Environ Model Assess (2017) 22:201–214
DOI 10.1007/s10666-016-9538-y

A Tool for Classification and Regression Using Random Forest Methodology: Applications to Landslide Susceptibility Mapping and Soil Thickness Modeling

Daniela Lagomarsino[1] · **V. Tofani**[1] · **S. Segoni**[1] · **F. Catani**[1] · **N. Casagli**[1]

Received: 21 May 2014 / Accepted: 20 October 2016 / Published online: 20 January 2017

Abstract Classification and regression problems are a central issue in geosciences. In this paper, we present Classification and Regression Treebagger (ClaReT), a tool for classification and regression based on the random forest (RF) technique. ClaReT is developed in Matlab and has a simple graphic user interface (GUI) that simplifies the model implementation process, allows the standardization of the method, and makes the classification and regression process reproducible. This tool performs automatically the feature selection based on a quantitative criterion and allows testing a large number of explanatory variables. First, it ranks and displays the parameter importance; then, it selects the optimal configuration of explanatory variables; finally, it performs the classification or regression for an entire dataset. It can also provide an evaluation of the results in terms of misclassification error or root mean squared error. We tested the applicability of ClaReT in two case studies. In the first one, we used ClaReT in classification mode to identify the better subset of landslide conditioning variables (LCVs) and to obtain a landslide susceptibility map (LSM) of the Arno river basin (Italy). In the second case study, we used ClaReT in regression mode to produce a soil thickness map of the Terzona catchment, a small sub-basin of the Arno river basin. In both cases, we performed a validation of the results and a comparison with other state-of-the-art techniques. We found that ClaReT produced better results, with a more straightforward and easy application and could be used as a valuable tool to assess the importance of the variables involved in the modeling.

Keywords Classification and regression · Random forest · Feature selection · Landslide susceptibility maps

✉ Daniela Lagomarsino
daniela.lagomarsino80@gmail.com

V. Tofani
veronica.tofani@unifi.it

S. Segoni
samuele.segoni@unifi.it

F. Catani
filippo.catani@unifi.it

N. Casagli
nicola.casagli@unifi.it

[1] Earth Sciences Department, University of Firenze, Via La Pira 4, 50121 Florence, Italy

1 Introduction

Methods for predicting a response variable given a set of features are needed in numerous scientific fields related to geosciences, including geomorphology [1], geochemistry [2, 3], pedology [4], hydrology [5–7], atmospheric physics [8, 9], hydrogeology [10, 11], engineering geology [12, 13], and environmental mapping [14, 15].

In particular, regression is used to predict continuous values, whereas classification is used to predict which class a data point is part of. Numerous techniques were developed to implement classification and regression procedures: discriminant analysis [16–18], logistic regression [19, 20], multivariate analysis [21–23], fuzzy linear regression [24], artificial neural network (ANN) [25, 26], and random forest (RF) [27, 28]. Several authors have compared these methods. In some cases, models based on classical statistics have been proven very effective for solving relatively simple problems [29], whereas more sophisticated techniques give better results for more complex

problems [30]. In other studies [31, 32], such methods seem quite equivalent and produce similar results. King et al. [33] reported that it was not possible to identify the best algorithm for classification, as any comparative study is by its nature limited and the best algorithm for a particular dataset depends on the features of that dataset.

This issue is of paramount importance when modeling physical problems: Once a given technique has been selected, results are still sensitive to model configuration [34]. In particular, several studies pointed out that the selection of conditioning variables is essential for classification/regression problems (see, e.g., [54, 55]).

As an instance, concerning landslide susceptibility maps (LSMs), there is an extensive literature on different statistical implementation techniques and on the comparison of their performances [35–37]. Moreover, the selection of the landslide conditioning variables (LCVs) is intensely debated: Many authors have discussed about the number [30, 38–42] and the type of the LCVs [35, 43–46]. The development of tools that allow to automatically perform LSM is already studied. For example, Akgun et al. [47] presented MamLand, a Matlab program, based on a fuzzy algorithm, for the assessment of LSM.

Conversely, in other fields of geosciences, nothing has been developed to automate the processes of feature selection and modeling. For example, to assess the spatial distribution of some physical properties of soil, such as soil thickness, the scientific literature accounts for several techniques or procedures [48–53], but none of them has generated a tool that could be applied to a wide selection of case of studies.

The objective of this work is filling this gap and presenting a tool, Classification and Regression Treebagger (ClaReT) that automates the classification/regression procedures, including the selection of the optimal set of explanatory variables based on an objective and quantitative standard.

We chose to base the tool on the RF method because it is a flexible environment for testing model parameters as it permits management of large amounts of data. In several applications [5, 32], it has shown numerous advantages, such as the possibility of using both categorical and numerical variables and the capability of accounting for interactions and nonlinearities between variables. RF can be considered an established technique in several fields of geosciences, including landslide susceptibility studies [38, 54–56]. There are several software and tools for random forest implementation, but their use in some fields of geosciences could be troublesome, due to the high quantity of data to manage. ClaReT overcomes this limitation, since it is able to manage and elaborate large amounts of spatial data such as those related to distributed geoenvironmental modeling at regional scale.

ClaReT is developed in Matlab and automatically ranks the variables according to their importance, selects the optimal configuration for regression or classification problems, applies that configuration to an input dataset, and provides a validation of the results. Furthermore, the simple graphical user interface (GUI) simplifies the model implementation, which can be carried out even without knowing Matlab language and environment.

As a test, we applied ClaReT to two case studies. In the first one, we used ClaReT in classification mode to identify the better subset of LCVs and to obtain a LSM of the Arno river basin (9100 km^2). In the second case study, we used ClaReT in regression mode to produce a soil thickness map in the Terzona catchment, a small (24 km^2) sub-basin of the Arno river basin.

In both cases, we performed a validation of the results and a comparison with other state-of-the-art techniques. We used the case studies also to discuss some additional features of ClaReT: In the first case study, we examined the ranking and visual display of the parameters' importance, which could be a valuable help in understanding the hierarchy of the variables used to explain a given physical problem. In the second test site, we discussed the implementation of models to real case studies, explaining why ClaReT resulted to be more straightforward and easy to apply than other state-of-the-art methods.

2 Materials and Methods

2.1 Random Forest Technique

RF is a nonparametric multivariate technique based on a machine-learning algorithm [27, 28]. It consists of a combination of tree predictors, each grown on a bootstrapped subsample of the training data. The data excluded from the construction of the model are called out of bag (OOB). For random forest construction, at each node, the best split is selected among a random subset of the predictors [57, 58].

Several studies investigated the influence of the number of trees on the stability of the model: Catani et al. [38] and Diaz-Uriarte and De Andres [59] proved that, beyond a certain value (200 and 1000, respectively), the error evaluated considering only OOB data (out of bag error) is independent of the number of trees.

The random forest technique has several advantages [5, 38, 47].

1. It allows the employment of both categorical and numerical variables.
2. No assumption is required about the statistical distribution of the data.
3. It is capable of accounting for interactions and nonlinearities between variables.
4. It can be considered robust with respect to changes in the composition of the dataset.

5. Prediction with OOB data avoids overfitting.
6. It allows exploring a large number of explanatory variables because it intrinsically emphasizes only those variables of high explanatory power at each node split.
7. It is robust with respect to noise features.

These properties are very useful in applications using variables that have mutual nonlinear interactions, that may be affected by measured errors or that have a mixed (numerical and categorical) nature.

Furthermore, this method allows measuring variable importance by assessing how much the prediction error increases if the values of that variable are permuted across the OOB observations [27, 47, 60].

In this work, we used the random forest treebagger (RFtb), a RF implementation developed in Matlab.

2.2 ClaReT

ClaReT is a tool for classification and regression developed in Matlab and based on RFtb.

Explanatory variables selection influences the performance of regression and classification models. ClaReT uses the RFtb peculiarities to objectively assess the relative importance of each predictor variable throughout *OOBPermutedVarDeltaError*, a Matlab function that measures the increase in prediction error if the values of that variable are permuted across the OOB observations. This measure is computed for every tree, then averaged over the entire ensemble, and divided by the standard deviation over the entire ensemble (http://www.mathworks.it/it/help/stats/treebagger.oobpermutedvardeltaerror.html).

