Assessment of Shallow Landslide Initiation Areas Using stochastic Modelling: The Vernazza Torrent Case Study, Liguria, Italy

Elmar Schmaltz[1], Hans-Joachim Rosner[2] and Michael Märker[1]

[1]Heidelberg Academy of Science and Humanities, Germany
elmar.schmaltz@student.uni-tuebingen.de
[2]Institute of Geography, University of Tübingen, Germany

Full paper double blind review

Abstract

The objective of this study is the assessment of potential failure zones of landslides in unstable areas. For this purpose, two different stochastic classification models were used: A boosted decision tree approach with TreeNet (TN), and a bagging decision tree approach with Random Forests (RF). Both topographic and soil parameters were considered as predictor variables for training and testing the models. We assume that several predictor variables will lead to misclassification and incorrectness, especially soil parameters. Hence, the misclassification of these particular predictors should be avoided, using the strategy of tree boosting. The investigated area is the hydrological basin of Vernazza in Cinque Terre, Northwest Italy. A disastrous flash flood on the 25th of October 2011 with numerous landslides caused fatalities and economic losses amounting to millions of Euros. We mapped landslide areas in the field and checked the resulting maps with high resolution remote sensing images. Furthermore, the relevant soil parameters were collected based on a geostatistical approach. We measured topographic parameters, and physical and hydrological soil characteristics such as maximum shear strength under saturated and unsaturated conditions, and hydraulic conductivity (Ksat), and attributed random points in three distinguished classes: i) initiation areas, representing the most likely failure areas for possible landslides, ii) transport areas which were considered as a mix of classes 1 and 3, and iii) stable areas, such as valley bottom, ridges, and unconditionally stable areas. We ran both models with a training dataset (0.8 of the total points Ntot) and a test dataset (0.2 of Ntot) and each with 2000 grown decision trees. We validated the models with a Receiver Operating Characteristic (ROC) curve integral. The regionalized results of the TreeNet dataset yielded potential susceptible landslide areas of a total area of 1.74 km², which is 29.74% of the total area. In contrast, the Random Forests model classified a much greater susceptible area (84.27% of the total area). The results show that Treenet is outperforming RF. The latter misclassifies especially the soil related variables, whereas TreeNet yields robust model results.

1 Introduction

GIS-based stochastic modelling approaches have successfully been used in several environmental studies like erosion prediction (MÄRKER et al. 2011, SIDORCHUK 2005, MEI et al. 2008), landscape reconstruction (VOGEL & MÄRKER 2010, CASTILLA-RHO et al. 2014,

 ISBN 978-3-87907-558-4.
 ISSN 2308-1708, doi:10.1553/giscience2015s61.

SCHMALTZ et al. 2015), and landslide assessment (VORPAHL et al. 2012). Especially for determining the assessment of slope stability or landslide susceptibility, different approaches can be applied. These are basically: expert knowledge based approaches, statistical models, non-deterministic models, and mechanical approaches (WU et al. 2014). To overcome the high level of subjectivity from expert evaluation, quantitative or semi-quantitative methods were developed (GHOSH et al. 2011, WU et al. 2014). Particularly, simple and multivariate statistical methods have been applied successfully to assess and evaluate landslide hazards, (BERNKNOPF et al. 1988, DIEU et al. 2011). To obtain accurate results in landslide assessment, a large amount of information concerning topography, soils, climate, and vegetation is required. In order to handle this amount of information, Geographical Information Systems (GIS) are used, improving the quality of a spatial landslide susceptibility assessment (WU et al. 2014, DIKAU et al. 1996, CARRARA & GUZZETTI 1995). Statistical and machine-learning methodologies have made a huge progress since the last decades in geoinformatics and GIS-based analyses (MÄRKER et al. 2011). Several methods such as logistic regression (HOSMER & LEMESHOW 2000), artificial neural networks (KOHONEN 1984), and classification and regression trees (BREIMAN et al. 1984, DE'ATH & FABRICIUS 2000) have been applied in a wide range of geomorphologic studies in the past (MOORE et al. 1993, GESSLER et al. 1995, PARUELO & TOMASEL 1997, MERTENS et al. 2002, BRENNING 2005, GRIMM et al. 2008). Certainly, the application of several modelling approaches mentioned above might yield good results for shallow landslide assessments. However, the differentiation and determination of delimited landslide initiation areas, as an input parameter for stochastic prediction using classification trees, have not been applied yet. We applied the statistical methods in this study to investigate the hydrological basin of Vernazza, a medieval village of the Cinque Terre, situated at the coastline of the Mediterranean Sea in eastern Liguria, Northwest Italy. The popular touristic village was affected by a disastrous flash flood event on the 25th of October 2011, which

