Environ Earth Sci (2015) 73:937–949
DOI 10.1007/s12665-014-3442-z

ORIGINAL ARTICLE

Using maximum entropy modeling for landslide susceptibility mapping with multiple geoenvironmental data sets

No-Wook Park

Received: 1 May 2013 / Accepted: 13 June 2014 / Published online: 2 July 2014

Abstract The potential of using maximum entropy modeling for landslide susceptibility mapping is investigated in this paper. Although the maximum entropy model has been applied widely to species distribution modeling in ecology, its applicability to other kinds of predictive modeling such as landslide susceptibility mapping has not yet been investigated fully. In the present case study of Boeun in Korea, multiple environmental factors including continuous and categorical data were used as inputs for maximum entropy modeling. From the optimal setting test based on cross-validation, the effective feature type for continuous data representation was found to be a hinge feature and its combination with categorical data showed the best predictive performance. Factor contribution analysis indicated that distances from lineaments and slope layers were the most influential factors. From interpretations on a response curve, steeply sloping and weathered areas that consisted of excessively drained granite residuum soils were very susceptible to landslides. Predictive performance of maximum entropy modeling was slightly better than that of a logistic regression model which has been used widely to assess landslide susceptibility. Therefore, maximum entropy modeling is shown to be an effective prediction model for landslide susceptibility mapping.

Keywords Landslide · Maximum entropy · Validation · Prediction

N.-W. Park (✉)
Department of Geoinformatic Engineering, Inha University,
100 Inha-ro, Nam-gu, Incheon 402-751, Korea
e-mail: nwpark@inha.ac.kr

Introduction

Landslides are a major geological disaster causing loss of life and serious economic damage to facilities (Lee and Min 2001; Park 2011; Akgun 2012). Landslides triggered by heavy rain are a major problem in Korea and the frequency of landslide occurrences is expected to increase due to climate change (Chae et al. 2009). Mapping or predicting areas that are susceptible to future landslides is important for preventing landslide damage and future land-use planning.

Landslide susceptibility mapping requires both geoenvironmental factors related to landslide occurrence and quantitative prediction models. Owing to recent advances in computer resources, geographic information systems (GIS) have been used widely for the preparation and management of those factors. Many quantitative models have been proposed for integrating the causal factors and applied to landslide susceptibility mapping. As probabilistic models, the likelihood ratio (also called the frequency ratio) and weights of evidence models have been applied frequently due to their simplicity and easy links to GIS operators (Chung and Fabbri 1999; Lee and Min 2001; Lee et al. 2004; Lee and Sambath 2006). Logistic regression has also been applied widely to map landslide susceptibility (Atkinson and Massari 1998; Lee and Min 2001; Dai and Lee 2002; Lee 2005; Lee and Sambath 2006; Greco et al. 2007; Akgun 2012). Other models include evidential belief functions (Ghosh and Carranza 2010; Park 2011; Althuwaynee et al. 2012; Lee et al. 2012), fuzzy set theory (Ercanoglue and Gokceoglu 2002; Park et al. 2003), artificial neural networks (Lee 2007; Choi et al. 2010), and support vector machines (Yao et al. 2008; Ballabio and Sterlacchini 2012).

Such models applied to landslide susceptibility mapping have also been used for different geological predictive modeling tasks, such as mineral potential mapping (Carranza and Hale 2003; Porwal et al. 2004) and ground subsidence mapping (Kim et al. 2006). In addition to these fields, another active research field for which predictive models have been developed and applied is species distribution modeling in ecology. Various statistical and machine learning models have been applied to predict species distributions (Franklin 2009). Multivariate statistical models applied frequently to species distribution modeling are generalized linear and generalized additive models (Austin 2002; Guisan et al. 2002; Lehmann et al. 2002). Species distribution modeling has also been performed using machine learning algorithms, such as boosted regression trees (Leathwick et al. 2006) and random forests (Prasad et al. 2006). Species distribution modeling is very similar to landslide susceptibility mapping, in that known occurrences are used to model the target distribution and multiple environmental variables are involved for the modeling procedure. The above models that have been applied to species distribution modeling require both presence and absence data for modeling. If true absence data are not available, "pseudo-absence" data are generated and used as an alternative. Absence and pseudo-absence data in species distribution modeling correspond to "stable" and "pseudo-stable" data in landslide susceptibility mapping, respectively, which are not affected by past landslides. As discussed by Phillips et al. (2009) and Van Der Wal et al. (2009), pseudo-absence data should be used with caution because they affect directly the modeling results, and reliable pseudo-absence data are not always available.

Regarding this issue, a maximum entropy model that provides a general way for estimating the unknown target probability distribution was proposed to predict species distributions from presence-only data (Phillips et al. 2006). This has been used widely for species distribution modeling (Ward 2007; Wollan et al. 2008; Pineda and Lobo 2009; Tinoco et al. 2009) with competitive performance in some case studies, compared with other models using both presence and absence data (Elith et al. 2006; Elith and Graham 2009).

Landslide susceptibility mapping can be regarded as predictive modeling with presence-only data, because the only data available are past landslide occurrences. Despite its promising potential for the assessment of landslide susceptibility, however, the maximum entropy model has not been investigated and applied fully. Very few studies based on maximum entropy modeling have been reported for landslide susceptibility mapping (Felicísimo et al. 2012; Vorpahl et al. 2012). Furthermore, these studies focused solely on its comparison with other models, without a complete investigation of either the theoretical or practical aspects of the application of maximum entropy modeling to landslide susceptibility analysis.

The main objective of this paper is to investigate the potential and applicability of maximum entropy modeling to landslide susceptibility mapping. Following a brief overview of maximum entropy modeling, specific analysis steps for landslide susceptibility mapping are presented. A case study for the Boeun area in Korea is examined to illustrate the main objective.

Study area and data sets

The Boeun area, which suffered considerable landslide damage following heavy rain in August 1998, was chosen as the case study area (Fig. 1). Precipitation values measured at rain gauge stations in the study area between August 11 and August 12 ranged from 390 to 409 mm/day (Kim et al. 2000). Similar values and patterns of precipitation were observed throughout the study area and therefore, precipitation data were not considered in this study. The fact that some areas were susceptible to landslides when others were not, under similar rainfall conditions, implies that there are some causal factors related to landslide occurrence.

The geology around the Boeun area including the study area is covered mainly with meta-sediments of the age-unknown Ogcheon group, accompanied with a small exposure of the Paleozoic system and several igneous intrusive bodies (Kim et al. 1977). The Boeun granite, especially biotite granite, is distributed widely throughout the study area and there are two mica adamellite areas in northern and central parts. A few acidic dykes are intruded in the study area and a composite mass of quartz porphyry and felsite, trending N–S, is exposed linearly in the Boeun granite.

