional data was supplied to the training set. The tendency of the classification models to rely upon the amount of the available training data is an indication that they cannot guarantee accurate and reliable LSM especially in large study areas where obtaining enough dataset is a challenge in various aspects such as amount of human labor and time required to acquire such data.

6. Conclusion

In this research paper, we proposed the CIBD-CURE model, which integrated the CURE algorithm and CIBD for LSM using dataset from the Baota District, China. The main objective of this work was to develop an improved LSM clustering model which addresses the limitations of inability to identify clusters (subclasses) with arbitrary shapes and varying sizes, sensitivity to noise, inability to perform well in large study areas with large dataset and inability to process rainfall (uncertain) data, that affect the results of traditional clustering algorithms in LSM.

The data analysis showed that, CURE algorithm failed to process the rainfall data thus obtained low performance accuracy, while with CIBD the CIBD-CURE could successfully overcome that limitation and obtain an improved performance accuracy.

In the case of implementation, the CIBD-CURE was able to cluster the mapping units into 483 distinguished subclasses with varying shapes and sizes and obtained improved performance accuracy due to its the ability to process well the rainfall data, handle noise as well as work well with the large dataset from the study areas. Also, the LEPAM method was developed to classify the subclasses into five susceptibility levels.

Furthermore, in terms of validation and comparison, the advantageous features enabled the proposed model to obtain the best results of St = 0.8874, Sp = 0.8920, Ac = 0.8893, ka = 0.7744 and AUC = 0.857, superior to CURE, AHC-OLID, HC and KPSO models, and showed more reliable and guaranteed performance accuracy than the classification models.

With these results, CIBD-CURE model can be regarded as a significant model for LSM in the Baota District and other prone areas, though its application to other areas will depend on the local conditions of that area that may alter the landslide conditioning factors.

Declarations

Funding Statement: This study was financially supported by the National Natural Science Foundation of China (Grant: 41562019) and the Foundation of Science and Technology in Education Department of Jiangxi (Grant: GJJ209406, GJJ209407).

Author Contribution

Yimin Mao, Lan Xiaoji, Deborah Simon Mwakapesa and Y.A. Nanehkaran contributed to the study conception and design. The first draft of the manuscript was written by Deborah Simon Mwakapesa and I.K. Mensah. Material preparation, data collection and analysis were performed by Yimin Mao, Deborah Simon Mwakapesa, Ye Li and Wang Xiangtai. Yimin Mao, Lan Xiaoji, I.K. Mensah and Y.A. Nanehkaran worked on improving the manuscript. I.K. Mensah, Liu Wei and Chen Liang worked on formatting the manuscript. All authors read and approved the final manuscript.