To evaluate the best configuration, the model is initially applied to the training set with the entire set of features. Subsequently, the model is run without the feature that resulted to be the least important in terms of *OOBPermutedVarDeltaError*. This procedure is iterated, and after each run, the least important variable is excluded and a reduced configuration is defined. ClaReT runs each configuration 10 times, with the purpose of averaging any discrepancy due to random components of the RFtb. The variable set is applied to the test points, and the total error is calculated. Once the iteration process gets to calculate the error of the configuration composed by two variables, the total errors of all the configurations are compared and the configuration with the lowest error is selected.

ClaReT uses the optimal configuration to compute the predicted response for the entire dataset, providing the calculated value (for regression mode) or the membership class (for classification mode).

In regression mode, the model also returns standard deviations of the computed responses over the ensemble of the grown trees. In classification mode, it returns scores for all classes, compiling a matrix with one row per observation and one column per class. For each observation and each class, the score generated by each tree is the probability of this observation originating from this class (http://www.mathworks.it/it/help/stats/treebagger.predict.html).

The model performance is evaluated in terms of total error: Comparing the real values and the predicted ones, the tool provides the root mean square error (in the regression mode) and the misclassification error (in the classification mode).

In the specific case of binary classification, the most commonly used cutoff-independent performance technique is the receiver operating characteristic (ROC) curve [61, 62]. ClaReT builds ROC curves comparing the estimated membership class probability and the true class, showing the true positive rate versus the false positive rate for different thresholds of the classifier output. ClaReT calculates the area under the ROC curve (AUC), which can be used as a metric to assess the overall quality of the model [63]. This threshold-independent measure of discrimination between both classes can have values between 0.5 (no discrimination) and 1 (perfect discrimination) [64].

2.3 ClaReT Usage

ClaReT is constituted by a simple GUI. The GUI can be easily used to choose the methodology (classification or regression), to upload of the input data, and to visualize the results. Figure 1 shows ClaReT GUI, and Fig. 2 reports a flow diagram that summarizes the input files needed to perform the analysis and the modeling and the outputs provided.

The left block of the GUI (Fig. 1) is dedicated to feed and set the model. First, the method (either classification or regression) and the number of decision trees should be chosen. Then, two files containing the training and the test data have to be provided. These files must contain the same number of columns, each corresponding to an explanatory variable, except for the last column, which must contain the value of the membership class (0–1 for classification) or the value of the predicting response. Furthermore, the headers with explanatory variable names need to be provided, and categorical variables should be clearly distinguished from the others.

At the end of the elaborations, the "output feature selection" panel (central block of Fig. 1) summarizes some relevant information about the optimal configuration identified by the model. The information includes how many and which variables are encompassed in the optimal configuration and the error obtained applying this configuration to the test points. ClaReT outputs include also a series of files that can be used to characterize the configurations iteratively discarded by the automated process. These files are identified by the number

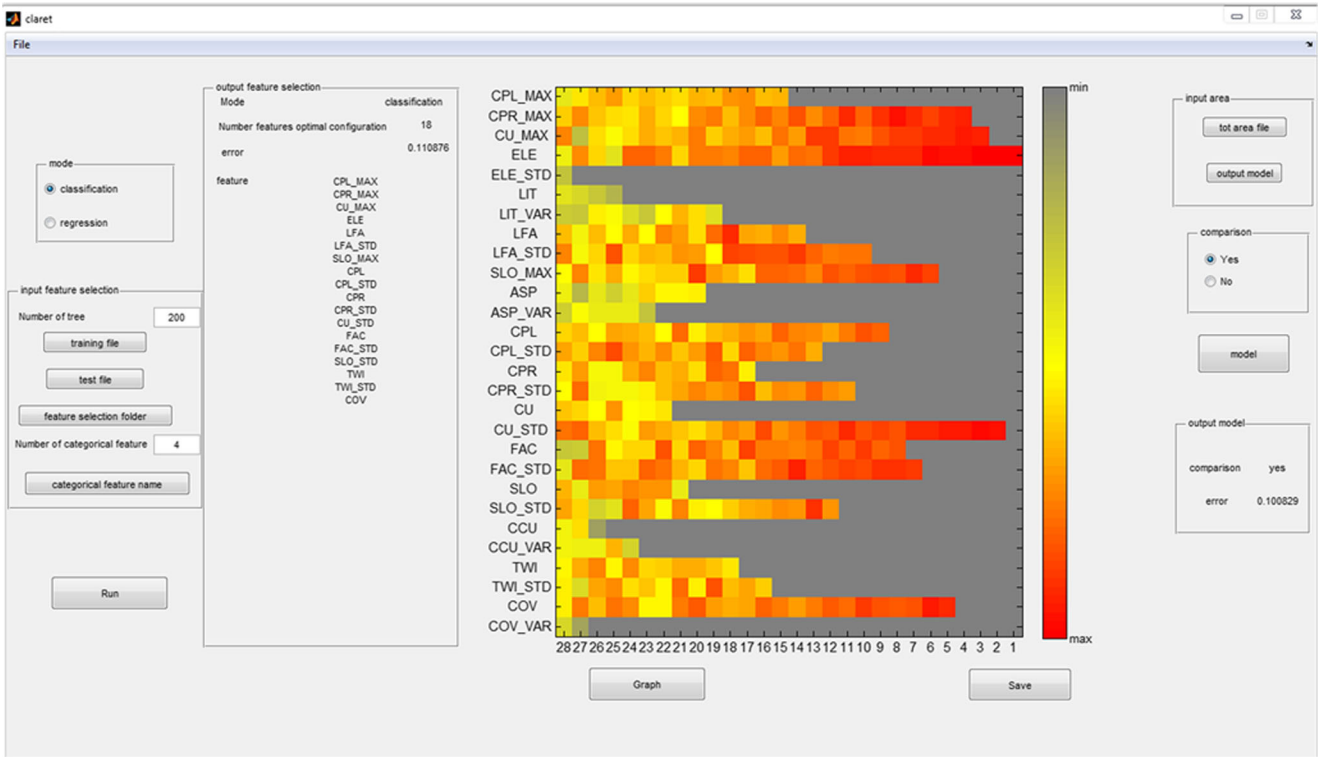

Fig. 1 ClaReT graphic user interface (GUI). The GUI is divided into three blocks. *Left block* is dedicated to input parameters and model settings; *central block* automatically provides information about the optimal configuration and ranks all input parameters according to their importance; *right block* is dedicated to model application to a full dataset, including options to carry out a quantitative validation

of explanatory variables used and contain the feature names, the mean *OOBPermutedVarDeltaError* with its standard deviation, and the mean rank. For each configuration, the software also saves a file with the mean and the maximum error obtained by applying the model to the test points. In another file, the rank of each feature in each configuration is reported. This result is also illustrated in a graph where the rank of each feature, at decreasing parameter number, is displayed according to a color ramp; the variables discarded are shown in gray (Fig. 1).

