International Environmental Modelling and Software Society (iEMSs)
2012 International Congress on Environmental Modelling and Software
Managing Resources of a Limited Planet, Sixth Biennial Meeting, Leipzig, Germany
R. Seppelt, A.A. Voinov, S. Lange, D. Bankamp (Eds.)
http://www.iemss.org/society/index.php/iemss-2012-proceedings

Application of support vector machines in landslide susceptibility assessment for the Hoa Binh province (Vietnam) with kernel functions analysis

Dieu Tien Bui[a,b,*], Biswajeet Pradhan[c], Owe Lofman[a], Inge Revhaug[a], Oystein B Dick[a]

[a]*Department of Mathematical Sciences and Technology, Norwegian University of Life Sciences, P.O. Box 5003 IMT, N-1432, Aas, Norway*
[b]*Faculty of Surveying and Mapping, Hanoi University of Mining and Geology, Dong Ngac, Tu Liem, Hanoi, Vietnam.*
[c]*Institute of Advanced Technology, Spatial and Numerical Modelling Laboratory, University Putra Malaysia, Serdang, Selangor Darul Ehsan 43400, Malaysia.*
** Bui-Tien.Dieu@umb.no;BuiTienDieu@gmail.com*

Abstract: The main objective of this study is to investigate the potential application of support vector machines (SVM) with kernel functions analysis for spatial prediction of landslides in the Hoa Binh province, Vietnam. A landslide inventory map that accounts for landslides that occurred during the last ten years was constructed using data from various sources. The landslide inventory was randomly divided into a training dataset 70% for building the models and the remaining 30% for the validation of the models. Ten landslide conditioning factors, such as slope angle, aspect, relief amplitude, lithology, soil type, landuse, distance to roads, distance to rivers, distance to faults and rainfall were prepared. During the model building process, four different SVM kernel functions (linear, polynomial, radial basic function, and sigmoid) were employed and four landslide susceptibility maps were constructed. Using the prediction rate method, the validation was performed by using landslide locations, which were not utilized during the model building. The validation results showed that the area under the curve (AUC) for landslide susceptibility maps produced by the SVM linear function, SVM polynomial function, SVM radial basic function, and SVM sigmoid function are 0.956, 0.956, 0.952, and 0.945 respectively. It indicates that the four landslide models seem to have performed well. Compared with the logistic regression (AUC =0.938) and Bayesian neural network model (AUC 0.903), the accuracy of the SVM landslide models in this study (using radial basic function and polynomial function) are slightly better. The result shows that SVM is a powerful tool for landslide susceptibility mapping at a regional scale. These maps can be very useful for natural hazards assessment and for land use planning.

***Keywords**:* Landslide susceptibility; Support vector machines; Remote sensing; GIS; Hoa Binh province; Vietnam

1. Introduction

Landslides are considered as one of the most common recurring natural hazards in Vietnam that have caused large loss of lives and property in recent years (Tien Bui et al., 2011b) . Landslides mainly occurred during heavy rainfall, especially in the tropical rain storms. Landslide susceptibility map preparation is considered as the

first important step for landslide hazard mitigation and management. Due to the complex nature of landslides, a reliable spatial prediction of landslide hazards are not easy (Ercanoglu et al., 2004).Therefore, various techniques and methods have been proposed and a review of these methods can be seen in the literature (Guzzetti et al., 1999). In recent years, support vector machines (SVM) approaches have been employed for landslide studies with results considered to outperform conventional methods (Ballabio et al., 2012). However, the performance of the SVM is heavily influenced by the selection of kernel functions. From the literature review, it seems that the effects of the kernel functions on landslide susceptibility models have been less analyzed. The main objective of this study is to investigate the potential application of SVM with kernel functions analysis for spatial prediction of landslides in the Hoa Binh province (Vietnam). In addition, the SVM landslide models were compared with models estimated from logistic regression and Bayesian neural network for the study area.