References

1. Akgun A (2012) A comparison of landslide susceptibility maps produced by logistic regression, multi-criteria decision, and likelihood ratio methods: a case study at İzmir. Turk Landslides 9(1):93–106
2. Arabameri A, Pal C, Rezaie S, Chakrabortty F, Saha R, Blaschke A, Thi T, Ngo PT (2021) Decision tree based ensemble machine learning approaches for landslide susceptibility mapping.Geocarto International,1–35
3. Bragagnolo L, da Silva R, Grzybowski J (2020) Artificial neural network ensembles applied to the mapping of landslide susceptibility. *Catena*, *184*, 104240
4. Cai M, Liang Y (2018) An improved CURE algorithm. International conference on intelligence science
5. Cruden DM (1991) A simple definition of a landslide. Bull Int Association Eng Geology-Bulletin de l'Association Int de Géologie de l'Ingénieur 43(1):27–29
6. Das I, Stein A, Kerle N, Dadhwal VK (2012) Landslide susceptibility mapping along road corridors in the Indian Himalayas using Bayesian logistic regression models. Geomorphology 179:116–125
7. de Souza RM, De Carvalho FdA (2004) Clustering of interval data based on city–block distances. Pattern Recognit Lett 25(3):353–365
8. Dou J, Yunus AP, Merghadi A, Shirzadi A, Nguyen H, Hussain Y, Yamagishi H (2020) Different sampling strategies for predicting landslide susceptibilities are deemed less consequential with deep learning. *Science of the total environment*, *720*, 137320
9. Guha S, Rastogi R, Shim K (1998) CURE: An efficient clustering algorithm for large databases. ACM Sigmod record 27(2):73–84
10. Hong H, Miao Y, Liu J, Zhu A-X (2019) Exploring the effects of the design and quantity of absence data on the performance of random forest-based landslide susceptibility mapping. CATENA 176:45–64
11. Hu J, Zhu H, Mao Y, Zhang C, Liang T, Mao D (2019) Using Uncertain DM-Chameleon Clustering Algorithm Based on Machine Learning to Predict Landslide Hazards. J Robot Mechatron 31(2):329–338
12. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. *biometrics*, 159–174
13. Lee J-H, Sameen MI, Pradhan B, Park H-J (2018) Modeling landslide susceptibility in data-scarce environments using optimized data mining and statistical methods. Geomorphology 303:284–298
14. Lee S, Hong S-M, Jung H-S (2017) A support vector machine for landslide susceptibility mapping in Gangwon Province. Korea Sustain 9(1):48
15. Li Y, Chen W (2020) Landslide susceptibility evaluation using hybrid integration of evidential belief function and machine learning techniques. Water 12(1):113
16. Mao Y-m, Mwakapesa DS, Li Y-c, Xu K-b, Nanehkaran YA, Zhang M-s (2021) Assessment of landslide susceptibility using DBSCAN-ADH and LD-EV methods.Journal of Mountain Science,1–14
17. Mao Y-m, Zhang M-s, Wang G-l, Sun P-p (2015) Landslide hazards mapping using uncertain Naïve Bayesian classification method. J Cent South Univ 22(9):3512–3520
18. Mao Y, Mwakapesa DS, Wang G, Nanehkaran Y, Zhang M (2021) Landslide susceptibility modelling based on AHC-OLID clustering algorithm. Adv Space Res 68(1):301–316
19. Mao Y, Mwakapesa DS, Xu K, Lei C, Liu Y, Zhang M (2021) Comparison of wave-cluster and DBSCAN algorithms for landslide susceptibility assessment. Environ Earth Sci 80(22):1–14
20. Mao Y, Zhang M, Sun P, Wang G (2017) Landslide susceptibility assessment using uncertain decision tree model in loess areas. Environ Earth Sci 76(22):1–15
21. Nohani E, Moharrami M, Sharafi S, Khosravi K, Pradhan B, Pham BT, Melesse M, A (2019) Landslide susceptibility mapping using different GIS-based bivariate models. Water 11(7):1402
22. Park S, Kim J (2019) Landslide susceptibility mapping based on random forest and boosted regression tree models, and a comparison of their performance. Appl Sci 9(5):942
23. Pham BT, Prakash I, Dou J, Singh SK, Trinh PT, Tran HT, Shirzadi A (2020) A novel hybrid approach of landslide susceptibility modelling using rotation forest ensemble and different base classifiers. Geocarto Int 35(12):1267–1292
24. Pokharel B, Althuwaynee OF, Aydda A, Kim S-W, Lim S, Park H-J (2021) Spatial clustering and modelling for landslide susceptibility mapping in the north of the Kathmandu Valley, Nepal. Landslides 18(4):1403–1419
25. Qian Y-T, Shi Q-S, Wang Q (2002) CURE-NS: A hierarchical clustering algorithm with new shrinking scheme. Proceedings. International Conference on Machine Learning and Cybernetics

