Intelligent Analysis of Landslide Data Using Machine Learning Algorithms

Natan Micheletti, Mikhail Kanevski, Shibiao Bai, Jian Wang, and Ting Hong

Abstract

Landslide susceptibility maps are useful tools for natural hazards assessments. The present research concentrates on an application of machine learning algorithms for the treatment and understanding of input/feature space for landslide data to identify sliding zones and to formulate suggestions for susceptibility mapping. The whole problem can be formulated as a supervised classification learning task. Support Vector Machines (SVM), a very attractive approach developing nonlinear and robust models in high dimensional data, is adopted for the analysis. Two real data case studies based on Swiss and Chinese data are considered. The differences of complexity and causalities in patterns of different regions are unveiled. The research shows promising results for some regions, denoted by good performances of classification.

Keywords

Landslide susceptibility mapping • Machine learning • Support vector machines

Introduction

Over the last decade, landslide hazard and risk have been one of the major research fields at the international level. The increasing of economic costs for the insurance companies resulting for landslide events and other natural hazards introduce the need of better knowledge and tools to treat and to study this phenomenon. Because of landslide susceptibility (LS) maps are useful tools for natural hazards assessment, a lot of research was carried out to find the efficient and precise methods for LS mapping. The use of statistical classification models instead of physical approaches is justified by the fact that the link between landslide events and predisposing and triggering factors is complex or even not well-known.

During last years, the number of studies about slope stability and susceptibility to landslides using machine learning algorithms increased considerably. Examples of publications about this subject are Yao et al. (2008) and Brenning (2005). For more information, a state-of-the-art about this topic is presented in Micheletti (2011). The most widely used methods include artificial neural networks (ANN), kernel methods (i.e. support vector machines, SVM) and logistic regression (LR).

Generally, landslide susceptibility analysis by using machine learning is formulated as a supervised classification learning problem. Despite of the increasing number of publications on this topic, many open questions still remain. The main challenges consist in choosing a suitable sampling technique, formulating the input-output spaces properly and selecting the relevant features for the learning task.

In this research we introduce a complete analysis, starting from raw data to landslide susceptibility mapping, using Support Vector Machines as a modelling tool using real datasets from different regions.

N. Micheletti (✉) • M. Kanevski
Institute of Geomatics and Risk Analysis, University of Lausanne, Lausanne, Switzerland
e-mail: Natan.Micheletti@unil.ch

S. Bai • J. Wang • T. Hong
Key Laboratory of Virtual Geographic Environments, Nanjing Normal University, Nanjing, China

C. Margottini et al. (eds.), *Landslide Science and Practice*, Vol. 3,
DOI 10.1007/978-3-642-31310-3_22,

Methodology

Supervised Learning

In a supervised learning task, the output is known for every observation. Hence, the learning task is to model the unknown dependence between input and output. To model landslide susceptibility a supervised classification task is formulated. In this case, estimation of the class of an unseen sample using a model constructed on examples provided by the user is aimed.

Support Vector Machines

Support Vector Machines, the workhorse of Statistical Learning Theory (Vapnik 1998), is a very attractive approach developing robust and stable models using high dimensional datasets. SVM can be adapted to different supervised classification learning task. In the present research, two-class SVM is applied. SVM is using Structural Risk Minimization principle of inference trying to minimize classification error and controlling the complexity of the model. Details about this approach and its application to environmental studies are presented in Cherkassky and Mulier (2007) and Kanevski et al. (2009).

The simplest case is a separable linear problem. Considering labelled samples provided by the user, SVM aims the separation of the classes in the input space by a hyperplane (linear model) of the form

$$f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b \tag{1}$$

where $\mathbf{x}$ is a vector describing position of data in high dimensional input space, $\mathbf{w}$ and b are the constants obtained after the solution of SVM optimization problem (Vapnik 1998).

The correct location of the classification samples in relation to the hyperplane is controlled by corresponding constraints. Further, the margin between the two classes is maximized to ensure a classification model with good generalization ability. In most cases the solution of the optimization problem is sparse and only support vectors (SV) – the samples located on margin borders – contribute to the solution with non-zero weights.

