Is the ROC curve a reliable tool to compare the validity of landslide susceptibility maps?

V. Vakhshoori and M. Zare

Department of Earth Sciences, College of Sciences, Shiraz University, Shiraz, Iran

ABSTRACT
This study investigates how effective the receiver operating characteristic (ROC) curve is for comparing the reliability of landslide susceptibility maps (LSMs). In this regard, we selected a basin prone to landslides in northern Iran and employed frequency ratio and weight-of-evidence methods for modelling. Modelling for each method was done by considering three fractions of training/test landslides (50/50, 60/40, and 70/30) and, in addition, three random seeds for each of these fractions, leading to produce 18 LSMs. Validation rates of LSMs were obtained through calculation of the area under the ROC curve. Moreover, Cohen’s kappa index was calculated to reveal the magnitude of the agreement between the maps. Results showed that all LSMs, despite having equal validation rates, were considerably different in terms of spatial prediction pattern. It is concluded that, although as a prevalent validation tool, ROC curve is only an indicator of the general reliability of geographical prediction maps and cannot reveal the uncertainty of spatial prediction patterns. Therefore, to reduce the conflicts between the maps and create a single reliable map, all LSMs were merged using the overlaying statistical functions that can extract the best possible zone pattern from all the overlapped patterns.

1. Introduction
To organize land-use plans and to define the areas where slope stabilization is necessary, landslide susceptibility maps (LSMs) are produced as a powerful tool (Akgun 2012). However, like spatial prediction map of any other natural hazards, LSMs need to be validated (Beguería 2006) to ascertain their reliability before using. Until the commencement of recent decade, validation of LSMs had not been paid enough attention (Beguería 2006). In many studies, LSMs validity had merely been presented by simple statistics such as landslide percentage per susceptible zones and model efficiency (the proportion of correctly classified events derived from a confusion matrix). Owing to the inefficiency of these simple statistics (Provost and Fawcett 1997; Provost et al. 1998), threshold-independent methods, like receiver operating characteristic (ROC), have been recommended for validation (Fielding and Bell 1997; Beguería 2006; Corominas et al. 2014). Accordingly, many recent studies (Nandi and Shakoor 2010; Akgun 2012; Pourghasemi, Pradhan, and Gokceoglu 2012; Zare et al. 2012; Pourghasemi et al. 2013; Conforti et al. 2014; Pham et al. 2016; Youssef et al. 2016; Tsangaratostes et al. 2017; Wang et al. 2017) have utilized ROC curve to demonstrate and compare the reliability of their created LSMs. Undoubtedly, the aim of LSMs is predicting the landslide-prone areas, and ROC method is used to indicate how reliable these areas have been predicted. However, in some studies (Mohammady et al. 2012; Park et al. 2012; Vakhshoori and Zare 2016), while two separate...
LSMs produced for an area are given the same or very close reliability by ROC method, they show different spatial patterns: either ROC method is not efficient enough to show that differences or these visually seen differences are not considerable. In this study, it is investigated how statistically different equally reliable LSMs of an area could be, why the ROC curve cannot reveal this discrepancy, and which approach can be adopted to reduce this type of uncertainty. The study is carried out in the continuation of our previous study (Vakhshoori and Zare 2016) but with a different purpose. The aim of the previous study was creating LSMs by different models and comparing their reliability by ROC method, whereas the purpose of this study is to investigate the reliability of ROC method itself.

2. Study area

Generally, the regions between Alborz mountain ranges and Caspian Sea in northern Iran can be named the most susceptible areas to landslide, which encompass Gillan, Golestan, and Mazandaran provinces. The selected area in this study is Qaemshahr basin in the middle part of Mazandaran province. This catchment covers a local-scale area of 990 km² between latitudes 36° 00' 04'' to 36° 30' 00'' N and longitudes 52° 30' 50'' to 52° 55' 06'' E (Figure 1). Elevation gradually rises from 11 m below m.s.l on the northern flat regions near the Caspian Sea to 3709 m above m.s.l on the southern steep highlands. Most of the valleys in the area show the main geographical direction of south-north and lead the surface streams to flow towards the northern areas where they form the main river called Babol River. If considering the study area in three geological parts, the northern part is mostly covered by urban areas and agricultural lands, the southern part by mountains and some pasture lands, and the central part mainly by dense forests facing annual mean precipitation of 900–1000 mm (based on the long-term precipitation data from the 15 climatology stations of Mazandaran province). The precipitation gradually decreases to annual average of 600 mm towards the northern and, likewise, the southern parts. In terms of temperature, the annual average of 16 °C has been recorded in the northern part, 14 °C in the middle part, and from 12 to 4 °C (declining by altitude increment) in southern part. In addition, the climatic conditions gradually vary from semi-humid in the north to very humid climate in the south. From the lithological point of view, the basin

Figure 1. Geographical position of Qaemshahr basin and the location of occurred landslides.
encompasses 28 separable units (Figure 2). These units can be categorized into three groups considering both their age and the mentioned geological parts. In the Southern parts (Mesozoic era), the prevailing materials are sandstone and quartzite sandstone, siltstone, claystone, shale, and conglomerate. The central areas belong to the Triassic to Tertiary periods and comprise limestone, dolomite, sandstone, siltstone, and marl. The northern part (Quaternary) almost includes pluvial and fluvial fans and terraces, and alluvial floodplains. In this part, the predominant lithological materials are conglomerate, silty marl, siltstone, and sandstone.