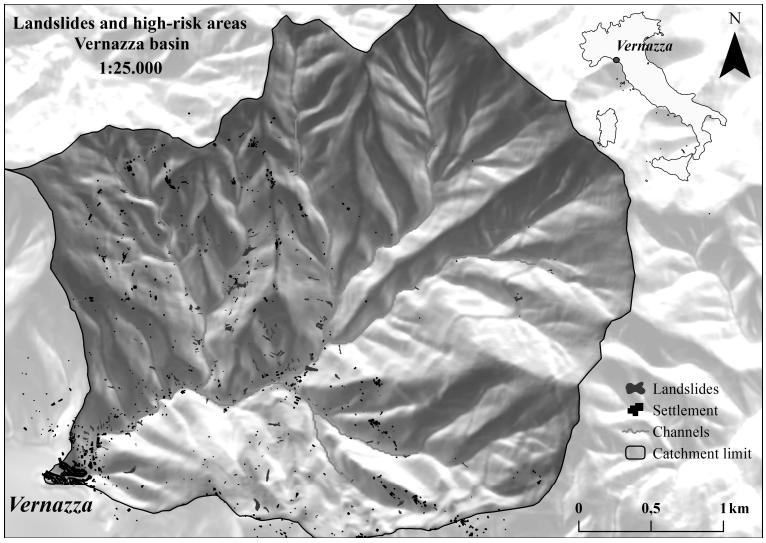

Fig. 1: Study area of Vernazza basin with the mapped shallow landslides that occurred at or after the 25th of October 2011 flash flood event

caused significant damages, including seven fatalities and an economic loss amounting to millions of Euros. A specific problem in Cinque Terre is the changing land use over the last decades, such as the abandonment of formerly cultivated vineyard terraces, caused by much higher income potentials from the tourism and the service sectors.

As shown in Fig. 1, landslides occur primarily in abandoned vineyard terraces with a typical coverage of shrubs and bushes. Moreover, landslides are found on geological formations with sandstone, linearly distributed in the centre of the basin from NNW to SSE.

The objective of this study is the implementation of physical and hydro-pedological soil data in stochastic models, to obtain additional information about susceptible landslide areas. We concentrated on the landslide initiation areas characterized by the failure edges of the mapped landslides that occurred since the 25th of October 2011. We compare two stochastic modelling approaches and focus especially on their performance using topographic predictors, as well as with physically and hydro-pedologic predictors.

2 Input Data and Pre-Processing

The collection and pre-processing of the data was performed in the Vernazza study area during field trips in March, August, and October 2014. Land use conditions and landslide areas were mapped, both in the field, and with remote sensing data based on World View and Geo Eye images. We also measured physical and hydro-pedological soil parameters. Saturated hydraulic conductivity – Ksat – was measured in 25cm, 50cm, and 100cm depth with a compact constant head permeameter (Amoozemeter; AMOOZEGAR 1989). Additionally, we conducted infiltrability measurements with a Hood-Infiltrometer IL-2700. To estimate the maximum shear resistance under saturated and unsaturated conditions, shear strength measurements were taken with a light Torvane shear strength device. To acquire supporting information about the pedogenic properties of the soil, samples were collected, and analysed considering texture, skeleton, pH, and water content. In total, 33 locations were sampled.