For the non-separable case (noisy data) the SVM was extended by allowing some misclassifications and keeping solution simple - linear. This is done by introducing the slack variables ξ. A hyper-parameter C balances the trade-off between margin maximization and classification error (empirical risk). The optimization problem can be formulated as follows:

$$\begin{cases} \min & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{L}\xi_i \\ \text{Subject to} & y_i(\langle \mathbf{w}, \mathbf{x} \rangle + b) \geq 1 - \xi_i \end{cases} \tag{2}$$

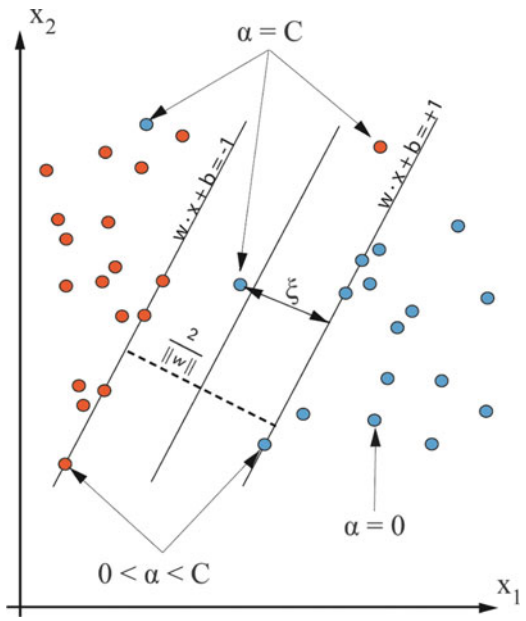

Fig. 1 Support vector machine scheme

The optimization problem (minimization with constraints) is solved in a classical way by introducing the Lagrange multipliers α. Figure 1 presents a two dimensional illustration of the two-class (red and blue dots) classification problem. Well classified points ($\alpha = 0$) do not contribute to the solution. "Normal" support vectors correspond to $0<\alpha<C$ and "atypical" or noisy SV to $\alpha = C$.

The prediction for a sample $\mathbf{x}$ is formulated as follows:

$$Sign(f(\mathbf{x})) = sign\left(\sum_{i=1}^{L} y_i \alpha_i \langle \mathbf{x}_i, \mathbf{x} \rangle + b\right) \tag{3}$$

where $y_i = \{+1, -1\}$ for two-class problem (e.g. $+1$ = red and -1 = blue) and only points with $\alpha_i \neq 0$ contribute to the solution.

To solve non-linear problems, a kernel function $K(x_i, x)$ can be introduced. The latter replaces the dot product $\langle x_i, x \rangle$ and maps data into a higher dimensional feature space, where a linear decision boundary can be constructed. The most commonly used is the Gaussian RBF kernel, which introduces another hyper-parameter, the so-called kernel bandwidth σ:

$$K(\mathbf{x}_i, \mathbf{x}) = e^{-\frac{(x_i - x)^2}{2\sigma^2}} \tag{4}$$

The goal of SVM training is to select the optimal values for the regularization parameter C and the kernel bandwidth σ. There are many algorithms to do it but the most usual procedure is to split data into training, validation and testing subsets and to use the validation data with a grid search. The optimal parameters minimize validation error (see Kanevski et al. 2009).

Landslide Susceptibility Mapping Using SVM in Vaud, Switzerland

Vaud Region

The Swiss case study of this research is canton Vaud, Switzerland. The location of Vaud in the Swiss territory is illustrated in Fig. 2.

The three main geological regions of Vaud (Prealpes, Plateau and Jura) were selected as separate study zones. This is done because of triggering factors and landslide nature differ considerably between these zones. Hence, the patterns of these regions which we want to reproduce probably also differ significantly. Figure 3 shows the canton Vaud and its three main geological and morphological sectors.

The Plateau zone is located in the middle part of canton Vaud. Its rock tender lithology is easily affected by surface processes as hydrological action and wind erosion. Plateau features countless moraine formations, generated during last ice age. After the glaciers melting, hydrological processes influenced the surface morphology. Today the region is characterized by gentle hills and an articulated hydrographical network. Landslide phenomena in Plateau are linked to fluvial erosion and thus they are mainly located along the rivers. Generally, these surface sliding zones are strictly influenced by local hydrological conditions, as preferential flow channels, presence of aquifers, soil infiltration, etc.

On the North-West side of canton Vaud we can find the Jura zone. Its structure is characterized by a large scale folded tectonic, featuring folds of more than a hundred kilometres length. This structure follows a NE-SW orientation. Valleys and local morphology are strictly controlled by these folds, following their direction. Jura's landslides are mainly associated to tender rocks and quaternary forms. The karst nature of Jura hydrogeology prevents the development of large landslide phenomena, unless extreme rainfalls occur.