2. Study area and data

The study area is the mountainous Hoa Binh province (Fig.1) in the North West part of Vietnam. Its area is around 4,660 km^2, situated between longitudes 104°48'E and 105°50'E, and latitudes 20°17'N and 21°08'N. The elevation is in the range from 0 to 1,510 meter descending from Northwest to Southeast. More than 38 geologic formations outcropped in the province. The main characteristics are limestone, conglomerate, aphyric basalt, sandstone, silty sandstone, and black clay shale. The rainy season is normally from May to October with a total rainfall that accounts for 84-90% of the total yearly rainfall. The high frequency and intensity of the rain especially during tropical rainstorms is considered to be the most important landslide triggering factor.

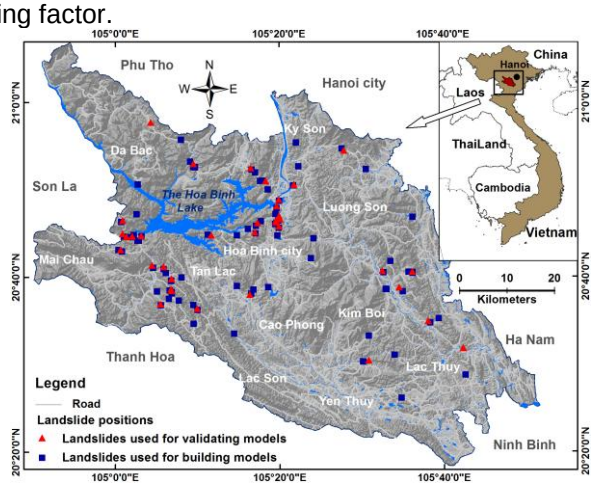

Figure 1. Landslide inventory of the study area.

Landslides that have occurred in the past and present are keys to the spatial prediction of landslide hazard in the future (Guzzetti et al., 1999). The landslide inventory map is therefore, the first step in landslide modeling. In this study, we used the landslide inventory map (Fig.1) prepared by Tien Bui et al. (2011a) to analyze the relationships between landslide occurrence and landslide conditioning factors. The landslide inventory map included 118 landslides depicted as polygons. The size of the smallest landslide is about 380 m^2, the largest is 14,340 m^2, and the average is 3,440 m^2.

A total of ten landslide conditioning factors were selected for this study: slope angle, aspect, relief amplitude, lithology, soil type, land use, distance to roads, distance to rivers, distance to faults and rainfall. This selection is based on the spatial relationship between landslide occurrence and landslide conditioning factors carried out by Tien Bui et al.(2011a). The classes in detail for the ten landslide conditioning factors are shown in table 1.

A digital elevation model (DEM) with a spatial resolution of 20x20 m was generated using national topographic maps (scale of 1:25,000). Based on the DEM, three derivative factors (slope angle, aspect and relief amplitude) were extracted. The

lithology and distance-to-faults maps were constructed based on the geological and mineral resources map of Vietnam (scale of 1:200,000). The land use map (scale of 1:50,000) was compiled from the national status land use database. The soil type map (scale of 1:100,000) was compiled from the National Pedology Map. The distance-to-roads and the distance-to-rivers maps were constructed by buffering the road and river network that undercut slopes. The road and river network was extracted from the national topographic map in a scale of 1:50,000. The rainfall map was constructed using the value of maximum rainfall of eight days (seven rainfall day plus last day of rainfall larger than 100 mm) for the period from 1990 to 2010, using the inverse distance weighed method.

Table 1. Landslide conditioning factors and their classes for this study.

Landslide conditioning factors	Class
Slope angle (^{o})	(1)0–10; (2)10–20; (3)20–30; (4)30–40; (5)40–50;(6)> 50
Aspect	(1)Flat; (2)N; (3)NE; (4)E; (5)SE; (6)S; (7)SW; (8)W; (9)NW
Relief amplitude (m)	(1)0–50; (2)50–100; (3)100–150; (4)150–200; (5) 200–250; (6)250–532.
Lithology	(1)Group 1; (2)Group 2; (3)Group 3; (4)Group 4; (5)Group 5; (6)Group 6; (7) Group 7
Land use	(1)Populated area; (2)Orchard land; (3)Paddy land; (4)Protective forest land; (5)Natural forest land; (6)Productive forest land; (7)Water; (8)Annual crop land; (9)Non tree rocky mountain; (10)Barren land; (11)Specially used forest land; (12)Grass land
Soil type	(1)Eutric Fluvisols; (2)Degraded soil; (3)Limestone Mountain; (4)Ferralic Acrisols; (5)Rhodic Ferralsols; (6)Humic Acrisols; (7)Dystric Fluvisols; (8)Dystric Gleysols; (9)Luvisols; (10)Humic Ferralsols; (11)Populated Area; (12)Water, (13)Gley Fluvisols.
Rainfall (mm)	(1)362–470; (2) 470–540; (3) 540– 610; (4) 610–950
Distance to roads (m)	(1)0–40; (2)40–80; (3) 80–120; (4) >120
Distance to rivers (m)	(1)0–40; (2)40–80; (3) 80–120; (4) >120
Distance to faults (m)	(1)0–200; (2)200–400; (3) 400–700; (4)700–1,000; (5) > 1,000