This case study is based on a GIS-based database, including landslide locations and several causal factors relevant to landslide occurrence. Past landslide locations were detected using change detection analysis of aerial photographs taken in 1996 and 1999, aided by field verification (Lee et al. 2004). A total of 481 landside scars were detected and the topographically highest scarps were used as landslide triggers or occurrences in this case study (Fig. 1). The main type of landslide that occurred in the study area was a rainfall-triggered debris flow (Kim et al. 2000; Lee et al. 2004). The geology of the study area consists mainly of granite and therefore, the landslide type is related to the coarse-grained granite residuum that is distributed widely throughout the study area (Kim et al. 2000).

Three continuous data layers and three categorical data layers (Table 1; Fig. 2) were chosen as causal factors by

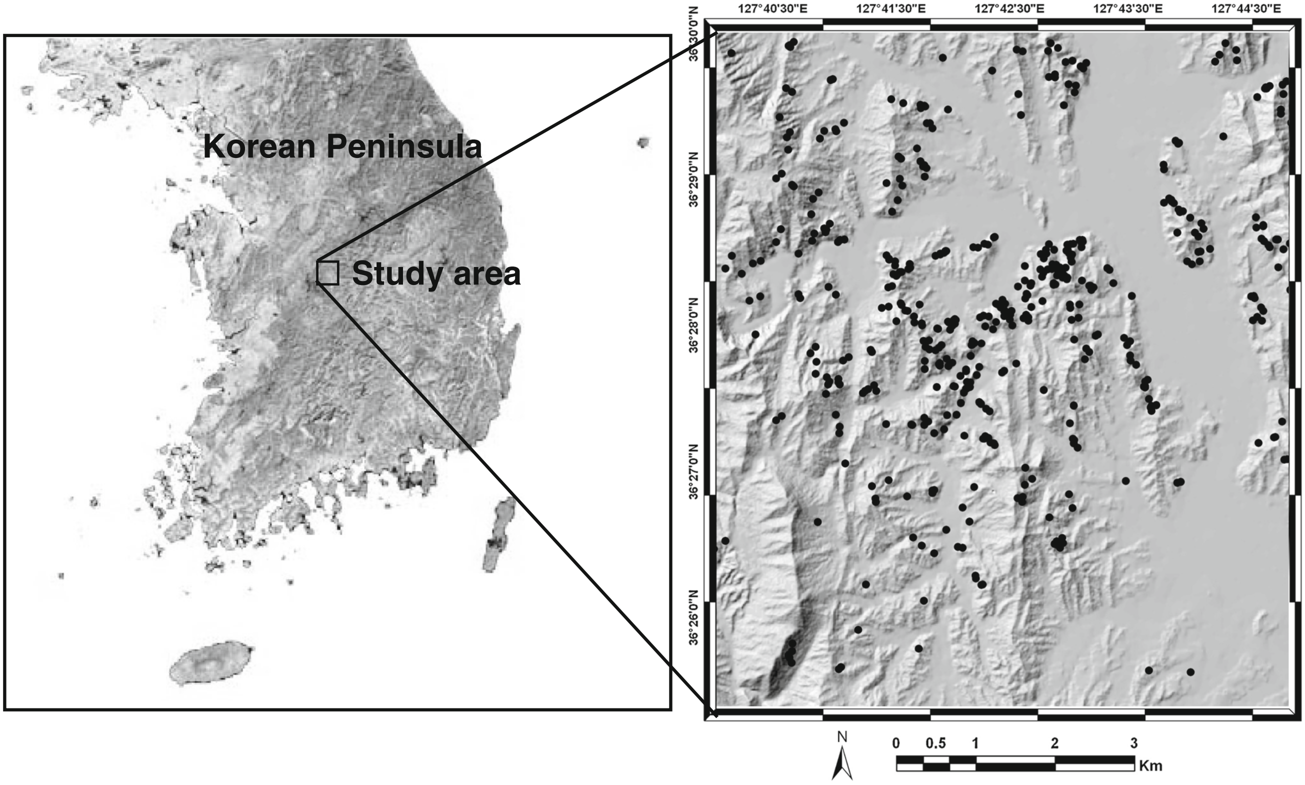

Fig. 1 Location of the study area and landslide inventory map. Landslide locations are denoted by *black dots* and the *background is a shaded* relief map

Table 1 Description of the data sets used in this case study

Factor	Source (scale)
Elevation	Digital topographic map (1:5,000)
Slope	Digital topographic map (1:5,000)
Distance from lineaments	Aerial photos and visual interpretation
Forest type	Digital forest map (1:25,000)
Soil drainage	Digital soil map (1:25,000)
Lithology	Digital geology map (1:50,000)

considering previous work undertaken in the study area (Park et al. 2003). To generate topographic data layers, a digital elevation model (DEM) was made from 1:5,000 scale digital topographic maps. Elevation and slope layers in a raster format were then extracted from the DEM. From lineaments extracted using remote sensing imagery, distances from lineaments were computed as continuous scale data. In Lee et al. (2004), the distances were treated as categorical data with buffer zones of 50 m intervals. These continuous scale data were generated because the selection of the buffering interval in the previous study was somewhat subjective and the categorization procedure resulted inevitably in loss of information.

Among the various attributes provided by the 1:25,000 scale digital forest and soil maps, the forest type and the soil drainage, which were the most influential factors reported by Park et al. (2003), were used for the forest and soil layers, respectively. As other categorical data, lithology information was extracted from a 1:50,000 scale digital geological map. In Korea, a 1:50,000 scale geological map is provided as the finest scale map and thus, only the overall lithology attributes of the study area are available. By considering the different scales of the original input data (Table 1), all data layers consisted of 290,068 pixels with a 15 m spatial resolution. Thus, the study area encompasses an area of about 65 km^2.

Basic principle of maximum entropy modeling

The maximum entropy principle was based originally on statistical mechanics and information theory, according to the concept that the best approximation of an unknown probability distribution is one with maximum entropy subject to certain known constraints (Jaynes 1957; Sivia and Skilling 2006).

Phillips et al. (2006) proposed the maximum entropy model specifically designed for species distribution