Prealpes are located in the South-East of Vaud. This mountainous zone features a complex morphology with steep slopes. The geological context is characterized by tablecloths juxtaposed during the Alps formation. The Prealpes slope stability is strictly correlated to its lithology and tectonic adduction. Moreover, glacial and hydrological influences remodelled the surface, resulting in an even more complex situation. Prealpes feature many huge landslides. Even if some triggering factors are well known, tectonic and geological complexity make predictions in this zone really difficult.

Data and Feature Extraction

To compute SVM classification, an input space including relevant features for the task must be constructed. The Vaud lithotype database derived from a geological map has been used. To introduce this categorical information, a binary formulation is necessary, indicating with 1 the presence of a given lithology.

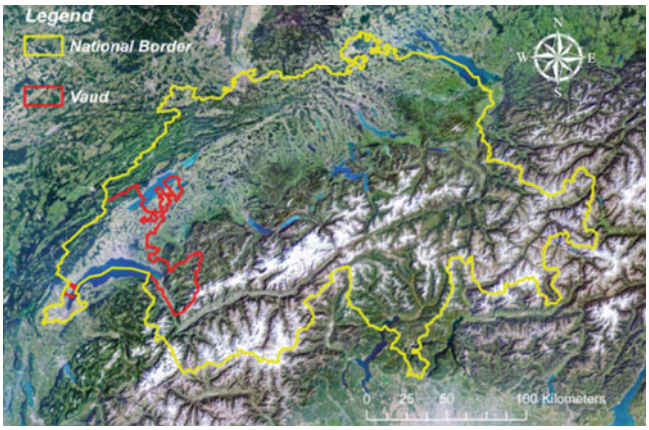

Fig. 2 Vaud location

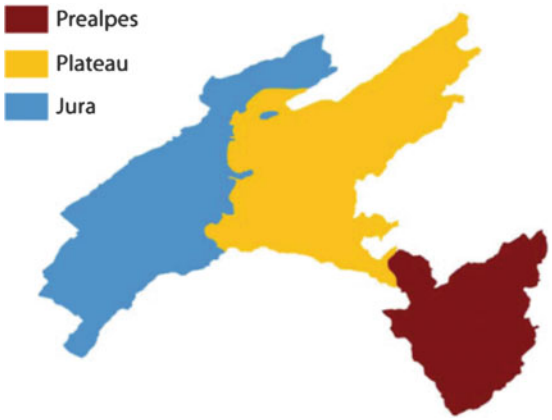

Fig. 3 Main geological regions of Vaud

A digital elevation model (DEM) of 25 × 25 m resolution allowed the extraction of different topographic features. Firstly, the most used features for slope analysis were created. The latter are aspect, slope, plan curvature, profile curvature, topographic wetness index (TWI) and surface curvature. Then the hypothesis that some variables could be important not only at small spatial scale, but also at a larger spatial scale has been formulated. The efficiency of this approach was demonstrated in other environmental study concerning modelling of wind fields in complex topography (Foresti et al. 2011). Therefore, some Gaussian filters were applied to some commonly useful features describing soil curvatures. We chose two filter of different size: 21 and 101 pixels. Finally, the continuous features in the input space are: aspect, slope, DEM height, slope filtered with Gaussian filter 101 pixels-sized (G. 101), topographic wetness index, plan curvature, profile curvature, plan curvature G. 21, plan curvature G. 101, profile curvature G. 21, profile curvature G. 101, surface curvature, surface curvature G. 21.

Finally, the Vaud landslide database is used to extract the target labels. The output has a binary formulation, -1 indicating an alleged stable sample and $+1$ indicating a sliding one.

Sampling and Experimental Setup

Both linear and non-linear SVM analyses were carried out. Linear SVM needs only tuning of the C hyper-parameter, while non-linear SVM requires the selection of C as well as of σ. For both algorithms, three subsets are needed to perform the analysis. Firstly, a training set to build the classifier (SVM training) is necessary. Then, a validation set is use to select the best pair of hyper-parameters. Here a grid search (scanning of different C and σ values and looking for a minimum on a validation surface error) was applied. We allowed four values for C (1, 10, 100 and 1,000) and we looked for the best σ between 1 and 10. Finally, a test set allows the evaluation of the model (model assessment or the estimation of generalization error).

In addition, two-class SVM needs samples of the two classes to be performed. Usually, alleged stable examples can be randomly selected from the region of study. On the other hand, the sampling of positive sample is more complicated. Many questions arise from this task: the heterogeneity of landslides zones (where to sample?) and the non-independence of training, validation and test set in the case of random sampling, just to mention two of them.

