Systematic sample subdividing strategy for training landslide susceptibility models

Maher Ibrahim Sameen[1], Biswajeet Pradhan[1,2*], Dieu Tien Bui[3], Abdullah M. Alamri[4]

[1] The Centre for Advanced Modelling and Geospatial Information Systems (CAMGIS), School of Information, Systems & Modelling, Faculty of Engineering and Information Technology, University of Technology Sydney, NSW 2007, Australia

[2] Department of Energy and Mineral Resources Engineering, Sejong University, Choongmu-gwan, 209 Neungdong-ro, Gwangjin-gu, Seoul 05006, Korea

[3] Geographic Information System Group, Department of Business and IT, University of South-Eastern Norway, Gullbringvegen 36, N-3800 Bø i Telemark, Norway

[4]Dept. of Geology & Geophysics, College of Science, King Saud Univ., P.O. Box 2455, Riyadh 11451, Saudi Arabia

*Email. Biswajeet24@gmail.com or Biswajeet.Pradhan@uts.edu.au (Corresponding author)

Abstract

Current practice in choosing training samples for landslide susceptibility modelling (LSM) is to randomly subdivide inventory information into training and testing samples. Where inventory data differ in distribution, the selection of training samples by a random process may cause inefficient training of machine learning (ML)/statistical models. A systematic technique may, however, produce efficient training samples that well represent the entire inventory data. This is particularly true when inventory information is scarce. This research proposed a systemic strategy to deal with this problem based on the fundamental distribution of probabilities (i.e. Hellinger) and a novel graphical representation of information contained in inventory data (i.e. inventory information curve, IIC). This graphical representation illustrates the relative increase in available information with the growth of the training sample size. Experiments on a selected dataset over the Cameron Highlands, Malaysia were conducted to validate the proposed methods. The dataset contained 104 landslide inventories and 7 landslide-conditioning factors (i.e. altitude, slope, aspect, land use, distance from the stream, distance from the road and distance from lineament) derived from a LiDAR-based digital elevation model and thematic maps acquired from government authorities. In addition, three ML/statistical models, namely, k-nearest neighbour (KNN), support vector

machine (SVM) and decision tree (DT), were utilised to assess the proposed sampling strategy for LSM. The impacts of model's hyperparameters, noise and outliers on the performance of the models and the shape of IICs were also investigated and discussed. To evaluate the proposed method further, it was compared with other standard methods such as random sampling (RS), stratified RS (SRS) and cross-validation (CV). The evaluations were based on the area under the receiving characteristic curves. The results show that IICs are useful in explaining the information content in the training subset and their differences from the original inventory datasets. The quantitative evaluation with KNN, SVM and DT shows that the proposed method outperforms the RS and SRS in all the models and the CV method in KNN and DT models. The proposed sampling strategy enables new applications in landslide modelling, such as measuring inventory data content and complexity and selecting effective training samples to improve the predictive capability of landslide susceptibility models.

Keywords: Landslide modelling; Sampling strategy; GIS; Hellinger distance; Outlier detection

1. Introduction

Landslide is a destructive natural geohazard that threatens human lives and affects infrastructures and the economy (Schlogel et al., 2015; Korup et al., 2012; Pradhan and Sameen, 2017). Although landslides cannot be fully avoided, building mitigation strategies and tools to reduce their impacts have been a useful research and industrial objective. As a result, several researchers and industrial authorities have been developing methods to predict landslides before they occur for improved planning and risk mitigations. According to international scientific indexing, landslide susceptibility modelling (LSM) (also known as spatial prediction) is a common procedure for landslide occurrence prediction. Guzzetti et al. (2006) defined landslide susceptibility as the propensity of an area to generate landslides. In this method, historical landslide events are collected and analysed thoroughly to prepare inventory datasets. The mapping units (e.g. grid cells and slope units) and scales are defined at this stage (Huabin et al., 2005; Rotigliano et al., 2012). Similarly, geospatial information including those collected by remote sensing and field surveys is used to obtain conditioning factors that can be used as predictors (also known as landslide predisposing factors, slope instability factors and general features) (Hussin et al., 2016). Thereafter, models are built using these inventory datasets and conditioning factors via machine learning (ML), statistical

or expert-based techniques (Pradhan et al., 2017). The created models are then used to simulate future scenarios.

In practice, two main factors decide the accuracy of the LSM. First, the quality and quantity of landslide inventory records in a dataset. This factor includes the process of selecting training/testing samples. Second, the selection process of the hyperparameters, geometric properties and other configurations of the learning algorithm affects its performance. Most existing works have focused on the model development process, which involves adjusting hyperparameters or obtaining hybrid models created by combining several weak learners (Althuwaynee et al., 2014; Dehnavi et al., 2015; Aghdam et al., 2016; Althuwaynee et al., 2016; Hong et al., 2018; Pradhan and Sameen, 2018; Fanos and Pradhan, 2019). Other researchers have investigated different approaches to improving the performance of LSM (Nefeslioglu et al., 2008; Yilmaz, 2010; Hussin et al., 2016; Hong et al., 2018).

Despite the efforts of many researchers to improve the existing LSM, the existing models often make unclear conclusions for the practitioners. The same model performs differently, and researchers arrive at inconsistent conclusions whilst analysing the same models on different datasets. In practice, the identification of a competitive algorithm for landslide modelling is a complex task mainly due to a large number of models. With all the provided information about the existing models, users have difficulty determining in advance whether a selected model will work appropriately. In landslide modelling, efforts have been made to identify algorithms that consistently perform well; however, sometimes less sophisticated methods perform better in some cases due to a different dataset or other hidden factors. It is then important to determine whether these less sophisticated methods can be optimised or training datasets should be transformed into other forms to obtain different scenarios.

To address this issue, this study develops a systematic strategy for subdividing inventory samples into training and testing datasets. It is based on minimising the Hellinger distance calculated between the distribution of a training subset and the distribution of the entire inventory set. A new graphical representation of landslide inventory datasets is also proposed to improve the understanding of the landslide inventory nature. This process improves the learning and generalisation of the selected models. This study overcomes the previous shortcomings as it

provides tools to investigate and select effective training samples for training ML and statistical models for LSM development.

This study uses three standard ML and statistical different algorithms, namely, decision tree (DT), support vector machine (SVM) and the *k*-nearest neighbour (KNN) classifier, to analyse and test the proposed methods. Before presenting the details of the current work, the next section discusses the related literature.

2. Related literature

Numerous algorithms have been utilised to prepare LSM. They are often categorised into expert-based (Sezer et al., 2017; Ercanoglu et al., 2008; Althuwaynee et al., 2014; Hasekioğulları and Ercanoglu, 2012; Zhu et al., 2014; Pourghasemi et al., 2012; Ahmed, 2015), statistical (i.e. bivariate and multivariate) (Demir et al., 2015; Youssef et al. 2015; Mohammady et al., 2012; Pradhan et al., 2010; Cui et al., 2016; Zare et al., 2013; Erener et al., 2016; Pradhan and Lee, 2010; Umar et al., 2014; Youssef et al., 2015; Mousavi et al., 2015; Pradhan, 2010), ML (Yeon et al., 2010; Pradhan, 2013; Chen et al., 2016; Park et al., 2014; Ren and Wu, 2014; Pradhan, 2013) and hybrid models (Althuwaynee et al., 2016; Sangchini et al., 2016). To improve these base models, researchers have tried various approaches. Methods such as integrating several individual models (Wen et al., 2016), optimising spatial resolution of conditioning factors (Schlögel et al., 2016), selection of suitable training samples (Nefeslioglu et al., 2008), optimising model's hyperparameters (Jebur et al., 2014; Dou et al., 2015, Pradhan and Lee, 2010) and others have focused on the development of new models (Hoang and Tien Bui, 2016).

To apply the above-mentioned modelling techniques, two basic datasets are needed: (1) landslide inventory and (2) landslide conditioning factors. The quality and characteristics of these datasets play a key role in training powerful models that can generalise to the spatial domain of the investigated areas. The following sections explain the effects of landslide inventory dataset and different sampling methods on the performance of LSM.

The quality and quantity of the landslide inventory dataset play a major role in defining the accuracy of the LSM. In addition, the preparation procedure and pre-processing of inventory datasets remarkably affect the accuracy of LSM (Nefeslioglu et al., 2008, Rotigliano et al., 2011, Rotigliano et al., 2012, Conoscenti et al. 2016; Hong et al., 2018). The problems related to training

samples and their impacts on modelling performance are broadly discussed. This section reviews some related literature to identify the shortcomings and potential solutions that can be implemented to resolve the existing problems.

Landslide inventory is usually created by analysing historical aerial photographs, satellite images, or field surveys using global navigation satellite systems (Mezaal et al., 2017; Pradhan et al., 2017). Hussin et al. (2016) suggested different strategies for mapping landslides such as (1) mapping the mass or scarp centre (a single pixel), (2) mapping all the pixels that correspond to a landslide boundary, (3) using the pixels that correspond to the mapped scarp and (4) the ones close to the scarp area denoting the line of landslide crown and (5) using the pixels that correspond to the landslide boundary with an added buffer zone. The latter strategy is widely used throughout literature and is found suitable for training ML and statistical models (e.g. LR, conditional probability (CP) and artificial neural networks (ANN)) (Nefeslioglu et al., 2008; Yılmaz, 2010). Yılmaz (2010) analysed the effects of different sampling strategies on the predictive performance of CP and ANN. The study suggested that adding a zoning buffer to a landslide boundary improves the accuracy of the mentioned models. By contrast, Regmi et al. (2014) found that creating landslide samples from the scarp centre yields the best accuracy amongst other strategies when the LR model is used.

