

Classification of landslide susceptibility in the development of early warning systems

Dominik Gallus, Andreas Abecker

Research Center for Information Technologies (FZI), Haid-und-Neu-Str. 10-14, 76131 Karlsruhe, Germany, gallus@abecker@fzi.de

Daniela Richter

Institute for Photogrammetry and Remote Sensing (IPF), University of Karlsruhe, Englerstr. 7, 76128 Karlsruhe, Germany, Daniela.Richter@ipf.uni-karlsruhe.de

Abstract. Statistical classification techniques complemented by GIS yield good results in predicting landslide hazard/ susceptibility. In this work, several well-known classification methods are applied to data from distinct alpine areas in Vorarlberg, Austria. It is shown that kernel methods (Support Vector Machines – SVM – and Gaussian Processes) outperform classic techniques for this task. As a further result, hazard maps for the study areas are generated, which can be used as input for early warning systems focussing on landslide hazard.

Keywords: landslides, classification, early warning system

1 INTRODUCTION

In populated mountainous regions, natural disasters resulting from large mass movements (in particular, slope movements and avalanches) cause huge human and material damages. According to the International Disaster Database, more than 10.000 people have been affected by disastrous mass movements in Austria's alpine regions in the last 50 years, placing this type of natural hazard among the most catastrophic in this part of Europe, next to large-scale flood events (EM-DAT¹ 2007). Following heavy rainfall periods in 1999/2000, more than 250 avalanches, landslides and other slope movements have been registered in the administrative region Vorarlberg, resulting in estimated damages of about 180 million €².

Statistical classification methods from machine learning/ pattern recognition have been shown to yield good results at the task of assessing landslide hazard/ susceptibility. In combination with GIS, hazard maps resulting from fitting classifiers on data can be produced with relative ease, reducing necessary time and effort, at the same time introducing a principled and objective methodology.

This paper is structured as follows: Chapter 2 provides a short overview on basic concepts of statistical classification. Chapter 3 shortly reviews related work. In chapter 4, the study areas and input data are described. Chapter 5 presents the methods used (logistic regression; Gaussian Process models; SVM). Chapter 6 discusses experimental results. Chapter 7 concludes with an outlook, complemented by ideas for further research.

¹<http://www.em-dat.net>

²http://www.vlr.gv.at/vorarlberg/finanzen_abgaben/finanzen/landesbudget/weitereinformationen/rechnungsabschluss/rechnungsabschluss2005.htm

2 CLASSIFICATION

2.1 The Classification task

Classification can be defined as the procedure of building a mapping from a set of objects X into a set of classes C based on certain characteristics of the objects, often referred to as features or attributes.

More formally, given a model M (classifier) and a set of objects x_i with known class labels t_i , $T_1 := \{(x_i, t_i) \mid i = 1 \dots N\}$ ³ (the training set),

the task of classification consists of building a mapping $h : X \rightarrow C$ based on the values of the objects' features $\bar{\theta}_i$, which, given a new object $x_{n+1} \in X$, will produce the correct class label $t_{n+1} \in C$.

Since in general, $T_1 \subset X$ ⁴, the mapping h is an approximation of a hypothetical true mapping, subject to statistical limitations. Thus, the performance of the classifier will vary, depending on the size of the training set N , the (dis-)similarity of the elements in T_1 and in a (test) set of unseen objects T_2 ⁵ and the number and choice of the objects' features.

2.2 Classifier performance

A common measure for the performance of the classifier is the misclassification error, the proportion of wrongly classified objects. The misclassification error can be measured on the training set (the training error) or the test set. Because T_1 is limited, the training error tends to be biased towards sets of objects following a distributional structure similar to that of T_1 – hence it is also called apparent misclassification error. Since in general T_1 and T_2 will differ, the training error will be too optimistic with respect to a hypothetical true misclassification error, resulting from training on an infinite training set.

One common technique for producing a less biased measure of classification performance is (k-fold) cross-validation: (1) Divide the training set T_1 into k disjoint subsets, each holding $|T_1|/k$ objects; (2) train on $(k-1)$ training sets, and test on the remaining k -th set; (3) repeat k times, each time using a different permutation of $(k-1)$ training sets, and a different test set; (4) estimate the effective classifier performance of the classifier as the average of the k test errors.

Another common method for evaluating classification performance in the case of binary classification⁶ is plotting the ROC (receiver operating characteristics) curve which is a plot of the proportion of objects of the class c_1 correctly classified as c_1 (*tpr*, “true positives”) vs. the proportion of objects of the class c_2 wrongly classified as c_1 (*fpr*, “false positives”) for each value of *tpr*. It describes the ability of the classifier to correctly recognize objects belonging to c_1 (*sensitivity*) in relation to its (in-)ability to discriminate these objects from objects belonging to c_2 (*1-specificity*).