3. Landslide susceptibility mapping using support vector machines

3.1 Support vector machines

Support vector machines (SVM) is a supervised learning algorithm that is based on statistical learning theory (Vapnik, 1998). Given a training dataset that contains a set of landslide conditioning factors as inputs and landslide locations as output values, the goal of the SVM training algorithm is to find an optimal hyper-plane that separate the dataset into two classes one with landslides and one with no-landslides. The process of maximizing the separation will result into two parallel hyper-planes known as boundary planes. The distance between them is called the margin and the observations lying near the boundary planes are called the support vectors (Vapnik, 1998).

Assume we have a training dataset (X_i, y_i) with $X_i \in R^n, y_i \in \{1,-1\}$. X_i represents an input vector of ten landslide conditioning factors. The two classes $\{1,-1\}$ denote landslide and no-landslide. The optimal separating hyper-plane decision function w and b can be obtained by solving the following optimization function:

$$\underset{W,b,\xi}{Minimize}: \frac{1}{2}W^T W + C\sum_{i=1}^{l}\xi_i \quad (1)$$

$$Subject\ to\ \ y_i(w^T\phi(X_i)+b)\geq 1-\xi_i \quad (2)$$

where w is a coefficient vector that determines the orientation of the hyper-plane, b is the offset of the hyper plane from the origin, ξ_t is the positive slack variables that allows for penalized constraint violation. C is the penalty parameter that controls the trade-off between the maximum margin and the minimum error.
Using Lagrange multiplier (α_i), the dual is:

$$Maximize: \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j \phi(X_i)\phi(X_j) \quad (3)$$

$$Subject\ to\ \sum_{i=1}^{l}\alpha_i y_j = 0\ and\ 0\leq\alpha_i\leq C \quad (4)$$

The decision function can be written as:

$$f(x) = sign(\sum_{i=1}^{l} y_i\alpha_i K(X_i, X_j) + b) \quad (5)$$

where $K(X_i, X_j) = \phi(X_i)^T\phi(X_j)$ is the kernel function.

Table 2. Kernel functions and their parameters used in this study.

Kernel	Formula	Kernel parameters
Linear kernel function (LN)	$K(X_i, X_j) = X_i^T X_j$	
Radial basis function (RBF)	$K(X_i, X_j) = exp(-\gamma\|x_i - x_j\|^2)$	γ
Polynomial function (PL)	$K(X_i, X_j) = (\gamma X_i^T X_j + 1)^d$	γ , d
Sigmoid kernel function (SIG)	$K(X_i, X_j) = Tanh(\gamma X_i^T X_j + 1)^d$	γ

3.2 Performance assessment of the landside susceptibility models

Using several statistical evaluation criteria such as true positive (TP), false positive (FP), true negative (TN), false negative (FN). The overall accuracy of the trained landside susceptibility model is calculated as (TP+TN)/N, with N as the total number of training pixels. The reliability of the landslide susceptibility model is estimated using Cohen's Kappa index (κ) (Guzzetti et al., 2006) as follows.

$$\kappa = \frac{P_{obs} - P_{exp}}{1 - P_{exp}} \quad (6)$$

Where P_{obs} = (TP+TN)/ the proportion of pixels that is correctly classified as landslide or non-landslide. P_{exp} = ((TP+FN)(TP+FP)+(FP+TN)(FN+TN))/Sqr(N) is the proportion of pixels for which the agreement is expected by chance.
According to Landis and Koch (1977), the strength of agreement between the model and the reality is as follows: ≤0 (poor); 0-0.2 (slight); 0.2-0.4 (fair); 0.4-0.6 (moderate); 0.6-0.8 (substantial); 0.8-1 (almost perfect).