The right block of the GUI (Fig. 1) can be used to apply the model, in its optimal configuration, to the entire dataset. The parameters' table relative to the entire study area can contain features that do not belong to the optimal set, as the software

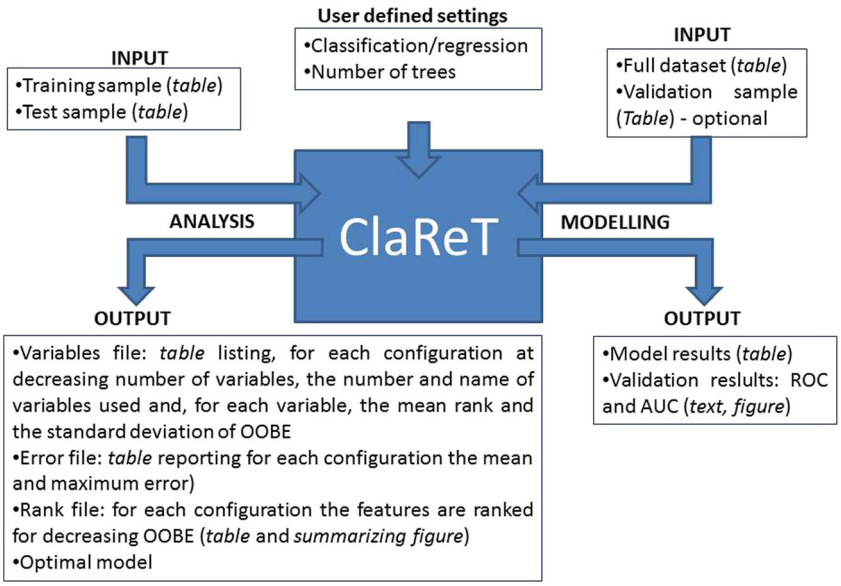

Fig. 2 ClaReT flowchart describing model inputs, outputs, and user-defined settings

automatically selects only the features required to apply the model. The outputs consist of two tables contained in distinct files. In the classification mode, these tables contain the predicted class and the membership class probability for each row of the input table. In the regression mode, the tables contain the predicted value of each observation and the standard deviations of the computed responses over the ensemble of the grown trees. If actual observations of the dependent variable are available, ClaReT carries out the comparison between these ones and the model results. The total error (the misclassification probability for classification trees or the mean squared error for regression trees) is then displayed in the "output model" panel. For binary classification problems, the related ROC curve is depicted and the AUC value is provided.

At present, the software can be requested contacting the corresponding author.

3 Case Study 1, Classification Mode: Application to Landslide Susceptibility Mapping

3.1 Description of the Area

The selected test site is the hydrographic basin of the Arno River in Central Italy. The area is 9100 km^2 wide, and it is located in the northern Apennines, a complex thrust-belt system composed of different tectonic units and sedimentary basins [38]. The relief is characterized by a succession of NW–SE ridges, made up of Mesozoic/Tertiary flysches and calcareous units, and Pliocene-Quaternary sedimentary basins with cohesive and granular soils.

Landslides are very common in the study area. The geological settings and the lithological characteristics of the area affect the typology and occurrence of landslides, which are mainly constituted by slow-moving rotational slides [65–67].

The landslide inventory of the Arno river basin counts about 27,500 landslides [67]. The most represented landslide types are earth slides and solifluctions. The majority of the landslides are reactivations of dormant slides, and the frequency of first-time landslides is very low. Consequently, landslide susceptibility chiefly depends on the presence or absence of known instability [38].

3.2 Landslide Susceptibility Mapping

To account for the variability of physiographic settings within the study area, the Arno basin was split into three homogeneous domains (area 1, area 2, and area 3), in accordance with the different lithological and geomorphological characteristics. In particular, area 1 is mainly a hilly region characterized by low to medium slope angles and relatively low elevations (Fig. 3). Granular and cohesive soils are prevailing on the other types of lithologies. Area 2 is similar to area 1 from a geomorphological point of view: It has a heterogeneous geology, with the presence of cohesive and granular soils, flysches, and calcareous rocks. Area 3 is dominated by the Apennines and is characterized by flysch formations, with higher slope angles and elevations than other domains.

This division is necessary to test the model in almost homogenous zones and to verify the dependence of the performance on geomorphological and geological characteristics. Inside each domain, a training subset and a test subset have been selected (Fig. 3). Training and test data were selected carefully from the entire database of mass movements, to have a representative sample of the total population. Moreover, these areas were checked in the field for accuracy and completeness of the landslide inventory. The pixels of the areas represented in Fig. 3 were randomly separated to obtain two distinct datasets of test and training. In Table 1, the number of training points, the number of total pixel of each area, and the percentage of training points with respect to the entire area are reported. To train the model, we used about 10% of the points for each zone.

For each domain, ClaReT was applied independently, using pixels with a dimension of 100 × 100 m as computational unit. To speed up the computation and the preparation of input parameters and to limit the analysis to the portion of the territory in which landslides may potentially occur, alluvial plains were excluded from the analysis [38, 68].

In this work, we have considered two kinds of input parameters: morphometric attributes and attributes derived from thematic maps. The complete list of LCV parameters is shown in Table 2.

To derive the morphometric attributes, a 20 × 20 m DTM was available. For each morphometric variable, we considered the average value inside the 100 × 00 m cell. Concerning the thematic attributes (namely lithology and land cover), we used a 1:100,000 lithotechnical map and a 1:50,000 land cover map [51], and we estimated the most frequent value within 100 × 100 m cells. The variability of the features was taken into account as well: for each 100 × 100 m cell, we calculated the standard deviation for numeric variables and the variety for the categorical ones. Most of the considered features are dependent from each other, but the RF technique is capable of accounting for interactions between variables. Therefore, as explained in Sect. 2.1, preliminary studies on variables are not required.

3.3 Results

Feature selection results are shown in Fig. 4, in which the variable importance is plotted for each model configuration. In Fig. 4, white box points out the optimal configuration for each area, while the variables discarded at each iteration are displayed in gray.

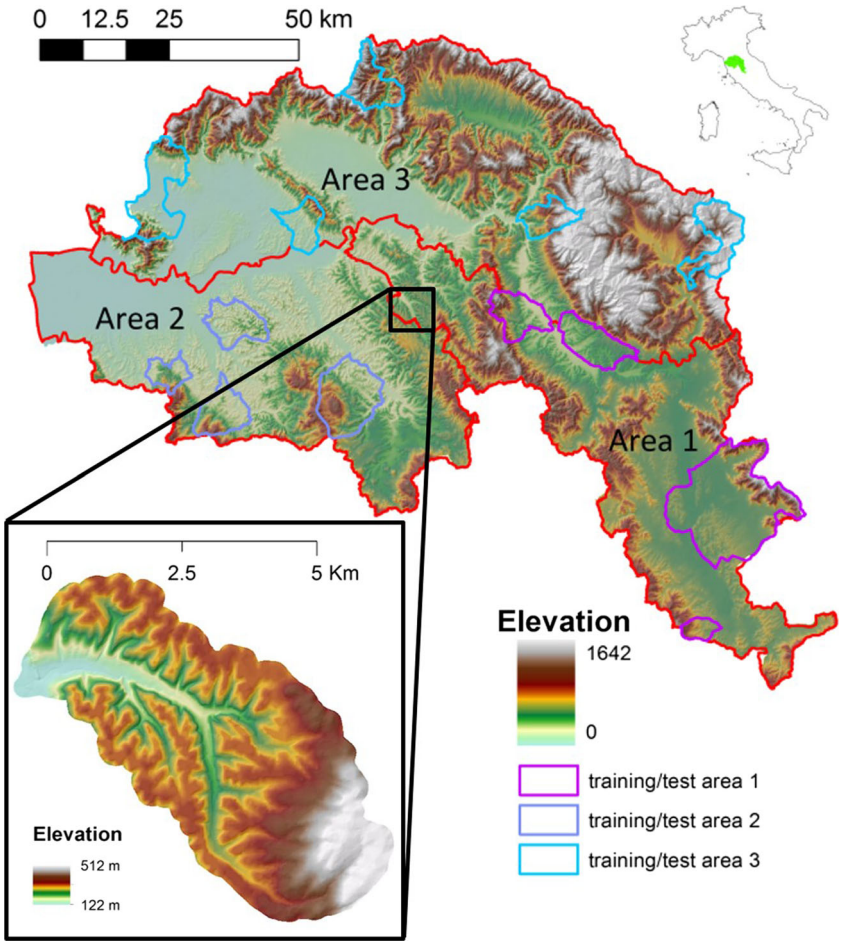

Fig. 3 Location of the Arno river basin, subdivision into three geomorphological domains, and training/test zones

For area 1, the full configuration (27 parameters) represents the best configuration, with a misclassification probability value equal to 0.11. Considering area 2, land use variety, combo curvature variety, and combo curvature were discarded; thus, the optimal configuration encompasses 25 features. The misclassification probability obtained applying this set to the test points is 0.007. For area 3, the optimal configuration encompasses 26 parameters: Land use variety and combo curvature were discarded, obtaining a misclassification error equal to 0.16.