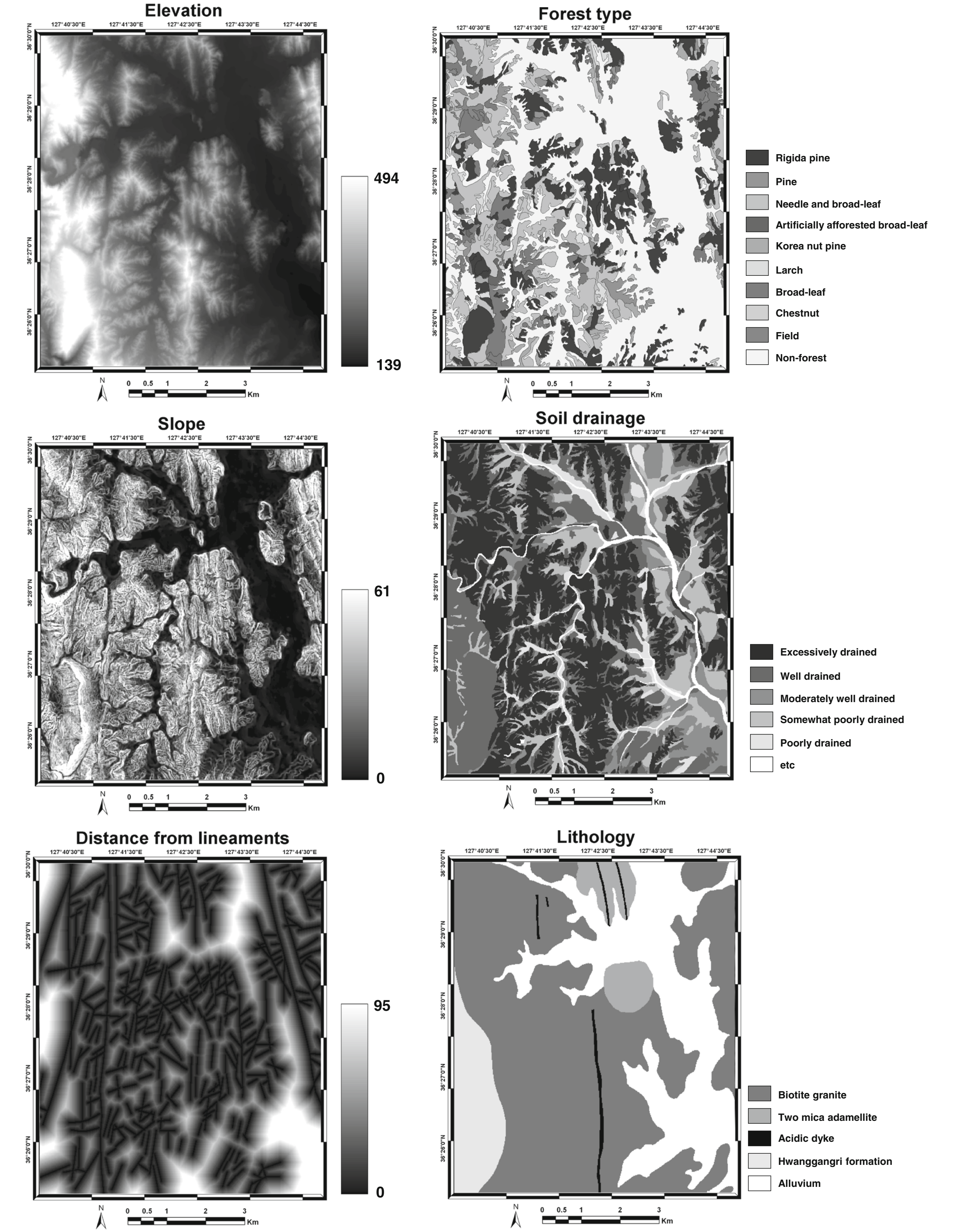

Fig. 2 Environmental factors used for maximum entropy modeling

modeling, when only presence data are available for modeling. The goal of maximum entropy modeling is to find the probability distribution (π) of target occurrences over the set locations $\mathbf{X}$ within the study area. Causal factors or features are used to define the moment constraints on the distribution π. The moment, such as the mean, is defined from the values of the causal factors at all presence locations. For example, the expected slope value of the estimated distribution should be close to the average slope value at all presence locations. There may be many possible distributions that satisfy the above constraints. By applying the maximum entropy principle, the most uniform distribution is chosen from among these many possible distributions (Phillips and Dudík 2008).

In this paper, only the salient aspects of the maximum entropy model for predictive modeling are given, synthesized from Phillips et al. (2006), Phillips and Dudík (2008), and Elith et al. (2011). Let $\mathbf{x}$ denotes a random site over the study area and $\pi(\mathbf{x})$ be the target probability distribution value at each location $\mathbf{x}$, which is non-negative and sums to one. If $y = 1$ denotes the target occurrence, $\pi(\mathbf{x})$ can be regarded as the probability that is found at location $\mathbf{x}$, given the target is present, as $P(\mathbf{x}|y = 1)$. The probability that the target is present at location $\mathbf{x}$, denoted as $P(y = 1|\mathbf{x})$, is expressed using $P(\mathbf{x}|y = 1)$ by applying Bayes' rule, as shown:

$$P(y = 1|\mathbf{x}) = \frac{P(y = 1)P(\mathbf{x}|y = 1)}{P(\mathbf{x})} = \frac{P(y = 1)\pi(\mathbf{x})}{1/|\mathbf{X}|} \qquad (1)$$

where $P(y = 1)$ is the prevalence of target occurrences and $|\mathbf{X}|$ is the number pixels or locations over the study area. $P(y = 1)$ cannot be determined exactly from occurrence-only data; therefore, in maximum entropy modeling, $\pi(\mathbf{x})$ in Eq. (1) is modeled using occurrence data, instead of directly estimating $P(y = 1|\mathbf{x})$.

As discussed in Phillips et al. (2006), $\pi(\mathbf{x})$ estimated by the maximum entropy principle is equal to a Gibbs probability distribution expressed as an exponential distribution. If n features ($f_i, i = 1, \ldots, n$) are considered, then the Gibbs probability distribution is defined as:

$$q_\lambda(\mathbf{x}) = \frac{\exp\left(\sum_{i=1}^{n} \lambda_i f_i(\mathbf{x})\right)}{Z_\lambda} \qquad (2)$$

where Z_λ is a normalization constant that ensures $q_\lambda(\mathbf{x})$ sums to one and λ is the vector of weights assigned to the features.

During the estimation of $q_\lambda(\mathbf{x})$, maximum entropy modeling tries to find the distribution closest to the constraints using l_1 regularization to avoid overfitting. Therefore, maximum entropy modeling aims to find the Gibbs distribution that maximizes the penalized log likelihood. If there are m occurrences in the study area, the difference between log likelihood and regularization, which should be maximized, is expressed as (Phillips and Dudík 2008):

$$\frac{1}{m}\sum_{i=1}^{m} \ln(q_\lambda(\mathbf{x}_i)) - \sum_{j=1}^{n} \beta_j|\lambda_j| \qquad (3)$$

where β_j is the regularization parameter for the jth feature f_j.

The first term in Eq. (3) is a log likelihood, which gets larger as the fit of the model to the data improves. The second term in Eq. (3) is used for regularization. Consequently, maximum entropy modeling finds the Gibbs distribution that not only fits the occurrence data well, but also generalizes well (Elith et al. 2011).

Procedures for landslide susceptibility mapping

All the modeling steps applied in this study are shown in Fig. 3. As with other machine learning models, the maximum entropy model requires some optimization procedures. Thus, before the generation of the landslide susceptibility map, optimal settings are first searched by predictive performance measures based on cross-validation. Using optimal settings, the landslide susceptibility map over the study area is generated, and the relationships between the input causal factors and landslide susceptibility are interpreted. For comparison purposes, the predictive performance of the maximum entropy model will be compared with that of logistic regression.