We proposed an object-based sampling strategy. All landslides have been labelled, and then divided into three groups: a training one, a validation one and a testing one. A random sampling is performed from these groups, achieving landslides independence in geographical space.

Different sizes of training and validation datasets have been created, featuring 10 subsets for each size to compute empirical confidence intervals of the results. A single set of 10,000 points is kept as test set. The performance measure used for model selection assessment is the Area Under the ROC Curve (AUC).

SVM-Performances

Five hundred and two thousand sized datasets are used to analyse SVM performances. Five hundred sized datasets are a benchmark for their good quality/computational time ratio, while 2,000-sized sets generally ensure a solid classification.

A first analysis of SVM performances considers the percentage of support vectors in the training set, presented in Table 1. This percentage is a good indicator of the level of overfitting of the model. The latter is very satisfactory for our case, since percentages are very low for such a complex task.

Table 1 Percentage of support vector in the training set for Gaussian SVM (with standard deviation)

Points	Plateau
500	43.72 % (5.63 %)
2,000	36.60 % (3.28 %)

Table 2 SVM performances: mean test AUC (with standard deviation)

Model	Points	Plateau
Gaussian SVM	500	0.87 (0.010)
Gaussian SVM	2,000	0.88 (0.011)
Linear SVM	500	0.84 (0.020)
Linear SVM	2,000	0.88 (0.015)

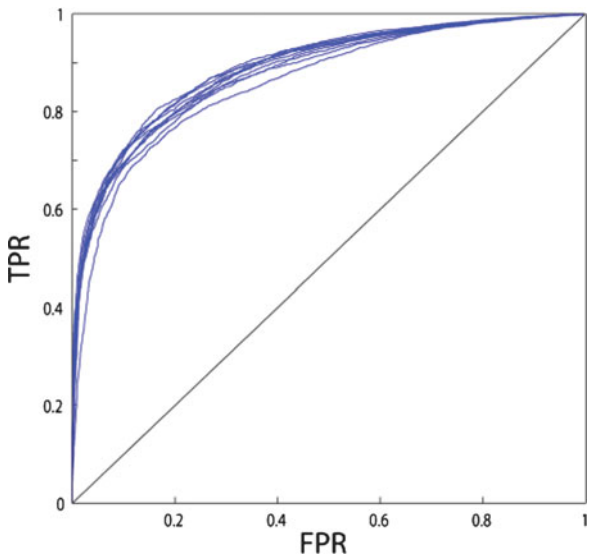

Fig. 4 ROC curves for 10 experiments with 2,000 samples (Gaussian SVM)

Table 2 summarizes the performances of linear and non-linear SVM algorithms. The evaluation of models is performed by using the test set. Mean and standard deviation on 10 models created with the same training and validation number of data provide a solid estimation of the performance efficiencies and their uncertainties. Very good results are found for Plateau sub-region, motivating the application of similar approach for other zones.

It is interesting to note how linear SVM has very similar performances to the Gaussian one for Plateau. The same is proven true for Jura. The hypothesis that, including enough training samples, the two-classes are linearly separable in these sub-regions can be advanced.

Finally, ROC curves for 2,000-sized datasets models are given in Fig. 4.

Performances analyses indicate that Plateau is an easy case study. The two classes are discriminated very well, resulting in a mean test AUC of 0.88 using 2,000 samples. Empirical observations underline that landslides are almost

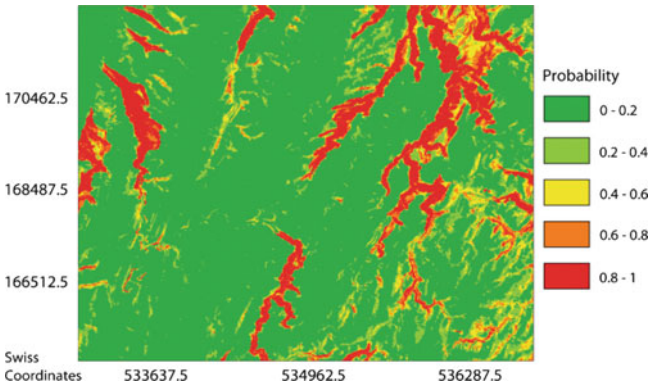

Fig. 5 Plateau ROI, landslide susceptibility mapping using two-class Gaussian SVM with 2,000 points, split n° 5