For each domain, the optimal parameter set was used to classify the entire dataset. The output of these elaborations was three tables (one for each domain) that were imported into a GIS and converted into raster maps. The union of the three maps represents the landslide susceptibility map of the Arno basin (Fig. 6).

Since a complete and homogeneous landslide database was available for the entire study area, we used ClaReT to compare the model results with the ground truth, building ROC curves and calculating AUC values (Fig. 5). We performed a distinct computation for each of the three domains, obtaining AUC values of 0.75, 0.59, and 0.69 for area 1, area 2, and area 3, respectively (Fig. 5).

Table 1 Number of training points for each area

	Number of training points	Total number of pixels	Percentage of training points
Area 1	30,839	194,529	15%
Area 2	24,603	187,053	13%
Area 3	27,962	344,299	8%

The percentage of training points with respect to the entire area is reported

3.4 Benchmark Model for Comparison: Discriminant Analysis

To have a benchmark for comparison, we carried out another landslide susceptibility assessment based on discriminant analysis [16], which is a more established technique with a longer tradition of applications to landslide susceptibility mapping than RF [69, 70] (Fig. 6).

First, we transformed categorical variables into dummy variables, as needed in discriminant analysis. We obtained nine additional classes for aspect, nine for combo curvature,

Table 2 Landslide conditioning variables (LCVs)

Parameter	Description	LCV	Acronym
Curvature	The second derivative of elevation	Mean	CU
		Max	CU_MAX
		Standard deviation	CU_STD
Planar curvature	The second derivative of elevation calculated orthogonally to the direction of the maximum slope	Mean	CPL
		Max	CPL_MAX
		Standard deviation	CPL_STD
Profile curvature	The second derivative of elevation calculated in the direction of the maximum slope	Mean	CPR
		Max	CPR_MAX
		Standard deviation	CPR_STD
Combo curvature	Categorical variable obtained by the combination of the values of plan and profile curvature	Most frequent	CCU
		Variety	CCU_VAR
Elevation	DEM value	Mean	ELE
		Standard deviation	ELE_STD
Lithology	Represented by eight classes	Most frequent	LIT
		Variety	LIT_VAR
Flow accumulation	The upslope contributing area	Mean	FLA
		Standard deviation	FLA_STD
Log flow accumulation	Logarithm of the flow accumulation	Mean	LFA
		Standard deviation	LFA_STD
Slope	The first derivative of elevation	Mean	SLO
		Max	SLO_MAX
		Standard deviation	SLO_STD
Aspect	Orientation in the space, represented by nine classes	Most frequent	ASP
		Variety	ASP_VAR
Topographic wetness index	Ln(flow accumulation/tan slope)	Mean	TWI
		Standard deviation	TWI_STD
Land cover	Represented by nine different classes	Most frequent	COV
		Variety	COV_VAR

For each parameter, the acronym used in the graphics is reported

nine for land use, and eight for lithology. This approach increased to 59 the number of LCVs taken into account. Then, the test areas were used to perform a stepwise analysis to select the features required for the best possible susceptibility assessment: We selected 33 features for area 1, 28 for area 2, and 29 for area 3. Considering these configurations, we developed the three LSMs, obtaining performances worse than using ClaReT: The AUC was 0.61, 0.53, and 0.61 in area 1, area 2, and area 3, respectively.

4 Case Study 2, Regression Mode: Application to Soil Thickness Mapping

4.1 Description of the Area

The study area for this application is the Terzona Creek basin (about 24 km^2), which is contained in the Arno river basin (inset of Fig. 1). The area is located in the Chianti region and is characterized by gentle hills made up of Pliocene and Quaternary terrains, while in the eastern sector, the bedrock is constituted by Paleocene and Eocene flysch, with rougher and higher (up to 512 m) reliefs. The land is sparsely urbanized and is covered by vineyards, olive groves, and small woods.

The study area is characterized by marked erosive processes, and soil is generally rather shallow. Typically, the thicker deposits are found in the valley floors (where they can reach the maximum value of 1.5 m) and in some hilltops, where a paleosol is present [48].

4.2 Soil Thickness Regression

The input data used for the soil thickness regression are curvature, planar curvature, profile curvature, combo curvature, elevation, lithology (as derived from a 1:10,000 detailed geological map), flow accumulation, logarithm of flow

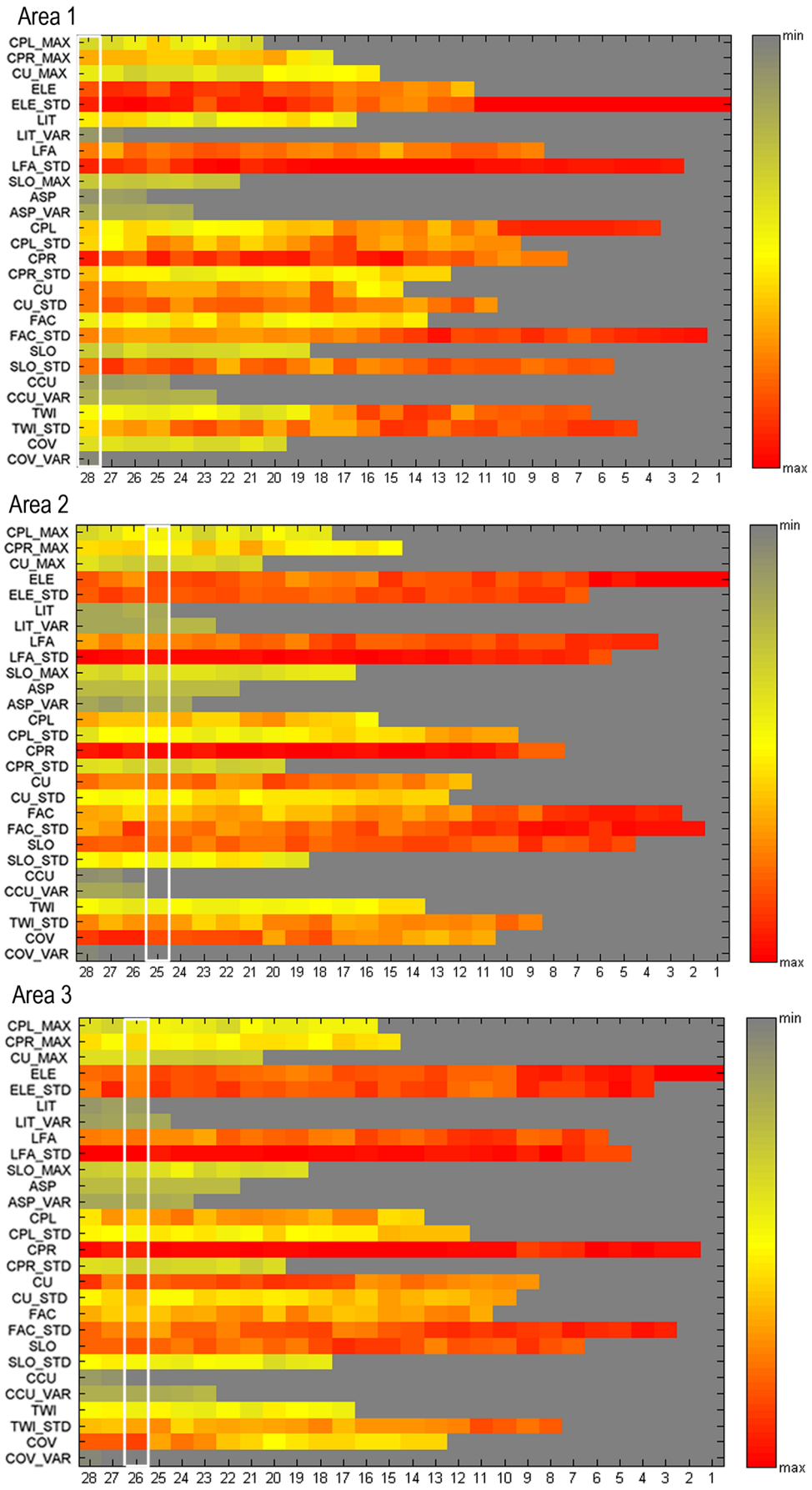

Fig. 4 Importance of LCVs for each iteration of feature selection. The *white box* highlights the best configuration. At each iteration, the least important parameter is excluded from the model, and it is shown in *gray*. The importance (i.e., explanatory power) of each parameter is graphically represented according to the color ramp shown on the right