3. Materials and methods

3.1. Landslide inventory and causative factors

The available data employed in this study include landslide inventory and landslide causative factors (consist of environmental and triggering factors) as the required data-set for landslide susceptibility zoning (van Westen et al. 2005). Among them, the most important data is the landslide inventory, which must be complete as much as possible regarding the fact that ‘today and the past are keys to the future’ (Carrara et al. 1991; Aleotti and Chowdhury 1999; Guzzetti et al. 1999; van Westen et al. 2008; Harp et al. 2011). In this study, the landslide inventory that had been collected up to 2005 (by the organization of Forest, Range, and Watershed management of Iran) was extended until 2016 with the aid of Google Earth®, which provides multi-temporal high-resolution satellite images in a three-dimensional view (Sato and Harp 2009). A multi-temporal detected landslide is shown in Figure 3. Overall, the landslide inventory map of the study area includes 135 landslides from which, about 63% are in the form of shallow slides, around 32% are flows (mud flow, debris flow, and often a combination of them), and approximately 15% are complex landslides. In terms of area, 36% of the landslides are smaller than 1000 m², just over 50% are in the range of 1000–10,000 m², 13% between 10,000 and 35,000 m², and only one landslide is bigger than 55,000 m². For each landslide, a 30 x 30 m pixel as the initiation point (Sterlacchini et al. 2011; Zêzere et al. 2017) was selected to be engaged in modelling. It is recommended that different LSMs should be created for different groups of landslides (Glade et al. 2006; van Westen et al. 2008; Corominas et al. 2014). In this study, there were 22 rock falls and rock topples happened on the uninhabited southern mountains that
were not considered along with those 135 landslides in the assessment because, according to our knowledge about the study area, they occurred in different conditions. For example, slopes are often steep to vertical where rock falls and rock topples occur (Highland and Bobrowsky 2008), while other kinds often happen on the gentle slopes. Likewise, the role of water as a triggering factor was different: the occurrence of rock falls and rock topples is not very dependent on the amount of rainfall. Whereas a high volume of rainfall is often needed to cause other types of landslides, a minimum amount of water in the role of freeze/thaw and lubrication factor can effectively influence the occurrence of rock falls and topples.

In the case of causative factors, different types of them can be employed based on the kind of landslides for which the assessment is done, the scale and the conditions of the study area, the availability of maps and information, and time and budget constraints (van Westen et al. 2008; Corominas et al. 2014). In the present study, we provided 10 causative factors including nine environmental factors and a triggering one with the aid of ArcGIS® 10. Considering the pixel size of the ASTER DEM, all the factors were prepared in the raster format with 30 × 30 m pixel size according

Figure 3. Example of multi-temporal detection of landslides by Google Earth® in 2009 (a), 2012 (b), and 2013 (c).
to the grid cells procedure (Guzzetti 2006) and were classified into different classes for modelling (Figure 4(a–i), except for the lithology factor that was given in Figure 2). Based on the data sources they were derived from, these factors are described in different groups in the following.

Using an ASTER digital elevation model (DEM) in 30 m resolution, the digital layers of slope degree, slope aspect, altitude, and stream network were prepared. Slope degree can be entitled the most powerful factor in landslide occurrence (Corominas et al. 2014), so that the gradient increment rises the shear stress in materials and changes the situation in favour of landslides. In the case of slope aspect, it consists of four main geographical and four sub-geographical directions plus a flat area class. The correlation between the different directions and landslides has still not been
completely comprehended, but it is said that different directions can reflect the soil moisture and vegetation density on the slopes (van Westen et al. 2008) and, in addition, could be affected by the toe erosion and deposition process of meandering rivers differently depending on the flow direction (Vakhshoori and Zare 2016). These underlain factors might contribute to landslides occurrence and, therefore, the slope aspect layer is used as a geomorphological variable in landslide susceptibility assessments (Dai and Lee 2002). Likewise, altitude variable in connection with other factors (for example, tectonic, precipitation, erosion, and weathering) can considerably affect the landslide occurrence (Rozos et al. 2008). The stream network is considered as an important linear factor that leads to the toe erosion of slopes. This factor was classified in different buffer zones using the Euclidean Distance tool in ArcGIS® 10.

Drawing upon the 1:100,000 geology map of Qaemshahr (produced by Geological Survey and Mineral Exploration of Iran), two other linear factors (i.e. the faults and the roads) as well as the layer of lithology were digitized. Similar to the layer of distance to stream network, the buffer zones around the active faults and the roads were mapped. The effect of active faults on the landslides is mainly related to their potential seismic activities, and therefore, the relatively wide buffers were considered around them (van Westen et al. 2008). The inverse effect of building the roads especially on the stability of steep slopes has been well known (Barnard et al. 2001; Borga et al. 2005; Kamp et al. 2008; Eker and Aydin 2016). Road construction on steep slopes, for example, disrupts the natural gradient leading to landslides occurrence. Lithology as a traditional factor used in landslide susceptibility assessments can indicate different properties of rocks and soils like their resistance to weathering and erosion, permeability, shear strength, and thus, their sensitivity to landslides (Varnes 1984).

By employing Landsat 8 satellite data (band 2–7 with a resolution of 30 m), land use/cover was prepared based on the supervised classification procedure of maximum likelihood classification. Land use/cover is often used in landslide susceptibility assessments as a static factor and can show valuable information such as human activities that contribute to landslides occurrence (e.g. deforestation and cultivation on the slopes that formed the most susceptible class of land use in this study), and inversely, the positive effect of vegetation like root reinforcement in slopes stability (van Westen et al. 2008). By means of the same satellite data (bands 4 and 5), normalized difference vegetation index (NDVI) was calculated in Raster Calculator tool of ArcGIS® 10 as follows:

\[
\text{NDVI} = \frac{\text{Band 5} - \text{Band 4}}{\text{Band 5} + \text{Band 4}} \tag{1}
\]

where Band 4 and Band 5 of Landsat 8 reflect the visible red and the near infrared measurements, and the NDVI value varies between −1 and +1. A high value of NDVI in an area indicates a high density of vegetation in that area, where the annual average precipitation and soil moisture are often high, and distance to the stream network is often short; therefore, a high NDVI value may indicate a high probability of landslides occurrence (Vakhshoori and Zare 2016).