3.3 Preparation of training and validation dataset

In this study, the ten landslide conditioning factor maps were converted into a pixel format with a spatial resolution of 20×20 m. In each map, the frequency ratio value for each individual attribute class was calculated. Each attribute class was then assigned a sequence number based on the ratio value. In the next step, the Max-Min normalization procedure was carried out to rescale in the range 0.1 to 0.9 using Eq(7):

$$v' = \frac{v - Min(v)}{Max(v) - Min(v)}(U - L) + L \quad (7)$$

where v' is the normalized data matrix; v is the original data matrix; U and L are the upper and lower normalization boundaries.

The landslide inventory map with 118 landslide polygons was randomly split into two parts: Part1 with 70% of the data (82 landslides with 684 landslide grid cells) used in the training phase of the landslide models. Part-2 is a validation dataset with 30% of the data (36 landslides with 315 landslide grid cells). A total of 684 landslide pixels in the part1 were assigned the value of 1, and the same amount of no-landslide pixels was randomly generated from the landslide-free area and assigned the value -1. Finally, an extracting process was carried out to extract the value of ten landslide conditioning factors to build a training dataset. This dataset contains a total of 1368 observations, and ten input variables, one target variable (landslide, no-landslide).

3.4 Training models and generation of landslide susceptibility maps

The performance of the SVM model is depended on the choice of kernel functions and their parameters. Table 2 shows SVM Kernel functions and their parameters used in this study. *C* is the regularization parameter, γ is the kernel width, and *d* the degree of the polynomial kernel. If *the* value of *C* is large, it will lead to few training errors. In contrast, a small value of *C* will generate a larger margin and increase the number of training errors. Parameters γ and *d* controls the degree of nonlinearity and degree of the polynomial kernel respectively.

In this study the grid-search method and 5-fold cross-validation were selected to be used to find the best kernel parameters. The training dataset was randomly split into 5 equal sized subsets. The merged four subsets were used to train models whereas the remaining subset was used as a test set. The cross-validation process was repeated five times for each of the five subsets. A grid space was set with $C = 2^{-5}, 2^{-4}, \ldots, 2^{10}$; $\gamma = 2^{10}, 2^{9}, \ldots, 2^{-4}$; $d = 1, \ldots, 8$. Table 3 shows overall accuracy and Cohen's kappa index of the trained landslide models. The best value of *C* for LN-SVM is 4 with the overall accuracy 87.8%. The best *C* and *γ* for RBF-SVM are found as 8 and 0.25 respectively with the overall accuracy 91.1%. In the case of PL-SVM, the best *C*,*γ,*and *d* are 1, 0.3536, 3 respectively, with the overall accuracy 91.1%.

Cohen's Kappa indexes are 0.756, 0.822, 0.823 and 0.727 for the four landslide models (Table 3). The Kappa values indicate that the strength agreement between the observed and the predicted values is substantial for LN-SVM and SIG-SVM. Whereas it is almost perfect for RBF-SVM, PL-SVM

Table 3. Overall accuracy and Cohen's Kappa index for the four SVM models.

No	Parameters	RBF-SVM	PL-SVM	LN-SVM	SIG-SVM
1	Overall accuracy (%)	91.08	91.15	87.79	86.33
2	Cohen's Kappa index	0.822	0.823	0.756	0.727

Once the landslide susceptibility models were successfully trained in the training phase, they were then used to calculate the landslide susceptibility indexes (LSI) for all the study pixels. The results were then converted into a GIS.

4. Validation and comparison of landslide susceptibility maps

4.1 Success-rate and prediction-rate curves

The four landslide susceptibility maps were validated by means of the success-rate and prediction-rate curves (Chung et al., 2003; Guzzetti et al., 2006). The success-rate results were obtained by comparing the four landslide susceptibility maps with the landslide pixels in the training dataset (Fig. 2a). And then areas under the success-rate curves (AUC) were estimated. The result show that RBF-SVM and PL-SVM have the highest area under the curve (AUC) values 0.961 and 0.957 respectively. They are followed by LN-SVM (0.940) and SIG-SVM (0.932). The success-rate measures the goodness of fit for the landslide models to the data. The AUC results indicate that the capacity of correctly classifying the areas with existing landslides is highest for RBF-SVM, followed by the PL-SVM, LN-SVM, SIG-SVM.