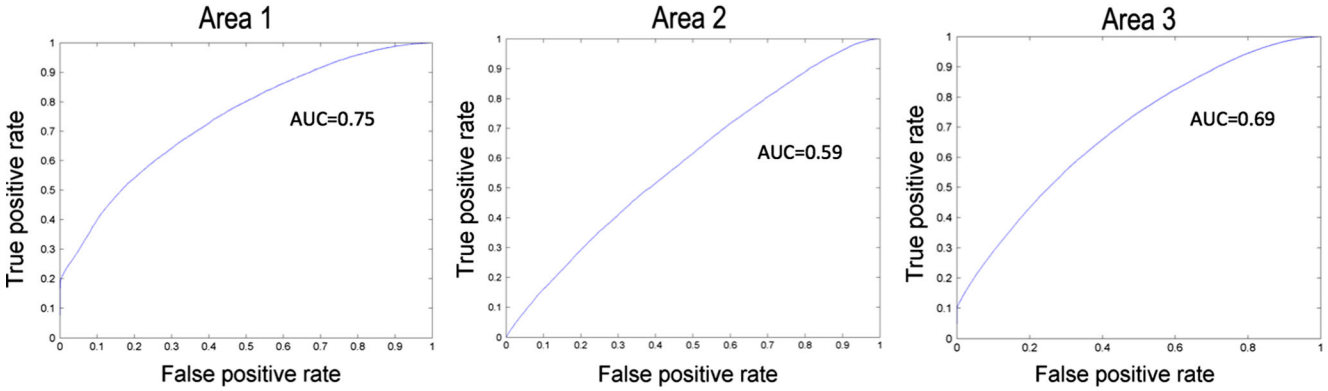

Fig. 5 ROC curves and AUC values obtained with ClaReT

accumulation, slope gradient, aspect, topographic wetness index, land cover (derived from the same 1:50,000 map used in the landslide susceptibility assessment), and geomorphological units (after a geomorphological survey, the area was subdivided into three domains according to the prevailing geomorphological features). Topographic attributes were derived from a 10-m resolution grid digital elevation model, and their values were calculated on a pixel-by-pixel basis.

To calibrate and validate the regression model, we used a database of direct soil thickness measurements performed by Catani et al. [48] for similar purposes. As in Catani et al. [48], soil thickness data were split into a calibration subset (55 measures) and a validation subset (162 measures).

By means of GIS analyses, at sample point locations, each soil thickness measure was associated with the pixel value for each input variable. Data were organized into tables, which were fed into ClaReT (Fig. 7).

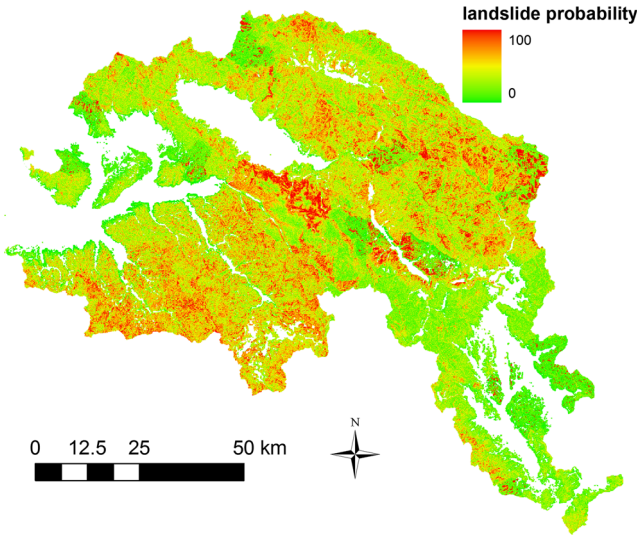

Fig. 6 Landslide susceptibility map of the Arno river basin. *White areas* represent flat zones

4.3 Results and Comparison with Other Models

After importing into a GIS the output table of ClaReT, it was possible to draw the distributed soil thickness map of the Terzona catchment (Fig. 8).

The soil thickness distribution is clearly influenced by the geomorphological features of the area: The thickest values are located in the main valleys, while mid-high values are in the flat hilltops occupied by paleosols. The shallowest soils can be found in the hillsides dominated by erosive processes like creep and landsliding and in the highest parts of the rocky reliefs. The validation procedure demonstrated that this distribution is in good agreement with ground truth: The mean absolute error is 9 cm, and the maximum error is −54 cm. Since the Terzona catchment has long been used as a test site for soil thickness modeling [48, 71], it was possible to compare the results obtained by ClaReT with those obtained by other state-of-the-art techniques (Fig. 9). The models used for comparison are the following.

- Z model (linear correlation with elevation) [49]. This approach is based on the assumption that soil thickness and elevation are linked by an inverse correlation since in higher portions of the relief, erosive processes prevail over depositional processes, while at lower altitudes, the opposite takes place.
- S model (linear correlation with slope gradient) [49]. This widely used approach is based on the assumption that soil thickness and slope gradient are inversely correlated since in steep areas, erosive processes prevail over depositional processes, while in flat areas, the opposite circumstance takes place.
- Sexp. As above, except for the fact that the correlation is based on an exponential law. This is also a widely used method to define spatially distributed soil thickness maps [50, 72].
- GIST [48]. This is the model that since now had held the better performances in Terzona basin [71]. It takes into

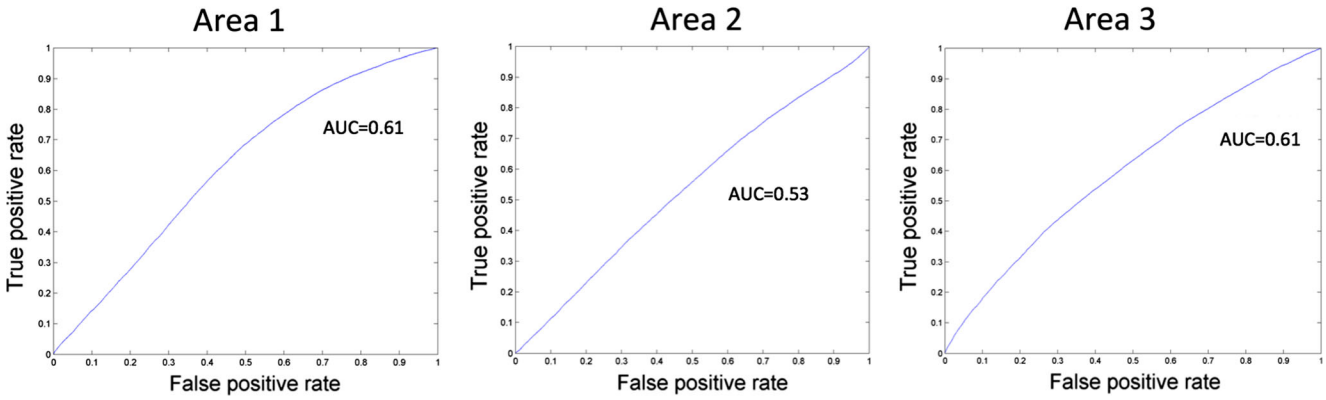

Fig. 7 ROC curves and AUC values obtained with the discriminant analysis

account three morphometric attributes (profile curvature, slope gradient, and relative position within the hillslope profile).

Figure 9 shows that the most effective soil thickness modeling is the one obtained by ClaReT: The distribution of errors is centered on almost correct occurrences, negligible errors are the majority, and the tails of the distribution (relevant underestimations or overestimations) are more contained than in the other models.