Eventually, using the long-term data of Mazandaran climatological stations, the annual mean precipitation map was produced as a triggering factor. Water plays an active role in triggering landslides by creating permanent moisture and saturating slopes, changing groundwater level, and lubricating rupture surface (Varnes 1984; Dahal and Hasegawa 2008). Considering these important roles, in the absence of meteorological stations data for a small-scale study, even a general estimation of rainfall using satellite data is sufficient (van Westen et al. 2008).

3.2. Methodology

The methodology of this study differs from a routine landslide susceptibility study in some aspects, although repeatable in other areas. The aim of this study is to show the inefficiency of ROC method in validation of LSMs; therefore, it is enough to prove that the same maps in terms of the area under
the curve (AUC) of ROC method are different in terms of showing the landslide-prone areas. To do that, first, the models should be repeated several times using two relatively similar methods in order to increase the chance of producing different LSMs with the same ROC-AUC. Second, those LSMs that are equal in terms of AUC should be proved discrepant by discovering their disagreement, which is done using Cohen’s Kappa index here. After proving that matter, the overlaying statistical functions are employed to merge the LSMs in order to reduce the uncertainty of their spatial patterns. The following sections briefly discuss all the employed methods and approaches.

3.2.1. Partitioning the landslide inventory
The prepared landslide inventory should be divided into training and test data-sets; the first for constructing the model and calculating its success rate, and the second for measuring the prediction rate of the model (Guzzetti 2006; Fell et al. 2008). Subdividing a landslide inventory can be done in different ways: spatial, temporal, or random (Chung and Fabbri 2003). In this study, the random procedure was chosen in order that the models could be repeated freely using different generated datasets. To that end, three different training/test ratio including 50/50, 60/40, and 70/30 were considered, and for each ratio, partitioning was repeated three times based on different random seeds. In this way, nine separate training and test inventories were provided. This manner was considered to ensure that if there was a discrepancy between LSMs with equal reliability, it is not due to either equal training/test ratio or the same random seed.

3.2.2. Modelling
To explore the relationship between the training inventory data-set and the landslide causative factors, different methods can be applied. Despite this possibility, the frequency ratio (FR) and weight of evidence (WofE) methods were chosen primarily due to the fact that they often generate LSMs with the same reliability, as a necessity for this study. Moreover, they always show almost a high accuracy (Lee and Pradhan 2006; Neuhäuser and Terhorst 2007; Akgun et al. 2008; Dahal et al. 2008; Yilmaz 2009; Pradhan, Lee, and Buchroithner 2010; Neuhäuser et al. 2011; Choi et al. 2012; Park et al. 2012; Vakhshoori and Zare 2016).

FR is a bivariate statistical method that discloses the correlation between occurred landslides and each class of the causative factors (Lee 2005). The weight (FR$_c$) of each class is calculated as the percentage of landslides within the class divided by its area percentage. After assigning the relevant weights to each class, the summation of the classes produces the landslide susceptibility index as follows:

$$LSI = \sum_{c=1}^{n} FR_c$$

WofE is a probabilistic method employed in different fields as well as landslide susceptibility modelling (Dahal et al. 2008; Pradhan, Oh, and Buchroithner 2010; Neuhäuser et al. 2011; Pourghasemi, Pradhan, Gokceoglu, Mohammadi, and Moradi 2012; Regmi et al. 2013). WofE is a log-linear form of Bayes’ rule that is

$$P(A|B) = P(B|A) \times P(A) / P(B)$$

The standard weight of a class in this method, $C / S(C)$, is calculated based on the presence (A) or absence (A) of landslides inside (B) or outside (B) of the class. C is the difference between the magnitude of positive (Wi$^+$) and negative (Wi$^-$) weights (Bonham-Carter 1994) which are

$$Wi^+ = \ln \left( \frac{P\{B|A\}}{P\{B|\bar{A}\}} \right) \quad \text{and}$$

$$Wi^- = \ln \left( \frac{P\{\bar{B}|A\}}{P\{\bar{B}|\bar{A}\}} \right)$$
and $S(C)$ is the standard deviation of $C$ which is expressed as

$$S(C) = \sqrt{S_{W_i^+}^2 + S_{W_i^-}^2}$$  \hspace{1cm} (6)$$

where $S_{W_i^+}^2$ and $S_{W_i^-}^2$ are the variances of $W_i^+$ and $W_i^-$ (Agterberg et al. 1990) calculated as follows:

$$S_{W_i^+}^2 = \frac{1}{6}P\{B|A\} + \frac{1}{6}P\{B|\overline{A}\}$$  \hspace{1cm} (7)$$

$$S_{W_i^-}^2 = \frac{1}{6}P\{\overline{B}|A\} + \frac{1}{6}P\{\overline{B}|\overline{A}\}$$  \hspace{1cm} (8)$$

where $P$ means the probability. The landslide susceptibility index of the WofE method is prepared by overlaying all reclassified LCFs.

Employing these methods (FR and WofE) with different scenarios of landslide inventory partitioning constructed 18 models totally. To name the models, we considered a three-part name for each one. For example, in the case of the model FR603, the two letters refer to the employed method (FR or WofE), the first double digit (60) is the percentage of training data-set, and the last single digit (here 3) shows which repetition of the random partitioning is used.

### 3.2.3. Visualization of the model outputs

The output of each model, landslide susceptibility index with continuous numerical values, should be converted into LSM with discrete susceptible zones. This visualization can be done through five main techniques including simple ranking, natural breaks, mean value and standard deviation intervals, equal interval classes, and equal area classes (Chung and Fabbri 1999, 2003; Fabbri and Chung 2008). The last method – recommended by Chung and Fabbri (2003) – is employed in this study since it allocates an equal area to each homonymous zone (that are identical in terms of susceptibility but are in different LSMs). Due to having equal area, these zones can be compared visually (Pradhan and Lee 2010) and, more importantly, quantitatively (that is discussed in Section 3.2.5) as a necessity for discovering the spatial pattern differences between the maps.