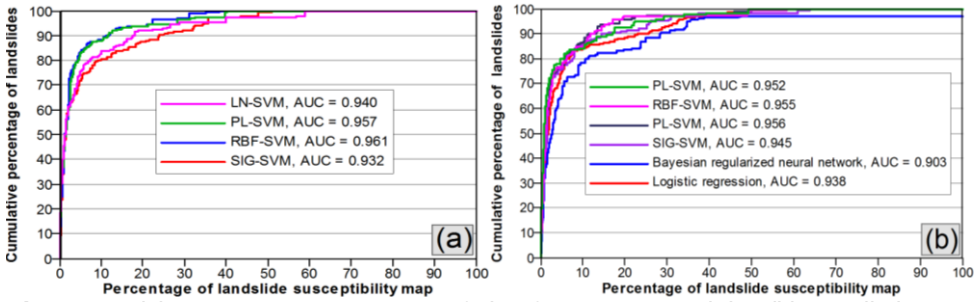

Figure 2. (a) Success rate curves of the four SVM models; (b) Prediction rate curves of the four SVM models and the Bayesian regularized neural networks and the logistic regression.

The success rate is not a suitable measure for the prediction capability of the landslide models because it is based on the landslide pixels that have already been used for building the model. The prediction rate may be used to estimate the prediction capability. In this study, the prediction-rate curves and area under the curves were obtained (Fig. 2b) by comparing the four susceptibility maps with the landslide pixels in the validation dataset. The results show that the highest prediction capability is for RBF-SVM and PL-SVM with AUC values of 0.955 and 0.956 respectively, followed by LN-SVM (0.952) and SIG-SVM (0.945). Compared with the results from the logistic regression (0.938), and Bayesian regularized neural networks (0.903), the prediction capability of the two RBF-SVM and PL-SVM models seems to be slightly better.

3.2. Reclassification of landslide susceptibility indexes and relative importance assessment of landslide conditioning factors

The landslide susceptibility indexes were reclassified into 5 classed based on the percentage of area (Pradhan et al., 2010) high (10%), moderate (10%), low (20%), very low (20%), and no (40%) (Fig. 3a). Landslide density analysis (Sarkar et al., 2008) was performed on the five landslide susceptibility classes. The results show that the landslide density gradually increases from the no to the high susceptibility class (Fig. 3b). The four landslide susceptibility maps are shown in Fig. 4.

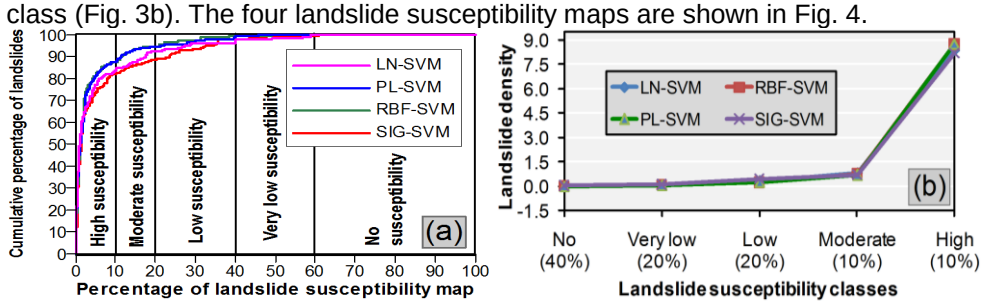

Figure 3. Cumulative percentage of landslides vs. percentage of landslide susceptibility map (a); landslide density plots of five landslide susceptibility classes (b).