5 Discussion

5.1 Application to Landslide Susceptibility

The application to landslide susceptibility mapping in a large area (Arno river basin, 9100 km^2) provided relevant outcomes. The geomorphological heterogeneity was handled subdividing the test area into three different domains analyzed and modeled independently; consequently, ClaReT identified three different regression models based on explanatory variable ensembles that show some differences.

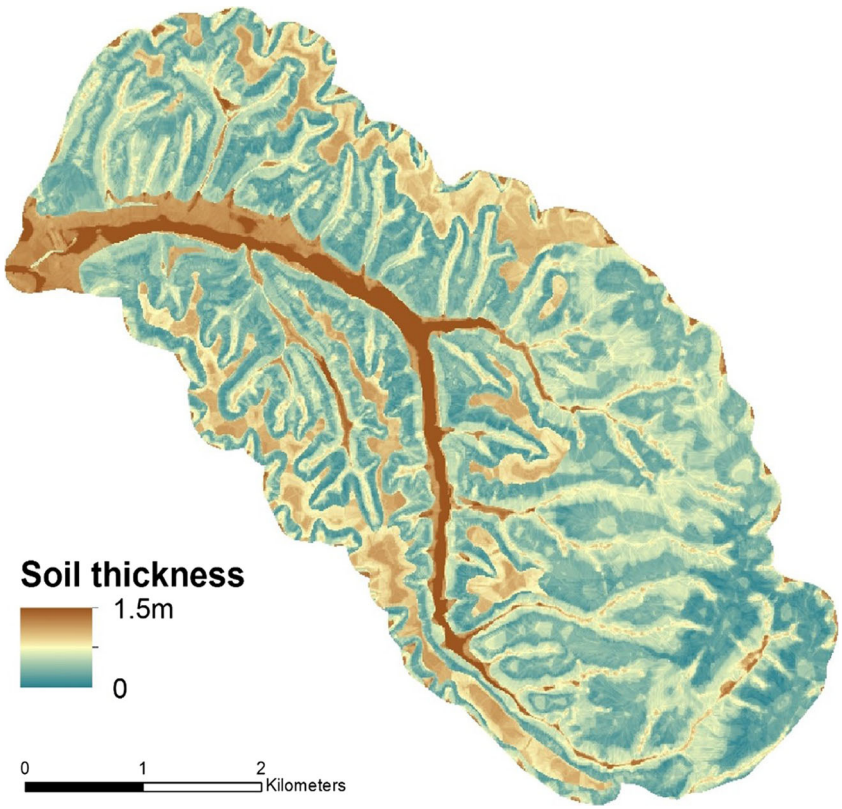

Fig. 8 Soil thickness map obtained using ClaReT in the Terzona catchment

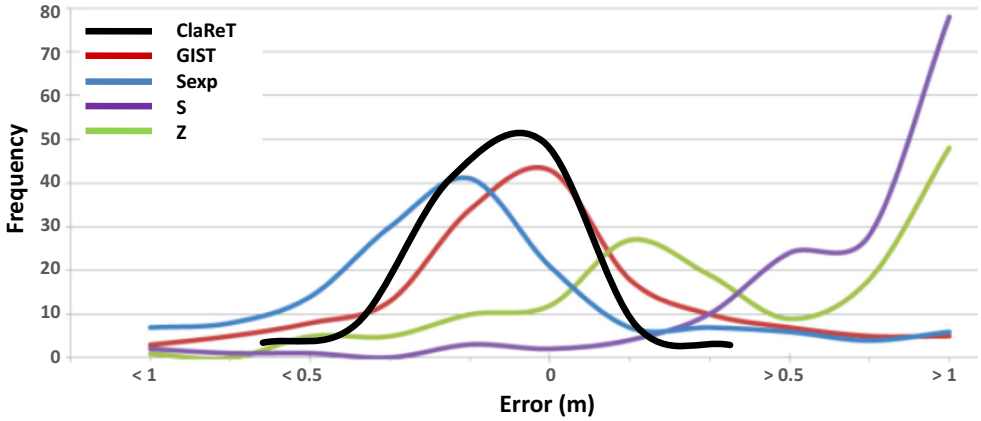

Fig. 9 Distribution of errors obtained with five different soil thickness models applied in the Terzona catchment

Elevation is the most important parameter in all three geomorphological domains, while the other relevant (i.e., with high rank) explanatory variables are different: In area 1 and area 3, the most important are flow accumulation, planar curvature, and topographic wetness index; in area 2, another variable appears in the highest ranks (slope gradient) and profile curvature is replaced by planar curvature. However, the ranks of almost all variables noticeably vary from a domain to another. This outcome demonstrates the importance of following an objective and quantitative criterion to select the input parameters to be used in susceptibility mapping.

The model performances are very different for the different domains. Area 2 shows the worst result (AUC = 0.59), whereas in the other two domains, AUC values are 0.75 and 0.69. These values could be considered medium-high if compared to the values that are reported in the landslide susceptibility literature. However, discriminant analysis outcomes have a similar trend (area 2 is worse than the other two), and in each zone, the RF technique shows better performance, in terms of AUC, with respect to discriminant analysis. Since discriminant analysis is a well-established technique, we can conclude that the proposed tool is effective and that the low AUC values obtained for area 2 could depend on other factors.

The reasons beyond this performance can be investigated making a comparison with the study of Catani et al. [38], which applied in the same test area the same susceptibility model based on treebagger random forest algorithms used in this work. Using identical pixel resolution (100 m) and landslide dataset, they obtained considerably better performances in terms of AUC (AUC = 0.88). The main difference between the two studies is the different sampling strategy: Catani et al. [38] used a random selection of 10% of the pixels to train their treebagger model, while in this work, we defined some training areas, and all their pixels were used to train the model. The training area is about 10% of the total area; therefore, the worst performance is not linked to the dimension of the training sample. Therefore, we conclude that the option to use a sampling strategy based on the selection of testing areas led to worst landslide susceptibility assessments than a random sampling over the entire study area. This is valid especially in cases where the landslide inventory has a good degree of homogeneity and completeness.

The comparison with the application of Catani et al. [38] leads to another outcome: In the present work, the approach of subdividing the study area into three geomorphological domains and to analyze each of them independently does not improve the susceptibility assessment. Catani et al. [38] analyzed the whole area with a single configuration and obtained better results. This is a proof of the robustness of the random forest technique: When the model is fed with enough and appropriate input parameters, the subjective judgment can be reduced.

5.2 Application to Soil Thickness Regression

The advantages of using ClaReT in a regression problem like soil thickness mapping were twofold.

First, the modeling based on the random forest treebagger technique allowed for results that proved to be better than any other model used in the same test site so far (Fig. 9).

Second, ClaReT had the advantage of a very fast and straightforward application. In fact, the only soil thickness model with comparable quality of the results was GIST (Fig. 9), which required extensive field works and subjective interpretations to be conveniently applied. In fact, according to the methodology described in [48], to correctly parameterize the model, it is necessary to devote a relevant part of the work toward the characterization of the typical soil catenas encountered in the area, the identification of different landsurface units, and the recognition of different typologies of hillslope morphology.

Conversely, to apply ClaReT, it was sufficient to feed the tool with state-of-the-art morphometric parameters and

thematic attributes (as lithology and land use), regardless of their correlations and mutual influences. The forward selection of input parameters implemented in ClaReT is sufficient to discard uninfluential or pejorative predictors and to give the right weight to each parameter. In this way, the use of ClaReT tool in soil thickness modeling made possible that the only necessary preliminary step was constituted by the GIS computations needed to calculate the input features.

6 Conclusions

ClaReT tool can be used to automate the feature selection process and to identify and apply the optimal classification/regression model. ClaReT is applicable to a wide range of regression or classification problems and can be used to manage large amounts of data with a reduced computation time. The GUI permits the use of RFtb without having to write Matlab code, and it assists the operator in the model implementation. The tool is based on a standardized procedure, making the classification and regression process reproducible.