For visualization with this method, the pixels of a landslide susceptibility index are sorted based on their values and then divided into different susceptibility zones so that a concomitant rise in terms of susceptibility is seen for zones that are comparatively more susceptible. However, it should be noted that there is a drawback with this method that may cause an inevitable systematic error in visualization. On occasion, a large number of pixels that are on the border of two consecutive zones get an equal value, and therefore, cannot be divided between the zones. Subsequently, software allocates them to one of the zones, causing a slight change in the zones’ area, and thus increasing the differences between the maps. Generally, this error rises by increasing the number of zones as well as the size of the pixels, and if the value range of a susceptibility index generated by a prediction method is limited. Although that may scarcely exceed a few per cent in normal conditions, this error should be considered when comparing the zones pattern of the LSMs.

The area earmarked for each zone in this study are, respectively, 40%, 20%, 20%, 10%, and 10% for very low, low, moderate, high, and very high susceptibility zones. As an illustration, the LSM and the landslide susceptibility index of the model FR703 are shown in Figure 5. After creating the LSMs, ROC method was used to validate them.

### 3.2.4. Validation by ROC method

The main advantage of ROC method as a threshold-independent curve is exactly its independency from the number of thresholds considered for calculations as well as their interspaces (Fawcett 2006). Assuming $n$ classes for a landslide susceptibility index, $n + 1$ thresholds are defined so that the value of the first threshold ($i = 1$) is lower than the minimum susceptibility observed in the
most stable class, and the value of the last threshold \((i = n + 1)\) is higher than the maximum susceptibility in the most sensitive class. Each threshold forms a confusion matrix in which four types of pixels are defined: true positive (TP), false positive (FP), true negative (TN), and false negative (FN) pixels. TP and FN pixels are landslides within the classes above and below the value of the threshold, respectively. On the contrary, TN and FP pixels are the stable pixels within the classes below and above the value of the threshold, respectively. Based on the numbers of these pixels for each threshold, two statistics are calculated, namely TPR (true positive rate) and FPR (false positive rate) as follows:

\[
TPR = \frac{TP}{TP + FN} \\
FPR = \frac{FP}{TN + FP}
\]

TPR and FPR are plotted, respectively, on the \(y\)-axis and the \(x\)-axis of ROC curve. They form the point of \((1, 1)\) on the curve for the first threshold \((i = 1)\), and the point of \((0, 0)\) for the last threshold \((i = n + 1)\). In the last step of validation, the AUC is calculated as follows:

\[
AUC = \sum_{i=1}^{n+1} \frac{1}{2} \sqrt{(x_i - x_{i+1})^2 \cdot (y_i + y_{i+1})}
\]

The AUC value can show the model success rate by engaging the training data-set and its prediction rate by the test data-set. The success rate indicates the model fitting rate that means how well the LSM separates the landslides among its susceptibility zones (Chung and Fabbri 1999). However, it cannot properly reveal the model performance in prediction of the future landslides (Lee et al. 2007; Pourghasemi, Pradhan, and Gokceoglu 2012); therefore, the prediction rate is calculated.

3.2.5. Cohen’s kappa index

To measure the quantity of the pixel-by-pixel agreement between different LSMs (especially those with equal ROC-AUC in order to fulfil the aim of this study), Cohen’s kappa index is employed.
This metric is calculated based on the observed agreement of maps \( P_{\text{obs}} \) and the expected agreement of them by chance \( P_{\text{exp}} \) as follows (Cohen 1960; Guzzetti et al. 2006):

\[
k = \frac{P_{\text{obs}} - P_{\text{exp}}}{1 - P_{\text{exp}}} \tag{12}
\]

where \( P_{\text{obs}} \) is written as

\[
P_{\text{obs}} = \frac{TP + TN}{N} \tag{13}
\]

and \( P_{\text{exp}} \) as

\[
P_{\text{exp}} = \frac{(TP + FN)(TP + FP) + (FP + TN)(FN + TN)}{N^2} \tag{14}
\]

where \( N \) is the total number of pixels in the map. The value of \( k \) could fall in the range of \(-1\) to \(+1\): from an utter disagreement to a perfect agreement.

### 3.2.6. Overlaying statistical functions

To reduce the discrepancies between the spatial prediction patterns of the maps, the overlaying statistical functions in the Cell Statistics tool of the ArcGIS® 10 software can be used. Although these functions may not raise the resultant ROC-AUC of the final integrated map, they are used to ensure that the best possible spatial prediction pattern is selected from a range of predicted patterns. Three useful kinds of these functions are Maximum, Majority, and Minimum that are used in this study. Maximum function selects the highest susceptibility values among the values observed in all overlapped pixels and assigns them to the output pixels, thus increasing the considered area of high and very high susceptible zones. This provides the assurance that all the possible susceptible areas are considered in the mentioned zones for more detailed studies and plans, but perhaps at the expense of budget and time overspending. Majority function produces a map of the most observed value in each pixel among the values of all overlapped pixels; hence, the resultant map shows an average spatial pattern of predictions without a great change in zones’ area. In the case of Minimum function, it selects the minimum susceptibility degrees as the output meaning that only those pixels form the most sensitive zones that are absolutely very susceptible to landslides because they must have been predicted very susceptible by all the maps. Selection from the described functions depends on different factors such as the scale and the conditions of the study area, the importance and the density of elements at risk, the financial constraints and the time limitations, and the type of action that is supposed to be taken based on these maps (e.g. detailed susceptibility mapping, landslides hazard and risk mapping, slope stability measures, land use management, or a combination of them).