26. Rdusseeun L, Kaufman P (1987) Clustering by means of medoids. Proceedings of the Statistical Data Analysis Based on the L1 Norm Conference, Neuchatel, Switzerland
27. Ren Y, Liu Y-H, Rong J, Dew R (2009) Clustering interval-valued data using an overlapped interval divergence. Proceedings of the Eighth Australasian Data Mining Conference-Volume 101
28. Roy J, Saha S, Arabameri A, Blaschke T, Bui DT (2019) A novel ensemble approach for landslide susceptibility mapping (LSM) in Darjeeling and Kalimpong districts, West Bengal, India. Remote Sens 11(23):2866
29. Tran T-H, Dam ND, Jalal FE, Al-Ansari N, Ho LS, Phong TV, Prakash I (2021) GIS-Based Soft Computing Models for Landslide Susceptibility Mapping: A Case Study of Pithoragarh District, Uttarakhand State, India. *Mathematical problems in Engineering*, *2021*
30. Wan S (2013) Entropy-based particle swarm optimization with clustering analysis on landslide susceptibility mapping. Environ Earth Sci 68(5):1349–1366
31. Wan S, Yen JY, Lin CY, Chou TY (2015) Construction of knowledge-based spatial decision support system for landslide mapping using fuzzy clustering and KPSO analysis. Arab J Geosci 8(2):1041–1055
32. Wang G, Chen X, Chen W (2020) Spatial prediction of landslide susceptibility based on GIS and discriminant functions. ISPRS Int J Geo-Information 9(3):144
33. Wang Q, Wang Y, Niu R, Peng L (2017) Integration of information theory, K-means cluster analysis and the logistic regression model for landslide susceptibility mapping in the Three Gorges Area, China. Remote Sens 9(9):938
34. Youssef AM, Pourghasemi HR, Pourtaghi ZS, Al-Katheeri MM (2016) Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at Wadi Tayyah Basin, Asir Region, Saudi Arabia. Landslides 13(5):839–856
35. Zhang M-s, Liu J (2010) Controlling factors of loess landslides in western China. Environ Earth Sci 59(8):1671–1680
36. Zhao S, Zhao Z (2021) A comparative study of landslide susceptibility mapping using SVM and PSO-SVM models based on Grid and Slope Units. *Mathematical problems in Engineering*, *2021*