The physical and hydro-pedological data were pre-processed with ArcGIS, SAGA, and R to use them as input data for the stochastic modelling. To obtain a proper regionalization of the restricted hydro-pedological dataset, it is necessary to select a suitable interpolation method. The dataset consists of 41 spatial data points for the infiltration and shear stress at the topsoil (≤25cm), respectively, 33 for a depth of 50cm and 100cm. We applied a multivariate geostatistical interpolation method. Due to the very high diversity of the vegetation coverage and the geological setting, as well as the low quantity of spatial data points, mathematical interpolation methods such as Inverse Distance Weighting (IDW), cannot reproduce the natural conditions in this case. For this reason, Cokriging was used to obtain the best possible estimation of the infiltrability, shear stress, and saturated hydraulic conductivity, considering the land use and the geology. The correlation between the principal variable of interest and other, more easily measured auxiliary variables is the basis of the Cokriging technique (SHAHROKHNIA et al. 2004, ELDEIRY & GARCIA 2009, ODEH et al. 1995). A weighted spherical based model was applied to obtain the best fit of the Cokriging input data (Fig. 2). However, Cokriging is a quite sensitive method when the quantity of the sample point population is fairly low. Therefore, we performed an iterative cross-validation by not including 5% of the data set for each depth in every iteration. This was performed

for three cases of the Cokriging assumption in the interpolation process, considering: i) land use and geology, ii) only land use, iii) only geology. The combination of land use and geology yielded the best interpolation results for both parameters, Ksat for all three depths and shear stress (Table 1). The input of other parameters, like complex topographic indices, yielded non-satisfying interpolation results.

Table 1: Cross-validation results for the Cokriging interpolation performance of saturated hydraulic conductivity (Ksat) and shear stress under unsaturated (unsat.) and saturated (sat.) conditions in all three investigated depths

	Interpolation performance after cross-validation (R^2)								
	Ksat			**Shear stress unsat.**			**Shear stress sat.**		
Assumption / Input	***25***	***50***	***100***	***25***	***50***	***100***	***25***	***50***	***100***
Land use & geology	0,93	0,98	0,91	0,95	0,95	0,93	0,91	0,88	0,90
Land use	0,72	0,60	0,32	0,55	0,49	0,52	0,68	0,72	0,66
Geology	0,37	0,45	0,89	0,56	0,42	0,77	0,69	0,80	0,75

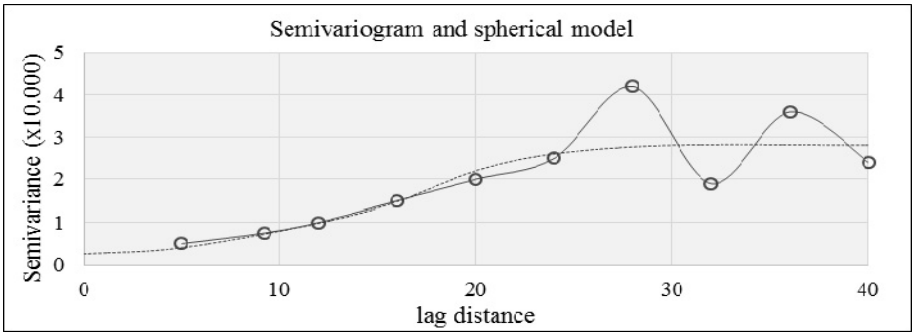

Fig. 2: Example of semivariogram from spherical model applied on maximum shear strength point dataset for 100cm depth under saturated conditions

A Topographic Wetness Index (TWI), calculated with BEVEN & KIRKBY'S (1979) equation used in TOPMODEL, as well as land use and geology were applied as input grids. Finally, nine different grids with infiltrability (for the topsoil ≤25cm), saturated hydraulic conductivity (for measurements deeper than 50cm) and maximum shear strength for each depth under saturated and unsaturated conditions were calculated with the gstat-package implemented in R.

The topographic data were derived from a Digital Elevation Model (DEM) with 5m resolution in SAGA. Certain relief variables describing the slope stability, such as elevation, slope, plan, and profile curvature were implemented as input parameters in the statistical prediction.