Landslide inventory preparation also includes collecting negative samples (pixels that correspond to non-landslide areas). In this regard, negative samples can be selected from randomly distributed circular zones, which have a specific diameter (e.g. equal to the mean width of the landslide source area), or from the grid cells without creating buffer zones (Conoscenti et al., 2016). However, no obvious conclusion is found to know which method will work best prior to model building and testing. After the inventory data are prepared, the standard process is to divide the data into training, validation and testing samples for the purposes of model development, optimisation and evaluation, respectively. In particular, constructing a training set from the entire inventory dataset is a challenge and greatly influences the performance of the models.

A set of strategies for selecting training samples with regard to LSM is used in literature. The basic strategy is random sampling (RS) (Pradhan and Lee, 2010; Conoscenti et al., 2016; Erener et al., 2017; Raja et al., 2017). The inventory dataset of N samples is randomly divided into training N_{train} and testing N_{test}. The training dataset is further divided into training and validation samples

N_{vald} in some cases. The latter is used to fine-tune the learning algorithms and to select the optimum settings. Although preserving the randomness in selecting training samples is an important property of any sampling strategy, RS does not guarantee selecting best subsets to train a model (Dhakal et al., 2000). The distribution of the training dataset should not vary too much compared with the entire set. The inconsistent results are mainly due to the differences in the training datasets created by a random process without investigating the properties of the created samples. To improve this approach, a stratified RS (SRS) is utilised (Marjanović et al., 2011; Dhakal et al., 2000). In this method, the observations belong to landslide/non-landslide targets that are kept as proportional as possible. Dhakal et al. (2000) found that SRS performs better than the purely random process of subdividing the inventory data into training and testing samples. Other methods include bootstrap resampling (Goetz et al., 2011), cross-validation (CV) (Goetz et al., 2015) and a few special strategies (e.g. region partitioning) (Hong et al., 2018).

In addition to the sampling strategy, the percentage of training/testing samples is found to be critical to LMS performance (Hjort and Marmion, 2008; Heckmann et al., 2014). The size of the training dataset should be determined by a systematic approach rather than a completely random process. The small size training dataset may not capture the spatial variability of the conditioning factors. By contrast, the training dataset of large size will more likely violate the independent observation assumptions because of spatial autocorrelation (Kalantar et al., 2018; Heckmann et al., 2014). In addition to the number of training samples, the quality of training samples also play a key role in determining the performance of the modelling algorithms. Methods such as Divergence, Transformed divergence, Bhatarchraya distance, Jeffries-Matusita distance, Wilk's Lambda, Hotelling's T-squared have been used to evaluate the quality of the training samples and subsequently select training/testing samples for modelling (Kalayeh and Landgrebe, 1983; Djouadi et al., 1990; Kavzoglu and Mather, 2002).

Overall, apart from developing models with robust architectures and optimal hyperparameters, training data are also keys to obtain accurate and generalisable LSM tools. Selecting training samples that capture the distribution of the entire set is an important property that should be incorporated into the process of sample selection. Conclusions drawn on the basis of randomly selected test samples cannot be generalised because the characteristics of the selected dataset may favour certain models unnoticeably. A systematic approach for analysing training samples can

provide insights into the data characteristics and the performance of models from certain families, which can save time and computing resources. It also helps simplify the selection of a model immediately for practitioners.

3. Methodology

3.1 Case study and data used

3.1.1 Study area

Cameron Highlands is located in the north-central part of Peninsular Malaysia and is geographically located between latitudes 101°24'00"E and 101°25'10"E and longitudes 4°30'00"N and 4°30'55"N (Figure 1). The area is a tropical rainforest region and is approximately 200 km from Kuala Lumpur. This area was selected due to its frequent landslide occurrences (Khan, 2010), which have caused considerable damage to environments and properties, and the availability of its landslide inventory data. The lithology of the area mainly consists of Quaternary and Devonian granite and schist. The granite in Cameron Highlands is classified as megacrysts biotite granite (Pradhan and Lee, 2010). The area mostly has hilly landforms (land slope ranges from 0° to 78°) where the lowest and highest altitudes are 1,153 m and 1,765 m, respectively.

3.1.2 Landslide inventory data

The selected site occupies a surface area of approximately 25 km^2. The landslide inventory data of this area were collected from the Department of Geological Survey. The data were prepared using multisource remote sensing images such as archived 1:10,000–1:50,000 aerial photographs, SPOT 5 panchromatic satellite images and high-resolution orthophotos acquired by airborne laser scanning systems. Old landslides were validated by visual interpretation of aerial photographs, whereas fresh landslides were validated by field surveys. A total of 104 landslide locations were mapped in the study area, as shown in Figure 1. Most of the landslides are shallow rotational and a few translational in type.

Landslide susceptibility modelling with machine learning and statistical methods also requires negative samples. In this research, the non-landslide points were selected as follows. We used landslide inventories as a guide to select these non-landslide points. The selection process was performed randomly, but any selected had to satisfy the following conditions. First, any non-

landslide sample should be at least 500 meters away from landslides. Second, the distance between any two non-landslide samples must be greater than 100 meters. As a result, 208 sample points were prepared for further analysis including landslide (labelled 1) and non-landslide points (labelled 0).

Figure 1. Location of the study area (part of Cameron Highlands, Malaysia) and the landslide inventory map.

3.1.3 Landslide conditioning factors

LiDAR point clouds were used to construct a very high-resolution (0.5 m) digital elevation model (DEM) of the area (Table 1). Seven conditioning factors were prepared from the derived DEM and thematic maps acquired from government agencies (Figure 2). From DEM, several geomorphological factors such as altitude, slope and aspect were obtained. The land cover map was prepared from SPOT 5 satellite images (10 m spatial resolution) using the maximum likelihood classification method. Then, 10 classes of land cover types were identified, including water bodies, transportation, agriculture, residential and bare land. The overall accuracy of the classified map was 87.20% and the Kappa index was 0.863, verified with field surveys. Lastly, distances to road, river, and lineament were calculated based on the Euclidean distance method using the thematic layers. These conditioning factors and the landslide inventories were geometrically calibrated and organised in a geodatabase file in GIS.

Table 1. Information on LiDAR data collection mission.

Figure 2. Landslide conditioning factors.

3.2 Data pre-processing

The landslide inventory data were pre-processed before using them in modelling experiments (Figure 3). Firstly, landslide records with missing values due to outside data coverage were removed. Three observations were removed. Secondly, the outliers (i.e. infrequent observations) were detected by Grubbs outlier detection tests (Grubbs, 1969). Lastly, the values of landslide conditioning factors were normalised using the min–max method with the following expression:

$$X' = \frac{\max(X) - X}{\max(X) - \min(X)} \quad (1)$$

where X' is the normalised value; X is the original value; and $max(X)$ and $min(X)$ are the maximum and minimum values in the original data, respectively.

Figure 3. Flowchart of the research methodology used to provide the landslide susceptibility map.

3.3 Inventory information curve

Inventory information curve (IIC) is a graphical representation of the information contained in an inventory dataset (D). ICC is a plot of Hellinger distance calculated between the distribution of a training subset randomly drawn with incremental sizes (from a lower bound of 0.1 to an upper bound of 0.95) (P) and the distribution of the entire inventory set (P_D) versus training subset size (Equation 2). For a given inventory data set D that contains landslide and non-landslide samples, firstly, $K \leq D$ random subsets are drawn from D with landslide and non-landslide samples, with only landslide samples (D_l) and with only non-landslide samples (D_n). Secondly, for all the K subsets, the Hellinger distance is calculated between the subset and the entire set. Thirdly, the area under IIC (AUIIC) is calculated for the landslide curve, non-landslide curve, and their combined curve. Lastly, a line plot is produced showing the Hellinger distance curves plotted against the size of the training subset.

$$H^2(P, P_D) = \frac{1}{2}\int \left(\sqrt{p(x)} - \sqrt{p_D(x)}\right)^2 dx \quad (2)$$

3.4 Training sample selection

For a given inventory dataset ($D = \{(\overrightarrow{x_i}, y_i) | i = 1{:}N\}$), to select a training subset ($D_{train} \in D$), a randomly $n < D_{train}$ samples were selected, where n is the initial sample size. Thereafter, a variable named 'hellinger_distances' was instantiated to store the Hellinger distances calculated between a progressive subset from the original dataset and X. To reduce the effect of feature interdependencies, X is transformed by the min–max (Equation 1) method firstly and then independent component analysis method. The iterative process starts with the initial subset of size n and calculate the Hellinger distance between X_n and X. This process continues for N samples. Next, Hellinger distances were calculated for the instances of the initial subset. Accordingly, a lookup table with the indices of the original samples and their associated Hellinger distance is obtained. Thereafter, the observations in the lookup table are sorted on the basis of the ascending

value of the Hellinger distance. Lastly, the first N_s samples were selected to create the training subset, whereas the remaining samples are returned as testing samples. The exact procedure is presented in Algorithm 1.