Figures

Figure 1

Examples of social-economic and environmental impacts of landslides

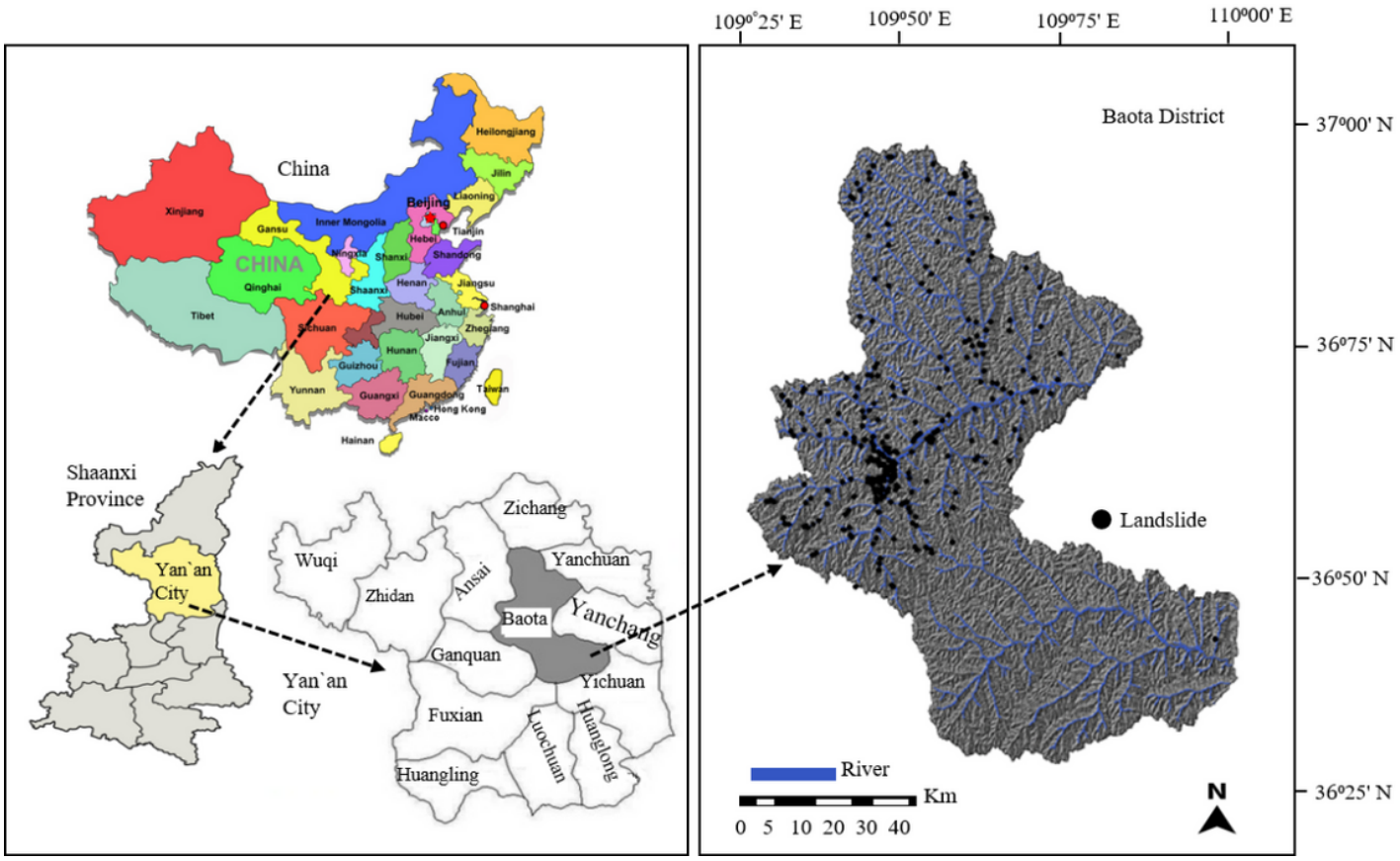

Figure 2

The Study area: Baota District, Shaanxi Province, China

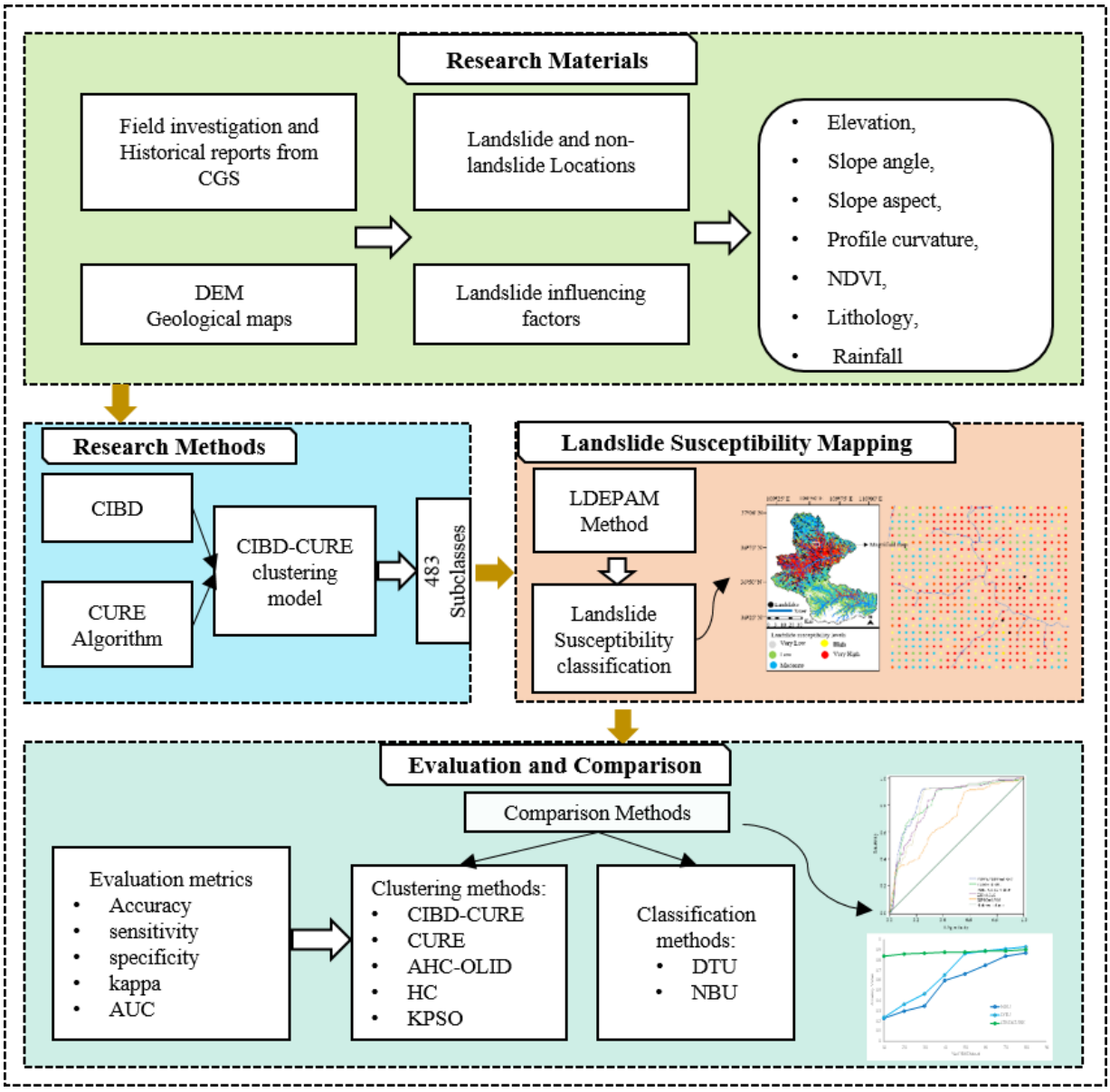

