A data-driven model for stability condition prediction of soil embankments based on visual data features

Joaquim Tinoco[1], A. Gomes Correia[2], Paulo Cortez[3], and David G. Toll[4]

[1]PhD, ISISE - Institute for Sustainability and Innovation in Structural Engineering/ALGORITMI Research Center, School of Engineering, University of Minho, Guimarães, Portugal. Email: jtinoco@civil.uminho.pt

[2]Full Professor, ISISE - Institute for Sustainability and Innovation in Structural Engineering, School of Engineering, University of Minho, Guimarães, Portugal. Email: agc@civil.uminho.pt

[3]Associate Professor, ALGORITMI Research Center/Department of Information Systems, University of Minho, Guimarães, Portugal. Email: pcortez@dsi.uminho.pt

[4]Full Professor, School of Engineering and Computing Sciences, University of Durham, Durham, UK. Email: d.g.toll@durham.ac.uk

ABSTRACT

Keeping large-scale transportation infrastructure networks, such as railway networks, operational under all conditions is one of the major challenges today. The budgetary constraints for maintenance purposes and the network dimension are two of the main factors that make the management of a transportation network such a challenging task. Accordingly, aiming to assist the management of a transportation network, a data-driven model is proposed for stability condition prediction of embankment slopes. For such purpose, the highly flexible learning capabilities of Artificial Neural Networks (ANN) and Support Vector Machines (SVM) were used to fit data-driven models for Earthwork Hazard Category (EHC) prediction. Moreover, the data-driven models were created using visual information that is easy to collect during routine inspections. The proposed models were addressed following two different data modeling strategies: nominal classification and regression. Moreover, to overcome the problem of imbalanced data (since typically good conditions are much common than bad ones), three training sampling approaches were explored: no resampling, SMOTE and Oversampling. The achieved modeling results are presented and discussed, comparing the predictive performance of ANN and SVM algorithms, as well as the effect of the sampling approaches. A comparison between nominal classification and regression strategies was also carried out. Moreover, aiming a better understanding of the proposed data-driven

models, a detailed sensitivity analysis was applied, allowing to quantify the relative importance of each model input, as well as measuring their global effect on the prediction of embankments stability conditions.

INTRODUCTION AND BACKGROUND

Transportation infrastructures play a key role in our modern society. Indeed, countries often invest in keeping or enhancing their transportation network, aiming at a safer and more functional infrastructure. Particularly for developed countries that already have a very well complete transportation network, the main challenge today is how to keep it operational under all conditions. From the point of view of a transportation network management, a key issue is how to identify the critical parts of the network that require budget allocation for their maintenance or repair.

Therefore, and in order to optimize the available budget, it is important to have a set of tools that help decision makers to identify such critical network parts and thus make the best decision about how to invest the available budget. In the framework of transportations networks, in particular for railways, slopes are perhaps the element for which their failure can have the strongest impact at several levels, including potentially major economic damage and loss of life. As such, it is important to develop ways to identify potential problems before they result in failures.

Although there are some models and systems to detect slope failures, most of them were developed for natural slopes, presenting some constraints when applied to engineered (human-made) slopes. Moreover, they have limited applicability at network level as most of the existing systems were developed based on particular case studies or using small databases. Furthermore, another aspect that can limit its applicability is related with the information required to feed them, such as data taken from complex tests or from expensive monitoring systems. Some approaches found in the research literature for slope failure detection are identified next.

Pourkhosravani and Kalantari (2011) summarizes the current methods for slope stability evaluation, which were grouped into Limit Equilibrium (LE) methods, Numerical Analysis methods, Artificial Neural Networks and Limit Analysis methods. There are also approaches based on finite elements methods (Suchomel et al. 2010), reliability analysis (Sivakumar Babu and Murthy 2005; Husein Malkawi et al. 2000), as well as some methods making use of soft computing algorithms (Gavin and Xue 2009; Wang and Sassa 2005; Cheng and Hoang 2016; Ahangar-Asr et al. 2010; Lu and Rosenbaum 2003; Sakellariou and Ferentinou 2005; Cheng et al. 2012b; Yao et al. 2008; Kang et al. 2015; Kang et al. 2016b; Kang and Li 2016; Kang et al. 2016a; Kang et al. 2017; Das et al. 2011; Suman et al. 2016). More recently, a new flexible statistical system was proposed by Pinheiro et al. (2015), based on the assessment of different factors that affect the behavior of a given slope. By weighting the different factors, a final indicator of the slope stability condition is calculated. In the summer of 2016, Power et al. (2016) presented an evidence-based asset management policy, which contemplates the development of a risk-based prioritisation matrix for all earthwork assets and determination of quantitative likelihood of earthwork failure.

As mentioned above, the main limitations of most approaches so far proposed, from the point of view of the network management, are related with their applicability domain or dependency on information that is difficult to obtain. Moreover, the prediction of whether a slope will fail or not is often a complex multi-variable modelling problem that is characterized by a high dimensionality. Accordingly, aiming to overcome this limitation, in this original work we take advantage of the learning capabilities of flexible DM algorithms, such as the Artificial Neural Networks (ANNs) and Support Vector Machines (SVMs), which can automatically learn from the data through complex nonlinear mappings. These DM algorithms were used to fit a large database of soil embankments slopes in order to predict their stability condition according to a pre-defined classification scale based on four levels (classes). One of the premises underlying this work is to identify the real stability condition of a given slope based on information that can be easily obtained through visual routine inspections. Aiming at such goal, more than fifty variables related with data collected during routine inspections, as well as geometric, geological and geographic data, were used to feed the data-driven models. This type of visual information is sufficient from the point of view of the network management, allowing the identification of critical zones for which more detailed information can then be obtained in order to perform more detailed stability analysis, which is out of the scope of this study.

Besides showing that interesting predictive performances can be achieved by the data-driven models, in this study also detailed sensitivity analysis was applied over the proposed models, allowing to identify the key variables in the stability condition prediction of soil embankments slopes, as well as quantify their global effect on the target variable.

In summary, this proposal will allow to identify the stability condition level of a given soil embankment slope based on visual information that, in most of the cases, can be easily obtained during routine inspections. Such novel approach is intended to support railway network management companies to allocate the available funds in the priority assets according to its stability condition.

DATA CHARACTERIZATION

The proposed model to identify the stability condition, from this point referred to as EHC (Earthwork Hazard Category, (Power et al. 2016)), of soil embankments were developed using DM techniques and considering a database containing 25673 records.

The EHC system comprises 4 classes ("A", "B", "C" and "D") where "A" represents a slope with good stability condition and "D" a slope with bad stability condition. In other words, the expected probability of failure is higher for class "D" and lower for class "A". To fit the model for EHC prediction of soil embankments, a database was compiled containing information collected during routine inspections and complemented with geometric, geological and geographic data of each slope. The databases was gathered by Networ Rail workers and is concerned with the railway network of the UK. For each slope a class of the EHC system was defined by the NetworkRail Engineers based on their experience, which will be assumed as proxy for the real stability condition of the slope for year 2015.

Figure 1 depicts the distribution of EHC classes. From its analysis, it is possible to observe a high asymmetric distribution (imbalanced data) of the records for each EHC class. Indeed, more than 63% of the embankments are classified as "A" and only 2.5%

belongs to class “D”. Although this type of asymmetric distribution, where most of the slopes present a low probability of failure (class “A”), is normal and desirable from the safety point of view and slope network management, it can represent an important challenge for DM models learning, as detailed in next section.