³ $x_i \in X, t_i \in C$

⁴ $T_1 \neq X$

⁵ In general, $T_1 \cap T_2 = \emptyset$

⁶ $C := \{c_1, c_2\}$

One threshold-independent measure contained in the ROC curve is the AUROC (area under ROC) taking values between 0 (no discrimination) and 1 (perfect discrimination), computed on the training set, or the test set.

3 RELATED WORK

A review of literature on statistical classification of landslide hazard/ susceptibility reveals that the multivariate logistic regression, a variant of the generalized linear model, is most frequently chosen. ([1], [2], [3]). It is typically employed in an automatic stepwise model selection procedure - starting with a model including all available parameters, followed by their successive elimination according to a measure of parameter (in-) significance, or a measure of goodness of fit of the model penalized by model complexity. The procedure is aimed at reducing the parameter space in order to avoid overfitting – fitting too many model parameters to the training data, where some of the features might be uninformative with respect to the class label, irrelevant for the task of discrimination, or simply noise. At the same time, it can be used as an operational method to select “significant”, “essential” or even “interesting” features.

More recently, several methods from pattern recognition and machine learning have been proposed, including linear discriminant analysis ([4], [5]), neural networks ([6], [7]), and the Support Vector Machine ([8]). These yield good results at the task, comparable to or surpassing the performance of traditional techniques.

With respect to classification, spatial data may give rise to certain problems. In the context of spatial classification, the property of spatial autocorrelation of grid points (pixels) may lead to invalid significance statements ([3]). In the case of the logistic regression, this has led to variants which explicitly model spatial autocorrelations ([9]).

Another problem of spatial data is locality of feature occurrence. Depending on the choice of the training and test set, this can make classification of new data difficult, and estimated error rates on the training set may be hardly transferable out of its spatial (and also temporal) scope, due to the occurrence of different dispositive (e.g. geology) or triggering (precipitation, earthquake) factors. This is a major reason for the objective difficulty of the task of a prediction of landslide occurrence.

4 DATA

4.1 Study area

Three areas in Vorarlberg, Austria were studied in the project “Georisikokarte” conducted by the Dept. of Applied Geology (AGK), University of Karlsruhe, in cooperation with the Federal Government of Vorarlberg ([10]), namely (Figure 1):

1. Hochtannberg (HTB), 114 km², at Vorarlberg’s eastern border,
2. Walgau (WAL) with an area of 105 km², and
3. Walsertal (WST), 147 km² in size.



Figure 1: Study areas in Vorarlberg/ Austria

4.2 Data & Pre-processing

Digital data was produced within ArcGIS, ArcInfo 9.2. A digital elevation model (DEM), topographic and geologic data and orthophotographs covering the whole area of Vorarlberg were provided by the Land Surveying Office Feldkirch (Austria).

The DEM (spatial resolution of 5 m) was used to calculate various morphometric features, including slope, curvature, slope aspect and flow accumulation. The slope was computed as the rate of maximum change in z value from each cell (i.e., the first derivative of the surface). The curvature of a slope, referring to concavity/ convexity of a surface, was computed using the CURVATURE function, with output resulting from subtracting the profile component from the planar component. For the slope aspect (slope direction), the output is the direction of maximum rate of change in z value from each grid point. The flow accumulation grid was calculated by accumulating the slope for all grid points flowing into each down-slope grid point, and was included as indicator of erosion effects.

Digital data on geology and tectonics of Vorarlberg is based on the geologic map by R. Oberhauser, Geological Federal Institution Vienna (scale 1:100.000)⁷. The vector data includes a number of classes describing the geology and line features for tectonic faults. The data was used to create a grid of 25 m grid point size. Additionally, a raster of Euclidean distances to faults was calculated, with the shortest distance to a tectonic fault assigned to every grid point within the area of investigation.

⁷ http://www.vorarlberg.at/geokatalog_internet/index.htm

The Austrian land cover database for 1990, supplied by Umweltbundesamt GmbH (Corine Landcover nomenclature: Level 2, 25 ha minimum mapping unit, comparable scale: 1:100.000), was derived from satellite data (Landsat 5, TM) by means of computer-assisted visual photo-interpretation. For Vorarlberg, it consists of 12 classes, including built-up area (4), agricultural area (3), forests and natural area (3), wetlands (1) and water surfaces (1)⁸.