Figure 3

Flow chart of the study

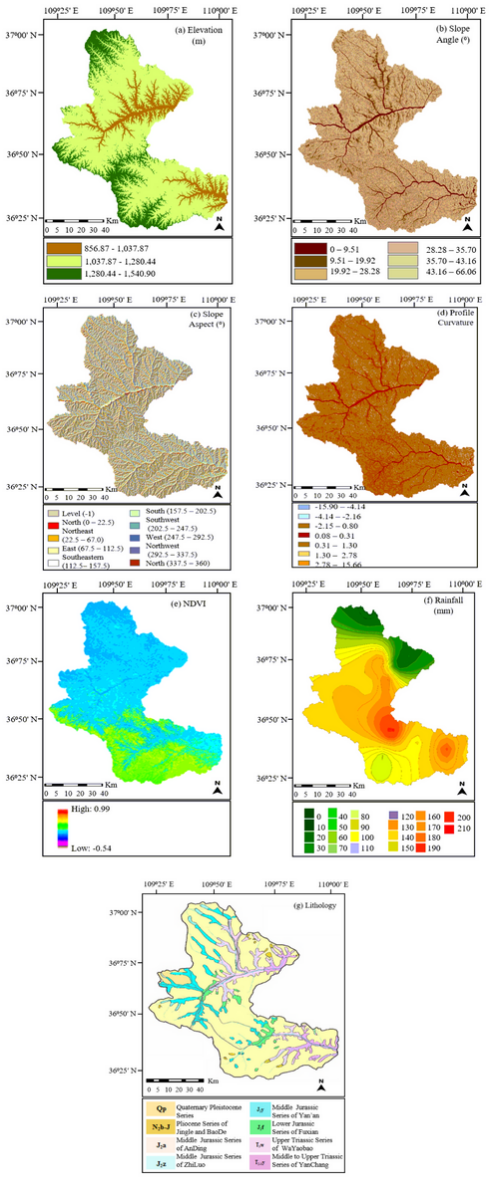

Figure 4

The thematic maps of the landslide conditioning factors

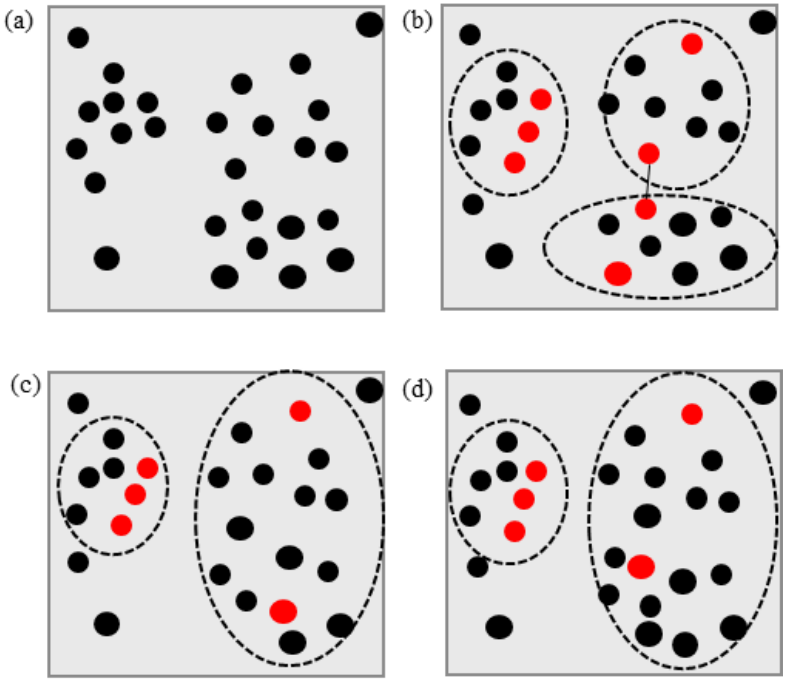

Figure 5

Clustering of points in the dataset by CURE algorithm (a) A random sample of data (b) 3 clusters with representative points (red points) (c) Merge clusters with closest representative points (d) Shrink the representative points