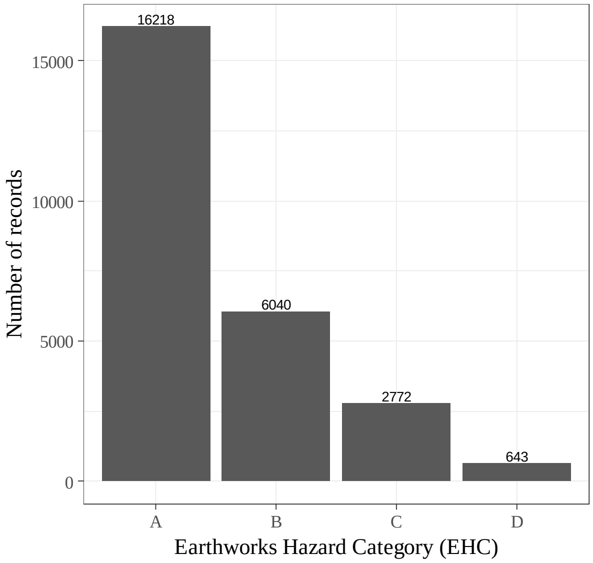

Fig. 1. Soil embankments data distribution by EHC classes.

For demonstrative purposes, Figure 2 shows the distribution of EHC across several input variables. Due to space limitations, the figure includes only the 20 most relevant input variables, as measured by a sensitivity analysis procedure (Cortez and Embrechts 2013) that was performed over the best data mining (DM) model (ANN model following an OVERed approach and according to a nominal classification strategy, Section “Model interpretation”). Nevertheless, it should be noted that DM models perform a multi-dimensional analysis, including interactions between inputs variables, which cannot be seen by the histograms of Figure 2.

The proposed models for EHC identification of soil embankments consider 53 variables normally collected during routine inspections as well as geometric, geographic and geological information. Bellow are listed all variables used during the present study as model inputs:

- Actual Angle1
- Actual Angle2
- Actual Angle3
- Actual Height1
- Actual Height2
- Actual Height3
- Actual Hyp1 (Hyp = Hypotenuse)
- Actual Hyp2
- Actual Hyp3
- Actual Slope To Track
- Adjacent Catch Area
- Adjacent Catch Gradient
- Adjacent Geology
- Adjacent Land Drainage
- Animal Activity
- Area
- Attitude Of Trees
- Ballast
- Catchment Geology

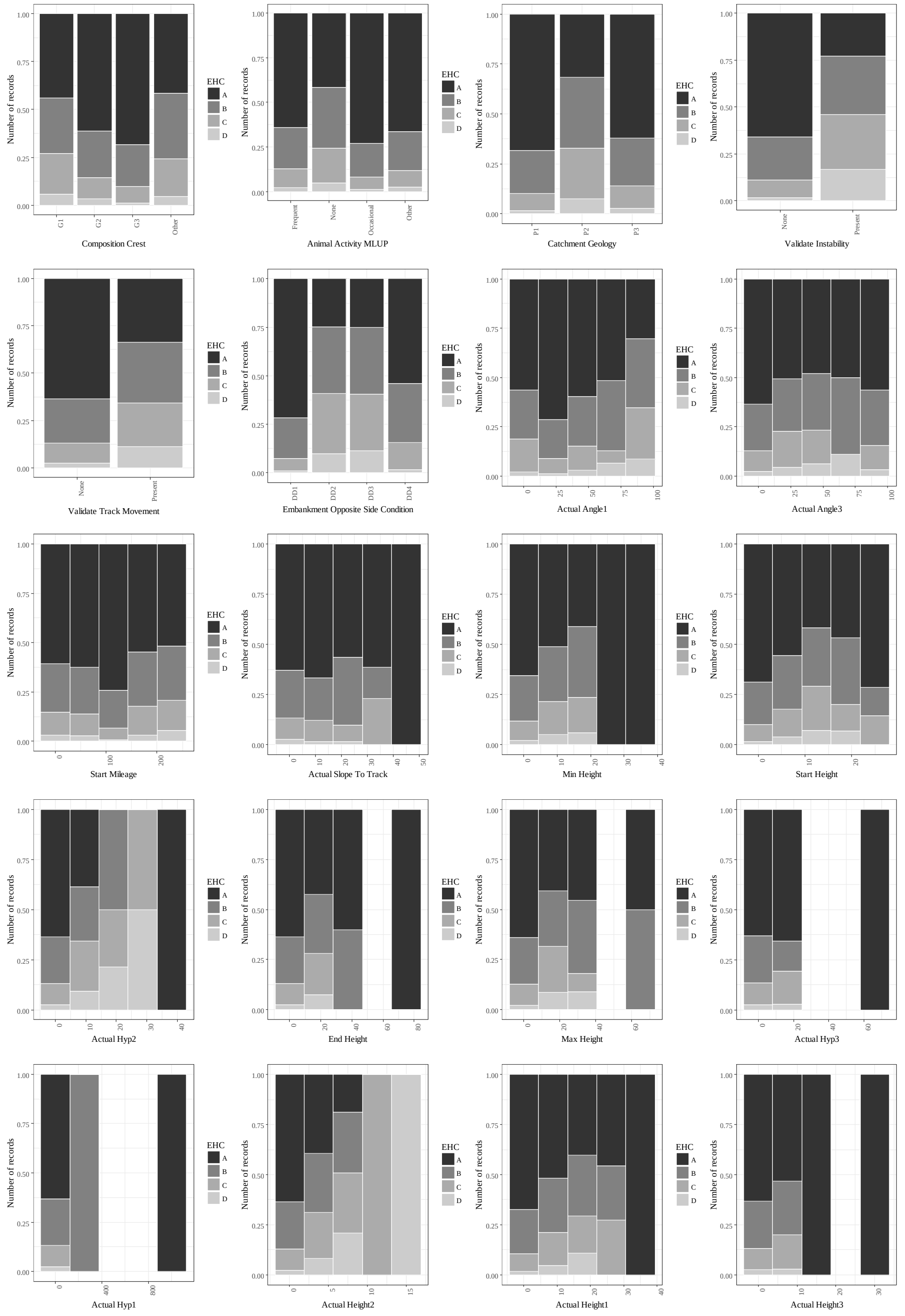

Fig. 2. Distribution of EHC across the 20 most relevant variables.

- Catchment Surface
- Class
- Composition Crest
- Composition Toe
- Construction Activity Toe

- Easting
- Engineer’s Line References (ELR)
- Embankment Opposite Side Condition
- Embankment Slope Toe Condition
- Embankment Sub Drainage
- End Height
- End Easting
- End Mileage
- End Northing
- Ground Cover
- Max Height
- Min Height
- Northing
- Operational Route
- Slope Angle Adjacent
- Slope Angle Height
- Slope To Track Separation
- Strategic Route (SR)
- Start Height
- Start Mileage
- Tree Cover
- Troughing
- Up or Down
- Validate Cracking
- Validate Instability
- Validate Mass Movement
- Validate Retaining Walls
- Validate Slope Form
- Validate Track Movement

METHODOLOGY

Modelling

To model EHC prediction of soil embankments two of the most popular DM algorithms, namely ANNs and SVMs were applied. Both algorithms had already been successful applied in different knowledge domains (Liao et al. 2012; Garg et al. 2014; Javadi et al. 2012) including in civil engineering (Tinoco et al. 2014a; Tinoco et al. 2014b; Gomes Correia et al. 2013; Miranda et al. 2011). There are also some examples of ANN and SVM applications in slope stability analysis (Wang et al. 2005; Yao et al. 2008; Cheng et al. 2012a).

ANN are learning machines that were initially inspired in functioning of the human brain (Kenig et al. 2001). The information is processed using iteration among several neurons. ANNs are considered nonlinear statistical data modelling tools where the complex relationships between inputs and outputs are modelled in order tp find patterns. This technique is capable of modelling complex non-linear mappings and is robust in exploration of data with noise. In this study the multilayer perceptron that contains only feedforward connections, with one hidden layer containing H processing units, was adopted.