In order to apply classification techniques to the study areas, training data was constructed based on results from the project "Georisikokarte". In context of the project, active landslides were mapped and an inventory of landslides for the study areas was constructed. Within the study area Hochtannberg, 107 landslides (total area of 0,79 km²) were mapped. Additional inventories were constructed for Walgau (262 slides with a total area of 0,62 km²) and Walsertal (field mapping until 2004, 761 slides with a total area of 3,54 km²).

In pre-processing, all data layers including feature layers and landslide inventories were converted into the ArcInfo ASCII grid format, with a size corresponding to the size of the corresponding study area, and a resolution of 25 m/ grid point.

5 CLASSIFICATION METHODS

In this work, four techniques were used: Logistic regression (standard and stepwise, i.e. without and with backward parameter elimination), Gaussian Process models and the SVM.

Each classifier was trained on a set of training sets $T_k = \{(x_i, t_i) \mid i=1 \dots n_k\}$, consisting of sets of grid points (pixels) x_i , each x_i having a m -dimensional feature vector ϕ_i taking on values sampled from a set of m thematic layers (ArcInfo grids)⁹, and a known class label t_i , indicating if the grid point was part of an area where landslide occurrence was registered ($t_1=1$) or not ($t_2=0$).

All calculations were performed within the statistical computing framework R¹⁰, using the additional packages *MASS*, *nnet*, and *kernlab*.

5.1 Logistic Regression

Logistic regression has been widely employed for statistical classification of landslide hazard. This is due to several desirable properties, including conceptual simplicity, straightforward interpretability in terms of probability estimation, and a decent performance at simple tasks. In addition, logistic regression readily lends itself to efficient and robust optimization schemes.

In terms of binary classification, given a set of m -dimensional feature vectors ϕ_i of grid points X_i with class labels t_i contained in the training set, the probability of X_i belonging to class c_1 is given by

⁸ <http://www.umweltbundesamt.at/umwelt/raumordnung/flaechennutzung/corine/>

⁹ see chapter 2

¹⁰ R: A language and environment for Statistical Computing, the R Development core team, R Foundation for Statistical Computing, Vienna, Austria, <http://www.r-project.com>

$$\begin{aligned}
 p(c_1 | \vec{\phi}_i) &= \frac{p(\vec{\phi}_i | c_1)p(c_1)}{p(\vec{\phi}_i | c_1)p(c_1) + p(\vec{\phi}_i | c_2)p(c_2)} \\
 &= \frac{1}{1 + e^{-\left(\ln\left(\frac{p(\vec{\phi}_i | c_1)p(c_1)}{p(\vec{\phi}_i | c_2)p(c_2)}\right)\right)}} = \frac{1}{1 + e^{-\vec{w}^T \vec{\phi}_i}} = \sigma(\vec{w}^T \vec{\phi}_i) = y(\vec{\phi}_i)
 \end{aligned}$$

with the logistic sigmoid function σ and the inner product $\mathbf{w}^T \phi_i$, defined as the logit, or log odds of x_i belonging to class c_1 , and class c_2 :

$$\vec{w}^T \vec{\phi}_i = \ln\left(\frac{p(\vec{\phi}_i | c_1)p(c_1)}{p(\vec{\phi}_i | c_2)p(c_2)}\right)$$

In order to find a maximum likelihood solution for \mathbf{w} , usually an efficient iterative higher order optimization scheme operating on the (log-) likelihood function is employed. This is true for both methods used in this work, *glm* and *multinom*, using the IRLS¹¹ and BFGS optimization schemes, respectively.

In the case of the stepwise regression procedure, the *stepAIC* method (package *MASS*) was used, which, by default, uses the AIC (Akaike Information Criterion) as criterion for model selection, defined as

$$AIC := -\ln p(\vec{t} | \vec{w}) + 2m$$

the first term being the negative logarithm of the likelihood function given the weight vector \mathbf{w} and a vector of class labels \mathbf{t} .

5.2 Gaussian Process models

Gaussian Process models are a class of non-parametric probabilistic discriminative models more recently introduced ([11], [12], [13]).

From the point of view of classification methodology, Gaussian Process models are interesting since they are expressed in the framework of Bayesian inference while allowing the use of a kernel (covariance function) for representation.