Algorithm 1
Procedure for selecting training samples.
1. D-original dataset, N is the size of D, N_s is the training subset size, n is the initial sample size.
2. Set hellinger_distances = {}
3. Extract X and y from D according to the features names and target column.
4. Scale X with the min-max method and transform it with ICA to obtain X_s
5. For $i := n$ to N do
    Set current_subset = $X[:i]$
    Calculate current_hellinger_distance using Equation 2, takes (current_subset, X) as inputs
    Append a new value (current_hellinger_distance) to hellinger_distances with key str (i)
6. For $i := 0$ to $n + 1$ do
    Set current_subset = $X[i:]$
    Calculate current_hellinger_distance using Equation 2, takes (current_subset, X) as inputs
    Append a new value (current_hellinger_distance) to hellinger_distances with key str (i)
7. Sort hellinger_distances ascending
8. Calculate indices of N_s training samples
9. Calculate the rest of indices, being testing indices
10. Create and Return X_{train} , y_{train} , X_{test} , y_{test} using training and testing indices and X, y

3.5 Supervised classification

3.5.1 Decision tree

DT is a member of supervised tree-based models and classifies a given training set into homogenous subgroups by using constructed rules or decisions (Friedl and Brodley, 1997; Quinlan, 2014). The main goal of DT during the learning process is to achieve the maximum information and minimum entropy in the generated model. DT consists of nodes that stand for circles and branches that stand for segments connecting the leaf nodes, as its name suggests. The most common algorithm to implement DT is J48 (i.e. a slightly modified version of C4.5). It generates a classification–DT for the given dataset by recursive partitioning of data (Zhao and Zhang, 2008). It examines all the possible tests to select the best option to split the dataset by measuring the information gain or gain ratio, which is calculated as follows:

$$GR = \frac{I(S,A)}{E(S,A)}, \tag{3}$$

$$E = -p_p \log_2 p_p - p_n \log_2 p_n, \quad (4)$$

where GR is the gain ratio, E is the entropy, I is the information gain, S is a training set, A is an attribute, p_p is the proportion of positive examples in S and p_n is the proportion of negative examples in S (Saghebian et al., 2014).

3.5.2 Support vector machines

SVM (Cortes and Vapnik, 1995; Vapnik, 1998) is a non-parametric ML algorithm that does not make any assumptions on training data distributions. In SVM, a hyperplane that acts as a decision surface is constructed. The margin of separation between two classes (landslides and non-landslides) is maximised via optimisation processes. This goal is achieved by mapping the original feature space into a surface that has high dimensionality. The aim of this transformation is to allow the target classes to be linearly separable. The standard procedure of these transformations is performed by a method called kernel trick. Kernel functions such as radial basis function (RBF) and logistic functions are often used (Equations 5 and 6). The learning process of the SVM is performed by constrained optimisation. In our implementation, SVM parameters were selected via a grid search and a 10-fold CV process, which yielded the following best local parameters. The kernel function was RBF, a penalty parameter (also known as C) of 500, and a kernel parameter (i.e. gamma) of 0.15. The best SVM was used throughout the experiments. This study also presented the impacts of these parameters on the proposed IIC and the performance of the SVM. Details and further information on SVM are provided in (Matkan et al., 2014; Melgani and Bruzzone, 2004; Zhan and Shen, 2005; Pradhan and Sameen, 2018).

$$K(x_i, x_j) = e^{\left(-\gamma \|x_i - x_j\|^2\right)}, \gamma > 0. \quad (5)$$

$$K(x_i, x_j) = \tanh\left(\gamma x_i^T x_j + r\right) \quad (6)$$

3.5.3 *k*-Nearest neighbours

KNN algorithm classifies an instance by a majority vote of its (k) neighbours. In other words, the instance is assigned the most frequent class amongst the selected neighbours. The parameter k is a positive integer and is often a small number. The selection of k depends on the data being used.

The rule of thumb is to select large k values when data are noisy; otherwise, small k values are suggested. For binary classification, an odd number of k can be helpful to avoid difficulties with tied votes. Proper validation methods, such as CV, should be used to select the best k for KNN.

3.6 LSM assessment

The predictive ability of the classification methods is assessed using an n-fold cross validation and the area under the receiving characteristic curves (ROCs) (i.e. AUC for short). ROCs are a graphical representation of model success and predictive accuracies, whereas AUC is a quantitative measure that summarises the model performance. ROC is plotted as a scatterplot with a landslide susceptibility percentage (horizontal axis) and cumulative landslide occurrence percentage (vertical axis). AUC is calculated by a trapezoidal formula as follows:

$$\mathrm{AUC} = \frac{1}{2}\sum_{\mathrm{i}=1}^{\mathrm{n}-1}(\mathrm{T}_{\mathrm{i}+1} - \mathrm{T}_{\mathrm{i}})(\mathrm{C}_{\mathrm{i}+1} + \mathrm{C}_{\mathrm{i}} - 2\mathrm{B}), \tag{7}$$

where T_{i} is the i^{th} percent landslide susceptibility, C_{i} is the i^{th} cumulative percentage of landslide occurrence, n is the number of the percent landslide susceptibility index value and B is the baseline value (i.e. B is usually equal to zero). The area between the baseline and the curve is computed by Equation 1 to determine the performance of the landslide susceptibility model.

4. Results and discussion

This section presents the main findings and discusses the implications of each of them. This research was implemented with open source libraries such as Numpy, Scikit-learn and Pandas in Python. The source code is available upon a reasonable request from the corresponding author. The ArcGIS Pro 2.4 was used to prepare data and represent modelling results.

4.1 Inventory Information Curve (**IIC)**

Figure 4 presents IIC of Cameron Highlands dataset. The large training subsets taken from the entire set approximates the original distribution effectively. This case is obvious in IIC as indicated by the decreasing Hellinger distance and is observed for three target scenarios: the complete dataset with landslide and non-landslide samples (blue solid line), the dataset with only landslide samples (red solid line) and the dataset with only non-landslide samples (green solid line). The graphical

representation of the inventory data provides insights into the landslide (positive) and non-landslide (negative) samples, and the entire dataset combines the observations from the two targets. The gradual decrease in Hellinger distance indicates that the general characteristics of the entire inventory data can be recovered with a relatively small number of training samples. However, the fluctuations in the curves imply that large amounts of data are needed to model the details precisely. The variations in the curves also suggest that the compositions of the individual subsets and the training set size considerably impact the Hellinger distance.

The shape of IIC captures the information content presented in the inventory dataset. The very steep curve at the beginning indicates that the small training subsets have relatively different distributions compared with the entire set. As the curves flatten towards the end of the plot and become close to each other, the training subsets represent the original distribution relatively accurately. AUIIC can be calculated to summarise the information into a single parameter. AUIIC ranges from 0 to 1 and is a useful measure to describe the characteristics of training samples and its differences from the entire inventory dataset. It is also useful for comparing the characteristics of different inventory datasets.

In our case, AUIIC for the landslide samples (0.216) is less than that for the non-landslide samples (0.275). The landslide samples are all collected from within landslide boundaries, whereas the non-landslide samples are collected from the rest of the area (presents different features, e.g. buildings, trees and water bodies). As a result, the distribution of different individual landslide subsets is varied less than that for non-landslide subsets. This condition also suggests using additional samples from non-landslide areas particularly their selection and subdivision into training and testing sets may affect the evaluation process of LSM. Thus, a systematic approach to selecting training and testing samples is a critical step in achieving improved learning and accurate evaluations in LSM.

Figure 4. Inventory information content curve for our dataset.

4.2 Performance of LSM using IIC-based sampling

The aim of these experiments was to determine the relationship of the variance error of the selected models (i.e. SVM, KNN and DT) with IIC curve. The models were trained with the same settings used to generate IIC curve. They were trained on random subsets with different sizes (10%–90%)

and tested on the entire dataset. The errors were calculated using the following expression $E = 1 - AUC$, where E is the classification.

Figure 5 presents IIC and error rate of SVM, KNN and DT. The error rate of the models decreases with the increase in training subset size. The average error of the models decreases from approximately 0.25 to 0.06 due to the increase in the training subset size from 0.1% to 0.9% of the entire dataset. The errors have a general decreasing curve; however, they present minimal variations when looking at the details. This general pattern and the details are represented relatively accurately by IIC. Therefore, IIC can be a useful tool to indicate the performance of the susceptibility models prior to their applications.

Figure 5. Error curves of the KNN, SVM, and DT models overlaid with IIC.

4.3 Impacts of hyperparameters

This research analysed the relationship between the performance of the investigated modelling methods with different hyperparameter configurations and the Hellinger distance using IIC and AUIIC. The details are explained in the following sections.

For KNN, the effect of the *k* parameter is examined, which controls the number of neighbours to use when classifying an instance. Figure 6a shows the error rate of KNN classification with different values of the *k* parameter. The results indicate that the error rate decreases by increasing the training samples, but KNN with $k = 1$ achieves the lowest error. The error curve of this scenario is the closest to the Hellinger distance curve. By contrast, the worst result is observed when $k = 2$. The selection of *k* depends on the size and distribution of training samples because different values of *k* are found best for different training subsets. Thus, to achieve accurate predictions of landslide susceptibility with KNN, the value of *k* should be carefully selected with the proposed IIC as a suitable guide.