Because the network’s performance is sensitive to H (a trade-off between fitting accuracy and generalisation capability), we adopt a grid search of $\{0, 2, 4, 6, 8\}$ under an internal (i.e. applied over training data) three fold cross validation during the learning phase to find the best H value. Such grid search only considered training data, dividing it into fitting (70%) and validation data (30%), where the validation error was used to select the best H. In other words, a different model (under a cross validation approach) was trained for each H value. Then the best model is selected according to the lowest validation error. After selecting the best H value, the ANN is retrained with the whole training data. The neural function of the hidden nodes was set to the popular logistic function $1/(1 + e^{-x})$. Hence, the general model of the ANN is given by Hastie et al.

(2009):

$$\hat{y} = w_{o,0} + \sum_{j=I+1}^{o-1} f\left(\sum_{i=1}^{I} x_i \cdot w_{j,i} + w_{j,0}\right) \cdot w_{o,i} \tag{1}$$

where $w_{j,i}$ represents the weight of the connection from neuron j to unit I (if $j = 0$, then it is a *bias* connection), o corresponds to an output unit, f is a logistic function and I is the number of input neurons. ANN optimization was done via the BFGS method (Venables and Ripley 2003). Method "BFGS" is a quasi-Newton method (also known as a variable metric algorithm), specifically that published simultaneously in 1970 by Broyden, Fletcher, Goldfarb and Shanno. This uses function values and gradients to build up a picture of the surface to be optimized (Cortez 2010).

SVMs are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships. They were initially proposed for classification tasks (Cortes and Vapnik 1995). Then it became possible to apply SVM to regression tasks after the introduction of the ϵ-insensitive loss function (Smola and Schölkopf 2004). The main purpose of the SVM is to transform input data into a high-dimensional feature space using non-linear mapping. The SVM then finds the best linear separating hyperplane, related to a set of support vector points, in the feature space. This transformation depends on a kernel function. In this work the popular Gaussian kernel (Hastie et al. 2009) was adopted. In this context, its performance is affected by three parameters: γ, the parameter of the kernel; C, a penalty parameter; and ϵ (only for regression), the width of a ϵ-insensitive zone (Safarzadegan Gilan et al. 2012). The heuristics proposed by Cherkassky and Ma (2004) were used to define the first two parameter values, C=3 (for a standardised output) and $\epsilon = \hat{\sigma}/\sqrt{N}$, where $\hat{\sigma} = 1.5/N \cdot \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$, y_i is the measured value, $\hat{y}_i$ is the value predicted by a 3-nearest neighbour algorithm and N is the number of examples. A grid search (similar to the one used for ANN) of $2^{\{-1,-3,-7,-9\}}$ was adopted to optimise the kernel parameter γ, under the same internal threefold cross-validation scheme adopted for ANN.

The problem of EHC prediction of soil embankments was initially approached following a nominal classification strategy. However, aiming to improve models performance, the problem was also addressed following a regression task by adopting a numeric regression scale instead of class labels, where A = 1, B = 2, C = 4, D = 10. The regression strategy requires the correct setting of the scale class values and for a given problem it is not clear what is the “best” scale. In this paper, the scale values were set using domain knowledge, putting higher distances for the more important classes (e.g. C and D). In addition, some other different regression scales (e.g. 1, 2, 20, 100) have been tested. However, the best performance was achieved using 1, 2, 4 10.

Moreover, in order to minimize the effect of the imbalanced data (Weiss and Provost 2003) (which is particular relevant for this application domain, see Figure 1), two resampling approaches were applied over the training data before fitting the models, namely Oversampling (Ling and Li 1998) and SMOTE (Chawla et al. 2002). When approaching imbalanced classification tasks, where there is at least one target class label with a smaller number of training samples when compared with other target class labels, the simple use of a DM training algorithm will often lead to data-driven models with better prediction accuracies for the majority classes and worst classification accuracies for the minority classes. Thus, techniques that adjust the training data in order to balance the output class labels, such as Oversampling and SMOTE, are commonly used with imbalanced datasets. In particular, Oversampling is a simple technique that randomly adds samples (with repetition) of the minority classes to the training data, such that the final training set is balanced. SMOTE is a more sophisticated technique that creates "new data" by looking at nearest neighbors to establish a neighborhood and then sampling from within that neighborhood. It operates on the assumptions that the original data is similar because of proximity. More recently, Torgo et al. (2015) adapted the SMOTE method for regression tasks. In the framework of SMOTE approach, and concerning to the nearest neighbour k value, several values were tried and at the end a $k = 3$ was adopted, which lead to the best overall performance. Figure 3 shows the flowchart of the methodology applied for EHC prediction of soil embankments slopes.

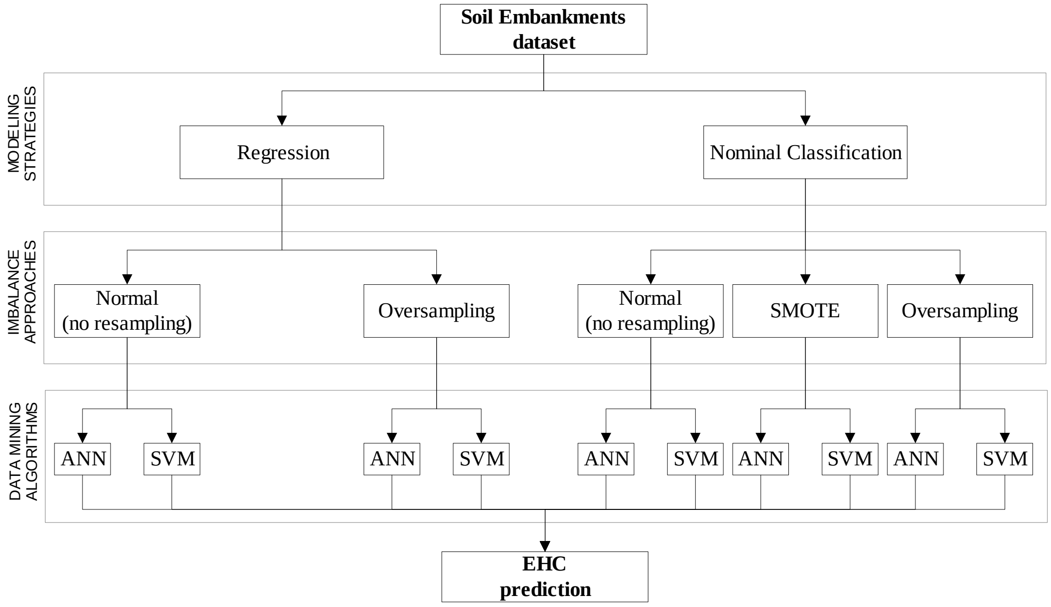

Fig. 3. Flowchart of the applied methodology.

All experiments were conducted using the R statistical environment (Team 2009) and supported through the rminer package (Cortez 2010), which facilitates the implementation of ANNs and SVMs algorithms, as well as different validation approaches such as cross-validation.

Models evaluation

For models evaluation and comparison we used four classification metrics: average utility score (AUS), recall, precision and F1-score.

A cost-benefit matrix (CBM) is used to compute the AUS (Baía and Torgo 2015), which averages all individual predictions in terms of their expected cost or benefit. This approach intends to calculate a metric more directly related to a particular real-world problem. In this work, CBM was set in order to reflects the ECH classification system and the characteristics of its slope identification tasks (Table 1). The assumption behind the adopted CBM was to penalise every misclassification but using different weights according to the "distance" of the misclassification and putting larger penalties to bad stability condition (the ones that are more important to be correctly classified). For example, if a particular soil slope was identified as class "A" (true condition), then the benefit is +1 if the model predicts the same class. For the same sample, the cost is -4 if the model predicts a class "C" and it doubles to -8 if the prediction is class "D". It should also be noted that the adopted CBM is not symmetrical. For example, predicting class "D" for a true observation of "A" leads to a cost of -8, which is half the cost when predicting class "A" for a true "D" slope condition.