In case of classification, given a model trained on $T = \{(x_i, t_i) \mid i=1 \dots n\}$, a predictive distribution for the class label t_{n+1} given a x_{n+1} is defined by

¹¹ Iterative reweighted least squares

$$p(t_{n+1} = 1 | \vec{t}) = \int p(t_{n+1} = 1 | y_{n+1}) p(y_{n+1} | \vec{t}) dy_{n+1}$$

with $p(t_{n+1}=1|y_{n+1})=\sigma(y_{n+1})$, the sigmoid logistic with argument y_{n+1} ,

Since the posterior probability $p(y_{n+1}|\mathbf{t})$ is non-Gaussian, this integral is analytically intractable, and is evaluated in approximation. Here, the Laplace approach to obtain a Gaussian model for $p(y_{n+1}|\mathbf{t})$ is used, resulting in the expressions for the moments:

$$E(y_{n+1} | \vec{t}) = \vec{k}^T (\vec{t} - \vec{\sigma})$$

$$Var(y_{n+1} | \vec{t}) = c - \vec{k}^T (W^{-1} + C)^{-1} \vec{k}$$

with \mathbf{k} a vector of values from a positive semi-definite kernel function k evaluated on the set of tuples $\{(x_i, x_{n+1}) | i=1..n\}$, $\vec{\sigma}$ a vector with the elements $\sigma(y_i)$, c the value of k evaluated at (x_{n+1}, x_{n+1}) , W a n -dimensional diagonal matrix with elements $\sigma(y_i)(1-\sigma(y_i))$, and C the n -dimensional covariance matrix of the Gaussian Process $y(x)$, equivalent to the Gram matrix of the kernel k .

Using the above result, the integral can be evaluated, resorting to an approximation to a convolution of a logistic sigmoid with a Gaussian, yielding

$$p(t_{n+1} = 1 | \vec{t}) \cong \sigma(\kappa(Var(y_{n+1} | \vec{t}))E(y_{n+1} | \vec{t}))$$

with

$$\kappa(Var(y_{n+1} | \vec{t})) = (1 + \pi Var(y_{n+1} | \vec{t}) / 8)^{-\frac{1}{2}}$$

(c.f.[14]).

Gaussian Process-based regression is reported to yield good performance on different tasks from the prediction of chemical properties of molecules in drug design to Wi-Fi localization. In the area of geostatistics, an equivalent form known as kriging ([15]) is used extensively for interpolation.

In R, the Gaussian Process model for regression and classification is implemented in the *gausspr* method, which is part of *kernlab* ([16]). In this work, the default Gaussian kernel *rbfdot* was used, in the functional form

$$k(\vec{x}, \vec{x}') = e^{(-\|\vec{x}-\vec{x}'\|^2 / 2\sigma^2)}$$

with the hyperparameter σ , which was estimated using the *sigest* estimation routine available in *kernlab*.

5.3 Support Vector Machines

The Support Vector machine ([17], [18]) has emerged in the 1990's as a novel method for classification and regression. Like the Gaussian Process models, it belongs to the class of kernel methods, i.e., makes use of a symmetric positive semi-definite kernel $k(\mathbf{x}, \mathbf{x}')$, evaluated at the elements of the training set.

The design of the SVM lends the model several properties making it an excellent tool for classification. One such property is that for the SVM, determination of model parameters (training) takes form of convex optimization. Thus, the training procedure is guaranteed not to produce a suboptimal solution wrt the global optimum¹².

An important distinction of the SVM to the simple linear model is the introduction of the concept of the margin. The SVM training procedure is aimed at this quantity, defined as the perpendicular distance between the decision boundary and the closest of the data points. Maximizing the margin leads to a particular choice of decision boundary, with a maximal separation of the classes. This property lends the model improved generalization capabilities with respect to unseen data.

Another property resulting in better generalization capabilities of the SVM is the adaption of a "soft" margin. With the soft margin property, the SVM is allowed to make errors in classification on the training set. Due to this property, the SVM is less sensitive to the presence of outliers in the training set, hence less prone to overfitting.

In the dual (kernel) formulation, the definition of the SVM is

$$y(\vec{x}) = \sum_{n=1}^N a_n t_n k(\vec{x}, \vec{x}_n) + b$$

with Lagrange multipliers a_n resulting from a quadratic optimization of the term

$$\tilde{L}(\vec{a}) = \sum_{n=1}^N a_n - \frac{1}{2} \sum_{n=1}^N \sum_{m=1}^N a_n a_m t_n t_m k(\vec{x}_n, \vec{x}_m)$$

s.t.

$$0 \leq a_n \leq C$$

$$\sum_{n=1}^N a_n t_n = 0$$

¹² This is not the case with all classifiers, e.g. artificial neural networks, which are notoriously subject to the problem of falling into local minima of the error function.

The dual formulation of the SVM introduces the kernel function, which can be interpreted as the inner product of the inputs in a feature space. Since it does not require an explicit definition of the mapping into feature space for each pair of arguments (x_i, x_i') , it is possible to substitute the inner product kernel, defined as

$$k(\vec{x}, \vec{x}') = \vec{\phi}(\vec{x})^T \vec{\phi}(\vec{x}')$$

with another symmetric, positive semi-definite kernel k to separate the data in some (possibly high or even infinite dimensional) induced space¹³.