The maximum depth of a tree in DT plays a major role in determining its prediction accuracy. It controls the maximum depth of the tree that will be created. Alternatively, the longest path is from the tree root to a leaf. The results indicate that DT performs nearly similar with a maximum depth of 2 being the worst (Figure 6b).

Two parameters, namely, the penalty parameter (C) and the kernel function, were explored for SVM. The C parameter determines the effect of the misclassification on the objective function in the optimisation process. By contrast, the kernel function is responsible for transforming the space of original features into a space of a higher dimension. Kernel functions are keys to separate the nonlinearly separable instances in SVM. Our experiments reveal that the C parameter has a greater influence than the kernel function on the performance of landslide prediction (Figures 6c and 6d). The best result is obtained when $C = 250$ or $C = 500$. No difference is found when linear or RBF kernels are used. The error of SVM is approximately 0.07 in both cases.

Figure 6. Impacts of hyperparameters of the selected models on LSM performance and its relationship with IIC.

4.4 Impact of noise and outliers

Noise and outliers considerably impact ML and statistical models. However, not all the methods are affected by the same magnitude. Thus, investigating the impacts of noise and outliers on such models is important to improve LSM. Several experiments were conducted to understand the effects of artificial noise and outliers. The Cameron Highlands dataset was modified by introducing Gaussian noise (mean = 0, standard deviation = 1) and 1% of the factors values by multiplying them by a random number between −5 and 5.

Figures 7a and 7b show IICs for the Cameron Highlands dataset and its modified version (noise added). The dataset with noise has a higher variance and is less regular than the curve of the original dataset. Adding noise to the dataset increases not only the variance but also AUIIC. Therefore, IIC has a strong relationship with the quality of the samples selected from the entire set. This information can be used to design effective samples for training LSM models and obtain insights into the dataset prior to the application of the algorithms.

This study also explored the effects of adding noise to landslide inventory datasets on the performance of modelling algorithms (SVM, KNN, and DT). Figures 7c and 7d present the performance of the models with the original and noisy datasets. The results indicate that DT is highly affected by noise, whereas SVM and KNN have less sensitivity to the noise added to the dataset. The error of DT increases from 0.06 to 0.13. By contrast, the errors calculated for SVM and KNN with noisy data are found to be 0.08 and 0.05, respectively. These results suggest that

IICs are useful tools to explore the quality of the inventory datasets before developing the models and can be used to decide the family of the models according to the shape of the curves. For datasets with few variations and low AUIIC, models that are sensitive to noise such as tree-based methods (DT and random forests) may not be a good choice. In those cases, models such as SVM or ensemble models are suggested to reduce the impacts of noise and outliers.

Figure 7. Impacts of noise and outliers on IIC and performance of KNN, SVM, and DT.

4.5 Comparison with other sampling methods

Various sampling strategies can be used in experimental studies regarding LSM. The most popular methods are RS, SRS and CV. RS divides the inventory data into training and testing subsets randomly with a given threshold. SRS is like RS, but the observations belong to landslide/non-landslide targets that are kept as proportional as possible. Instead, CV divides the inventory data into k subsets. One of the subsets is kept testing the model while the remaining subsets are merged and used to train the model. This process is repeated for several iterations. It aims at a better evaluation of modelling methods with limited data and it generally results in a less biased model compare to other methods. To show the validity of our sampling strategy based on IIC, its performance was compared with those of the standard techniques. Table 2 summarises the performance of different models based on AUC and a test dataset that is unseen during the training process. The parameters of the models and the subset size were kept the same during the experiments to ensure a fair comparison. The results suggest that the selection of training samples remarkably affects the performance of the models. With results regarding SVM, CV and our method are far better than RS and SRS. SVM with the same hyperparameters (C = 250, kernel = 'RBF') and a random seed (42) achieves improved predictive accuracy when CV and our methods are used to generate the training samples. The accuracy of SVM decreases to approximately 0.06 when RS and SRS are applied. For KNN, the results indicate that this model has less sensitivity to the sampling method used. However, our method is stable and can achieve the highest accuracy of 0.944 (±0.03). Similarly, analysing the results of DT reveals that this model is greatly affected by the sampling method. CV and ours achieve the best accuracies (0.933 and 0.978, respectively). Overall, the results of these experiments indicate the importance of utilising suitable sampling strategies whilst subdividing the inventory datasets into training and testing subsets. In addition,

CV and the proposed method can be an alternative to the classical RS methods, especially for SVM and DT.

Table 2. Training and testing accuracies of LSM with different sampling methods.

5. Conclusions

The quality of landslide training samples play an important role in the performance of the susceptibility models. Selection of training samples with random processes does not guarantee the best models. This research studies the selection of training samples with a novel method based on Hellinger distance measurements which calculate the difference in the probability distributions of training subsets and the entire dataset. Experimental analyses were conducted to understand the properties of IIC and its use in selecting training samples for applying ML and statistical models for landslide susceptibility assessment. Experiments were also conducted to explain the impacts of model's hyperparameters, noise and outliers on the performance of the selected models (KNN, SVM and DT), and the relationship between the error rate of these models and the shape of IICs. Furthermore, the proposed method was compared with other benchmark methods such as RS, SRS and CV.

Results from this research indicate that the proposed IIC as a graphical representation of landslide inventory data provides insights into the landslide and non-landslide samples, and the entire dataset. This research also provided a parameter (AUIIC) that describes the characteristics of training samples and its differences from the entire inventory dataset. AUIIC is an important parameter to compare different inventory datasets and determine suitable training samples for certain susceptibility models. Our experiments on various models (SVM, KNN, DT) showed that IIC is a useful tool to approximate the predictive performance of the susceptibility models prior to their applications. This feature provides a cost-effective solution for model selection in landslide susceptibility assessment. The use of IIC also helped in selecting appropriate hyperparameters of the landslide susceptibility models. In another experiment, we introduced artificial noise to the landslide inventory dataset. The results indicated that adding noise to the dataset introduces high variance in AUIIC. The results also suggested that DT was affected by the artificial noise larger than the SVM and KNN models. The performance of our sampling strategy also proved to be better

than RS, SRS, and CV for the selected models as shown in the comparative studies conducted in the research.

The performance of landslide susceptibility models is dependent on the number of training samples and their quality. This is more impactful when the training data is scarce. This is because the estimates of the first, second-order statistics cannot accurately represent all the information which is contained in the data. Inaccurate estimation of these data properties effect analysis of the data including modelling. Therefore, further studies should be conducted to improve our understanding on training data selection and estimation of the sample's quality. New improved methods should be developed to predict in a statistically reasonable way the required number of training samples for ML and statistical models.

Acknowledgements

This research is supported by the Centre for Advanced Modelling and Geospatial Information Systems (CAMGIS) in the University of Technology Sydney (UTS) under Grants 321740.2232335 and 321740.2232357; Grant 321740.2232424, and Grant 321740.2232452. This research is also supported by Researchers Supporting Project number (RSP-2019 / 14, King Saud University, Riyadh, Saudi Arabia. Thanks to two anonymous reviewers for their critical review which helped us to improve the manuscript.

References

Aghdam, I.N., Varzandeh, M.H.M., Pradhan, B., 2016. Landslide susceptibility mapping using an ensemble statistical index (Wi) and adaptive neuro-fuzzy inference system (ANFIS) model at Alborz Mountains (Iran). Environ. Earth Sci. 75(7), 553. https://doi.org/10.1007/s12665-015-5233-6

Ahmed, B., 2015. Landslide susceptibility modeling applying user-defined weighting and data-driven statistical techniques in Cox's Bazar Municipality, Bangladesh. Nat. Hazards 79(3), 1707-1737.

Althuwaynee, O.F., Pradhan, B., Lee, S., 2016. A novel integrated model for assessing landslide susceptibility mapping using CHAID and AHP pair-wise comparison. Int. J. Remote Sens. 37(5), 1190-1209.

Althuwaynee, O.F., Pradhan, B., Park, H.J., Lee, J.H., 2014. A novel ensemble bivariate statistical evidential belief function with knowledge-based analytical hierarchy process and multivariate statistical logistic regression for landslide susceptibility mapping. Catena, 114, 21-36.

Chen, W., Chai, H., Zhao, Z., Wang, Q., Hong, H., 2016. Landslide susceptibility mapping based on GIS and support vector machine models for the Qianyang County, China. Environ. Earth Sci. 75(6), 1-13.

Cortes, C., Vapnik, V., 1995. Support vector machine. Machine Learning 20(3), 273-297.

Dehnavi, A., Aghdam, I.N., Pradhan, B., Varzandeh, M.H.M., A new hybrid model using step-wise weight assessment ratio analysis (SWARA) technique and adaptive neuro-fuzzy inference system (ANFIS) for regional landslide hazard assessment in Iran, Catena, 135, 2015, 122-148. https://doi.org/10.1016/j.catena.2015.07.020

Djouadi, A., Snorrason, O., Garber, F.D., 1990. The quality of training sample estimates of the bhattacharyya coefficient. IEEE T. Pattern Anal. 12(1), 92-97.