Table 1. Cost-benefit matrix adopted for both rock and soil cuttings slopes studies.

Obs/Pred	A	B	C	D
A	1	-4	-8	-16
B	-2	1	-4	-4
C	-4	-2	1	-4
D	-8	-4	-2	1

The recall measures the ratio of how many cases of a certain class were properly captured by the model. In other words, the recall of a certain class is given by:

$$\frac{TruePositives}{TruePositives + FalseNegatives} \tag{2}$$

On the other hand, the precision measures the correctness of the model when it predicts a certain class. More specifically, the precision of a certain class is given by:

$$\frac{TruePositives}{TruePositives + FalsePositives} \tag{3}$$

The F1-score was also calculated, which represents a trade-off between the recall and precision of a class. The F1-score corresponds to the harmonic mean of precision and recall, according to the following expression:

$$2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{4}$$

For all four metrics, the higher the value, the better are the predictions. The AUS values can be negative (if on average, the predictions lead to a cost) and the ideal predictor will have an AUS of 1. The other metrics, recall, precision and F1-score can range from 0% to 100%.

The generalization capacity of the models was accessed through a 5-fold cross-validation approach under 20 runs (Hastie et al. 2009). This means that each modelling setup is trained $5 \times 20 = 100$ times. Also, the four prediction metrics are always computed on test unseen data (as provided by the 5-fold validation procedure).

Sensitivity analysis

As important as model performance is its interpretability. It is well known that data-driven models, namely those based on ANNs ao AVMs algorithms, are often viewed as complex models, which are difficult to be understood by human experts. Indeed, they are often treated as "black box" models. Yet, in critical domains it is important to extract human understandable data-driven knowledge. For such purpose, in this work a detailed sensitivity analysis (SA) (Cortez and Embrechts 2013) we applied over the proposed predictive models. SA is a simple method that is executed after the training phase and measures the model responses when a given input is changed, allowing to quantify the relative importance of each attribute, as well as its average effect on the target variable.

In particular, the global sensitivity analysis (GSA) method (Cortez and Embrechts 2011) was applied, which is able to detect interactions among input attributes. This is achieved by performing a simultaneous variation of F inputs (that can range from 1, one dimensional SA, denoted as 1-D, to I, I-D SA). Each input is varied through its range with L levels and the remaining inputs are kept fix to a b baseline value. In this work, the number of levels was set to 7 ($L = 7$), which allows an interesting detail level under a reasonable amount of computational effort; and b is the average input variable value.

First, the DM model is fitted to the whole dataset. Then, the GSA algorithm is applied to the fitted DM model, being the respective sensitivity responses stored. Next, using these responses, two important visualization techniques can be computed. The input importance bar plot shows the relative influence of each input variable (from 0% to 100%). The rational of SA is that the higher the changes produced in the output, the more important is the input. To measure this effect, and following the suggestion of Cortez and Embrechts (2011), the gradient metric was adopted:

$$g_a = \sum_{j=2}^{L} \left| \hat{y}_{a,j} - \hat{y}_{a,j-1} \right| / (L-1) \tag{5}$$

where a denotes the input variable under analysis, $\hat{y}_{a,j}$ is the sensitivity response for $x_{a,j}$. Having computed the gradient for all inputs, then the relative importance (R_a) is calculated using:

$$R_a = g_a / \sum_{i=1}^{I} g_i \cdot 100(\%) \tag{6}$$

To analyse the average impact of a given input x_a in the fitted model, the variable effect characteristic (VEC) curve can be used, which plots the attribute L level values (x-axis) versus the SA responses (y-axis). Between two consecutive $x_{a,j}$ values, the VEC plot performs a linear interpolation. To enhance the visualization analysis, several VEC curves can be plotted in the same graph. In such case, the x-axis is scaled (e.g. within [0,1]) for all x_a values.

RESULTS AND DISCUSSION

This section summarizes the main results achieved in EHC prediction of soil embankments through the application of soft computing techniques. Two different DM algorithms (ANN and SVM) were applied for EHC prediction under two distinct modelling strategies: nominal classification and regression. Moreover, in order to overcome the problem of imbalanced data, three training sampling approaches were explored: Normal (no resampling), OVERed (Oversampling) and SMOTEd (SMOTE). In the case of regression, two sampling approaches are compared: Normal (no resampling) and SMOTEd (SMOTE for regression). It should be noted that the different sampling approaches were applied only to training data, used to fit the data-driven models, and the test data (as provided by the 5-fold procedure) was kept without any change.

Models performance

Table 2 summarizes AUS, recall, precision and F1-score of all fitted models for EHC prediction of soil embankments, according to a nominal classification and regression strategies, as well as using SMOTE and Oversampling resampling approaches. For a better analysis and models comparison, Figure 4 compares recall, precision and F1-score metrics of all models in EHC prediction following a nominal classification strategy. The proposed models, particularly those based on ANNs algorithm, are able to identify very accurately soil embankments of class "A", observing a slightly decreasing on its performance for the other three classes. Considering F1-score as reference, for class "A" a value higher than 92% was achieved. Concerning to class "D" also a very promising performance is observed with an F1-score around 55%. Comparing ANN and SVM algorithms, its clear that the first one performs better, particularly for classes "C" and "D", where the probability of failure is higher.

Analysing the effect of the training sampling approaches (oversampling e SMOTE), it is observed some effectiveness for class "D" (minority class). For the other classes, the application of a sampling approach seems to be ineffective. Indeed, and considering F1-Score as reference, better results are achieved with no resampling. These results show that applying a training sampling approach allows to improve models performance in the identification of the minority classes but decreasing its response for the other classes. In fact, taking in account that these training sampling approaches are tailed to address learning problems related with the minority classes in imbalanced datasets, it is acceptable and expected to observe a slightly decrease in the majority classes performance. Comparing oversampling and SMOTE approaches, the first one seems to be more effective.

Figure 5 compares models performance based on recall, precision and F1-score metrics following a regression strategy. Also here, ANNs performs better than SVMs, particularly for class "C" and mainly for class "D". On the other hand, and similarly to the nominal classification strategy, a very high accuracy is observed in stability condition identification of soil embankments of class "A", with an F1-score higher than 91%. For the remaining classes, models performance decreases slightly, have achieving an F1-score around 58% for class "D", according to ANN algorithm. Following a regression strategy, the application of a resampling approach, i.e., SMOTE sampling, has a residual effect on models performance, even for minority classes.

Comparing both nominal classification and regression strategies based on AUS metric, Figure 6 shows that approaching the problem as a nominal classification is slightly more effective than following a regression strategy. Moreover, Figure 6 also

Table 2. Metrics in EHC prediction of soil embankments (best values in bold)

Strategy	Model	Approach	AUS	Recall				Precision				F1-score			
				A	B	C	D	A	B	C	D	A	B	C	D
Classification	ANN	Normal	**0.28**	94.14	68.53	62.22	45.29	91.05	**69.36**	**68.55**	69.18	**92.57**	**68.94**	65.23	54.74
		SMOTEd	0.18	86.31	72.75	48.85	**65.93**	93.88	59.34	63.03	35.07	89.94	65.36	55.04	45.79
		OVERed	0.24	86.25	67.59	**67.12**	**65.29**	94.33	60.12	57.87	50.05	90.11	63.64	62.15	56.66
	SVM	Normal	0.08	**95.03**	64.89	49.76	0.55	88.59	64.66	62.41	77.17	91.70	64.77	55.37	1.09
		SMOTEd	−0.12	76.39	72.23	47.26	37.94	94.17	49.13	45.13	33.22	84.35	58.48	46.17	35.42
		OVERed	−0.35	89.82	53.72	38.75	14.00	81.75	55.39	58.36	57.67	85.60	54.54	46.58	22.53
Regression	ANN	Normal	0.21	93.53	64.53	64.38	50.33	90.23	67.89	67.27	69.30	91.85	66.17	65.79	**58.31**
		SMOTEd	0.27	90.21	71.00	**67.91**	40.43	92.60	64.40	65.37	77.92	91.39	67.54	**66.62**	53.24
	SVM	Normal	0.06	86.40	82.60	36.94	0.08	93.58	55.34	60.79	**100**	89.85	66.28	45.95	0.16
		SMOTEd	−0.08	73.01	**84.59**	50.21	3.71	**95.91**	46.55	59.86	89.66	82.91	60.05	54.61	7.13