The direct solution of a quadratic programming problem is computationally demanding. For the above quadratic optimization problem, efficient procedures like the SMO procedure ([19]) have been developed. In practice, the procedure is found to have a scaling with N between linear and quadratic.

A suitable model for the prediction of class probabilities can be obtained by fitting a logistic sigmoid to the output of a previously trained SVM. This method is based on an idea due to ([20]).

Within R, several packages implement SVM. In this work, the *ksvm* method in package *kernelab* was used, in the described variant of SVM (C-SVM) and the Gaussian kernel *rbfdot*. As in the case of Gaussian Process models, the hyperparameter σ was estimated using *sigest*.

6 RESULTS

6.1 Pre-processing

Four classifiers were compared on four independent data sets, drawing on methods described in chapter 5. In each case, the model was trained on a balanced set of 50% landslide and 50% non-landslide grid points, and subject to 5-fold cross-validation.

In order to determine the influence of training set size on classification performance and stability, training sets of decreasing size were used, corresponding to fractions of the full data set size.

For each dataset, a set of m thematic layers was used including morphometric data (elevation, slope, curvature, and aspect), geology, vegetation, flow accumulation and Euclidean distance to faults. The layers were derived from a DEM, a geologic, and a geotechnical map using standard GIS functionality. In course of further pre-processing, real-valued features (elevation, slope, curvature, flow accumulation, distance to faults) were standardized. Non-continuous valued features (slope orientation, geology, and vegetation) were transformed to a number of binary features equivalent to the number of values assumed on the dataset.

In the case of the third study area (WST), the data (>11000 landslide/ non-landslide pixels), was split into two (roughly) equal sized datasets corresponding to the north and the south part of the study area in order to obtain training sets of comparable size.

Table 1 summarizes the size of each training set, the number of features and the overall size of the study area.

¹³ This is often referred to as the kernel trick.

Study area	Size of dataset (pixels)	# of Features	Landslide area (pixels)	Non-landslide area (pixels)	Landslide/non-landslide pixels
HTB	2400	59	1200	176277	0.0067
WAL	1960	61	980	166235	0.0058
WST_N	4950	60	2475	117592	0.0206
WST_S	6320	51	3160	110938	0.0276

Table 1: Training data

6.2 Classification performance

A summary of the results (Tables 2-5) shows that the kernel-based classifiers consistently outperform logistic regression (both with and without parameter elimination). This is true for all combinations of study area and training set size. The SVM shows overall best results, achieving a classification performance over 80% in three out of four cases, performing best on the third dataset with full training set size (84.5%).

Gaussian Process models turn out competitive, with classification performance of less than 2% worse than the SVM on average.

Conversely, results of the logistic regression in both variants set the method apart from the SVM and Gaussian Process models. The difference in performance ranges from 4.5% in the best case (dataset 2) up to 12.2% on the first dataset with full training set size.

Reducing the size of the training set seems to have less of an effect on the classification performance here, with relatively constant results except for a notable decrease in case of dataset 2. For dataset 1 (HTB), a comparison of training error vs. test error for training sets of decreasing size shows signs of overfitting on the part of logistic regression, with a divergence of error rates on training (lower error) and test set. A corresponding trend, albeit less distinct, can be observed in the case of the stepwise variant.

In part, the inferior performance of the logistic regression might be considered a result of the relative simplicity of the model, which might be interpreted as a “soft” version of a threshold-based discriminant $\mathbf{w}^T \boldsymbol{\phi}$ in the logit space. By definition of the logit:

$$w^T \vec{\phi}_i = \ln \frac{p(\vec{\phi}_i | C_1)p(C_1)}{p(\vec{\phi}_i | C_2)p(C_2)} = \sum_{j=1}^M \ln \frac{p(\phi_{ij} | C_1)p(C_1)}{p(\phi_{ij} | C_2)p(C_2)}$$

the logit space might appear as a natural feature space for the task of probabilistic discrimination. However, the relative simplicity might give rise to certain problems.

First off, a linear separation based on logit threshold might not exist in the (one-dimensional) logit space, which might turn out sub-optimal for discrimination, e.g. due to outliers in the likelihood, $p(\boldsymbol{\phi}|C_k)$.

Another source of error might be introduced by the model's assumption of independence of $p(\phi_{ij}|C_k)$ and $(\phi_{i,j'}|C_k)$, for j and j' ¹⁴. This assumption does not hold in our case, with at least one example of (non-linear) functional dependence for the features (the morphometric parameters).