As a test, we applied ClaReT to two case studies. In the first one, we used ClaReT in classification mode to identify the better subset of LCVs and to obtain LSM of the Arno river basin. In the second case study, we used ClaReT in regression mode to produce a soil thickness map of the Terzona catchment, a small sub-basin of the Arno. In both cases, we also performed a validation of the results and a comparison with other state-of-the-art techniques, finding that ClaReT produced better results with a more straightforward and easy application.

ClaReT can also be a valuable tool to perform analyses on the dataset and on the variables used. As an instance, an important feature of ClaReT is the ranking and visual display of the parameters' importance, which may assist in understanding the hierarchy of the variables that explain any physical problem. The selection of conditioning variables is essential for classification/regression problems [73, 74], and ClaReT gives the opportunity to manage a large number of variables with an automatic selection based on objective and quantitative criteria.

However, even if ClaReT automates and standardizes the process of classification and regression, it is important to keep in mind that the result always depends on the decisions of the operator, as the input data arrangement is a fundamental step in classification/regression problems. In landslide susceptibility mapping, for example, the choice of the training set and of the working scale influences the model performance [38, 75].

References

1. Adediran, A. O., Parcharidis, I., Poscolieri, M., & Pavlopoulos, K. (2004). Computer-assisted discrimination of morphological units on north-central Crete (Greece) by applying multivariate statistics to local relief gradients. *Geomorphology, 58*, 357–370.
2. Grunsky, E. C. (1986). Recognition of alteration in volcanic rocks using statistical analysis of lithogeochemical data. *Journal of Geochemical Exploration, 25*(1–2), 157–183.
3. Zhao, J., Wang, W., & Cheng, Q. (2014). Application of geographically weighted regression to identify spatially non-stationary relationships between Fe mineralization and its controlling factors in eastern Tianshan, China. *Ore Geology Reviews, 57*, 628–638.
4. Mertens, M., Nestler, I., & Huwe, B. (2002). GIS-based regionalization of soil profiles with classification and regression trees (CART). *Z. Pflanzenernähr. Bodenk., 165*, 39–43.
5. Loos, M., & Elsenbeer, H. (2011). Topographic controls on overland flow generation in a forest—an ensemble tree approach. *Journal of Hydrology, 409*(1–2), 94–103.
6. Gharari, S., Hrachowitz, M., Fenicia, F., & Savenije, H. H. G. (2011). Hydrological landscape classification: investigating the performance of HAND based landscape classifications in a central European meso-scale catchment. *Hydrology and Earth System Sciences, 15*, 3275–3291. doi:10.5194/hess-15-3275-2011.
7. Khan, U., Tuteja, N. K., & Sharma, A. (2013). Delineating hydrologic response units in large upland catchments and its evaluation using soil moisture simulations. *Environmental Modelling and Software, 46*, 142–154.
8. Turco, M., Zollo, A. L., Ronchi, C., De Luigi, C., & Mercogliano, P. (2013). Assessing gridded observations for daily precipitation extremes in the alps with a focus on Northwest Italy. *Natural Hazards and Earth System Sciences, 13*, 1457–1468.
9. Mercogliano, P., Segoni, S., Rossi, G., Sikorsky, B., Tofani, V., Schiano, P., Catani, F., & Casagli, N. (2013). Brief communication: a prototype forecasting chain for rainfall induced shallow landslides. *Natural Hazards and Earth System Sciences, 13*, 771–777.
10. Steinhorst, R. K., & Williams, R. E. (1985). Discrimination of ground-water sources using cluster analysis, MANOVA, canonical analysis and discriminant analysis. *Water Resources Research, 21*, 1149–1156.
11. Szucs, P., & Horne, R. N. (2009). Applicability of the ACE algorithm for multiple regression in hydrogeology. *Computational Geosciences, 13*, 123–124. doi:10.1007/s10596-008-9112-z.
12. Carrara, A. (1983). Multivariate models for landslide hazard evaluation. *Mathematical Geology, 15*(3), 403–426.
13. Dong, J. J., Tung, Y. H., Chen, C. C., Liao, J. J., & Pan, Y. W. (2011). Logistic regression model for predicting the failure probability of a landslide dam. *Engineering Geology, 117*, 52–61.
14. Rennó, C. D., Nobre, A. D., Cuartas, L. A., Soares, J. V., Hodnett, M. G., Tomasella, J., & Waterloo, M. J. (2008). HAND, a new terrain descriptor using SRTM-DEM: mapping terra-firme rainforest environments in Amazonia. *Remote Sensing of Environment, 112*, 3469–3481. doi:10.1016/j.rse.2008.03.018.
15. Vannametee, E., Babel, L. V., Hendriks, M. R., Schuur, J., de Jong, S. M., Bierkens, M. F. P., & Karssenberg, D. (2014). Semi-automated mapping of landforms using multiple point geostatistics. *Geomorphology, 221*, 298–319. doi:10.1016/j.geomorph.2014.05.032.
16. Lachenbruch, P. A., & Goldstein, M. (1979). Discriminant analysis. *Biometrics, 35*, 69–85.
17. Press, S. J., & Wilson, S. (1978). Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association, 73*, 699–705.
18. Flury, B., & Riedwyl, H. (1990). *Multivariate statistics: a practical approach*. London: Chapman and Hall.