Dou, J., Bui, D.T., Yunus, A.P., Jia, K., Song, X., Revhaug, I., Xia, H., Zhu, Z., 2015. Optimization of causative factors for landslide susceptibility evaluation using remote sensing and GIS data in parts of Niigata, Japan. PloS One 10(7), e0133262.

Ercanoglu, M., Gokceoglu, C., 2002. Assessment of landslide susceptibility for a landslide-prone area (north of Yenice, NW Turkey) by fuzzy approach. Environ. Geol. 41(6), 720-730.

Ercanoglu, M., Kasmer, O., Temiz, N., 2008. Adaptation and comparison of expert opinion to analytical hierarchy process for landslide susceptibility mapping. B. Eng. Geol. Environ. 67(4), 565-578.

Erener, A., Sivas, A.A., Selcuk-Kestel, A.S., Düzgün, H.S., 2017. Analysis of training sample selection strategies for regression-based quantitative landslide susceptibility mapping methods. Comput. Geosci. 104, 62-74.

Fanos, A.M., Pradhan, B., 2019. A spatial ensemble model for rockfall source identification from high resolution LiDAR data and GIS. IEEE Access. 7, 74570 – 74585. Doi: 10.1109/ACCESS.2019.2919977

Friedl, M.A., Brodley, C.E., 1997. Decision tree classification of land cover from remotely sensed data. Remote Sens. Environ. 61(3), 399-409.

Goetz, J.N., Brenning, A., Petschko, H., Leopold, P., 2015. Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling. Comput. Geosci. 81, 1-11.

Goetz, J.N., Guthrie, R.H., Brenning, A., 2011. Integrating physical and empirical landslide susceptibility models using generalized additive models. Geomorphology 129(3-4), 376-386.

Grubbs, F.E., 1969. Procedures for detecting outlying observations in samples. Technometrics, 11(1), pp.1-21.

Guzzetti, F., Reichenbach, P., Ardizzone, F., Cardinali, M., Galli, M., 2006. Estimating the quality of landslide susceptibility models. Geomorphology 81(1-2), 166-184.

Hasekioğulları, G.D., Ercanoglu, M., 2012. A new approach to use AHP in landslide susceptibility mapping: a case study at Yenice (Karabuk, NW Turkey). Nat. Hazards 63(2), 1157-1179.

Hong, H., Pradhan, B., Sameen, M.I., Kalantar, B., Zhu, A., Chen, W., 2018. Improving the accuracy of landslide susceptibility model using a novel region-partitioning approach. Landslides 15(4), 753-772.

Hong, H., Liu, J., Zhu, AX., Shahabi, H., Pham, B.T., Chen, W., Pradhan, B., Tien Bui, D., A novel hybrid integration model using support vector machines and random subspace for weather-triggered landslide susceptibility assessment in the Wuning area (China), Environ. Earth. Sci. 2017, 76: 652. https://doi.org/10.1007/s12665-017-6981-2

Huabin, W., Gangjun, L., Weiya, X., Gonghui, W., 2005. GIS-based landslide hazard assessment: an overview. Prog. Phys. Geog. 29(4), 548-567.

Hussin, H.Y., Zumpano, V., Reichenbach, P., Sterlacchini, S., Micu, M., van Westen, C., Bălteanu, D., 2016. Different landslide sampling strategies in a grid-based bi-variate statistical susceptibility model. Geomorphology, 253, 508-523.

Jebur, M.N., Pradhan, B., Tehrany, M.S., 2014. Optimization of landslide conditioning factors using very high-resolution airborne laser scanning (LiDAR) data at catchment scale. Remote Sens. Environ. 152, 150-165.

Kalantar, B., Pradhan, B., Naghibi, S.A., Motevalli, A., Mansor, S., 2018. Assessment of the effects of training data selection on the landslide susceptibility mapping: a comparison between support vector machine (SVM), logistic regression (LR) and artificial neural networks (ANN). Geomat. Nat. Haz. Risk. 9(1), 49-69.

Kalayeh, H.M., Landgrebe, D.A., 1983. Predicting the required number of training samples. IEEE T. Pattern Anal. (6), 664-667.

Kavzoglu, T., Mather, P.M., 2002. The role of feature selection in artificial neural network applications. Int. J. Remote. Sens. 23(15), 2919-2937.

Kavzoglu, T., Sahin, E.K., Colkesen, I., 2015. An assessment of multivariate and bivariate approaches in landslide susceptibility mapping: a case study of Duzkoy district. Nat. Hazards 76(1), 471-496.

Korup, O., Görüm, T., Hayakawa, Y., 2012. Without power? Landslide inventories in the face of climate change. Earth Surf. Proc. Land. 37(1), 92-99.

Marjanović, M., Kovačević, M., Bajat, B., Voženílek, V., 2011. Landslide susceptibility assessment using SVM machine learning algorithm. Eng. Geol. 123(3), 225-234.

Matkan, A.A., Hajeb, M., Sadeghian, S., 2014. Road extraction from Lidar data using support vector machine classification. Photogramm. Eng. Remote S. 80(5), 409-422. doi: 10.14358/pers.80.5.409.

Melgani, F., Bruzzone, L., 2004. Classification of hyperspectral remote sensing images with support vector machines. IEEE T. Geosci. Remote. Sens. 42(8), 1778-1790.

Mezaal, M.R., Pradhan, B., Sameen, M.I., Mohd Shafri, H.Z., Yusoff, Z.M., 2017. Optimized neural architecture for automatic landslide detection from high-resolution airborne laser scanning data. Appl. Sci. 7(7), 730.

Mohammady, M., Pourghasemi, H.R., Pradhan, B., 2012. Landslide susceptibility mapping at Golestan Province, Iran: A comparison between frequency ratio, Dempster–Shafer, and weights-of-evidence models. J. Asian Earth Sci. 61(15), 221-236. https://doi.org/10.1016/j.jseaes.2012.10.005

Nefeslioglu, H.A., Gokceoglu, C., Sonmez, H., 2008. An assessment on the use of logistic regression and artificial neural networks with different sampling strategies for the preparation of landslide susceptibility maps. Eng. Geol. 97(3), 171-191.

Pourghasemi, H.R., Pradhan, B., Gokceoglu, C., Moezzi, K.D., 2012. Landslide susceptibility mapping using a spatial multi criteria evaluation model at Haraz Watershed, Iran. In Terrigenous Mass Movements (pp. 23-49). Springer Berlin Heidelberg.

Pradhan, B., 2010. Landslide susceptibility mapping of a catchment area using frequency ratio, fuzzy logic and multivariate logistic regression approaches. J. Indian. Soc. Remote. Sens. 38(2), 301-320.

Pradhan, B., 2013. A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS. Comput. Geosci. 51, 350-365.

Pradhan, B., Lee, S., 2010. Landslide susceptibility assessment and factor effect analysis: backpropagation artificial neural networks and their comparison with frequency ratio and bivariate logistic regression modelling. Environ. Modell. Softw. 25(6), 747-759.

Pradhan, B., Sameen, M.I., 2017. Laser scanning systems in landslide studies. In Laser Scanning Applications in Landslide Assessment (pp. 3-19). Springer, Cham.

Pradhan, B., Sameen, M.I., 2018. Manifestation of SVM-Based Rectified Linear Unit (ReLU) Kernel Function in Landslide Modelling. In Space Science and Communication for Sustainability (pp. 185-195). Springer, Singapore.

Pradhan, B., Seeni, M.I., Kalantar, B., 2017. Performance evaluation and sensitivity analysis of expert-based, statistical, machine learning, and hybrid models for producing landslide susceptibility maps. In Laser Scanning Applications in Landslide Assessment (pp. 193-232). Springer, Cham.

Pradhan, B., Seeni, M.I., Nampak, H., 2017. Integration of LiDAR and QuickBird data for automatic landslide detection using object-based analysis and random forests. In Laser Scanning Applications in Landslide Assessment (pp. 69-81). Springer, Cham.

Pradhan, B., Sezer, E.A., Gokceoglu, C., Buchroithner, M.F., 2010. Landslide susceptibility mapping by neuro-fuzzy approach in a landslide-prone area (Cameron Highlands, Malaysia). IEEE T. Geosci. Remote. Sens. 48(12), 4164-4177.

Quinlan, J. R. (2014). C4. 5: programs for machine learning. Elsevier.

Raja, N.B., Çiçek, I., Türkoğlu, N., Aydin, O., Kawasaki, A., 2017. Landslide susceptibility mapping of the Sera River Basin using logistic regression model. Nat. Hazards 85(3), 1323-1346.

Regmi, N.R., Giardino, J.R., McDonald, E.V., Vitek, J.D., 2014. A comparison of logistic regression-based models of susceptibility to landslides in western Colorado, USA. Landslides 11(2), 247-262.

Ren, F., Wu, X., 2014. GIS-based landslide susceptibility mapping using remote sensing data and machine learning methods. In Cartography from Pole to Pole (pp. 319-333). Springer Berlin Heidelberg.

Rotigliano, E., Agnesi, V., Cappadonia, C., Conoscenti, C., 2011. The role of the diagnostic areas in the assessment of landslide susceptibility models: a test in the Sicilian chain. Nat. Hazards 58(3), 981-999.