Optimal setting search

Tests on optimal settings for modeling focus on two aspects that affect the predictive performance and processing time significantly: the best feature selection for continuous data representation and the number of background samples. Categorical data are used directly as their original types in maximum entropy modeling. For continuous data representation, however, the maximum entropy model uses features that are a set of transformations of the original continuous factors. Consequently, when using continuous data for maximum entropy modeling, a greater number of features than input continuous factors are used. The selection of proper feature types is important in terms of both model performance and processing time. As shown in Eq. (3), using too many features for continuous data representation tends to increase the complexity of the target model, and additional regularization is necessary to control the effects of the greater complexity of features. Thus, the present study compares the resulting performance of various feature combinations, especially for continuous data

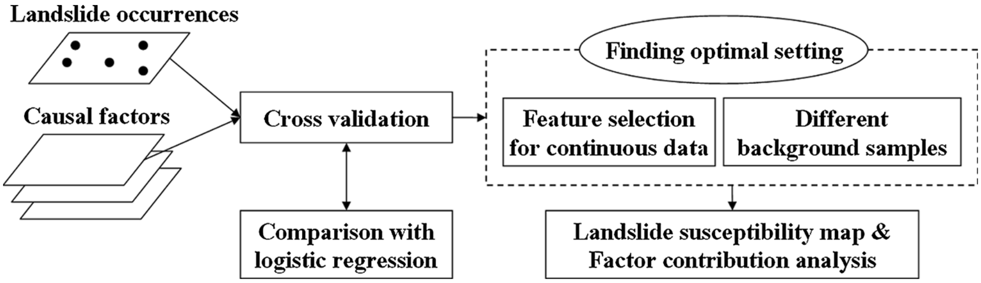
Landslide occurrences
Causal factors
Cross validation
Finding optimal setting
Feature selection for continuous data
Different background samples
Comparison with logistic regression
Landslide susceptibility map & Factor contribution analysis

Fig. 3 Schematic diagram of the processing flow adopted in this study

representation. The considered feature types are linear, product, quadratic, and hinge. The linear feature is the original continuous factor itself. The product and quadratic features are the products of any possible two continuous factors and the square of any factor, respectively. The hinge feature is a linear feature truncated at a given threshold (Phillips and Dudík 2008).

Background data or samples are used intrinsically during the modeling procedure because only past landslide occurrences are available. The background data, which are defined as all locations or random samples within the study area, are used to compare the differences between the probability of the presence sites and that of the other sites (i.e., background locations) (Phillips and Dudík 2008; Elith et al. 2011). If the numbers of background samples used for modeling are too small, the proper discrimination of the characteristics at landslide occurrence sites from those at background sites may be failed. Conversely, using too many background samples (e.g., the full data set from the entire study area) requires unnecessary processing time. In this study, predictive performance is compared according to varying numbers of background samples for landslide susceptibility mapping. The following nine different background sizes are considered for modeling: 481, 1,000, 2,500, 5,000, 10,000, 20,000, 40,000, 80,000, 160,000, and 290,000. A total of 481 background samples, corresponding to the number of past landslide occurrences, are chosen first and then a geometric progression of background sizes considered. The final size considered of 290,000 is similar to the total size of the study area (290,068).

Factor contribution analysis and generation of landslide susceptibility maps

Once the optimal settings have been determined, the landslide susceptibility map is generated and interpretation of the results is performed. In addition, a quantitative analysis of factor contribution on susceptibility mapping is also carried out. A jackknife technique is used to estimate the factor contribution to the modeling. In this approach, each factor in turn is excluded intentionally and a model constructed using the remaining factors. Then, the predictive performance from the model created using all factors is compared with that of the model constructed using the remaining factors. Consequently, the contribution of the excluded factor can be examined. A response curve is also used to derive relationships between each causal factor and the prediction modeling.

Comparison with other model and cross-validation

The final step is to compare the predictive performance of the maximum entropy model with that of the conventional model. Logistic regression, which has been used widely for assessing landslide susceptibility, is chosen for this comparison. For a fair comparison, the same background samples used for the maximum entropy modeling are also used as pseudo-absence data for logistic regression.

For all predictive performance comparisons, fivefold cross-validation is applied to restrict the use of landslide occurrences. All landslide occurrences are divided randomly into five groups with an equal number of landslide occurrences (about 96 occurrences). Each group is set aside to evaluate the maximum entropy model constructed using the remaining 80 % of all landslide occurrences (about 385 occurrences). This procedure is repeated five times by changing the validation group. In each validation procedure, the predicted values over the entire study area are sorted in ascending order, and then the relative rank values at the validation locations are recorded. Thus, through this approach, the relative susceptibility rank values at all landslide locations are obtained. These values are then used to compute the cumulative portion of the landslide occurrences within each relative susceptibility level. After constructing the prediction rate curves (Chung and Fabbri 2003), the area under the curve (AUC) values is computed using the trapezoidal method and used for predictive performance comparisons.

Optimal setting search results

Implementation of maximum entropy modeling was done using the Maxent software (version 3.3.3k), but entire validation procedures, such as the construction of prediction rate

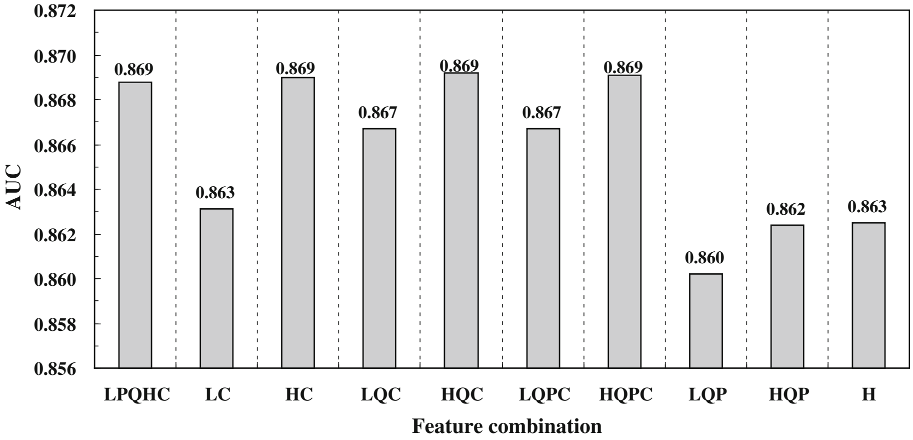

Fig. 4 Comparison of predictive performance (AUC values) for various feature combinations (*L* linear, *P* product, *Q* quadratic, *H* hinge, *C* categorical data). The number above each *bar* denotes the AUC value

curves and the computation of AUC values, were implemented using Fortran programming.