The importance of a certain factor was estimated by excluding the factor and then calculated the overall accuracy of the model (Table 4). It could be observed that the highest accuracy was obtained when all of the ten factors are used, with LN-SVM, RBF-SVM, PL-SVM. However, for the case of SIG-SVM, the soil type factor might have caused slightly noise by reducing the model accuracy 0.2%. Distance to roads, rainfall, distance to rivers, land use and slope angle are the most importance factors for LN-SVM. In the case of RBF-SVM, the most importance factors are distance to roads, soil type, slope angle, land use and distance to rivers. Whereas distance to roads, land use, distance to rivers, slope angle and soil type are most important for PL-SVM. And distances to roads, land use, distance to rivers, slope angle and distance to faults are most important for SIG-SVM. It could be observed that LN-SVM includes both rainfall and land use as important repressors. This may be helpful for a possible scenario analysis, including future climate and land use scenarios.

Table 4. Accuracy of the trained SVM models for landslide susceptibility using all conditioning factors and without one of the factors.

No	Conditioning factors	Overal accuracy (%)			
		LN-SVM	RBF-SVM	PL-SVM	SIG-SVM
1	Minus slope angle	87.4	89.4	89.3	84.6
2	Minus lithology	87.6	90.6	90.3	85
3	Minus rainfall	86.2	90.4	90.0	85.7
4	Minus landuse	86.9	89.5	89.0	82.9
5	Minus soil type	87.6	89.3	89.5	86.5
6	Minus aspect	87.7	90.5	90.7	85.5
7	Minus distance to roads	80.4	82.9	83.2	76.8
8	Minus distance to rivers	86.8	90.1	89.1	83.1
9	Minus distance to faults	87.8	90.5	90.1	84.9
10	Minus relief amplitude	80.3	90.4	90.8	85.7
	All	**87.8**	**91.1**	**91.1**	**86.3**

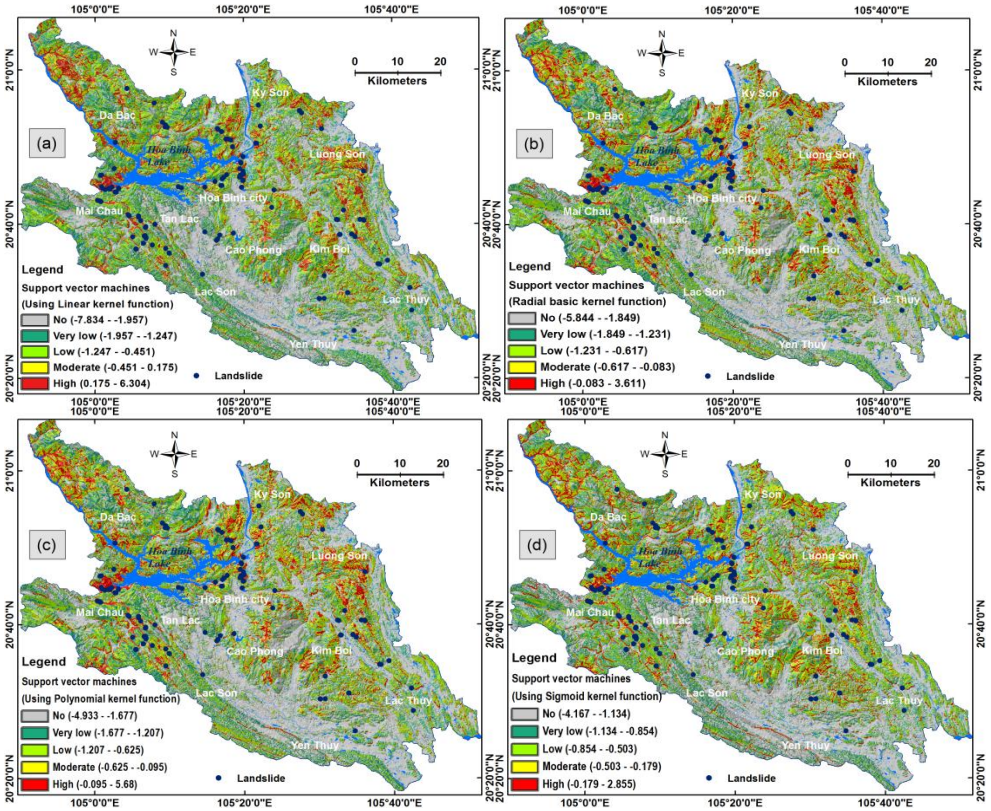

Figure 4. Landslide susceptibility zonation maps: (a) LN-SVM; (b) RBF-SVM; (c)PL-SVM; (d) SIG-SVM.