### 4. Results and discussion

Considering Table 1 (the last row), the average success rates of both FR and WofE methods are the same by 0.83, and so are the average prediction rates by 0.79, showing the equally good performance of the methods. Furthermore, considering the validation rates of the models constructed by different training/test ratio (Table 1), there is no conclusive finding in this study to confirm which ratio could have better results and it appears that none of them meaningfully reduces/increases the validity of models. Nevertheless, the training/test ratio of 50/50 is preferred because of having two advantages. First, it provides the opportunity of verification of the models with a larger number of validation data and, in addition, mapping with a fewer number of modelling data; therefore, it can more confidently show the consistency of the models performance. Second, it probably can reduce the
occurrence of an unbalanced distribution of landslides randomly divided between the modelling and validation data-sets, as evidenced by the comparatively lower differences between the success rate and the prediction rate of the models performed by this ratio (Table 1). However, the second advantage should be proved to be statistically meaningful by more repetitions of the models in a study for this purpose.

To quantitatively measure the differences between the LSMs, Cohen’s kappa index was calculated for each of their pairs (Table 2). Considering the fact that the homonymous zones had an equal area in the maps, Cohen’s kappa index shows their differences in terms of spatial patterns. In the case of the differences between each zone individually (columns 2–6 of Table 2), the average Kappa values indicated that the middle zones (i.e. low, medium, and high susceptible zones) have undergone more pattern changes compared to the very low and very high susceptible zones in each pair of compared LSMs. This matter is also corroborated by the standard deviation map (by overlaying the 18 LSMs) shown in Figure 6 which visually demonstrates the lower change rates of susceptibility in the very low and very high susceptible regions. The relatively more consistency of the very low and the very high susceptible zones can be justified with the fact that they have only one potential competitor (i.e. the low and high susceptible zones, respectively). Notwithstanding the seemingly low rate of spatial pattern changes, if the very high susceptible zone of each LSM was supposed to be considered for a detailed study, about 30% of this zone equivalent to 30 km², which was predicted not to be very high susceptible in the other map, might have been neglected. In the case of the overall differences between the maps, since none of the kappa values (this time the last column of Table 2) was higher than 0.61 for the LSMs created by the same random seeds but different methods, it can be deduced that all of their pairs were discrepant in terms of spatial prediction pattern, even those with exactly

| Training proportion | Random partitioning | FR | WofE |
|----------------------|---------------------|----|------|
| 50%                  | 1                   | 0.84 | 0.83 | 0.83 | 0.84 |
|                      | 2                   | 0.83 | 0.83 | 0.82 | 0.82 |
|                      | 3                   | 0.81 | 0.79 | 0.80 | 0.77 |
| 60%                  | 1                   | 0.83 | 0.83 | 0.83 | 0.82 |
|                      | 2                   | 0.86 | 0.74 | 0.84 | 0.73 |
|                      | 3                   | 0.85 | 0.80 | 0.85 | 0.80 |
| 70%                  | 1                   | 0.83 | 0.82 | 0.83 | 0.81 |
|                      | 2                   | 0.83 | 0.76 | 0.83 | 0.76 |
|                      | 3                   | 0.84 | 0.79 | 0.84 | 0.80 |
| **Average**          |                     | 0.830 | 0.798 | 0.830 | 0.794 |

Table 2. Kappa values for individual susceptible zones (VL: very low, L: low, M: moderate, H: high, VH: very high) and the whole area of each pair of maps.

| LSMs                | VL  | L   | M   | H   | VH  | $P_{obs}$ | $P_{exp}$ | $k$   |
|---------------------|-----|-----|-----|-----|-----|-----------|-----------|-------|
| FR501 vs. WofE501   | 0.78| 0.39| 0.44| 0.36| 0.73| 0.68      | 0.26      | 0.57  |
| FR502 vs. WofE502   | 0.78| 0.40| 0.48| 0.45| 0.78| 0.70      | 0.26      | 0.59  |
| FR503 vs. WofE503   | 0.86| 0.52| 0.49| 0.37| 0.68| 0.72      | 0.26      | 0.62  |
| FR601 vs. WofE601   | 0.76| 0.37| 0.42| 0.39| 0.73| 0.67      | 0.26      | 0.56  |
| FR602 vs. WofE602   | 0.81| 0.44| 0.48| 0.41| 0.75| 0.71      | 0.26      | 0.60  |
| FR603 vs. WofE603   | 0.82| 0.48| 0.50| 0.40| 0.72| 0.71      | 0.26      | 0.61  |
| FR701 vs. WofE701   | 0.82| 0.45| 0.47| 0.45| 0.75| 0.71      | 0.26      | 0.61  |
| FR702 vs. WofE702   | 0.75| 0.35| 0.40| 0.39| 0.71| 0.66      | 0.26      | 0.54  |
| FR703 vs. WofE703   | 0.71| 0.30| 0.37| 0.33| 0.67| 0.63      | 0.26      | 0.50  |
| WofE603 vs. WofE703 | 0.83| 0.55| 0.65| 0.60| 0.85| 0.78      | 0.26      | 0.71  |
| FR502 vs. FR601     | 0.78| 0.42| 0.55| 0.52| 0.82| 0.72      | 0.26      | 0.62  |
| **Average**         | 0.79| 0.42| 0.48| 0.42| 0.74| 0.70      | 0.26      | 0.59  |
equal ROC-AUC such as FR603 and WofE603, or FR702 and WofE702. These discrepancies were not merely limited to the maps produced by the different methods: even the maps of equal AUC created by an individual method but different random seeds, for example, FR502 versus FR601, or WofE603 versus WofE703, proved unequal in at least 29% of the pixels. As an illustration, Figure 7 shows the discrepancies between some of the maps. These results show that, while ROC curve is used as the main verification tool to compare the reliability of the LSMs, even having an equal ROC-AUC cannot guarantee that the LSMs predict the future landslides equally. In fact, AUC of the ROC graph provides a rough estimate of the general performance of the models. To substantiate this statement, two ROC curves (A and B) are given in Figure 8 as an illustration. In this hypothetical, exaggerated example, while different shapes of the curves indicate different geographical patterns of the susceptible zones in two maps, they both have an equal AUC, exactly 0.77, owing to the fact that the big AUCs between some thresholds counteract the small AUCs between the other ones. Therefore, in the case of landslide susceptibility models, where the geographical prediction of events is important, ROC method should not be completely trusted for comparing the validity of the models. This matter is the case notably when the created maps differ from each other by merely a few per cent in terms of AUC. It means that not all of the maps with a slightly bigger AUC than that of others necessarily conform to the correct pattern of landslide-prone areas.