First, the change of predictive performance according to the change of feature types for continuous data was tested. The resultant AUC values from cross-validation are shown in Fig. 4. The AUC values from the combination of complex features for continuous data with categorical data (LPQHC, HQC, and HQPC in Fig. 4) were the highest among the various feature combinations. However, the combination of hinge features for continuous data with categorical data also generated the same best predictive performance. Even though many complex feature types for continuous data were used for maximum entropy modeling, non-zero coefficient values were assigned to only a few feature types through regularization, as shown in Eq. (3). The linear features, which are the original continuous data, gave the worst results, when combined with categorical data. The hinge features, which are basis functions for piecewise linear splines, are very similar to the nonlinear smooth functions in generalized additive models (Elith et al. 2011). Thus, the superiority of hinge features implies that there are nonlinear relationships between the continuous data and landslide occurrences. The above characteristics from the hinge features enabled the proper modeling of the nonlinear relationships between the continuous data and landslide susceptibility. Consequently, the use of only hinge features for continuous data produced the best predictive performance when combined with categorical data. If both complex and simpler models show similar performances, simpler models are generally preferred from a modeling viewpoint. Therefore, the above combination is the best for the data sets in the study area. The predictive performance using only features for continuous data without categorical data (LQP, HQP, and H in Fig. 4) gave the worst results in this case study, indicating that categorical data such as the forest type should be combined with continuous data for landslide susceptibility mapping. From these test results, the combination of the hinge features of continuous data with categorical data was used for subsequent modeling procedures.

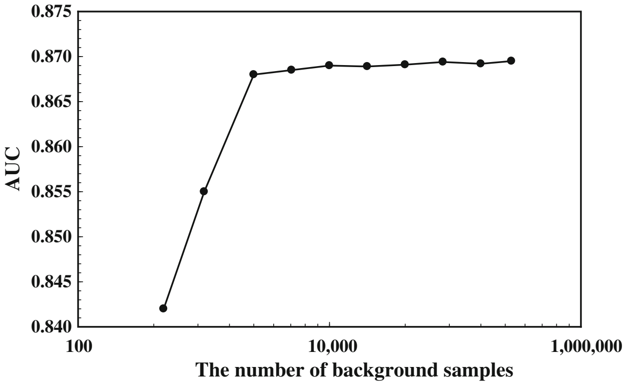

Fig. 5 Comparison of predictive performance (AUC values) for varying numbers of background samples

As the next step, the effect of the number of background samples was tested. As shown in Fig. 5, the background sample test indicated that more than 10,000 background samples (3.45 % of the entire study area) produced a similarly high predictive performance. The smallest background sample (i.e., 481) produced the worst prediction performance, which means that if the number of background samples is too small, they cannot represent the background environment accurately for comparison with the characteristics at landslide locations. In terms of processing time, the use of 10,000 background samples was the best choice in this case study.

Factor contribution analysis results

Based on the optimal setting search tests, landslide susceptibility analysis was performed using the combination of hinge features for continuous data and categorical data with 10,000 randomly chosen background samples. Before

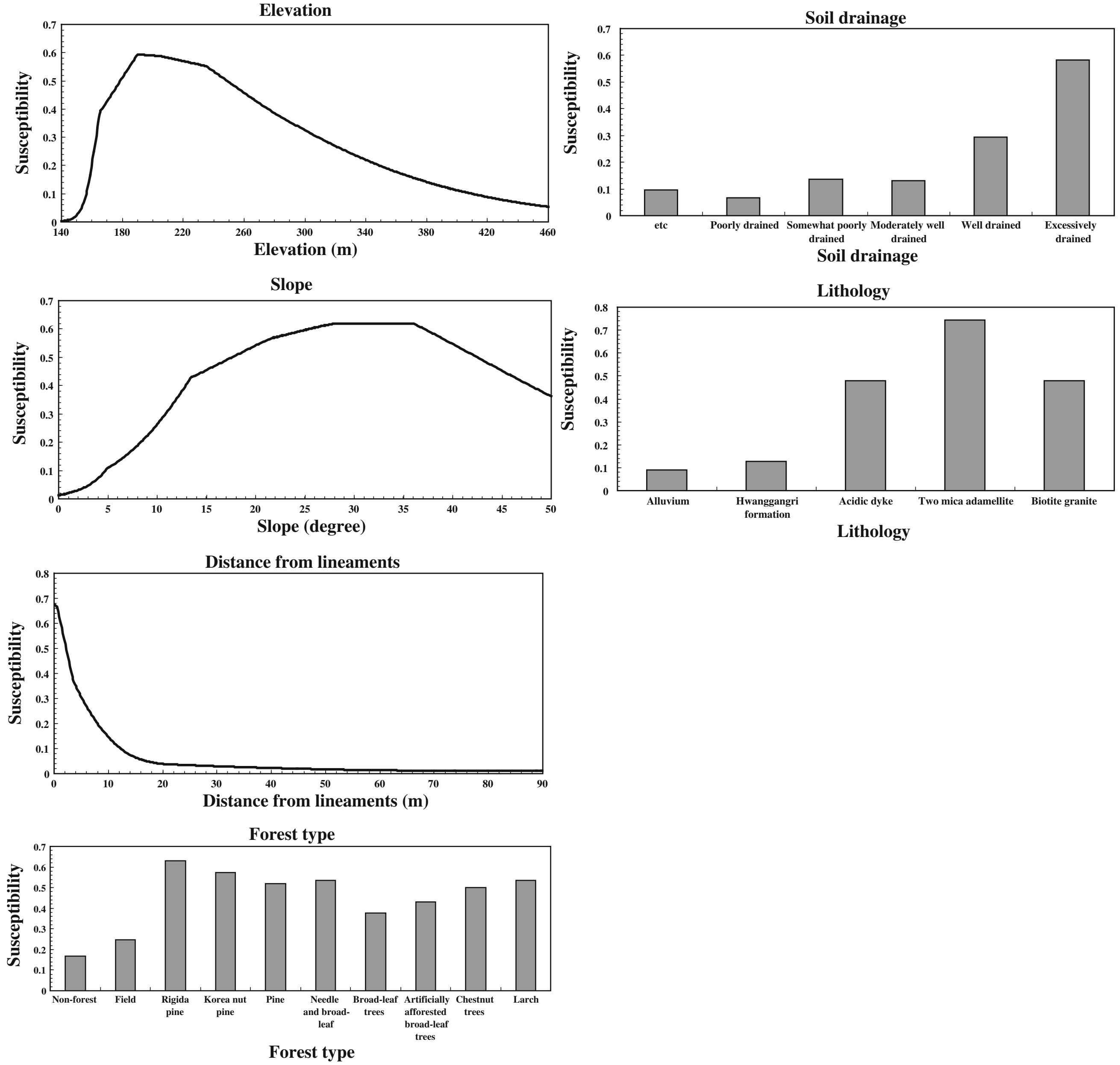

Fig. 6 Response curves for each factor

generating the susceptibility map, the manner in which each environmental factor affected the prediction result was investigated based on a response curve. The response curve shows the changes of the modeling output within the range or value for the factor. This curve was generated using only the considered factor.