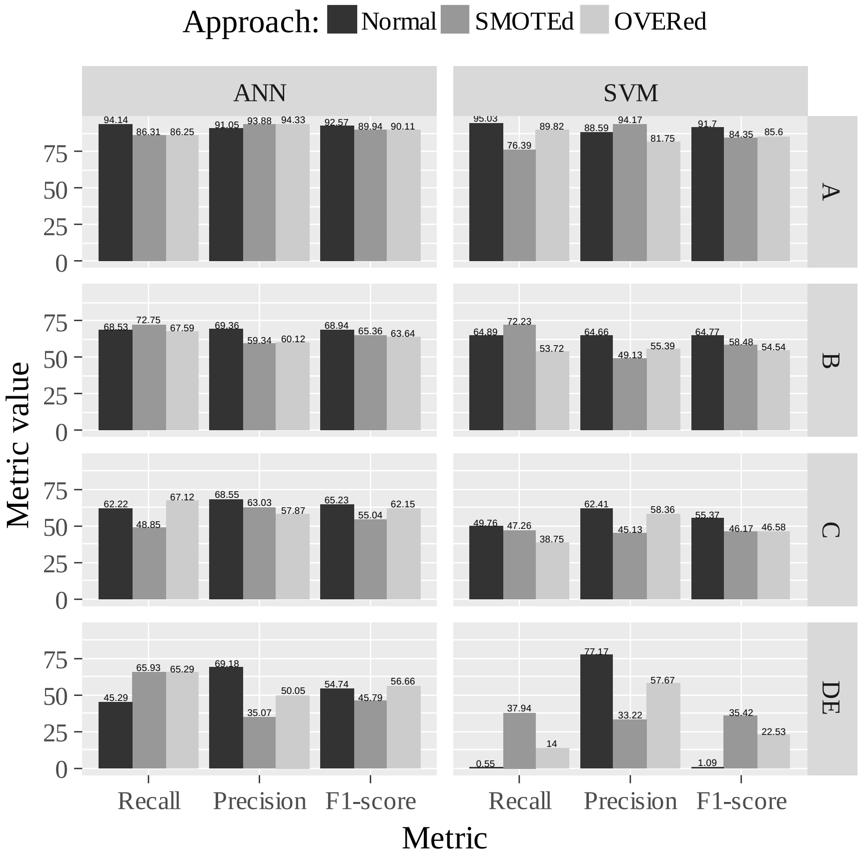

Fig. 4. Models comparison based on recall, precision and F1-score, according to a nominal classification strategy in EHC prediction of soil embankments.

illustrates the higher performance of ANNs in stability condition identification of soil embankments when compared with SVMs. Furthermore, keeping in mind that the ideal predictor has an AUS of 1, the highest value of 0.28 (achieved by ANN algorithm with no

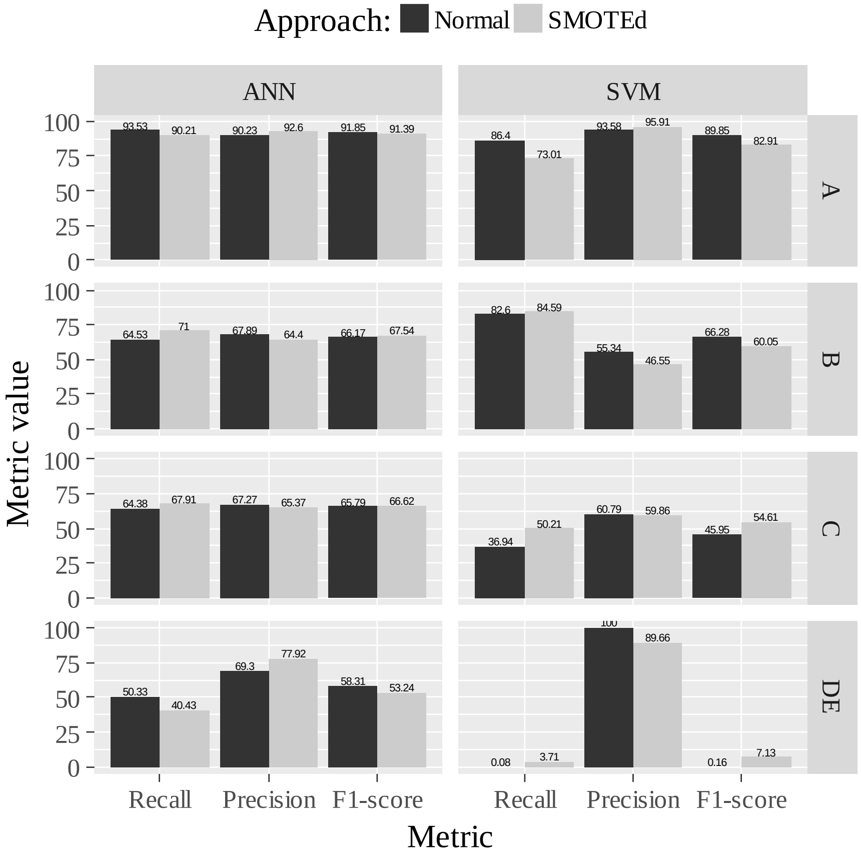

Fig. 5. Models comparison based on recall, precision and F1-score, according to a regression strategy in EHC prediction of soil embankments.

resampling), shows that, on average, models performance require further improvements.

Figures 7 and 8 show the relation between observed and predicted EHC values according to the best fits, following a nominal classification and regression strategies respectively. From its analysis, it is clear the superior performance of ANNs in stability condition identification of soil embankments when compared to SVMs. Indeed, SVM algorithm has some difficulty to correctly identify soil embankments of class "D". Besides that, it is also possible to observe that soil embankments of class "A" are very well identified by all models, particularly those based on ANN algorithm. Moreover, soil embankments of class "D" are better identified when a resampling approach is applied, as depicted in Figure 7a and Figure 7b.

In overall, one can conclude that the best model for stability condition identification of soil embankments is those based on ANN algorithm, by applying a oversampling approach and following a nominal classification strategy. Although the best metrics values (e.g. AUS or F1-score) are related to ANN model with no resampling and following a nominal classification strategy, comparing Figure 7a and Figure 7b the last one (ANN with oversampling) is more efficient in class "D" identification (more than 63%), which is a key point within the problem domain due to their highest probability