5. Concluding remarks

In this paper, we investigated the potential application of support vector machines for landslide susceptibility assessment at the Hoa Binh province (Vietnam). Ten landslide conditioning factors (slope angle, aspect, relief amplitude, lithology, soil type, land use, distance to roads, distance to rivers, distance to faults and rainfall) were used in this analysis. The landslide inventory with 118 landslide-polygons that occurred during the last ten years was used. 70% of the landslide inventory was used for building susceptibility models, whereas the remaining 30% was used for validating and assessing the prediction capability of the models.

Four kernel functions were included in the analysis, linear function, radial basis function, polynomial function, and sigmoid function. Four landslide susceptibility

maps were constructed. Using the success-rate and the prediction-rate methods, the landslide susceptibility maps were validated and compared. The largest area under the success-rate curve (AUC) is for the RBF-SVM (0.961), followed by PL-SVM (0.956), LN-SVM (0.940), and SIG-SVM (0.932). It indicates that RBF-SVM and PL-SVM have a better goodness of fit to the training data. The highest area under the prediction-rate curve (AUC) is for RBF-SVM (0.954) and PL-SVM (0.955), followed by LN-SVM (0.952), SIG-SVM (0.945). Compared to logistic regression (AUC =0.938) and Bayesian regularized neural networks (AUC =0.903), the prediction capability of RBF-SVM and PL-SVM performed slightly better.

The reliability of the four susceptibility models was assessed using the Cohen's Kappa index (κ). κ values are of 0.822, 0.823 for RBF-SVM, PL-SVM respectively, indicating almost perfect agreement. Whereas κ values for LN-SVM, SIG-SVM are of 0.756, and 0.722 indicating that the strength of agreement between the observed and predicted values are substantial.

Based on the aforementioned results, we conclude that RBF-SVM, PL-SVM models have almost equal accuracies and they may be somewhat better than logistic regression and Bayesian regularized neural networks. As a final conclusion, the results show that SVM is a powerful tool for landslide susceptibility mapping at medium scale. These maps can be very useful for natural hazards assessment and for land use planning.

ACKNOWLEDGMENTS

This research was funded by the Norwegian Quota scholarship program. The data analysis and write-up were carried out as a part of the first author's PhD studies at the Geomatics Section, Department of Mathematical Sciences and Technology, Norwegian University of Life Sciences, Norway.

REFERENCES

Ballabio, C., Sterlacchini, S., 2012. Support Vector Machines for Landslide Susceptibility Mapping: The Staffora River Basin Case Study, Italy. *Mathematical Geosciences 44*(1) 47-70.

Chung, C.J.F., Fabbri, A.G., 2003. Validation of spatial prediction models for landslide hazard mapping. *Natural Hazards 30*(3) 451-472.

Ercanoglu, M., Gokceoglu, C., 2004. Use of fuzzy relations to produce landslide susceptibility map of a landslide prone area (West Black Sea Region, Turkey). *Engineering Geology 75*(3-4) 229-250.

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*(1-4) 181-216.

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

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. *Biometrics 33* 159–174.

Pradhan, B., Lee, S., 2010. Regional landslide susceptibility analysis using back-propagation neural network model at Cameron Highland, Malaysia. *Landslides 7*(1) 13-30.

Sarkar, S., Kanungo, D., Patra, A., Kumar, P., 2008. GIS based spatial data analysis for landslide susceptibility mapping. *Journal of Mountain Science 5*(1) 52-62.

Tien Bui, D., Lofman, O., Revhaug, I., Dick, O., 2011a. Landslide susceptibility analysis in the Hoa Binh province of Vietnam using statistical index and logistic regression. *Natural Hazards 59* 1413–1444.

Tien Bui, D., Pradhan, B., Lofman, O., Revhaug, I., Dick, O.B., 2011b. Landslide susceptibility mapping at Hoa Binh province (Vietnam) using an adaptive neuro-fuzzy inference system and GIS. *Computers & Geosciences. Doi 10.1016/j.cageo.2011.10.031*.

Vapnik, V.N., 1998. *Statistical Learning Theory*. Wiley-Interscience