Figure 6 gives the response curves for six environmental factors used for landslide susceptibility mapping. The relationships between landslide occurrence and topographic factors are as follows. In the elevation map, most landslides occurred in the range of elevation between 180 and 260 m, in which most mountain areas are located. However, landslide susceptibility decreased in the highest areas with few surficial deposits. Landslide susceptibility increased with increasing slope angle, as expected. With an increase of slope angle, the shear stress in soil or unconsolidated material generally increases as well. However, the decrease of susceptibility for slope values in excess of 35° is related to the reduction of surficial deposits in those areas.

In the case of the forest type, needle-leaf trees, such as Korea nut pine, Rigida pine, and pine, exhibited relatively higher susceptibility values. The root systems of those tree types are relatively less extensive than those of broad-leaf trees. Therefore, areas covered with those trees are much more susceptible to landslides.

As for soil drainage, landslide susceptibility increased in accordance with improved drainage. This result is in agreement with other Korean case studies, such as Lee and Min (2001) and Lee et al. (2004). When there is heavy rain, well-drained soils can control the water flow and thus, contain more water. The soil materials of the excessively drained areas were mainly granite residuum, which consists of rocky sandy loam and sandy loam. These materials have relatively coarse grains; thus, during heavy rain, the soil can contain more water because of the additional space between the grains. The characteristics of these soil layers are also related to the geology of the study area, being mainly granite areas.

In the lithology map, most landslides occurred in the two mica adamellite areas, and granite areas generally exhibited relatively higher susceptibility values. Deep weathering was considerably well progressed in these granite areas and therefore, landslide susceptibility was relatively high. High susceptibility in acidic dykes can also be explained by the fact that the top layer covering the acidic dykes consists of deeply weathered rocks or soils. The highest susceptibility in the two mica adamellite areas was also related to their topography, i.e., erosion basin. With regard to the distance from lineaments, most landslides occurred very close to lineaments owing to an increase in the degree of weathering. From these relationships between geological factors and landslide susceptibility, it is concluded that weathering has a dominant effect of the degree of susceptibility within the study area.

To investigate the factor with the strongest effect on the prediction result, a jackknife-based test was implemented. The test results in Table 2 are summarized as a decrease of AUC values (i.e., loss of performance) by comparing the prediction based on all factors with that when one factor had been excluded intentionally. The larger the decrement, the greater the influence of the excluded factor. The relative decrease of AUC values as a percentage (RD) was also computed to quantify the factor contribution as:

$$\mathrm{RD}_i = 100 \cdot \frac{[\mathrm{AUC}_{\mathrm{all}} - \mathrm{AUC}_i]}{\mathrm{AUC}_{\mathrm{all}}} \qquad (4)$$

where $\mathrm{AUC}_{\mathrm{all}}$ and AUC_i denote the AUC values computed from the prediction using all factors and the prediction when the i-th factor has been excluded, respectively.

The most influential factor was the distance from lineaments, which afforded the largest decrease of AUC values (RD = about 4.4 %) when excluded in the predictive modeling. This result can be explained from the response curve of that factor in Fig. 6. With increasing distances from the lineaments, the susceptibility values decreased drastically and constant values were reached at a distance greater than about 20 m. Therefore, locations very close to lineaments with large susceptibility values could be separated from other locations, and the greatest contribution to prediction could be obtained. Overall, the contributions of three continuous data were strong, but those of the three categorical data sets were relatively very weak. The forest type was the most influential factor among the three categorical data sets and the next dominant factor was lithology. These results can be explained by the proportion of classes in the categorical data layers. The soil drainage and lithology layers have relatively small numbers of classes, which means that they only provide overall patterns of soil drainage and lithology classes within the study area. For example, well-drained or excessively drained soils consist mainly of granite residuum originated from granite, which occupy large portions of the study area. Therefore, these two categorical data sets provided overall information, such that weathered granite areas with excessively drained soils are susceptible to landslides. Conversely, the forest-type map includes 11 classes and therefore, some forest types with high susceptibility in small areas could be separated from other types. Consequently, a relatively higher contribution for prediction was observed among the three categorical data layers.

Table 2 Jackknife test results in terms of the decrease of AUC values when each factor is excluded in maximum entropy modeling

Factor	Decrease of AUC	Relative decrease of AUC (%)
Elevation	0.0076	0.875
Slope	0.0218	2.509
Distance from lineaments	0.0382	4.396
Forest type	0.0048	0.552
Soil drainage	0.0002	0.023
Lithology	0.0009	0.104

However, these lesser contributions of categorical data sets did not mean that the categorical data sets were useless for landslide susceptibility mapping. As discussed in section on feature selection results, all these categorical data sets did affect the final prediction result when combined with continuous data sets. In addition, a certain class that showed high susceptibility in a relative sense could still be extracted from the soil drainage and lithology data sets, such as the excessively drained soils and the two mica adamellite areas.

Landslide susceptibility mapping and comparison with logistic regression

Finally, the landslide susceptibility map in the study area was generated using both hinge features of continuous data

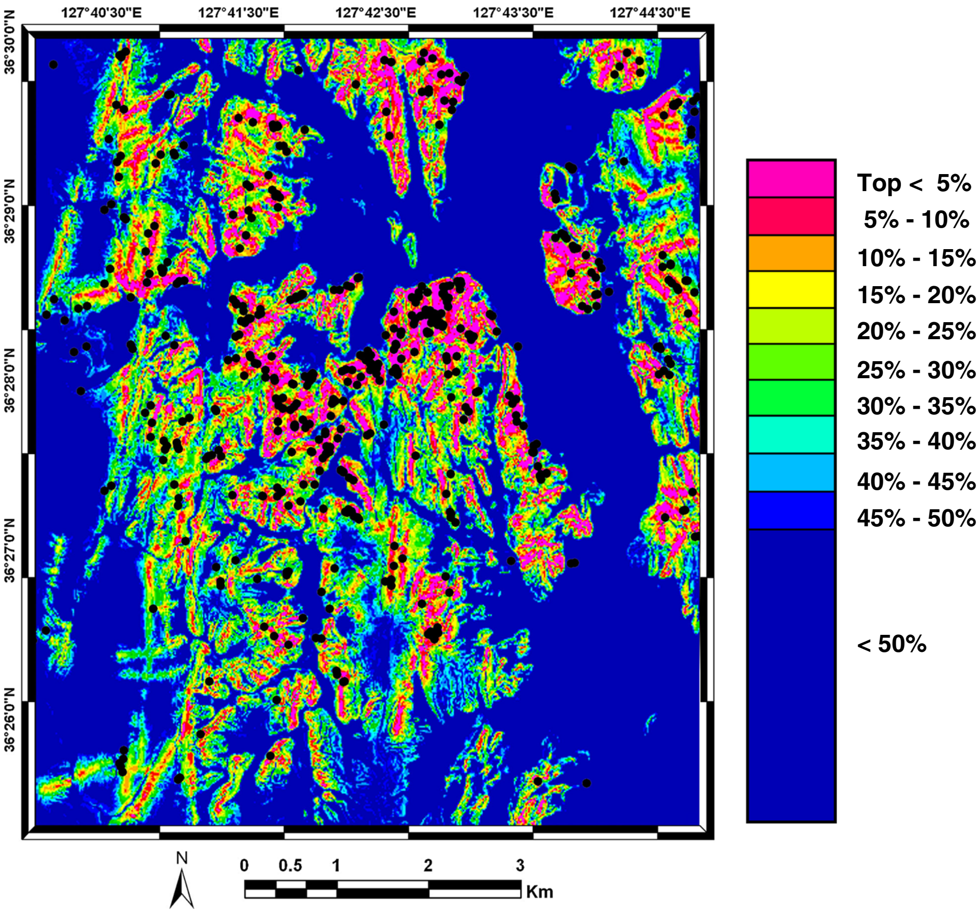

Fig. 7 Landslide susceptibility map in the study area using both hinge features of continuous data and categorical data. *Black dots* denote landslide locations