Generally, many variables that might lead to different LSMs for the same area can be mentioned, for example, the employed landslide inventory and the way it is partitioned as well as the partitioning ratio, combination of the causative factors and their scale and quality, and the applied prediction methods. Part of the LSMs discrepancy in this study was related to the drawback of the ‘equal are classes’ technique (described in Section 3.2.3), but it was responsible for only up to 1% of the differences that does not justify the great disparity between the LSMs. Therefore, other possible sources of uncertainty were responsible for that differences: changing the predictive methods and the random seeds of landslide data-set partitioning for modelling. Regarding the fact that all the other conditions of modelling were the same in this study, it is considerable that even applying two very similar
Figure 7. Example of difference between the same parts of the LSMs of FR502 (a) and WofE502 (b) shown in the difference image (c), and between the LSMs of FR502 (d) and FR601 (e) with associated difference image (f).
methods (FR and WofE) or using different random seeds for landslide inventory partitioning can lead to around 30% differences between the maps. It should thus be noted in the assessments that the slightest possible variations in different factors may lead to great changes in the results, while ROC method is unable to measure all of them. Therefore, it seems crucial to use an index like Cohen’s kappa, besides the ROC curve, for revealing the LSMs differences and, in addition, integrate the created LSMs for obtaining a more reliable spatial prediction pattern if they were discrepant.

In this study, we merged all the 18 produced LSMs (each one consists of five susceptibility zones with the same area but different patterns) by means of the overlaying statistical functions of Maximum, Majority, and Minimum (described in Section 3.2.6). The resultant maps are shown in Figure 9 and the relevant information (the percentages of the landslides and the area of each zone) are given in Table 3. Although their ROC-AUC by an equal value of 80% (all landslides were used for calculation) were a few per cent lower than that of some single maps, the LSMs produced in this

![Figure 8](image)

*Figure 8. An illustration of two ROC curves with equal AUC but different shapes.*

![Figure 9](image)

*Figure 9. Produced LSMs by the overlaying statistical functions of Maximum (a), Majority (b), and Minimum (c).*
way are more reliable in terms of spatial prediction pattern. These maps can be suitable in different situations as described in Section 3.2.6. It is seen from Table 3 that Maximum LSM accommodates above 87% of the landslides in the high and very high susceptible zones; however, the area of these zones together is more than 38% of the whole area that appears to be a large landslide-prone area for such a large catchment like Qaemshahr (990 km²). The LSM produced by Majority function encompasses over 69% of the landslides in its high and very high susceptible zones, and regarding the slight changes it made to the considered area for the zones, it was judged the most reasonable option in this study. Nevertheless, the Minimum LSM is also a good option based on which the landslide hazard and risk zoning can be done to show the area in urgent need of slope stability measures. The very high susceptible zone of this map encircled a large number of landslides (more than 40%) in only about 3.5% of the study area (equivalent to about 35 km²). It should be noted that this 3.5% of the area shows those pixels that were predicted to be the most susceptible ones by, without exception, all the 18 LSMs and, therefore, are of paramount importance in further assessments.

5. Summary and conclusion

This research was accomplished to demonstrate the specific uncertainty of LSMs in terms of spatial prediction pattern and the ineffectiveness of ROC method in revealing that uncertainty. In this regard, by applying two different methods (FR and WofE) and different scenarios of landslide inventory partitioning, different models were constructed producing 18 separate LSMs. All pairs of created LSMs showed very close to equal reliability rates by ROC method, while they were very different in terms of spatial prediction pattern, as evidenced by the results of Cohen’s kappa index. Therefore, it is concluded that the same ROC results of the maps by no means show their equality. In other words, there is a spatial-dependent uncertainty in the prediction maps that ROC method is not able to reveal. In fact, ROC curve could merely estimate a general validity for the produced LSMs; therefore, it is by no means certain that a map with a ROC-AUC a few per cent lower than that of the other maps is less reliable in predicting the future landslides. It is recommended that an additional validation statistic like Cohen’s kappa index that can measure the agreement between maps should be calculated besides ROC curve. Consequently, if the quantity of disagreement between the maps was high, they are better to be combined using one of the overlaying statistical functions like Maximum, Majority, and Minimum (depending on different conditions) so as to ensure that the mentioned uncertainty decreases. Furthermore, landslide susceptibility models could be repeated several times, using different random landslides inventory partitioning or applying different methods, before integration in order to extract the most reliable spatial prediction pattern. Therefore, the resultant map of the merged LSMs can be employed more reliably as the basis for land-use planning and slope stability measures. At the end of this paper, it is suggested that carrying out similar studies at different scales and by using other methods may be helpful.

Disclosure statement

No potential conflict of interest was reported by the authors.
References

Agterberg FP, Bonham-Carter GF, Wright DF. 1990. Statistical pattern integration for mineral exploration. In: Gabor G, Merriam DF, editors. Computer applications in resource estimation prediction and assessment for metals and petroleum. 1st ed. Oxford: Pergamon Press; p. 1–21.