Rotigliano, E., Cappadonia, C., Conoscenti, C., Costanzo, D., Agnesi, V., 2012. Slope units-based flow susceptibility model: using validation tests to select controlling factors. Nat. Hazards 61(1), 143-153.

Saghebian, S.M., Sattari, M.T., Mirabbasi, R., Pal, M., 2014. Ground water quality classification by decision tree method in Ardebil region, Iran. Arab. J. Geosci. 7(11), 4767-4777.

Sangchini, E.K., Emami, S.N., Tahmasebipour, N., Pourghasemi, H.R., Naghibi, S.A., Arami, S. A., Pradhan, B., 2016. Assessment and comparison of combined bivariate and AHP models with logistic regression for landslide susceptibility mapping in the Chaharmahal-e-Bakhtiari Province, Iran. Arab. J. Geosci. 9(3), 1-1

Schlögel, R., Marchesini, I., Alvioli, M., Reichenbach, P., Rossi, M., Malet, J.P., 2016. The role of method of production and resolution of the DEM on slope-units delineation for landslide susceptibility assessment-Ubaye Valley, French Alps case study. In EGU General Assembly Conference Abstracts (Vol. 18, p. 15505).

Sezer, E.A., Nefeslioglu, H.A., Osna, T., 2017. An expert-based landslide susceptibility mapping (LSM) module developed for Netcad Architect Software. Comput. Geosci. 98, 26-37.

Thiery, Y., Malet, J.P., Sterlacchini, S., Puissant, A., Maquaire, O., 2007. Landslide susceptibility assessment by bivariate methods at large scales: application to a complex mountainous environment. Geomorphology 92(1), 38-59.

Tien Bui, D., Ho, T.C., Pradhan, B., Pham, B-T., Nhu, V-H., Revhaug, I., 2016. GIS-based modeling of rainfall-induced landslides using data mining-based functional trees classifier with AdaBoost, Bagging, and MultiBoost ensemble frameworks. Environ. Earth Sci. 75(14), 1102. https://doi.org/10.1007/s12665-016-5919-4

Vapnik, V., 1995. The nature of statistical learning. Springer, New York

Vapnik, V., 1998. Statistical learning theory. New York: Wiley.

Wen, Z., He, B., Xu, D., Feng, Q., 2016. A method for landslide susceptibility assessment integrating rough set and decision tree: A case study in Beichuan, China. In Geoscience and Remote Sensing Symposium (IGARSS), 2016 IEEE International (pp. 4952-4955). IEEE.

Yeon, Y.K., Han, J.G., Ryu, K.H., 2010. Landslide susceptibility mapping in Injae, Korea, using a decision tree. Eng. Geol. 116(3), 274-283.

Yilmaz, I., 2010. The effect of the sampling strategies on the landslide susceptibility mapping by conditional probability and artificial neural networks. Environ. Earth Sci. 60(3), 505-519.

Youssef, A.M., Al-Kathery, M., Pradhan, B., 2015. Landslide susceptibility mapping at Al-Hasher area, Jizan (Saudi Arabia) using GIS-based frequency ratio and index of entropy models. Geosci. J. 19(1), 113-134. https://doi.org/10.1007/s12303-014-0032-8

Zare, M., Porghasemi, H.R., Vafakhah, M., Pradhan, B., 2013. Application of weights-of-evidence and certainty factor models and their comparison in landslide susceptibility mapping at Haraz watershed, Iran. Arab. J. Geosci. 6(8), 2873-2888. https://doi.org/10.1007/s12517-012-0610-x

Zhan, Y., Shen, D., 2005. Design efficient support vector machine for fast classification. Pattern Recognition 38(1), 157-161.

Zhao, Y., Zhang, Y., 2008. Comparison of decision tree methods for finding active objects. Adv. Space Res. 41(12), 1955-1959.

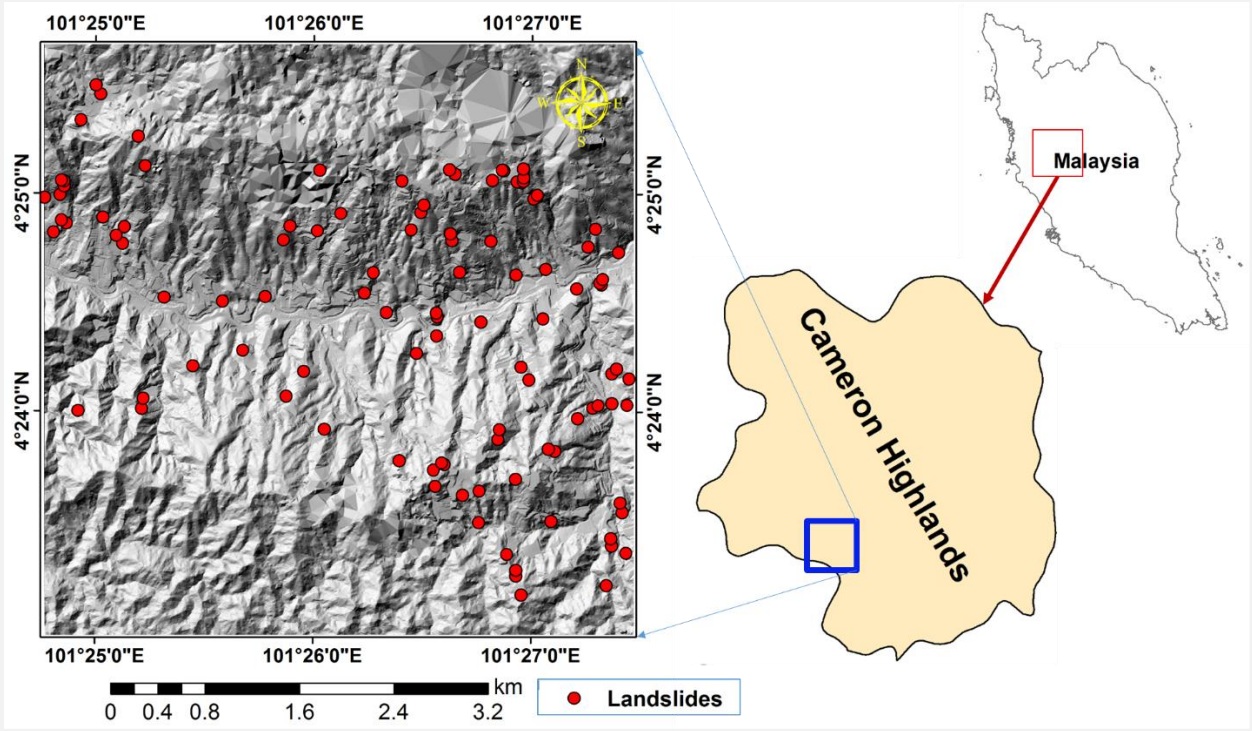
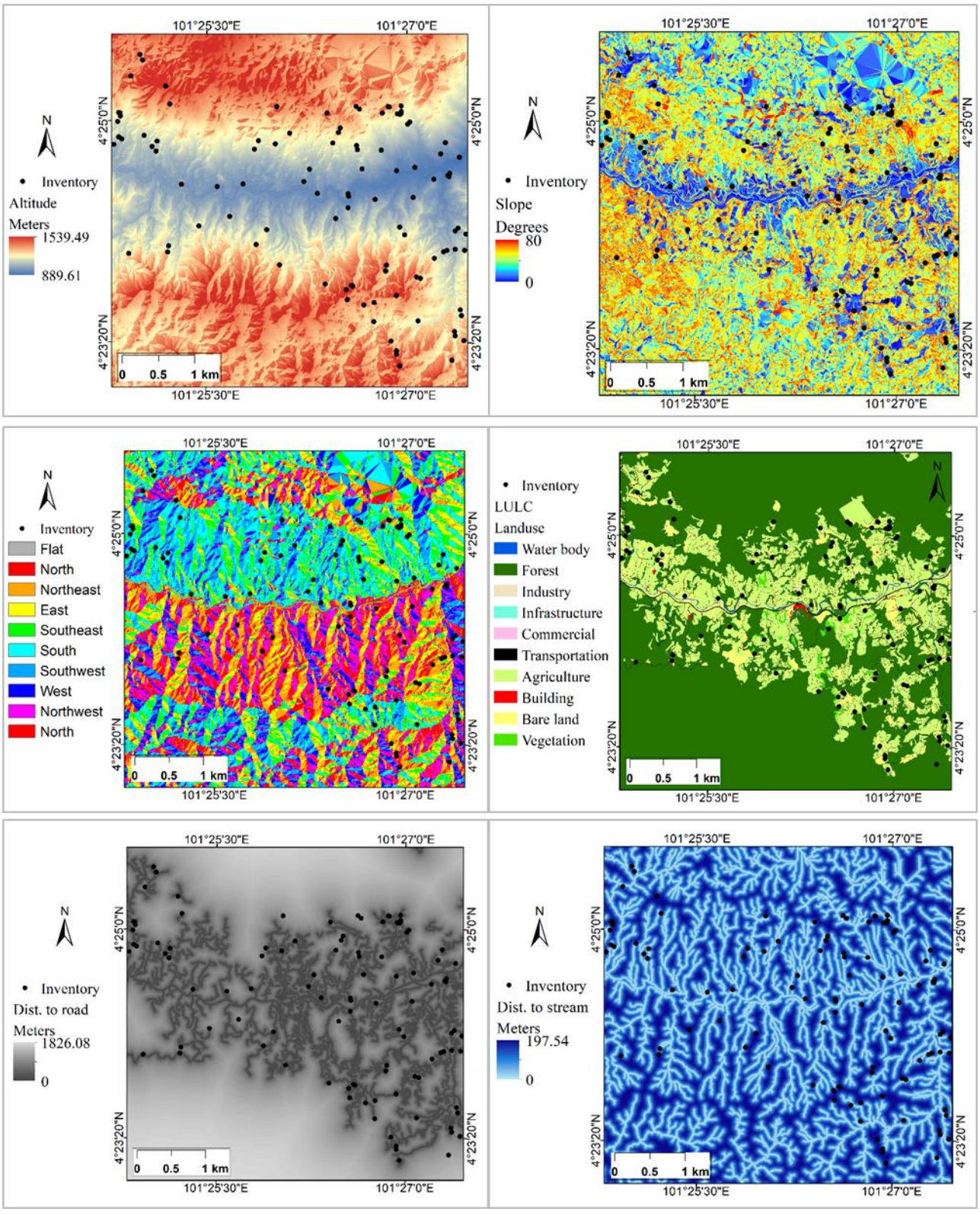