Another assumption of independence is implicit in the formulation of the likelihood function, defined as

$$p(\vec{t} | \vec{w}) = \prod_{i=1}^N y_i^{t_i} (1 - y_i)^{1-t_i}$$

In this case, independence of the grid points is assumed for i, i' ¹⁵. However, this assumption might not hold in the presence of spatial correlation, resulting in an approximation to the surface subject to optimization.

Finally, an examination of the logistic regression model in the standard variant reveals missing regularization. As a consequence, the method is more prone to overfitting, which is confirmed by the results.

Classifier	Test error [%]				
	Training	1920	960	480	240
	sample size [pixel]				
Glm		33.5	34.5	35.1	35.4
Glm_stepAIC		32.3	33.5	34.4	35.2
Gausspr		23.2	25.2	27.7	29.2
Ksvm		21.3	23.8	25.5	27.6

Table 2: Classification performance for study area 1 (HTB)

Classifier	Test error [%]				
	Training	1568	784	392	194
	sample size [pixel]				
Glm		24.3	24.8	25.1	30.2
Glm_stepAIC		24.1	24.6	25.8	31.4
Gausspr		21.8	22.3	23.6	25.4
Ksvm		19.8	21.2	21.7	25.7

Table 3: Classification performance for study area 2 (WAL)

¹⁴ $j, j' \in \{1 \dots M\}, j \neq j'$

¹⁵ $i, i' \in \{1 \dots N\}, i \neq i'$

Classifier	Test error [%]				
	Training sample size [pixel]	3960	1980	990	495
Glm		25.2	25.6	25.2	26.4
Glm_stepAIC		25.3	25.6	25.3*	27.1
Gausspr		17.5	19.7	21.2	24.7
Ksvm		15.5	19.8	21.9	24.3

Table 4: Classification performance for study area 3a (WST_N)

Classifier	Test error [%]				
	Training sample size [pixel]	5056	2528	1264	632
Glm		25.6	25.8	26.3	26.1*
Glm_stepAIC		26	26	26.6	26.8
Gausspr		20.8	21.8	22.8	24.3
Ksvm		19.9	20.7	21.8	22.8

Table 5: Classification performance for study area 3b (WST_S)

6.3 ROC curves

Using the *ROCR* package, ROC curves were drawn for each combination of study area and classifier for the full training set size. Fig. 1-4 show a set of plots for the four classifiers for the study area HTB. AUROC values are given in Table 7, reflecting performance in Table 2 (first column). Visual inspection of Fig. 1-4 and a comparison of the AUROC for the four classifiers confirm previous observations, with the kernel methods emerging as better models in terms of classification performance on unseen test data.

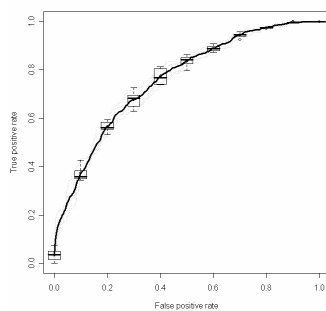


Fig. 2: HTB/ glm

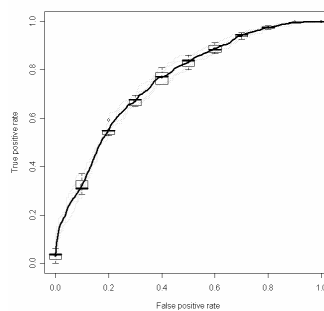


Fig. 3: HTB/ glm_stepAIC

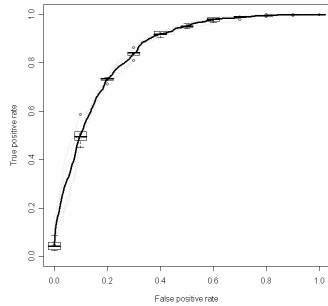


Fig. 4: HTB/ gausspr

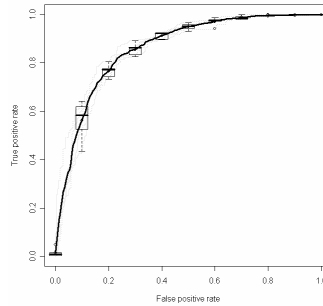


Fig. 5: HTB/ ksvm

Classifier	AUROC on test set [%]					Avg.
	Run 1	Run 2	Run 3	Run 4	Run 5	
Glm	76.8	74.3	76.3	77.3	74.4	75.8
Glm_stepAIC	76.3	74.2	76.1	76.0	74	75.3
Gausspr	86.5	84.6	84.4	85.1	83.3	84.9
Ksvm	88.4	86.3	87.5	85.9	85.8	86.8