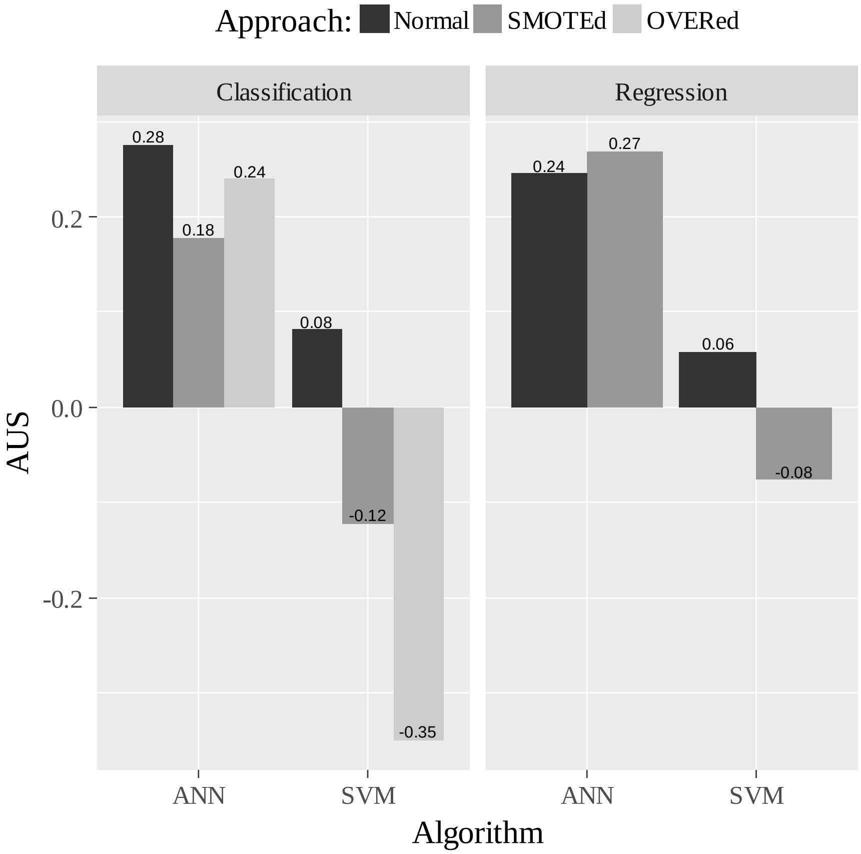

Fig. 6. Comparison of models performance in EHC prediction of soil embankments, based on AUS metric.

of failure. Moreover, and according to this model, it is also interesting to observe that when the stability condition of a given slope is not correctly identified, such slope is classified as belonging to the nearest class. For example, almost all soil embankments of class “D” not identified as it are classified as belong to class “C”.

Models interpretation

As importance as model performance is its interpretability, in particular when are involved data-driven models, namely those based on ANNs or SVMs algorithms. Due to its mathematical complexity, such models are difficult to understand and are usually treated as “black box”. Therefore, it is important to “open” such models in order to understand what have been learned by them. In this work a GSA methodology (Cortez and Embrechts 2013) was applied aiming to identify the key parameters (input importance bar plot, 1-D SA) and their average influence on the output response (VEC curves).

Figure 9 shows the relative importance of the twenty more relevant variables according to the ANN model with oversampling and following a nominal classification strategy, which was identified as the best model for stability condition identification of soil embankments and will be used from this point for model interpretation. Thus, and according to this model, three of the most relevant variables in EHC prediction of soil

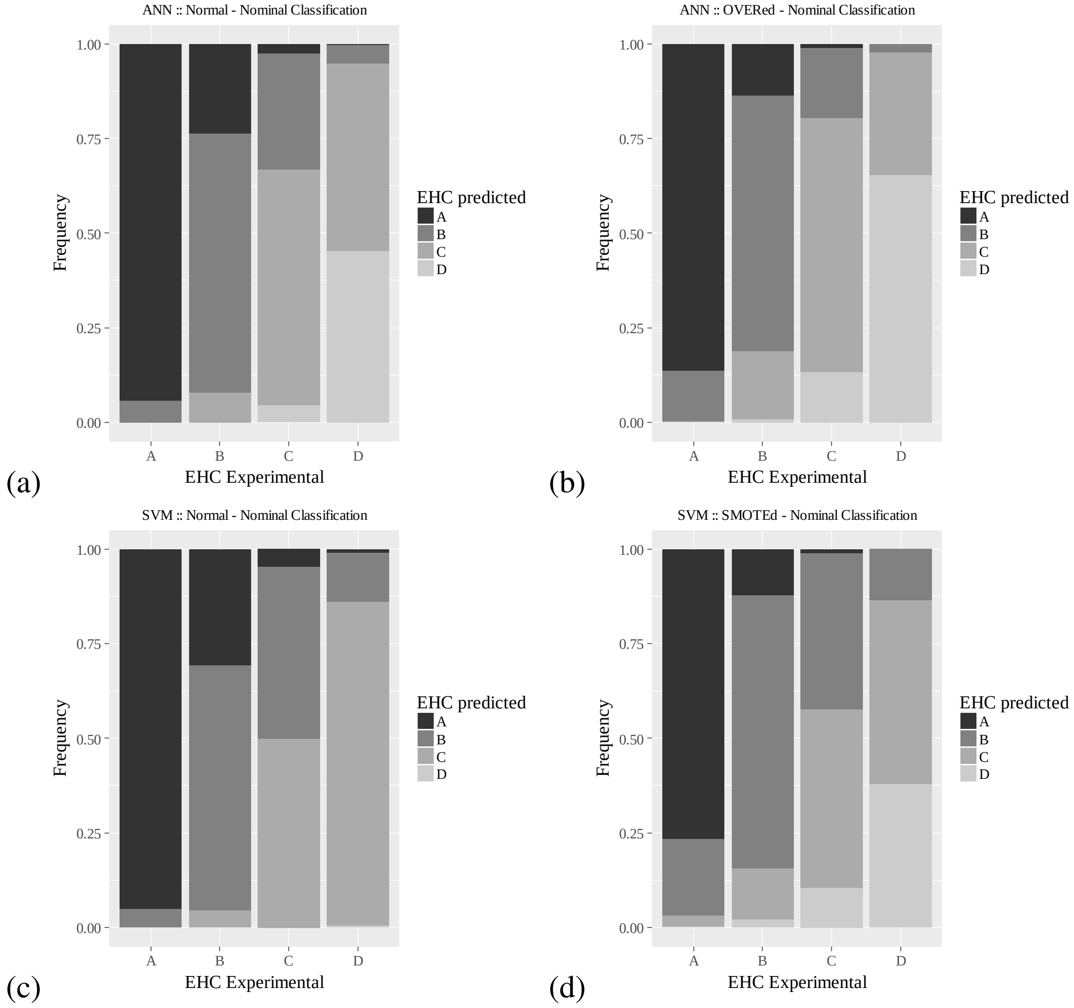

Fig. 7. Models performance comparison according to a nominal classification strategy in EHC prediction of soil embankments: (a) ANN model with no resampling; (b) ANN model following an OVERed approach; (c) SVM model with no resampling; (d) SVM model following an SMOTEd approach.

embankments are related with the height of the slope, summing more than 20% of the total influence. Moreover, "Embankment Opposite Side Condition" as well as "Validate Track Movement" also play an important role in EHC prediction of soil embankments.

In addition to the relative importance of each model attribute, it was also measured the effect of the key variables in stability condition prediction of soil embankments. Figure 10a plots the influence of "Actual Height3" in the probability of each EHC class. As expected, increasing slope height, the probability of a soil embankment be classified as "A" decreases. Indeed, if the slope height increase until 10 meters, the probability of such slope be classified as "A" decrease more than 0.7 points. On Figure 10b is depicted the effect of the second most relevant variable in EHC prediction of soil embankments ("Actual Height1"). In this case, the influence of "Actual Height1" in the probability of class "A" is not so pronounced. For example, while for an "Actual Height3" higher than