Akgun A. 2012. A comparison of landslide susceptibility maps produced by logistic regression, multi-criteria decision, and likelihood ratio methods: a case study at Izmir, Turkey. Landslides. 9(1):93–106.

Akgun A, Dag S, Bulut F. 2008. Landslide susceptibility mapping for a landslide-prone area (Findikli, NE of Turkey) by likelihood-frequency ratio and weighted linear combination models. Environ Geol. 54(6):1127–1143.

Aleotti P, Chowdhury R. 1999. Landslide hazard assessment: summary review and new perspectives. Bull Eng Geol Environ. 58(1):21–44.

Barnard PL, Owen LA, Sharma MC, Finkel RC. 2001. Natural and human-induced landsliding in the Garwhal Himalaya of northern India. Geomorphology. 40:21–35.

Beguería S. 2006. Validation and evaluation of predictive models in hazard assessment and risk management. Nat Hazards. 37(3):315–329.

Bonham-Carter GE. 1994. Geographic information systems for geoscientists: modelling with GIS. 1st ed. Ottawa: Pergamon Press.

Borga M, Tonelli F, dalla Fontana G, Cazorzi F. 2005. Evaluating the influence of forest roads on shallow landsliding. Ecol Modelling. 187(1):85–98.

Carrara A, Cardinali M, Detti R, Guzzetti F, Pasqui V, Reichenbach P. 1991. GIS techniques and statistical models in evaluating landslide hazard. Earth Surf Processes Land. 16(5):427–445.

Choi J, Oh H-J, Lee H-J, Lee C, Lee S. 2012. Combining landslide susceptibility maps obtained from frequency ratio, logistic regression, and artificial neural network models using ASTER images and GIS. Eng Geol. 124:12–23.

Chung C-JF, Fabbri AG. 1999. Probabilistic prediction models for landslide hazard mapping. Photogramm Eng Remote Sens. 65(12):1389–1399.

Chung C-JF, Fabbri AG. 2003. Validation of spatial prediction models for landslide hazard mapping. Nat Hazards. 30(3):451–472.

Cohen J. 1960. A coefficient of agreement for nominal scales. Educ Psychol Meas. 20(1):37–46.

Conforti M, Pascale S, Robustelli G, Sdao F. 2014. Evaluation of prediction capability of the artificial neural networks for mapping landslide susceptibility in the Turbolo River catchment (northern Calabria, Italy). Catena. 113:236–250.

Corominas J, van Westen C, Frattini P, Cascini L, Malet J-P, Fotopoulou S, Catani F, Van Den Eeckhaut M, Mavrouli O, Agliardi F. 2014. Recommendations for the quantitative analysis of landslide risk. Bull Eng Geol Environ. 73(2):209–263.

Dahal RK, Hasegawa S. 2008. Representative rainfall thresholds for landslides in the Nepal Himalaya. Geomorphology. 100(3):429–443.

Dahal RK, Hasegawa S, Nonomura A, Yamanaka M, Dhakal S, Paudyal P. 2008. Predictive modelling of rainfall-induced landslide hazard in the Lesser Himalaya of Nepal based on weights-of-evidence. Geomorphology. 102(3):496–510.

Dai FC, Lee CF. 2002. Landslide characteristics and slope instability modeling using GIS, Lantau Island, Hong Kong. Geomorphology. 42(3):213–228.

Eker R, Aydin A. 2016. Evaluation of forest roads conditions in terms of landslide susceptibility in Göyaka and Karduz Forest Districts (Düzce-Turkey). Eur J Forest Eng. 2(2):54–60.

Fabbri AG, Chung C-J. 2008. On blind tests and spatial prediction models. Nat Resour Res. 17(2):107–118.

Fawcett T. 2006. An introduction to ROC analysis. Pattern Recogn Lett. 27(8):861–874.

Fell R, Corominas J, Bonnard CH, Cascini L, Leroi E, Savage WZ. 2008. Guidelines for landslide susceptibility, hazard and risk zoning for land-use planning. Eng Geol. 102(3–4):85–98.

Fielding AH, Bell JF. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ Conserv. 24(01):38–49.

Glade T, Anderson MG, Crozier MJ. 2006. Landslide hazard and risk. West Sussex: Wiley.

Guzzetti F. 2006. Landslide hazard and risk assessment [dissertation]. Bonn: Universität-sund Landesbibliothek Bonn.

Guzzetti F, Carrara A, Cardinali M, Reichenbach P. 1999. Landslide hazard evaluation: a review of current techniques and their application in a multi-scale study, Central Italy. Geomorphology. 31(1):181–216.

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

Harp EL, Keefer DK, Sato HP, Yagi H. 2011. Landslide inventories: the essential part of seismic landslide hazard analyses. Eng Geol. 122(1):9–21.

Highland L, Bobrowsky PT. 2008. The landslide handbook: a guide to understanding landslides. Virginia: U.S. Geological Survey.

Kamp U, Growley BJ, Khattak GA, Owen LA. 2008. GIS-based landslide susceptibility mapping for the 2005 Kashmir earthquake region. Geomorphology. 101(4):631–642.
Lee S. 2005. Application of logistic regression model and its validation for landslide susceptibility mapping using GIS and remote sensing data. Int J Remote Sens. 26(7):1477–1491.

Lee S, Pradhan B. 2006. Landslide hazard mapping at Selangor, Malaysia using frequency ratio and logistic regression models. Landslides. 4(1):33–41.

Lee S, Ryu J-H, Kim I-S. 2007. Landslide susceptibility analysis and its verification using likelihood ratio, logistic regression, and artificial neural network models: case study of Youngin, Korea. Landslides. 4(4):327–338.

Mohammady M, Pourghasemi HR, 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. 61:221–236.