We distinguished the landslide area in three different classes: i) "initiation" class, which represents the areas most prone to failure, ii) "slipping" class, which can be considered as

areas with mixed conditions of class 1 and 3, and iii) deposition areas, stable and moderate stable areas, as well as inconsistently stable areas. For each class, 200 random points were created according to the several class areas (In total: Ntot=600). The derived topographic parameters and regionalized soil parameters were attributed to these prediction points.

To avoid a doubled inclusion in the performance of the model, the TWI was excluded from the training dataset, since it was used in the geostatistical interpolation of the soil parameters. However, land use and geology classification were included, to analyse the effect on model performance.

3 Methodology

Applications of classification and regression trees can be distinguished in two well-known methods (LIAW & WIENER 2002): boosting (SHAPIRE et al. 1998) and bagging (BREIMAN 1996). With boosting, successive trees give an extra weight to misclassified or incorrectly predicted variables. Finally, a weighted vote is taken for the predicted points. In contrast, successive trees do not depend on earlier grown trees in the bagging approach. Finally, a simple majority vote is taken for prediction (LIAW & WIENER 2002).

3.1 Random Forests

Random Forests append an additional layer of randomness to bagging. In addition, Random Forests change how the classification or regression trees are constructed. However, in Random Forests, each node is split using the best among a subset of predictor variables, randomly chosen at the respective node (BREIMAN 2001, LIAW & WIENER 2002). Compared to many other classification routines – including discriminant analysis, support vector machines, and neural networks – this method turns out to perform well and is robust against overfitting (BREIMAN 2001). The error ratios of the training data can be obtained by "out-of-bag" data (OOB) by predicting the data, which is not in the bootstrap sample using the grown tree within the bootstrap sample. Further, the OOB is aggregated as an estimation of the error rate (BREIMAN 2002). Variable importance is hard to define accurately, due to the interaction of possibly important variables. Random Forests estimates the variable importance by evaluating the quantification of misclassified or incorrect predictors and the increase of the prediction error of the respective variable.

For the two models (train and test) performed with Random Forests, we use the following settings: random separation of the entire dataset (Ntot=600) into the training fraction and the test fraction (Ntrain=480, 0.8 of Ntot, and Ntest=120, 0.2 of Ntot) with 2000 as a maximum number of trees.

3.2 Boosted Decision Trees (TreeNet)

The second method, TreeNet (TN), is also based on classification trees, but uses a stochastic gradient boosting technique (Salford Systems implementation: TN, cf. FRIEDMAN 1999, also called boosted regression trees: ELITH et al. 2008). The method of boosted decision trees applied in TreeNet is based on FRIEDMAN'S stochastic gradient boosting (FRIEDMAN

1999). Gradient boosting constructs additive regression models by sequentially fitting a simple parameterized function to current 'pseudo' residuals by least squares for each iteration (VOGEL & MÄRKER 2010). The pseudo residuals are the gradient of the loss function being minimized, with respect to the model values at each training data point, evaluated at the current step (FRIEDMAN 1999). Practically, the method derives several hundreds to thousands of small trees. Each of the small trees typically contains six nodes, which were also used in our study. Each tree is devoted to contributing a small portion of the overall model, whereas the final model prediction is constructed by adding up each of the individual tree contributions. A big advantage of this methodology is its robustness against data errors in the input variables. Especially for our study, since the soil parameters are supposed to be heterogeneous and might lead to misclassifications.

Similar to the Random Forest approach, two models with the following input parameters and settings were performed: The entire dataset (Ntot=600) used for the modelling was separated into a training fraction and a test fraction (Ntrain=480, 0.8 of Ntot, and Ntest=120, 0.2 of Ntot). The separation between train and test data was performed by random selection. The maximum number of trees to use was set to 2000. The TreeNet model uses a regression model with the Huber-M loss function.

4 Results and Discussion

We found significant differences between the performances of the two compared models RF and TN with the predicted variables. RF yielded an average balanced error rate (OOC) of misclassifications of 0.4923 in all three classes, respectively 49.23%. In contrast, 54.61% of the total point data was successfully predicted.