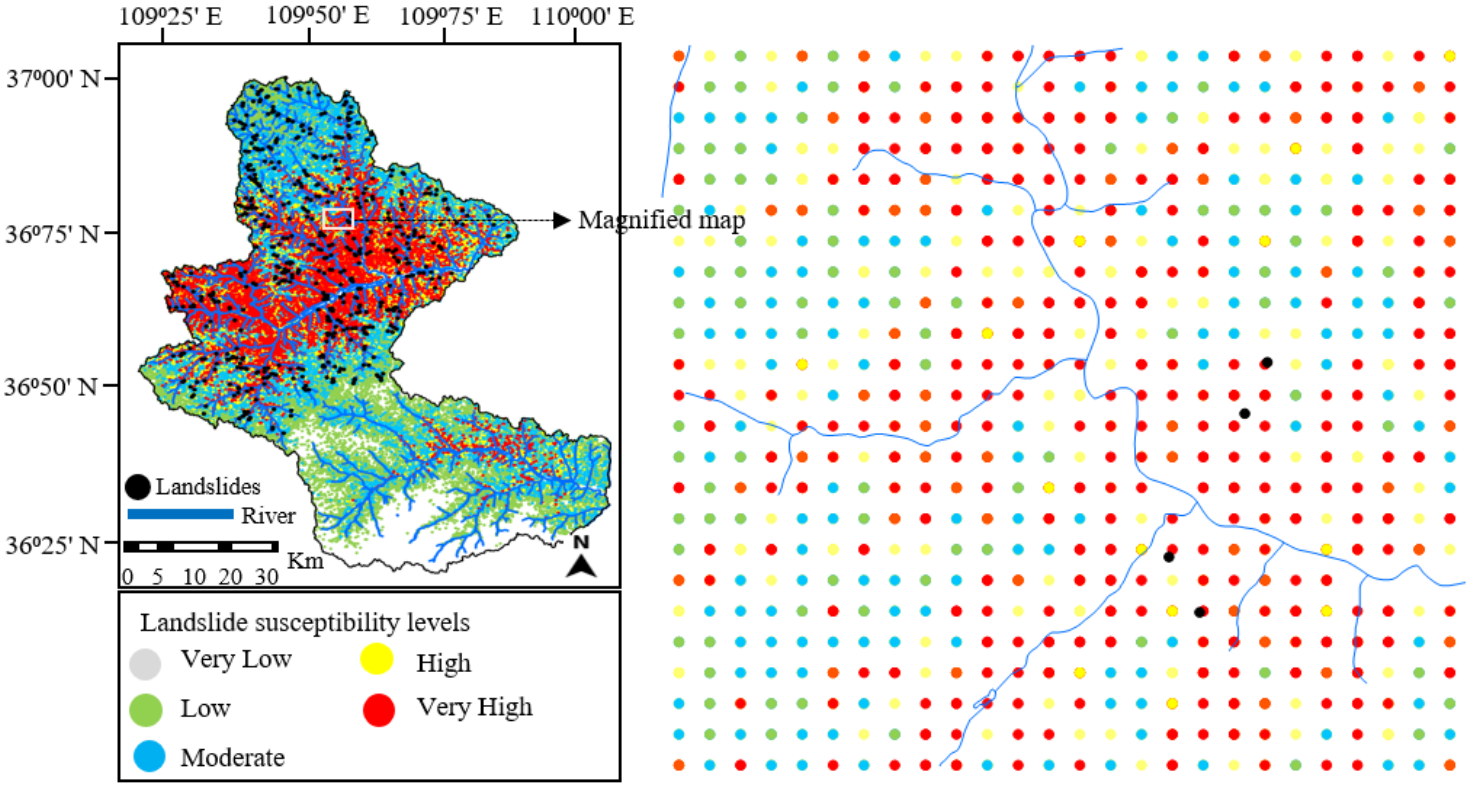

Figure 6

Landslide susceptibility map based on CIBD-CURE method

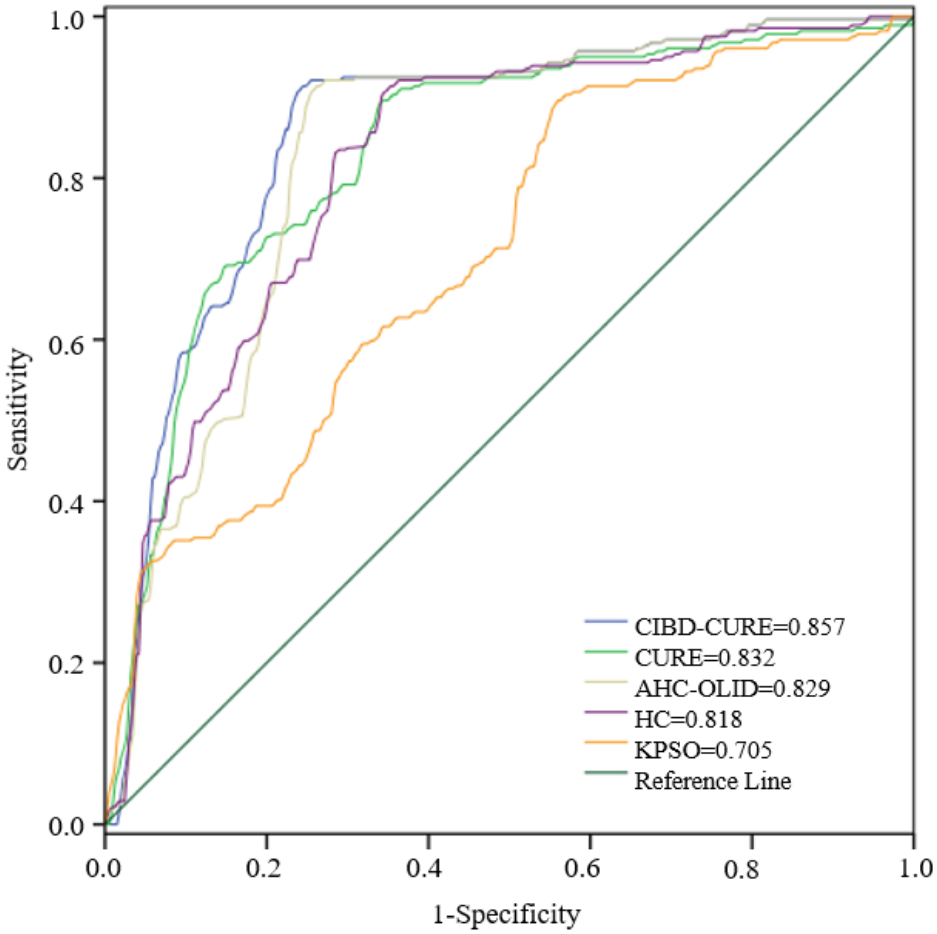

Figure 7