Figure 2. Location of the study area (part of Cameron Highlands, Malaysia) and the landslide inventory map.


101°25'30"E
101°27'0"E
4°25'0"N
4°23'20"N
N
Inventory
Altitude
Meters
1539.49
889.61
0 0.5 1 km
Inventory
Slope
Degrees
80
0
Inventory
Flat
North
Northeast
East
Southeast
South
Southwest
West
Northwest
North
Inventory
LULC
Landuse
Water body
Forest
Industry
Infrastructure
Commercial
Transportation
Agriculture
Building
Bare land
Vegetation
Inventory
Dist. to road
Meters
1826.08
0
Inventory
Dist. to stream
Meters
197.54
0

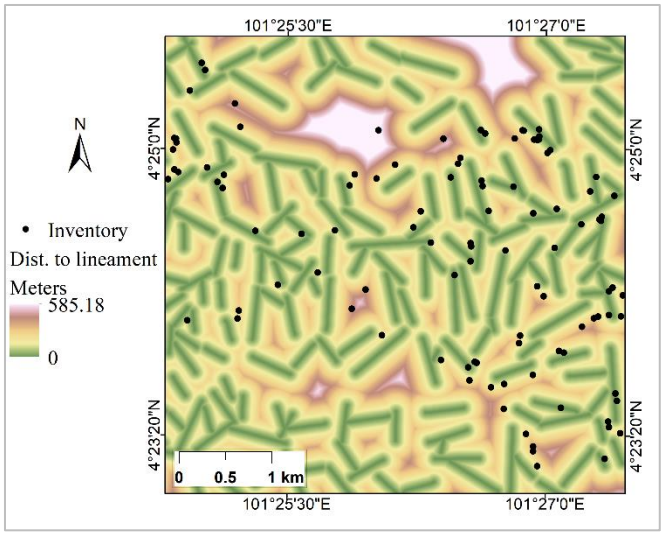


Figure 2. Landslide conditioning factors.

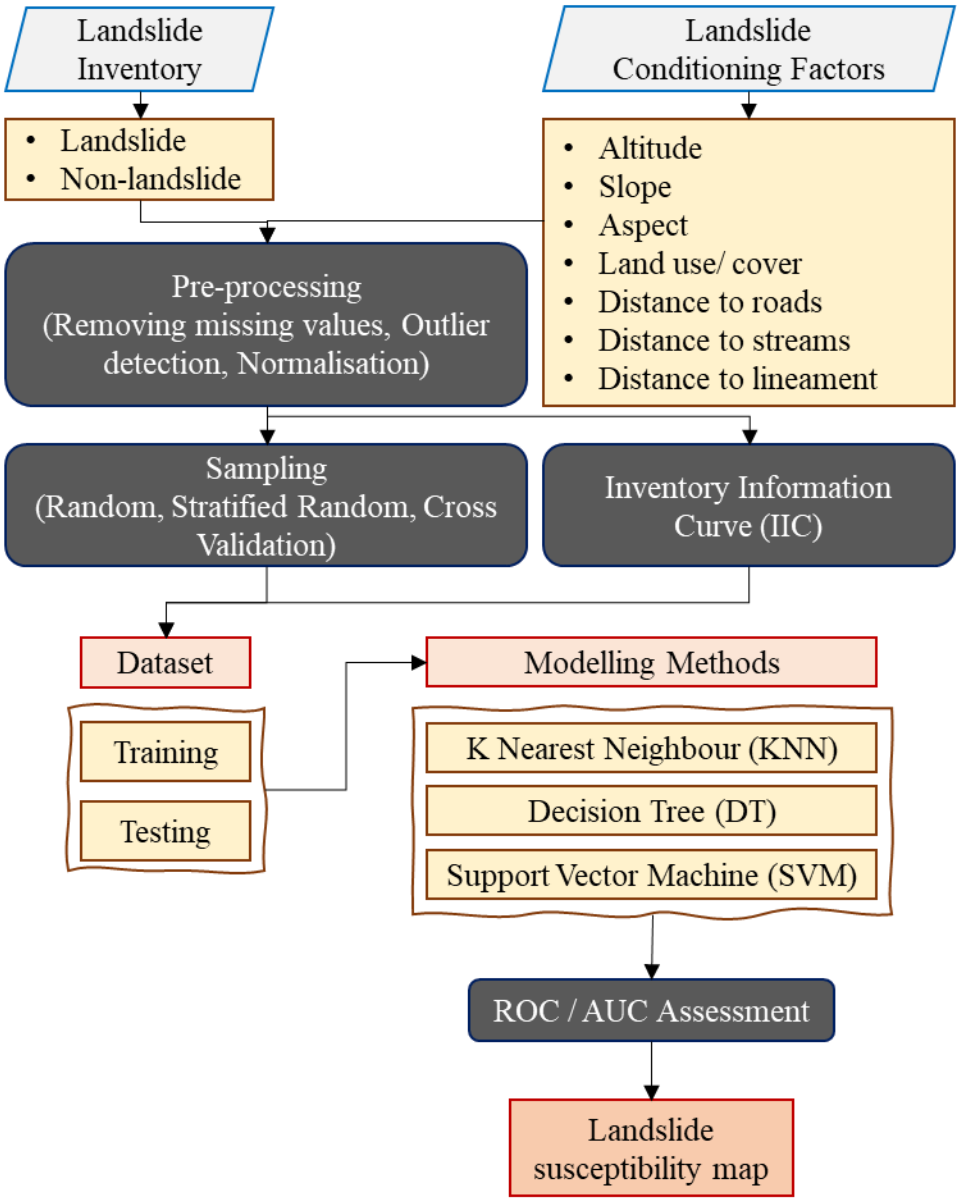


Figure 3. Flowchart of the research methodology used to provide the landslide susceptibility map.

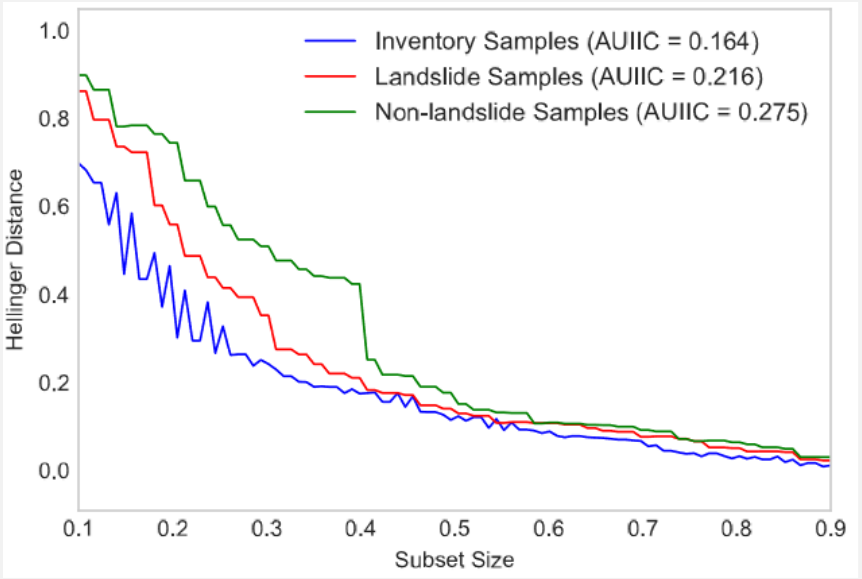


Figure 4. Inventory information content curve for our dataset.