Table 7: AUROC for different classifiers on study area 1 (HTB)/ full dataset

Hazard maps

After classification, the trained models were used to generate hazard maps for each study area. For this task, all available training data (i.e., all 5 subsets) was used, and the trained model was applied to the whole study area, containing the total number of pixels (Table 1). Hazard maps were generated as achromatic 8 bit raster images, the value of each pixel being the output of a sigmoid (in the case of the logistic regression) or the output of a sigmoid fitted to the output of the respective model, scaled by the maximum 8 bit value. To facilitate post-processing using GIS functionalities, each map was exported in a number of raster formats, including PNG, TIFF, and ArcInfo ASCII Grid using the *rgdal* package. Fig. 6-8 show (inverted) hazard maps resulting from ArcGIS post-processing of *rgdal* output based on results generated for HTB, WAL and WST by the (best-performing) SVM model.

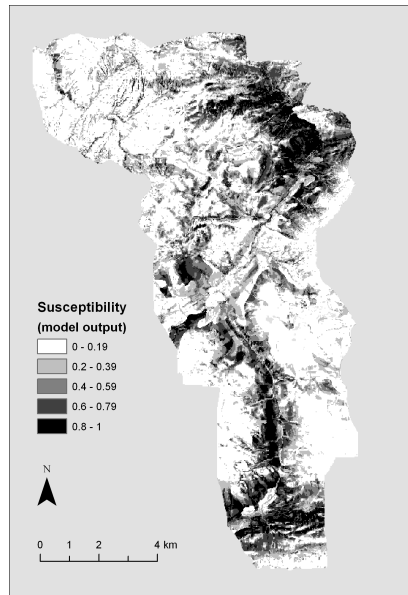


Figure 6: post-processed hazard map (SVM), study area HTB

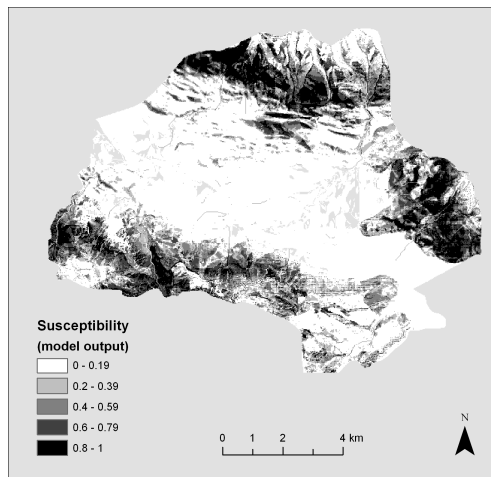


Figure 7: post-processed hazard map (SVM), study area WAL

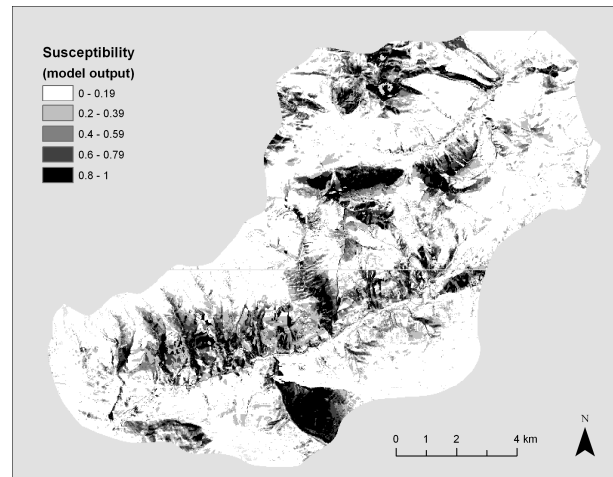


Figure 8: post-processed hazard map (SVM), study area WST

7 CONCLUSIONS/ OUTLOOK

In this work, several different classification techniques were applied to the task of an assessment of landslide hazard/ susceptibility. Comparison of classification performance on unseen test data shows that kernel methods (Gaussian Process model and the SVM) consistently outperform logistic regression, with misclassification errors for independent test datasets and different test set sizes hinting at robustness of the result.

The inferior performance of the logistic regression model can be explained by the relative simplicity of the model, which does not reflect (non-linear) dependencies inherent in the data. In addition, the lack of a built-in regularization mechanism on the part of the logistic regression hints at inferior generalization capabilities.

In contrast to the logistic regression, both kernel methods have the property of explicitly modelling a correlation between the data points by means of the kernel/ covariance function. Also, both kernel methods exhibit generalization properties superior to logistic regression in both variants.

SVM in particular are known to have improved generalization capabilities by design, resulting from the maximization of margin and the “soft” margin property. The Gaussian Process model is regularized by the Gaussian Process prior, a Gaussian with mean $\mathbf{0}$ and covariance function k .