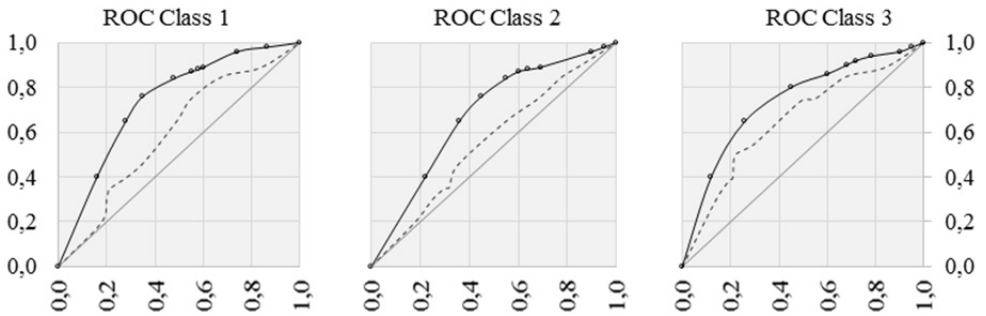

Fig. 3: ROC validation of each class for the train (black, continuous) and test (red, dashed) model in RF

The gains in the ROC validation demonstrate the much lower robustness of Random Forests compared with the training and test dataset performed with TreeNet (Fig. 3 and 4). Even though the training dataset in TN yields good results with a prediction success of 0.94 (class I), 0.87 (class 2), and 0.91 (class 3), the test data does not show this high accuracy in the ROC intervals (Fig. 4). This can be explained with the low number of point data in the test dataset (40 for each class). The test data tended to overfit after 365 trees.

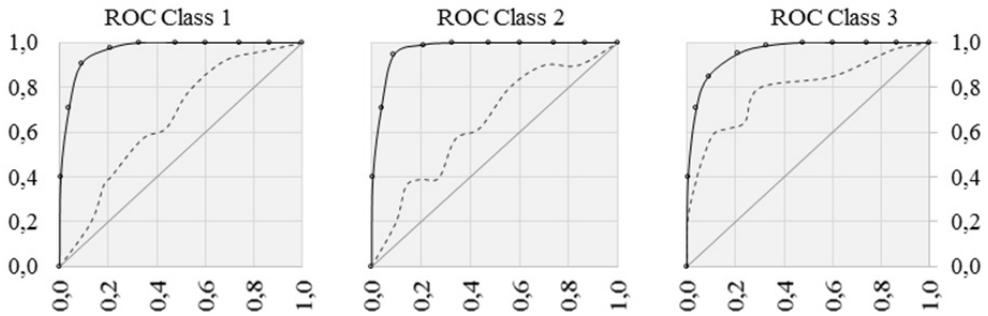

Fig. 4: ROC validation of each class for the train (black, continuous) and test (red, dashed) model in TN

Considering the variable importance (Fig. 5), substantial differences of predictors in TN and RF can be distinguished. In RF, soil parameters do not play an elementary role in the decision process of the model. In contrary, in TN the variable importance of soil parameters like MSS100_sat and MSS50_sat (maximum shear strength under saturated conditions in a depths of 100cm, respectively 50cm) is significantly higher. As expected, geology and land use are not important for both models.