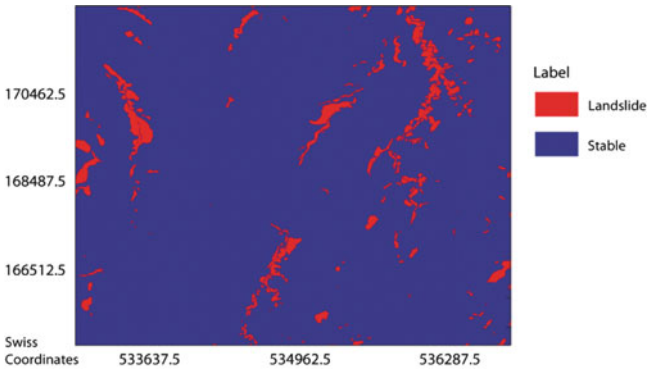

Fig. 6 Plateau region of interest: known landslides

only present on riverside slopes in this sub-region, thus the model can detect them easily. A trustful and accurate landslide susceptibility mapping can be advanced.

In next section, landslide susceptibility mapping in Vaud's Plateau provides a visual explanation of these considerations.

Landslide Susceptibility Mapping

The SVM decision function is continuous and unbounded. Therefore, a probabilistic mapping can be more accessible to readers and most importantly to stakeholders, who are the biggest users of such maps. Some regions of interest (ROI) have been chosen for predictions, in order to have a more suitable visualization of quality of the results. In addition, a visualization of labels in the same ROI can be useful to visually estimate the quality of predictions.

Landslide probabilistic maps for a Plateau ROI provide outstanding results (Fig. 5). A very strong discrimination between landslide susceptibility zones and stables ones can be observed which can be confirmed by comparing the predictions with the known landslide locations (Fig. 6). Very high confidence in uncertainties illustrates how the model's performances are solid and trustful.

Introduction to Landslide Data Analysis in Gansu Province, China

Baiyhue Catchment Area in Gansu

The Baiyhue catchment area in China's north-western province Gansu is the subject of a work-in-progress landslide analysis and modelling using SVM. The goal of this study is to assist regional planning and risk management.

The lithology of the mountainous 432 km^2 large study area comprises loess deposits, mudstones and other weak formation. These soil property, associated to the steepness of regional valleys and heavy rainfall events, causes strong soil erosion, landslides and viscous debris and mud flows on slopes. The large volume of material delivered to the valley also causes problems.

Data and Feature Extraction

Data for SVM classification for Baiyhue catchment area has three different sources.

Twelve geological formations were derived from geological maps and used to define fine lithological units, grouping rock types that present similar compositional and mechanical characteristics.

A digital elevation model was generated from a triangulated irregular network model. The DEM was used to produce continuous variables such as elevation, slope angle, aspect, terrain roughness, shape of the slope parameters, plan curvature and profile curvature.

Land cover information is interpreted from TM 5 (Path 127/Row 39, dated 7/2000) satellite imagery using various image processing and enhancement techniques. The interpreted images were digitally processed to further modify the boundaries by a supervision classification task. The accuracy of the land-use interpretation was checked in the field. Eleven vegetation types were recognized.

Finally, drainages, roads and faults digitized from topographic maps have been used. Distances from these objects are included as features.

A detailed landslide-inventory map of the study area was constructed by interpretation of aerial photographs. Extensive field studies were used to check the size and shape of landslides.

SVM-Analyses

Like for Swiss case study, SVM algorithm is used to perform a supervised classification for landslide susceptibility mapping.

Data was divided in three subsets for model construction, selection and assessment. Training dataset includes 1,820