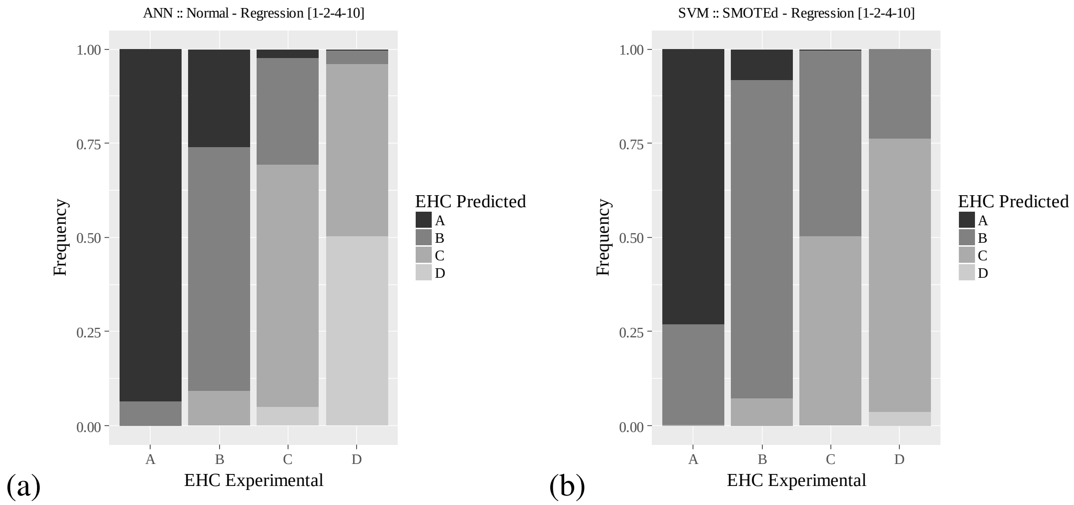

Fig. 8. Models performance comparison according to a regression strategy in EHC prediction of soil embankments: (a) ANN model with no resampling; (b) SVM model following an SMOTEd approach.

20m the probability of class “A” decrease to lower than 25%, the probability of class “A” for an “Actual Height1” is always higher than 27%.

FINAL REMARKS

This study is the first known attempt to predict EHC (Earthwork Hazard Category), assessed by four class (“A”, “B”,“C” and “D”), of soil embankments through the application of soft computing techniques and considering as model attributes information usually collected during routine inspections (visual information). A very promising predictive performance was observed, with F1-score values higher than 92% for class “A”, around 66% for classes “B” and “C” and close to 57% for class “D”. Moreover, it was shown that Artificial Neural Networks (ANN) perform better EHC prediction when compared with Support Vector Machines (SVM). In addition, the application of a resampling approach (aiming to overcome the problem of imbalanced data), namely the Oversampling method, allows to improve the predictive performance, particularly for the minority class “D”. Finally, a sensitivity analysis approach was used to open the proposed ANN model, revealing that three of the most relevant inputs for EHC prediction of soil embankments (accounting for 20% of the influence) are related with the height of the slope, which was expected from a geotechnical point of view.

As a final observation, the overall performance achieved in EHC prediction of soil embankments based on soft computing techniques, open good expectations for pursuing in further developments. In particular, and taking into account the high number of variables used as models inputs, in future works it is intended to reduce the number of variables trough the application of some feature selection (e.g., through the application of genetic algorithms). This will allow reducing models complexity and eventually improving their performance. Moreover, important contributions can also be taken to support further developments by analysing what have been done so far, namely the

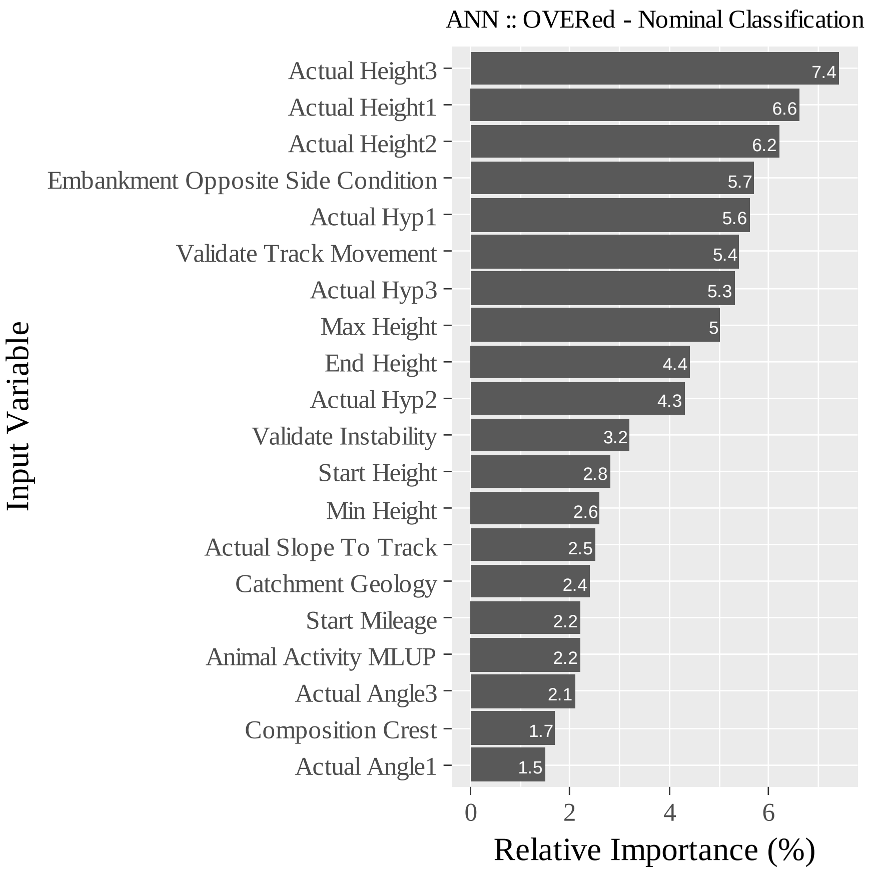

Fig. 9. Relative importance bar plot for each variable according to ANN model with oversampling and following a nominal classification strategy.

different strategies/approaches applied in order to overcome the different particularities of the problem at hands.

ACKNOWLEDGEMENTS

This work was supported by FCT - “Fundação para a Ciência e a Tecnologia”, within Institute for Sustainability and Innovation in Structural Engineering (ISISE), project UID/ECI/04029/2013 as well Project Scope: UID/CEC/00319/2013 and through the post-doctoral Grant fellowship with reference SFRH/BPD/94792/2013. This work was also partly financed by FEDER (Fundo Europeu de Desenvolvimento Regional) funds through the Competitivity Factors Operational Programme - COMPETE and by national funds through FCT within the scope of the project POCI-01-0145-FEDER-007633. This work has been also supported by COMPETE: POCI-01-0145-FEDER-007043. A special thanks goes to Network Rail that kindly make available the data (basic earthworks examination data and the Earthworks Hazard Condition scores) used in this work.

REFERENCES

Ahangar-Asr, A., Faramarzi, A., and Javadi, A. A. (2010). “A new approach for pre-

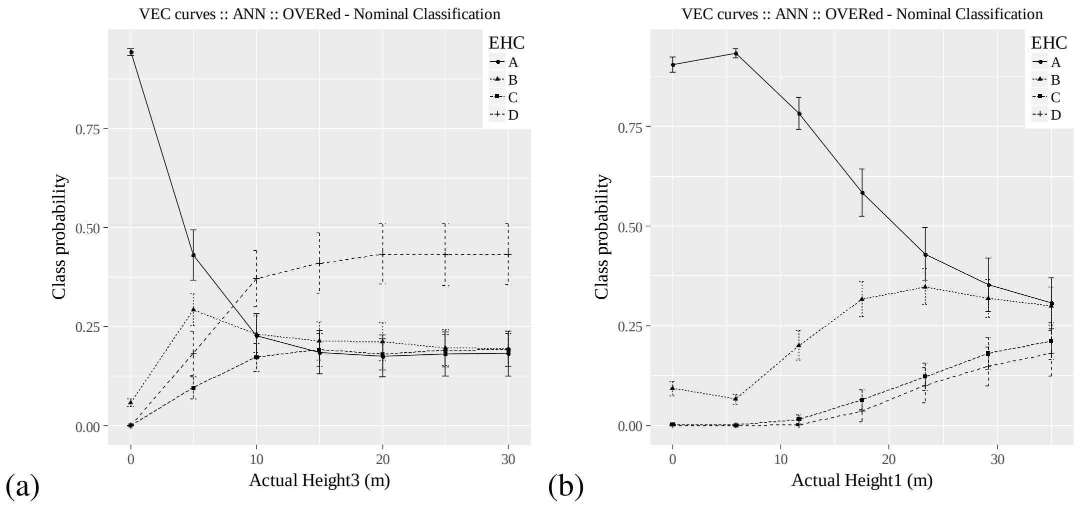

Fig. 10. VEC curves according to ANN model with oversampling and following a nominal classification strategy: (a) “Actual Height3”, (b) “Actual Height1”

.

diction of the stability of soil and rock slopes.” *Engineering Computations*, 27(7), 878–893.