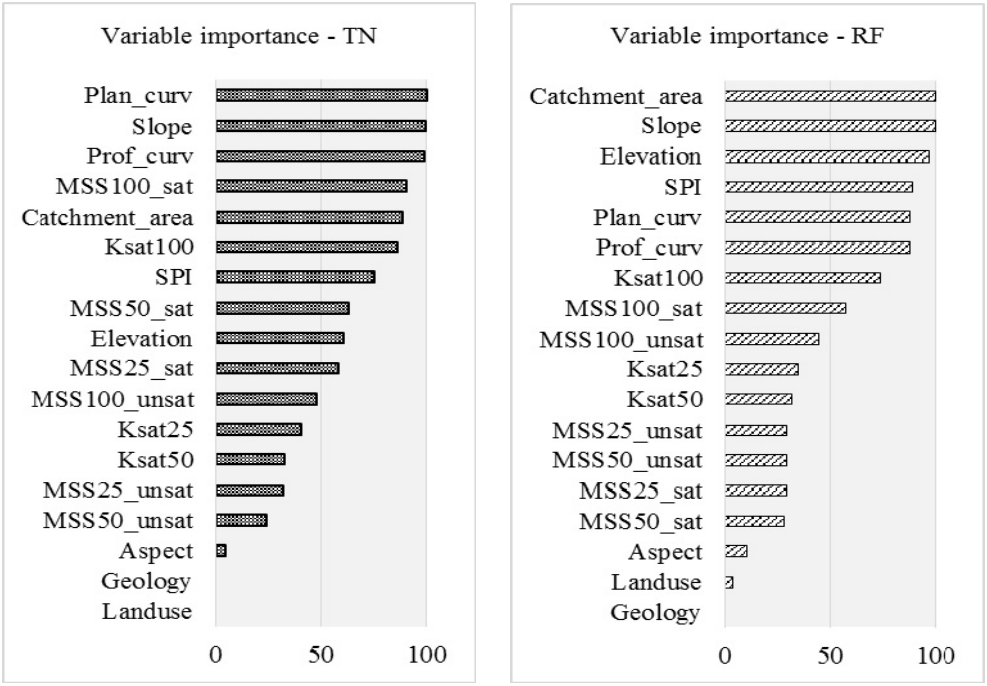

Fig. 5: Comparison of variable importance between TN and RF

After regionalization of the predicted point values, the dataset was interpolated and quantified in GIS. Table 2 shows the quantification of the several predicted areas according to the total area of the basin.

Table 2: Comparison of areal quantification of TN and RF results after regionalization

Model/issue	Class	Area [km²]	Percentage of total area (~5.8km²)
TreeNet	Initiation	1.74	29.74%
	Slipping	0.59	10.09%
	Stable	3.52	60.17%
Random Forests	Initiation	4.93	84.27%
	Slipping	0.84	14.36%
	Stable	0.08	1.37%
Landslides	–	0.03	0.51%

According to the high variable importance of topographic parameters and underrating of soil parameters, steep slopes appear as high-risk (initiation) areas. Even ridges appear as slipping areas in the RF model.

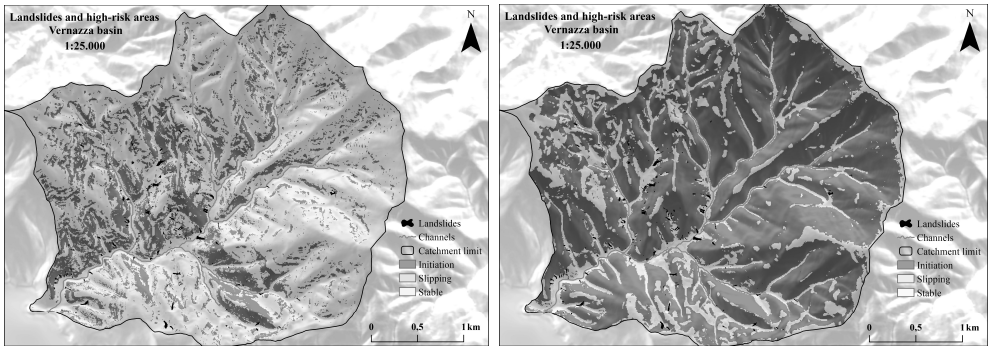

Fig. 6: TN (left) and RF (right) results with initiation, slipping, and stable areas

The linear structure of the sandstone lithology imprints through the predicted high-risk (initiation) areas from TreeNet and is well noticeable. Forested areas in the North-eastern half of the catchment are less hazardous, which fits with the low number of mapped landslides in forest areas. Even though RF yields the greatest areas of initiation, not all landslide initiation edges are within class 1 (87.4%). On the contrary, all initiation edges of the mapped landslides are located within class 1 areas of the TN model. The results of the TreeNet model show that potential failure zones are good and significantly representable in GIS with certain soil parameters as predictor variables.