and categorical data with 10,000 background samples. Relative landslide susceptibility levels throughout the study area were generated as the landslide susceptibility map. Thus, the final susceptibility map was visualized with 200 classes at a 0.5 % interval, as shown in Fig. 7. This visualization procedure was used because the main objective of this case study was to express relative susceptibility levels within the study area. In the landslide susceptibility map, the highly susceptible areas are found in the northern and central parts of the study area, where the forest and lithology types are needle-leaf trees and two mica adamellite, respectively. Overall, the steeply sloping areas that are also located near lineaments showed high susceptibility. Flat areas consisting of alluvium and non-forest types showed the lowest susceptibility values.

Landslide susceptibility analysis is related to the prediction of unknown future events. For this susceptibility map to be useful for landslide hazard prevention, predictive performance should also be conveyed for its interpretation. The prediction rate curve, which was used for computing the AUC values, can be used for the interpretation on the landslide susceptibility map in the study area with respect to the prediction of future landslides. Figure 8 shows the prediction rate curve based on fivefold cross-validation with the same data sets used for generating the susceptibility map shown in Fig. 7. From this prediction rate curve, it could be interpreted that the top 5 and 10 % classes in Fig. 7 could contain approximately 34.5 and 57.2 % of unknown future landslides, respectively.

To test the potential of the maximum entropy modeling, a quantitative comparison with logistic regression was finally carried out. For a quantitative comparison, the same validation procedure that has been applied to the maximum entropy modeling was also applied to logistic regression. The prediction rate curve with the AUC value for the logistic regression model is given in Fig. 8. The top 5 and 10 % classes in the logistic regression model contain 32.8 and 55.0 % of the landslides, respectively. The AUC value from the entropy modeling (0.869) was slightly greater than that from logistic regression (0.861). The interesting result is that the AUC value from logistic regression is very similar to that from the maximum entropy modeling using the linear feature for continuous data (0.863). As mentioned before, to represent continuous data using the linear feature means that the original scale value of the continuous data is used for modeling. The logistic regression model, which is a special form of generalized linear models, quantifies the linear relationships in a logistic

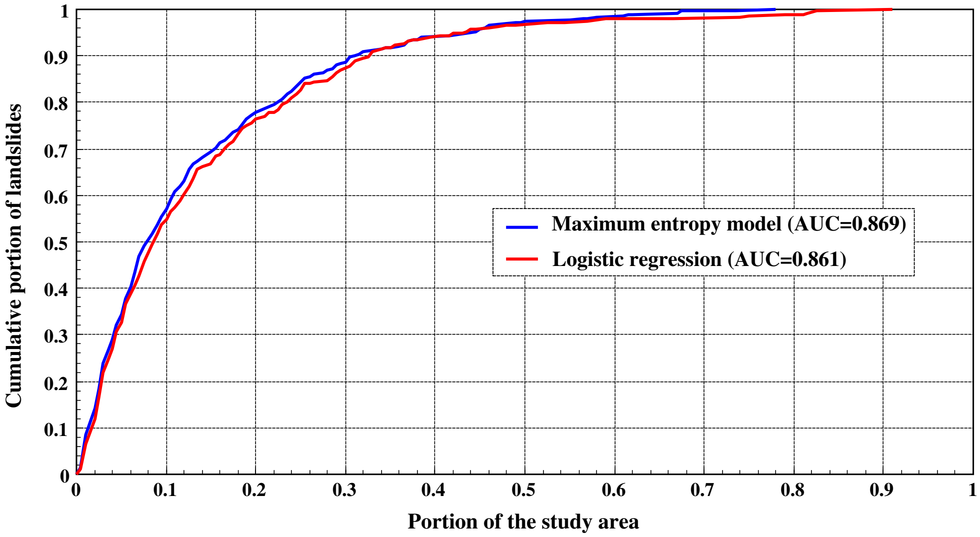

Fig. 8 Prediction rate curves based on fivefold cross-validation for the maximum entropy model and the logistic regression model

space. Therefore, it may not properly fit nonlinear relationships. Conversely, the maximum entropy model enables the fitting of complex relationships using various features. In the case study, the hinge feature used for continuous data representation can represent well the nonlinear relationships. This notable characteristic of the maximum entropy modeling resulted in the improvement of predictive performance.

Conclusions

Landslide susceptibility mapping can be regarded as an important preliminary step for assessing the risk of future landslides. To generate a reliable landslide susceptibility map, a consistent framework capable of integrating multiple environmental factors effectively is required. This study tested the applicability of maximum entropy modeling, which has been used widely for species distribution modeling, but which has not been investigated fully for landslide susceptibility mapping.

Based on a case study in the Boeun area of Korea, the maximum entropy modeling showed its particular characteristics for landslide susceptibility mapping. From a modeling viewpoint, the hinge feature was the most appropriate for continuous data representation and its combination with categorical data showed the best predictive performance. The hinge feature can provide smoothed response functions such as those of generalized additive models. Even though the hinge feature was the best type for continuous data in this case study, the maximum entropy model can properly model nonlinear or correlated relationships between input continuous data layers using other feature types.

Unlike the black-box type of other machine learning algorithms such as neural networks, the maximum entropy models can provide useful information for interpretations. For example, factor contribution analysis, based on a jackknife test and a response curve, determined that the distance from lineaments was the most influential factor in the study area and the slope layer was the next most influential factor. The contributions of the three categorical data sets were less than those of the three continuous data sets in the study area. However, following interpretation of the response curves, each categorical layer was found to have a certain category class that was much more susceptible than others. For example, most landslides occurred in deeply weathered granite areas with excessively drained soils and needle-leaf trees.

From a comparison with logistic regression, the maximum entropy model showed better predictive performance. This improvement of predictive performance was attributed mainly to using the hinge features for continuous data that were the most influential factors among the data layers.

To increase the practical applicability to landslide susceptibility mapping of the major findings of this study, additional case studies should be performed considering different numbers of landslide occurrences and/or a greater number of data sets. Extensive case studies including quantitative comparisons with other models will be carried out in future work.