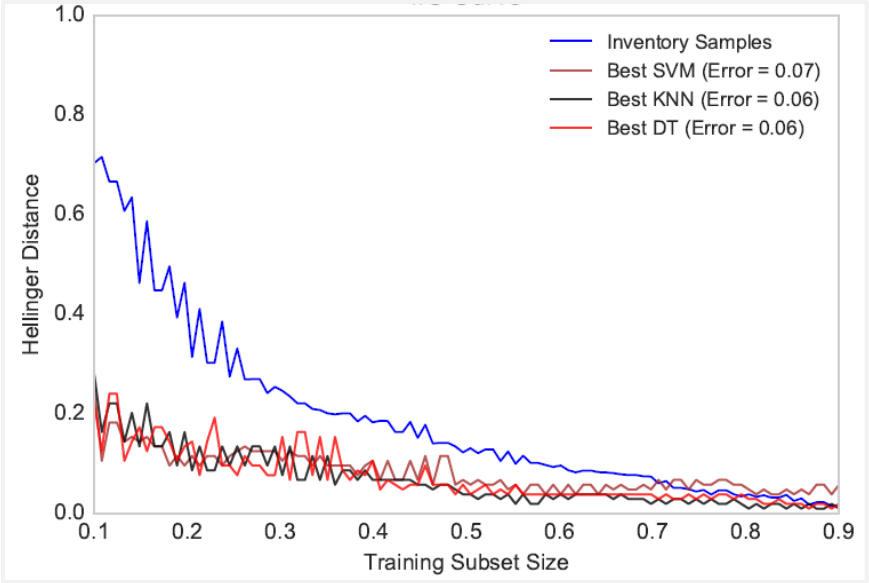


Figure 5. Error curves of the KNN, SVM and DT models overlaid with IIC.

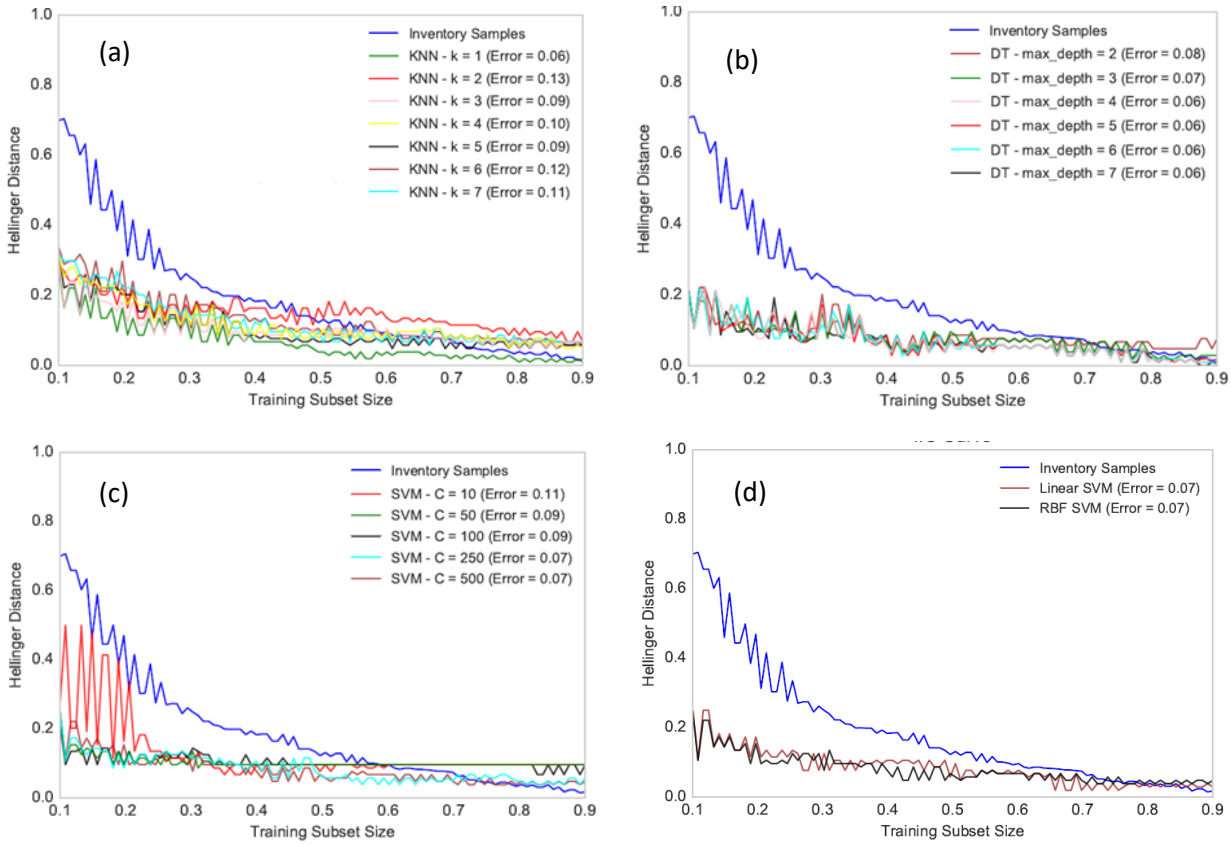


Figure 6. Impacts of hyperparameters of the selected models on LSM performance and its relationship with IIC for (a) KNN, (b) DT, (c) SVM penalty parameter, and (d) SVM kernel function.

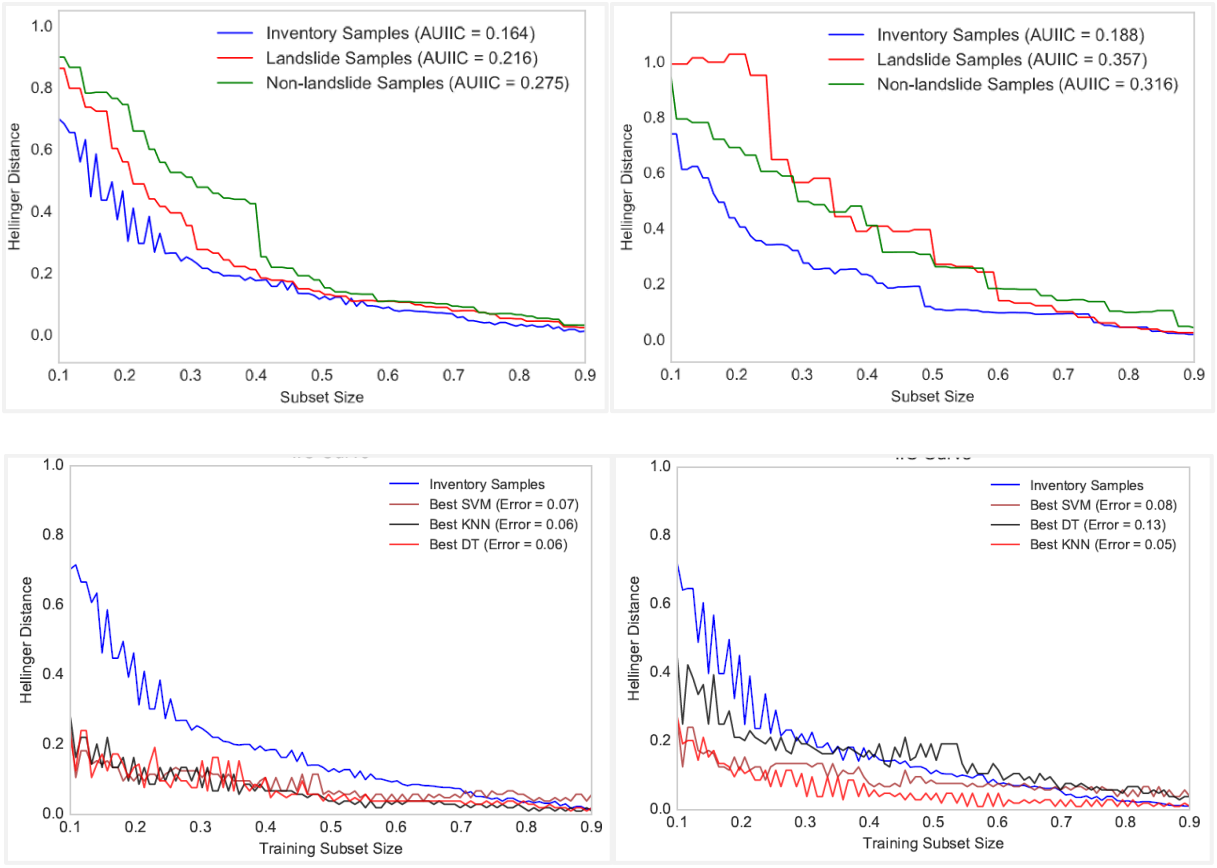


Figure 7. Impacts of noise and outliers on IIC and performance of KNN, SVM and DT.

Table 1. Information on LiDAR data collection mission.

Parameter	Value
Date	January 15, 2015
Average flight height	1,510 m
Point density	8 points per m^2
Frequency rate	25,000 Hz
Absolute vertical accuracy	0.15 m
Absolute horizontal accuracy	0.3 m

Table 2. Training and testing accuracies of LSM with different sampling methods.

Method	**Testing AUC**		
	SVM	KNN	DT
Random Sampling	0.884 (±0.08)	0.938 (±0.12)	0.916 (±0.07)
Stratified Random Sampling	0.887(±0.04)	0.942 (±0.06)	0.887 (±0.04)
Cross-validation (10-fold)	**0.943** (±0.04)	0.941 (±0.07)	0.933 (±0.05)
IIC (this work)	0.913 (±0.01)	**0.944** (±0.03)	**0.978** (±0.03)