5 Conclusion

In conclusion, it could be shown that: firstly, boosted decision trees are preferable to a Random Forest approach for the classification of potential failure spots in a catchment within unstable areas. In the range of the predictor variables we used, soil parameters as predictors have the tendency to be misclassified. Boosting avoids the misclassification of these predictors and yields significantly better results than unboosted approaches like Random Forests. Secondly, the boosted decision tree method with TreeNet has appeared as a robust approach for the quantification of hazardous slope areas of a landslide endangered basin, combining topographic parameters as well as soil parameters as predictor variables for the model. Several physical and hydro-pedological input parameters can provide information about soil saturation and soil failure under certain conditions in different depths. Hence, it could be a next step to simulate scenarios with different triggering parameters such as saturation and shear conditions considering rainfall or seismic activity. Thirdly, the combination of fieldwork for point analysis and the application of GIS in the field of quantifying hazardous slope areas encourages us to regionalize theses point patterns. This enables us to find areas with a high risk of landslide occurrence.

References

AMOOZEGAR, A. (1989), A compact constant-head permeameter for measuring saturated hydraulic conductivity of the vadose zone. Soil Science Society of America Journal, 53, 1356-1361.

BERNKNOPF, R. L., CAMPBELL, R. H., BROOKSHIRE, D. S. & SHAPIRO, C. D. (1988), A probabilistic approach to landslide hazard mapping in Cincinnati, Ohio, with applications for economic evaluation. Bull. Assoc. Eng. Geol., 25, 39-56.

BEVEN, K. J. & KIRKBY, M. J. (1979): A physically-based variable contributing area model of basin hydrology. – Hydrological Science Bulletin, 24, 43-69.

BREIMAN, L. (1996), Bagging predictors. Machine Learning, 24 (2), 123-140.

BREIMAN, L. (2001), Random forests. Machine Learning, 45 (1), 5-32.

BREIMAN, L. (2002), Manual on setting up, using, and understanding random forests. http://oz.berkeley.edu/users/breiman/Using_random_forests_V3.1.pdf (v. 3.1).

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. & STONE, C. J. (1984), Classification and Regression Trees. Chapman and Hall, Boca Raton.

BRENNING, A. (2005), Spatial prediction models for landslide hazards: review, comparison and evaluation. Natural Hazards and Earth System Sciences, 5, 853-862.

CARRARA, A. & GUZZETTI, F. (1995), Geographical Information Systems in Assessing Natural Hazards. Kluwer Academic Publisher, Dordrecht, The Netherlands, 353 p.

CASTILLA-RHO, J. C., MARIETHOZ, G., KELLY, B. F. J. & ANDERSEN, M. S. (2014), Stochastic reconstruction of paleovalley bedrock morphology from sparse datasets. Environmental Modelling & Software, 53, 35-52.

DE'ATH, G. & FABRICIUS, K. E. (2000), Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology, 81, 3178-3192.

DIEU, T. B., OWE, L., NGE, R. & OYSTEIN, D. (2011), Landslide susceptibility analysis in the Hoa Binh province of Vietnam using statistical index and logistic regression. Nat. Hazards, 59 (3), 1413-1444.

DIKAU, R., CAVALLIN, A. & JAGER, S. (1996), Databases and GIS for landslide research in Europe. Geomorphology, 15 (3-4), 227-239.

ELDEIRY, A. & GARCIA, L. A. (2009), Comparison of Regression Kriging and Cokriging Techniques to Estimate Soil Salinity Using Landsat Images. In: Proceedings of Hydrology Days, Colorado State University, March 2009, 27-38.

ELITH, J., LEATHWICK, J. R. & HASTIE, T. (2008), A working guide to boosted regression trees. Jour. Anim. Ecol., 77, 802-813.

FRIEDMAN, J. H. (1999), Stochastic gradient boosting. Technical Report. Department of Statistics, Stanford University, USA. http://www.salford-systems.com/treenet.html.