Acknowledgments This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (NRF-2012R1A1A1005024). This work was also partly supported by Inha University.

References

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 9:93–106

Althuwaynee OF, Pradhan B, Lee S (2012) Application of an evidential belief function model in landslide susceptibility mapping. Comput Geosci 44:120–135

Atkinson PM, Massari R (1998) Generalised linear modeling of susceptibility to landsliding in the central Apennines, Italy. Comput Geosci 24:373–385

Austin MP (2002) Spatial prediction of species distribution: an interface between ecological theory and statistical modeling. Ecol Model 157:101–118

Ballabio C, Sterlacchini S (2012) Support vector machines for landslide susceptibility mapping: the Staffora river basin case study, Italy. Math Geosci 44:47–70

Carranza EJM, Hale M (2003) Evidential belief functions for data-driven geologically constrained mapping of gold potential, Baguio district, Philippines. Ore Geol Rev 22:117–132

Chae BG, Cho YC, Song YS, Kim KS, Lee CO, Lee BJ, Kim MI (2009) Development of landslide prediction technology and damage mitigation countermeasures. Korea Institute of Geoscience and Mineral Resources, Korea (in Korean)

Choi J, Oh HJ, Won JS, Lee S (2010) Validation of an artificial neural network model for landslide susceptibility mapping. Environ Earth Sci 60:473–483

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

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

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

Elith J, Graham CH (2009) Do they? How do they? WHY do they differ? On finding reasons for different performances of species distribution models. Ecography 32:1–12

Elith J, Graham CH, Anderon RP et al (2006) Novel methods improve prediction of species distributions from occurrence data. Ecography 29:129–151

Elith J, Phillips SJ, Hastie T, Dudík M, Chee YE, Yates CJ (2011) A statistical explanation of Maxent for ecologists. Divers Distrib 17:43–57

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

Felicísimo A, Cuartero A, Remondo J, Quiros E (2012) Mapping landslide susceptibility with logistic regression, multiple adaptive regression splines, classification and regression trees, and maximum entropy methods: a comparative study. Landslides 9:175–189

Franklin J (2009) Mapping species distributions: spatial inference and prediction. Cambridge University Press, New York

Ghosh S, Carranza EJM (2010) Spatial analysis of mutual fault/fracture and slope controls on rock sliding in Darjeeling Himalaya, India. Geomorphology 122:1–24

Greco R, Sorriso-Valvo M, Catalano E (2007) Logistic regression analysis in the evaluation of mass movements susceptibility: the Aspromonte case study, Calabria, Italy. Eng Geol 89:47–66

Guisan A, Edwards TC, Hastie T (2002) Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecol Model 157:89–100

Jaynes ET (1957) Information theory and statistical mechanics. Phys Rev 106:620–630

Kim OJ, Lee DS, Lee HY (1977) Explanatory text of the geological map of Boeun sheet. Korea Institute of Geoscience and Mineral Resources, Korea (in Korean)

Kim KS, Kim WY, Chae BG, Cho YC (2000) Engineering geologic characteristics of landslide induced by rainfall -Boeun, Chungcheong Buk-Do-. J Eng Geol 10:163–174 (in Korean)

Kim KD, Lee S, Oh HJ, Choi JK, Won JS (2006) Assessment of ground subsidence hazard near an abandoned underground coal mine using GIS. Environ Geol 50:1183–1191

Leathwick JR, Elith J, Francis MP, Hastie T, Taylor P (2006) Variation in demersal fish species richness in the oceans surrounding New Zealand: an analysis using boosted regression trees. Mar Ecol Prog Ser 321:267–281

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 26:1477–1491

Lee S (2007) Landslide susceptibility mapping using an artificial neural network in the Gangneung area, Korea. Int J Remote Sens 28:4763–4783

Lee S, Min K (2001) Statistical analysis of landslide susceptibility at Yongin, Korea. Environ Geol 40:1095–1113

Lee S, Sambath T (2006) Landslide susceptibility mapping in the Damrei Romel area, Cambodia using frequency ratio and logistic regression models. Environ Geol 50:847–855

Lee S, Choi J, Min K (2004) Probabilistic landslide hazard mapping using GIS and remote sensing at Boun, Korea. Int J Remote Sens 25:2037–2052

Lee S, Hwang J, Park I (2012) Application of data-driven evidential belief functions to landslide susceptibility mapping in Jinbu, Korea. Catena 100:15–30

Lehmann A, Overton JM, Leathwick JR (2002) GRASP: generalized regression analysis and spatial prediction. Ecol Model 157:189–207

Park NW (2011) Application of Dempster–Shafer theory of evidence to GIS-based landslide susceptibility analysis. Environ Earth Sci 62:367–376

Park NW, Chi KH, Chung CF, Kwon BD (2003) GIS-based data-driven geological data integration using fuzzy logic: theory and application. Econ Environ Geol 36:243–255 (in Korean)

Phillips SJ, Dudík M (2008) Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography 31:161–175

Phillips SJ, Anderson RP, Schapire RE (2006) Maximum entropy modeling of species geographic distributions. Ecol Model 190:231–259

Phillips SJ, Dudík M, Elith J, Graham CH, Lehmann A, Leathwick J, Ferrier S (2009) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecol Appl 19:181–197

Pineda E, Lobo JM (2009) Assessing the accuracy of species distribution models to predict amphibian species richness patterns. J Anim Ecol 78:182–190

Porwal AK, Carranza EJM, Hale M (2004) A hybrid neuro-fuzzy model for mineral potential mapping. Math Geol 36:803–826

Prasad AM, Iverson LR, Liaw A (2006) Newer classification and regression tree technique: bagging and random forests for ecological prediction. Ecosystems 9:181–199

Sivia DS, Skilling J (2006) Data analysis: a Bayesian tutorial. Oxford University Press, New York

Tinoco BA, Astudillo PX, Latta SC, Graham CH (2009) Distribution, ecology and conservation of an endangered Andean hummingbird: the Violet-throated Metaltail (*Metallura baroni*). Bird Conserv Int 19:63–76

Van Der Wal J, Shoo LP, Graham C, Williams SE (2009) Selecting pseudo-absence data for presence-only distribution modeling: how far should you stray from what you know? Ecol Model 220:589–594

Vorpahl P, Elsenbeer H, Märker M, Schröder B (2012) How can statistical models help to determine driving factors on landslides? Ecol Model 239:27–39

Ward DF (2007) Modelling the potential geographic distribution of invasive ant species in New Zealand. Biol Invasions 9:723–735

Wollan AK, Bakkestuen Y, Kauserud H, Gulden G, Halvorsen R (2008) Modelling and predicting fingal distribution patterns using herbarium data. J Biogeogr 35:2298–2310

Yao X, Tham LG, Dai FC (2008) Landslide susceptibility mapping based on support vector machine: a case study on natural slopes of Hong Kong, China. Geomorphology 101:572–582