Nandi A, Shakoor A. 2010. A GIS-based landslide susceptibility evaluation using bivariate and multivariate statistical analyses. Eng Geol. 110(1):11–20.

Neuhäuser B, Damm B, Terhorst B. 2011. GIS-based assessment of landslide susceptibility on the base of the weights-of-evidence model. Landslides. 9(4):511–528.

Neuhäuser B, Terhorst B. 2007. Landslide susceptibility assessment using “weights-of-evidence” applied to a study area at the Jurassic escarpment (SW-Germany). Geomorphology. 86(1):12–24.

Park S, Choi C, Kim B, Kim J. 2012. Landslide susceptibility mapping using frequency ratio, analytic hierarchy process, logistic regression, and artificial neural network methods at the Inje area, Korea. Environ Earth Sci. 68(5):1443–1464.

Pham BT, Pradhan B, Bui DT, Prakash I, Dholakia M. 2016. A comparative study of different machine learning methods for landslide susceptibility assessment: a case study of Uttarakhand area (India). Environ Model Software. 84:240–250.

Pourghasemi HR, Pradhan B, Gokceoglu C. 2012. Application of fuzzy logic and analytical hierarchy process (AHP) to landslide susceptibility mapping at Haraz watershed, Iran. Nat Hazards. 63(2):965–996.

Pourghasemi HR, Pradhan B, Gokceoglu C, Mohammadi M, Moradi HR. 2012. Application of weights-of-evidence and certainty factor models and their comparison in landslide susceptibility mapping at Haraz watershed, Iran. Arab J Geosci. 6(7):2351–2365.

Pourghasemi HR, Pradhan B, Gokceoglu C, Moezzi KD. 2013. A comparative assessment of prediction capabilities of Dempster–Shafer and weights-of-evidence models in landslide susceptibility mapping using GIS. Geomatics Nat Hazards Risk. 4(2):93–118.

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 modeling. Environ Model Software. 25(6):747–759.

Pradhan B, Lee S, Buchroithner MF. 2010. Remote sensing and GIS-based landslide susceptibility analysis and its cross-validation in three test areas using a frequency ratio model. Photogramm Fernerkun. 2010(1):17–32.

Pradhan B, Oh H-J, Buchroithner MF. 2010. Weights-of-evidence model applied to landslide susceptibility mapping in a tropical hilly area. Geomatics Nat Hazards Risk. 1(3):199–223.

Provost FJ, Fawcett T. 1997. Analysis and visualization of classifier performance: comparison under imprecise class and cost distributions. In: Heckerman D, Mannila H, Pregibon D, Uthurusamy R, editors. Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining (KDD-97); Aug 14; Menlo Park (CA): AAAI Press. p. 43–48.

Provost FJ, Fawcett T, Kohavi R. 1998. The case against accuracy estimation for comparing induction algorithms. In: Shavlik J, editor. Proceedings of the ICML 1998; Jul 24; San Francisco (CA): Morgan Kaufmann. p. 445–453.

Regmi AD, Devkota KC, Yoshida K, Pradhan B, Pourghasemi HR, Kumamoto T, Akgun A. 2013. Application of frequency ratio, statistical index, and weights-of-evidence models and their comparison in landslide susceptibility mapping in Central Nepal Himalaya. Arab J Geosci. 7(2):725–742.

Rozos D, Pyrgiotis L, Skias S, Tsagaratos P. 2008. An implementation of rock engineering system for ranking the instability potential of natural slopes in Greek territory. An application in Karditsa County. Landslides. 5(3):261–270.

Sato H, Harp E. 2009. Interpretation of earthquake-induced landslides triggered by the 12 May 2008, M7.9 Wenchuan earthquake in the Beichuan area, Sichuan Province, China using satellite imagery and Google Earth. Landslides. 6(2):153–159.

Sterlacchini S, Ballabio C, Blahut J, Masetti M, Sorichetta A. 2011. Spatial agreement of predicted patterns in landslide susceptibility maps. Geomorphology. 125(1):51–61.

Tsagaratos P, Ilia I, Hong H, Chen W, Xu C. 2017. Applying Information Theory and GIS-based quantitative methods to produce landslide susceptibility maps in Nancheng County, China. Landslides. 14(3):1091–1111.

Vakhshoori V, Zare M. 2016. Landslide susceptibility mapping by comparing weight of evidence, fuzzy logic, and frequency ratio methods. Geomatics Nat Hazards Risk. 7(5):1731–1752.

van Westen CJ, Castellanos E, Kuriakose SL. 2008. Spatial data for landslide susceptibility, hazard, and vulnerability assessment: an overview. Eng Geol. 102(3):112–131.

van Westen C, van Asch TW, Soeters R. 2005. Landslide hazard and risk zonation – why is it still so difficult? Bull Eng Geol Environ. 65(2):167–184.
Varnes DJ. 1984. Landslide hazard zonation: a review of principles and practice. Paris: UNESCO.
Wang F, Xu P, Wang C, Wang N, Jiang N. 2017. Application of a GIS-based slope unit method for landslide susceptibility mapping along the Longzi River, Southeastern Tibetan Plateau, China. ISPRS Int J Geoinf. 6(6):172.
Yilmaz I. 2009. Landslide susceptibility mapping using frequency ratio, logistic regression, artificial neural networks and their comparison: a case study from Kat landslides (Tokat—Turkey). Comput Geosci. 35(6):1125–1138.
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.
Zare M, Pourghasemi HR, Vafakhah M, Pradhan B. 2012. Landslide susceptibility mapping at Vaz Watershed (Iran) using an artificial neural network model: a comparison between multilayer perceptron (MLP) and radial basic function (RBF) algorithms. Arab J Geosci. 6(8):2873–2888.
Zézere J, Pereira S, Melo R, Oliveira S, Garcia R. 2017. Mapping landslide susceptibility using data-driven methods. Sci Total Environ. 589:250–267.