GESSLER, P. E., MOORE, I. D., MCKENZIE, N. J. & RYAN, P. J. (1995), Soil-landscape modelling and spatial prediction of soil attributes. International Journal of Geographical Information Systems, 9, 421-432.

GHOSH, S., CARRANZA, E. J. M., VAN WESTEN, C. J., JETTEN, V. G. & BHATTACHARYA, D. N. (2011), Selecting and weighting spatial predictors for empirical modeling of landslide susceptibility in the Darjeeling Himalayas (India). Geomorphology, 131, (1-2), 35-56.

GRIMM, R., BEHRENS, T., MÄRKER, M. & ELSENBEER, H. (2008), Soil organic carbon Concentrations and stocks on Barro Colorado Island – digital soil mapping using Random Forest analysis. Geoderma, 146, 102-113.

HOSMER, D. W. & LEMESHOW, S. (2000), Applied Logistic Regression. 2nd Ed. Wiley, New York, 392 p.

LIAW, I. & WIENER, M. (2002), Classification and Regression by random Forest. R News, 2-3, 18-22.

MÄRKER, M., PELACANI, S. & SCHRÖDER, B. (2011), A functional entity approach to predict soil erosion processes in a small Plio-Pleistocene Mediterranean catchment in Northern Chianti, Italy. Geomorphology, 125, 530-540.

MEI, S.-L., DU, C.-J. & ZHANG, S.-W. (2008), Linearized perturbation method for stochastic analysis of a rill erosion model. Applied Mathematics and Computation, 200 (1), 289-296.

MERTENS, M., NESTLER, I. & HUWE, B. (2002), GIS-based regionalization of soil profiles with Classification and Regression Trees (CART). Journal of Plant Nutrition and Soil Science, 165, 39-44.

MOORE, I. D., GESSLER, P. E., NIELSEN, G. A. & PETERSON, G. A. (1993), Soil attribute prediction using terrain analysis. Soil Science Society of America Journal, 57, 443-452.

ODEH, I. O. A., MCBRATNEY, A. B. & CHITTLEBOROUGH, D. J. (1995), Further results on prediction of soil properties from terrain attributes: Heterotopic cokriging and regression-kriging. Geoderma, 67 (3-4), 215-226.

PARUELO, J. M. & TOMASEL, F. (1997), Prediction of functional characteristics of ecosystems: a comparison of artificial neural networks and regression models. Ecological Modelling, 98, 173-186.

SCHMALTZ, E., MÄRKER, M., ROSNER, H.-J. & KANDEL, A. W. (2015), The integration of landscape processes in archaeological site prediction in the Mugello basin (Tuscany/Italy). In: 21st Century Archaeology – Proceedings of the 41th Annual Conference on Computer Applications and Quantitative Methods in Archaeology, Paris, France, April 2014.

SHAHROKHNIA, M. A., SEPASKHAH, A. R. & JAVAN, M. (2004), Estimation of hydraulic parameters for Karoon River by Co-Kriging and residual Kriging. Iranian Journal of Science & Technology, Transaction B, 28 (B1), 153-163.

SHAPIRE, R., FREUND, Y., BARTLETT, P. & LEE, W. (1998), Boosting the margin: A new explanation for the effectiveness of voting methods. Annals of Statistics, 26 (5), 1651-1686.

SIDORCHUK, A. (2005), Stochastic components in the gully erosion modeling. Catena, 63 (2-3), 299-317.

VOGEL, S, & MÄRKER, M. (2010), Reconstructing the Roman topography and environmental features of the Sarno River Plain (Italy) before the AD 79 eruption of Somma-Vesuvius. Geomorphology, 115, 67-77.

VORPAHL, P., ELSENBEER, H., MÄRKER, M. & SCHRÖDER, B. (2012), How can statistical models help to determine driving factors of landslides? Ecological Modeling, 239, 27-39.

WU, Y., CHEN, L., CHENG, C., YIN, K. & TÖROK, A (2014), GIS-based landslide hazard predicting system and its real-time test during a typhoon, Zhejiang Province, Southeast China. Enigneering Geology, 175, 9-21.