ROC curves for CIBD-CURE, CURE, AHC-OLID, HC and KPSO LSM clustering models

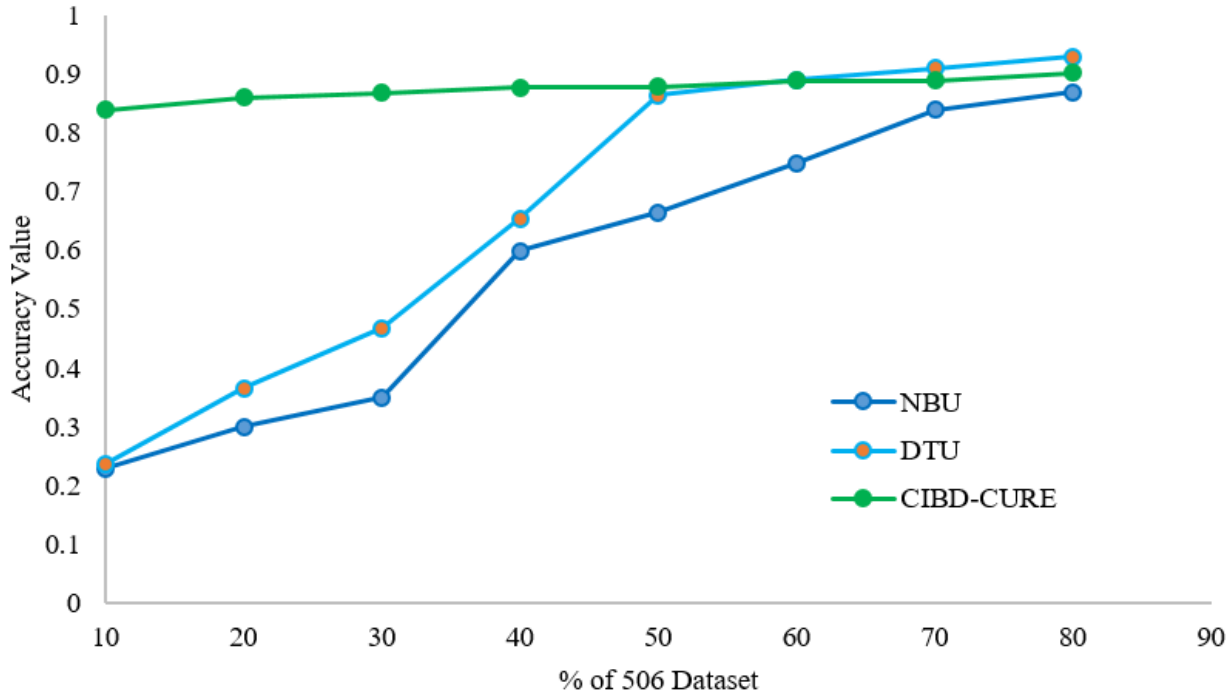

Figure 8

Performance evaluation of CIBD-CURE, DTU and NBU models