Our interpretation of the results suggests several opportunities for improvement departing from the logistic regression model. Generalized linear mixed models/ hierarchical models considering effects present in subsets of the data present a viable option. As an alternative, a systematic investigation of the potential of kernel methods with respect to improvements at our task – based on a deeper understanding of the nature of spatial data – might offer opportunities for further research.

Acknowledgements

The authors thank Michael Ruff, Walter Bauer and Peter Drexel (Land Surveying Office Feldkirch) for providing the landslide inventories and thematic data for Vorarlberg.

This work was partly funded by the German Federal Ministry of Education and Research in the project “development of suitable early warning systems” (EGIFF), a part of the GEOTECHNOLOGIEN program.

REFERENCES

1. Atkinson, P.M., and Massari, R., *Generalized linear modelling of susceptibility to landslides in the Central Apennines, Italy*. Computers and Geosciences, 1998. **24**(4).
2. Gorsevski, P.V., Gessler, P.E., and Foltz, R.B., *Spatial prediction of landslide hazard using logistic regression and GIS*, in *4th International Conference on Integrating GIS and Environmental Modeling (GIS/EM4)*. 2000: Banff, Alberta, Canada.
3. Ohlmacher, G.C., and Davis, J. C., *Using multiple logistic regression and GIS technology to predict landslide hazard in Northeast Kansas, USA*. Engineering Geology, 2003. **69**(3-4).
4. Gorsevski, P.V., Gessler, P.E., and Foltz, R.B., *Spatial prediction of landslide hazard using discriminant analysis and GIS*, in *GIS in the Rockies 2000 Conference and Workshop*. 2000: Denver, Colorado, USA.
5. Santacana, N., Baeza, B., Corominas, J., de Paz, A., and Marturia, J., *A GIS-based multivariate statistical analysis for shallow landslide susceptibility mapping in La Pobla de Lillet area (Eastern Pyrenees, Spain)*. Natural Hazards, 2003. **30**.
6. Lee, S., Ryu, J.-H., Min, K., and Won, J.-S., *Landslide susceptibility analysis using GIS and artificial neural network*. Earth Surface Processes and Landforms, 2003. **28**(12).
7. Ermini, L., Catani, F., and Casagli, N., *Artificial Neural Networks applied to landslide susceptibility assessment*. Geomorphology, 2005. **66**(1-4).
8. Brenning, A., *Spatial prediction models for landslide hazards: review, comparison and evaluation*. Natural Hazards and Earth System Sciences, 2005. **5**(6).
9. Gotway, C.A., and Stroup, W. W., *A generalized linear model approach to spatial data analysis and prediction*. Journal of Agricultural, Biological, and Environmental Statistics, 1997. **2**(2).
10. Ruff, M., Kühn, M., and Czurda, K. *Risikoanalyse für Massenbewegungen in den Ostalpen (Vorarlberg)*. in *15. Tagung Ingenieurgeologie*. 2005. Erlangen, Germany.
11. Williams, C.K.I., and Rasmussen, C.E. *Gaussian Processes for regression*. in *Neural Information Processing Systems*. 1995. Denver, Colorado, USA: MIT Press.
12. Williams, C.K.I., and Barber, D., *Bayesian Classification with Gaussian Processes*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998. **20**(12).
13. MacKay, D.J.C., *Introduction to Gaussian Processes*, in *Neural Networks and Machine Learning*, C.M. Bishop, Editor. 1998, Springer.
14. Bishop, C.M., *Pattern Recognition and Machine Learning*. Information Science and Statistics, ed. M. Jordan, Kleinberg, J., Schoelkopf, B. 2006: Springer.
15. Cressie, N.A.C., *Statistics for spatial data*. 1993, New York: John Wiley & Sons.
16. Karatzoglou, A., Smola J., Hornik, K., and Zeileis, A., *kernlab – An S4 Package for Kernel Methods in R*. Journal of Statistical Software, 2004. **11**(9).
17. Cortes, C., and Vapnik, V., *Support-Vector Networks*. Machine Learning, 1995. **20**(3).
18. Vapnik, V., *Statistical Learning Theory*. 1998, New York: John Wiley and Sons.
19. Platt, J.C., *Fast training of support vector machines using sequential minimal optimization*, in *Advances in Kernel Methods - Support Vector Learning*, B. Schoelkopf, Burges, C., and Smola, A., Editor. 1999, MIT Press: Cambridge.

20. Platt, J.C. *Probabilistic outputs for Support Vector Machines and comparison to regularized likelihood methods.* in *Advances in Large-Margin Classifiers*. 2000. Cambridge, Massachusetts, USA: MIT Press.