19. Hosmer, D. W., & Lemeshow, S. (1989). *Applied logistic regression*. Princeton, NJ: John Wiley & Sons.
20. Studenmund, A. H. (1992). *Using econometrics: a practical guide*. New York: Harper Collins.
21. Snedecor, G. W., & Cochran, W. G. (1980). *Statistical methods* (7th ed.). Ames, IA: The Iowa State University Press.
22. Neter, J., Wasserman, W., & Kutner, M. H. (1985). *Applied linear statistical models* (2nd ed.). Homewood, IL: Richard D. Irwin, Inc..
23. Myers, R. H. (1990). *Classical and modern regression with applications* (2nd ed.). Boston, Massachusetts: PWS-KENT Publishing Company.
24. Tanaka, H., Hayashi, I., & Watada, J. (1989). Possibilistic linear regression analysis for fuzzy data. *European Journal of Operational Research, 40*(3), 389–396.
25. Beale, R., & Jackson, T. (1991). *Neural computing: an introduction*. Bristol: Adam Hilger, Techno House.
26. Haykin, S. (1994). *Neural networks: a comprehensive foundation*. New York: Maxwell Macmillan International.
27. Breiman, L., Friedman, J. H., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Belmont: Wadsworth International Group.
28. Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32.
29. Razi, M. A., & Athappilly, K. (2005). A comparative predictive analysis of neural networks (NNs), nonlinear regression and classification and regression tree (CART) models. *Expert Systems with Applications, 29*(1), 65–74.
30. Pradhan, B., & Lee, S. (2010). Landslide susceptibility assessment and factor effect analysis: back propagation artificial neural networks and their comparison with frequency ratio and bivariate logistic regression modelling. *Environmental Modelling & Software, 25*, 747–759.
31. Kanungo, D. P., Arora, M. K., Sarkar, S., & Gupta, R. P. (2006). A comparative study of conventional, ANN black box, fuzzy and combined neural and fuzzy weighting procedures for landslide susceptibility zonation in Darjeeling Himalayas. *Engineering Geology, 85*, 347–366.
32. Kurt, I., Ture, M., & Kurum, A. T. (2008). Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. *Expert Systems with Applications, 34*(1), 366–374.
33. King, R. D., Feng, C., & Sutherland, A. (1995). Statlog-comparison of classification algorithms on large real-world problems. *Applied Artificial Intelligence, 9*(3), 289–333.
34. Segoni, S., Rossi, G., Rosi, A., & Catani, F. (2014). Landslides triggered by rainfall: a semiautomated procedure to define consistent intensity-duration thresholds. *Computational Geosciences, 63*, 123–131.
35. Guzzetti, F., Carrara, A., Cardinali, M., & Reichenbach, P. (1999). Landslide hazard evaluation: a review of current techniques and their application in a multiscale study, Central Italy. *Geomorphology, 31*, 181–216.
36. Carrara, A., Crosta, G. B., & Frattini, P. (2008). Comparing models of debris-flow susceptibility in the alpine environment. *Geomorphology, 94*, 353–378.
37. Yilmaz, I. (2009). Landslide susceptibility mapping using frequency ratio, logistic regression, artificial neural networks and their comparison: a case study from Kat landslides (Tokat-Turkey). *Computer & Geoscience, 35*, 1125–1138.
38. Catani, F., Lagomarsino, D., Segoni, S., & Tofani, V. (2013). Landslide susceptibility estimation by random forests technique: sensitivity and scaling issues. *Natural Hazards and Earth System Sciences, 13*(11), 2815–2831.
39. Lee, S., Choi, J., & Min, K. (2002). Landslide susceptibility analysis and verification using the Bayesian probability model. *Environmental Geology, 43*, 120–131.
40. Gorsevski, P. V., Gessler, P. E., Foltz, R. B., & Elliot, W. J. (2006). Spatial prediction of landslide hazard using logistic regression and ROC analysis. *Transactions in GIS, 10*, 395–415.
41. Costanzo, D., Rotigliano, E., Irigaray, C., Jiménez-Perálvarez, J. D., & Chacón, J. (2012). Factors selection in landslide susceptibility modelling on large scale following the GIS matrix method: application to the river Beiro basin (Spain). *Natural Hazards and Earth System Sciences, 12*, 327–340.
42. Felicísimo, A., Cuartero, A., Remondo, J., & Quirós, E. (2013). Mapping landslide susceptibility with logistic regression, multiple adaptive regression splines, classification and regression trees, and maximum entropy methods: a comparative study. *Landslides, 10*, 175–189.
43. Manzo, G., Tofani, V., Segoni, S., Battistini, A., & Catani, F. (2013). GIS techniques for regional-scale landslide susceptibility assessment: the Sicily (Italy) case study. *International Journal of Geographical Information Science, 27*, 1433–1452.
44. Lee, S., & Pradhan, B. (2007). Landslide hazard mapping at Selangor, Malaysia, using frequency ratio and logistic regression models. *Landslides, 4*, 33–41.
45. Van Den Eeckhaut, M., Reichenbach, P., Guzzetti, F., Rossi, M., & Poesen, J. (2009). Combined landslide inventory and susceptibility assessment based on different mapping units: an example from the Flemish Ardennes, Belgium. *Natural Hazards and Earth System Sciences, 9*, 507–521.
46. Pereira, S., Zêzere, J. L., & Bateira, C. (2012). Technical note: assessing predictive capacity and conditional independence of landslide predisposing factors for shallow landslide susceptibility models. *Natural Hazards and Earth System Sciences, 12*, 979–988.
47. Akgun, A., Sezer, E. A., Nefeslioglu, H. A., Gokceoglu, C., & Pradhan, B. (2012). An easy-to-use MATLAB program (MamLand) for the assessment of landslide susceptibility using a Mamdani fuzzy algorithm. *Computers & Geosciences, 38*, 23–34.
48. Catani, F., Segoni, S., & Falorni, G. (2010). An empirical geomorphology-based approach to the spatial prediction of soil thickness at catchment scale. *Water Resources Research, 46*, W05508. doi:10.1029/2008WR007450.
49. Saulnier, G. M., Beven, K., & Obled, C. (1997). Including spatially variable effective soil depths in TOPMODEL. *Journal of Hydrology, 202*, 158–172.
50. De Rose, R. C. (1996). Relationships between slope morphology, regolith depth, and the incidence of shallow landslides in eastern Taranaki hill country. *Zeitschrift fur Geomorphologie Supplementband, 105*, 49–60.
51. Tesfa, T. K., Tarboton, D. G., Chandler, D. G., & McNamara, J. P. (2009). Modeling soil depth from topographic and land cover attributes. *Water Resources Research, 45*, W10438. doi:10.1029/2008WR007474.
52. Tsai, C. C., Chen, Z. S., Duh, C. T., & Horng, F. V. (2001). Prediction of soil depth using a soil-landscape regression model: a case study on forest soils in southern Taiwan. *Proc. Natl. Sci. Counc. R.O.C., 25*(1), 34–49.
53. Ziadat, M. F. (2005). Analyzing digital terrain attributes to predict soil attributes for a relatively large area, soil Sci. *Soc. Am. J., 69*, 1590–1599.
54. Segoni, S., Lagomarsino, D., Fanti, R., Moretti, S., & Casagli, N. (2015). Integration of rainfall thresholds and susceptibility maps in the Emilia Romagna (Italy) regional-scale landslide warning system. *Landslides, 12*, 773–785.
55. Trigila, A., Iadanza, C., Esposito, C., & Scarascia-Mugnozza, G. (2015). Comparison of logistic regression and random forests techniques for shallow landslide susceptibility assessment in Giampilieri (NE Sicily, Italy). *Geomorphology, 249*, 119–136.
56. Youssef, A. M., Pourghasemi, H. R., Pourtaghi, Z. S., & Al-Katheeri, M. M. (2015). Landslide susceptibility mapping using

random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at Wadi Tayyah Basin, Asir region, Saudi Arabia. *Landslides*. doi:10.1007/s10346-015-0614-1.

57. Bachmair, S., & Weiler, M. (2012). Hillslope characteristics as controls of subsurface flow variability. *Hydrology and Earth System Sciences, 16*, 3699–3715.
58. Vorpahl, P., Elsenbeer, H., Märker, M., & Schröder, B. (2012). How can statistical models help to determine driving factors of landslides? *Ecological Modelling, 239*, 27–39.
59. Díaz-Uriarte, R., & De Andrés, S. A. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*. doi:10.1186/1471-2105-7-3.
60. Liaw, A., & Wiener, M. (2002). Classification and regression by random Forest. *R News, 2*, 18–22.
61. Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*, 861–874.
62. Frattini, P., Crosta, G., & Carrara, A. (2010). Techniques for evaluating the performance of landslide susceptibility models. *Engineering Geology, 111*, 62–72.
63. Swets, J. (1988). Measuring the accuracy of diagnostic systems. *Science, 240*, 1285–1293.
64. Brenning, A. (2005). Spatial prediction models for landslide hazards: review, comparison and evaluation. *Natural Hazards and Earth System Sciences, 5*, 853–862.
65. IAEG (1990). Suggested nomenclature for landslides. *IAEG Bulletin, 41*, 13–16.
66. Bertolini, G., Casagli, N., Ermini, L., & Malaguti, C. (2004). Radiocarbon data on Lateglacial and Holocene landslides in the northern Apennines. *Natural Hazards, 31*, 645–662.
67. Catani, F., Casagli, N., Ermini, L., Righini, G., & Menduni, G. (2005). Landslide hazard and risk mapping at catchment scale in the Arno River basin. *Landslides, 2*, 329–342.
68. Trigila, A., Frattini, P., Casagli, N., Catani, F., Crosta, G., Esposito, C. et al. (2013). Landslide susceptibility mapping at national scale: the Italian case study. In *Landslide Science and Practice* (pp. 287–295). Berlin: Springer.
69. Carrara, A., Crosta, G., & Frattini, P. (2003). Geomorphological and historical data in assessing landslide hazard. *Earth Surf. Process. Landforms, 28*, 1125–1142.
70. Baeza, C., & Corominas, J. (2001). Assessment of shallow landslide susceptibility by means of multivariate statistical techniques. *Earth Surf. Process. Landforms, 26*, 1251–1263.
71. Segoni, S., Rossi, G., & Catani, F. (2012). Improving basin-scale shallow landslides modelling using reliable soil thickness maps. *Natural Hazards, 61*, 85–101.
72. Godt, J. W., Baum, R. L., Savage, W. Z., Salciarini, D., Schulz, W. H., & Harp, E. L. (2008). Transient deterministic shallow landslide modeling: requirements for susceptibility and hazard assessments in a GIS framework. *Engineering Geology, 102*(3–4), 214–226.
73. Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*, 1157–1182.
74. Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics, 9*, 307. doi:10.1186/1471-2105-9-307.
75. Yilmaz, I. (2010). The effect of the sampling strategies on the landslide susceptibility mapping by conditional probability and artificial neural networks. *Environmental Earth Sciences, 60*, 505–519.