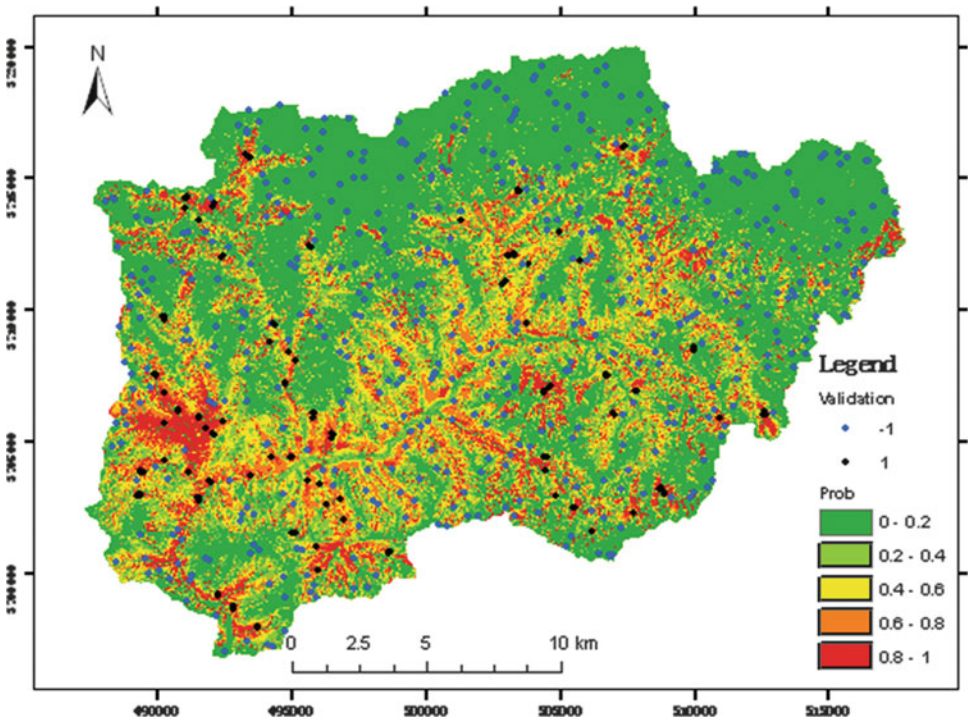

Fig. 7 Landslide susceptibility mapping using two-class Gaussian SVM, Baiyhue catchment area, Gansu, China

samples, while validation and test sets include 636 examples each. In the datasets the number of landslide and alleged stable point is the same. After the first analysis by applying SVM, an accuracy of 80.04 % is found on independent data.

A probabilistic map of landslide in Baiyhue area is presented in Fig. 7. Independent data used to assess the model is added to the map (+1 for landslide samples, −1 for stable ones).

In general, results are satisfactory, identifying well the valley slopes known as landslide susceptible. However, more efforts to assess the model are still in progress, as additional data and a more detailed analysis of results.

Conclusions

In conclusion, SVM analysis in Plateau, Vaud, Switzerland confirms prior knowledge acquired by empirical observations. Good performances achieved illustrate how Support Vector Machines could be useful for landslide susceptibility mapping. The other Vaud sub-regions can be analysed with the same methodology. The interested reader can find insight about this case study in Micheletti (2011). An important question in landslide susceptibility mapping is the feature space construction and feature selection, especially because different regions can be feature specific. The first results in this direction are quite promising and can be found in Micheletti (2011).

Besides being a work-in-progress, SVM analysis in the Baiyhue catchment area of Gansu, China already produced interesting new results, as illustrated by solid performances and a trustful landslide susceptibility mapping.

In general, results show that SVMs can be successfully adopted in landslide susceptibility mapping. Future perspective for such approach include, besides feature selection, analysis of different regions and application of other machine learning algorithms (artificial neural network, random forest, self-organizing maps, general regression neural networks, etc.) for comparison and decision-oriented mapping.

Acknowledgments This research was partly supported by Sino-Swiss cooperation project EG 42-032010, Swiss National Science Foundation, project "GeoKernels: kernel-based methods for geo- and environmental sciences, Phase II: 200020-121835/1" and National Natural Science Foundation of China (Nos. 40801212).

We would like to thank A. Pedrazzini and M. Jaboyedoff for their important contribution in data gathering and the indispensable knowledge in the field of landslides they provided to the current research. We also are grateful to L. Foresti, G. Matasci and M. Volpi for all interesting discussion and valuable help.

References

Brenning A (2005) Spatial prediction models for landslide hazards: review comparison and evaluation. Nat Hazard Earth Syst Sci 5:835–862

Cherkassky V, Mulier F (2007) Learning from data: concepts, theory and methods. John Wiley & Sons, Inc., Hoboken, New Jersey

Foresti L, Tuia D, Kanevski M, Pozdnoukhov A (2011) Learning wind fields with multiple kernels. Stoch Environ Res Risk Assess 25(1):55–66

Kanevski M, Pozdnoukhov A, Timonin V (2009) Machine learning for spatial environmental data: theory, applications and software. EPFL Press, Lausanne

Micheletti N (2011) Landslide susceptibility mapping using adaptive support vector machines and feature selection, M.S. thesis, University of Lausanne, Switzerland

Vapnik V (1998) Statistical learning theory. Wiley, New York

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