Baía, L. and Torgo, L. (2015). “Forecasting the correct trading actions.” *Proceedings of EPIA 2015*, L. 9273, ed., Springer, 560–571.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). “Smote: synthetic minority over-sampling technique.” *Journal of artificial intelligence research*, 16(1), 321–357.

Cheng, M.-Y. and Hoang, N.-D. (2016). “Slope collapse prediction using bayesian framework with k-nearest neighbor density estimation: Case study in taiwan.” *Journal of Computing in Civil Engineering*, 30(1), 04014116.

Cheng, M.-Y., Roy, A. F., and Chen, K.-L. (2012a). “Evolutionary risk preference inference model using fuzzy support vector machine for road slope collapse prediction.” *Expert Systems with Applications*, 39(2), 1737–1746.

Cheng, M.-Y., Wu, Y.-W., and Chen, K.-L. (2012b). “Risk preference based support vector machine inference model for slope collapse prediction.” *Automation in Construction*, 22(Mar), 175–181.

Cherkassky, V. and Ma, Y. (2004). “Practical selection of svm parameters and noise estimation for svm regression.” *Neural Networks*, 17(1), 113–126.

Cortes, C. and Vapnik, V. (1995). “Support vector networks.” *Machine Learning*, 20(3), 273–297.

Cortez, P. (2010). “Data mining with neural networks and support vector machines using the r/rminer tool.” *Advances in Data Mining: Applications and Theoretical Aspects, 10th Industrial Conference on Data Mining*, P. Perner, ed., Berlin, Germany, LNAI 6171, Springer, 572–583 (July).

Cortez, P. and Embrechts, M. (2011). “Opening black box data mining models using sensitivity analysis.” *2011 IEEE Symposium on Computational Intelligence and Data*

Mining, CIDM 2011, Paris, France, IEEE, 341–348 (April).

Cortez, P. and Embrechts, M. (2013). “Using sensitivity analysis and visualization techniques to open black box data mining models.” *Information Sciences*, 225(Mar), 1–17.

Das, S. K., Biswal, R. K., Sivakugan, N., and Das, B. (2011). “Classification of slopes and prediction of factor of safety using differential evolution neural networks.” *Environmental Earth Sciences*, 64(1), 201–210.

Garg, A., Garg, A., Tai, K., and Sreedeep, S. (2014). “An integrated srm-multi-gene genetic programming approach for prediction of factor of safety of 3-d soil nailed slopes.” *Engineering Applications of Artificial Intelligence*, 30, 30–40.

Gavin, K. and Xue, J. (2009). “Use of a genetic algorithm to perform reliability analysis of unsaturated soil slopes.” *Geotechnique*, 59(6), 545–549.

Gomes Correia, A., Cortez, P., Tinoco, J., and Marques, R. (2013). “Artificial intelligence applications in transportation geotechnics.” *Geotechnical and Geological Engineering*, 31(3), 861–879 doi:10.1007/s10706-012-9585-3.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag New York, second edition edition.

Husein Malkawi, A. I., Hassan, W. F., and Abdulla, F. A. (2000). “Uncertainty and reliability analysis applied to slope stability.” *Structural Safety*, 22(2), 161–187.

Javadi, A. A., Ahangar-Asr, A., Johari, A., Faramarzi, A., and Toll, D. (2012). “Modelling stress–strain and volume change behaviour of unsaturated soils using an evolutionary based data mining technique, an incremental approach.” *Engineering Applications of Artificial Intelligence*, 25(5), 926–933.

Kang, F., Han, S., Salgado, R., and Li, J. (2015). “System probabilistic stability analysis of soil slopes using gaussian process regression with latin hypercube sampling.” *Computers and Geotechnics*, 63(Jan), 13–25.

Kang, F. and Li, J. (2016). “Artificial bee colony algorithm optimized support vector regression for system reliability analysis of slopes.” *Journal of Computing in Civil Engineering*, 30(3), 04015040.

Kang, F., Li, J.-s., and Li, J.-j. (2016a). “System reliability analysis of slopes using least squares support vector machines with particle swarm optimization.” *Neurocomputing*, 209(Oct), 46–56.

Kang, F., Li, J.-S., Wang, Y., and Li, J. (2017). “Extreme learning machine-based surrogate model for analyzing system reliability of soil slopes.” *European Journal of Environmental and Civil Engineering*, 1–22.

Kang, F., Xu, Q., and Li, J. (2016b). “Slope reliability analysis using surrogate models via new support vector machines with swarm intelligence.” *Applied Mathematical Modelling*, 40(11), 6105–6120.

Kenig, S., Ben-David, A., Omer, M., and Sadeh, A. (2001). “Control of properties in injection molding by neural networks.” *Engineering Applications of Artificial Intelligence*, 14(6), 819–823.

Liao, S., Chu, P., and Hsiao, P. (2012). “Data mining techniques and applications. A decade review from 2000 to 2011.” *Expert Systems with Applications*, 39(12), 11303–11311.

Ling, C. X. and Li, C. (1998). “Data mining for direct marketing: Problems and solutions.” *KDD’98 Proc. Fourth Int. Conf. on Knowledge Discovery and Data Mining*, AAAI Press, New York, 73–79.

Lu, P. and Rosenbaum, M. (2003). “Artificial neural networks and grey systems for the prediction of slope stability.” *Natural Hazards*, 30(3), 383–398.

Miranda, T., Gomes Correia, A., Santos, M., Sousa, L., and Cortez, P. (2011). “New models for strength and deformability parameter calculation in rock masses using data-mining techniques.” *International Journal of Geomechanics*, 11, 44–58.

Pinheiro, M., Sanches, S., Miranda, T., Neves, A., Tinoco, J., , Ferreira, A., and Gomes Correia, A. (2015). “A new empirical system for rock slope stability analysis in exploitation stage.” *International Journal of Rock Mechanics and Mining Sciences*, 76(Jun), 182–191 http://dx.doi.org/10.1016/j.ijrmms.2015.03.015.

Pourkhosravani, A. and Kalantari, B. (2011). “A review of current methods for slope stability evaluation.” *Electronic Journal of Geotechnical Engineering*, 16(Jan), 1245–1254.

Power, C., Mian, J., Spink, T., Abbott, S., and Edwards, M. (2016). “Development of an evidence-based geotechnical asset management policy for network rail, great britain.” *Procedia Engineering*, 143(Sep), 726–733.

Safarzadegan Gilan, S., Bahrami Jovein, H., and Ramezanianpour, A. (2012). “Hybrid support vector regression–particle swarm optimization for prediction of compressive strength and rcpt of concretes containing metakaolin.” *Construction and Building Materials*, 34(Sep), 321–329.

Sakellariou, M. and Ferentinou, M. (2005). “A study of slope stability prediction using neural networks.” *Geotechnical & Geological Engineering*, 23(4), 419–445.

Sivakumar Babu, G. and Murthy, D. (2005). “Reliability analysis of unsaturated soil slopes.” *Journal of geotechnical and geoenvironmental engineering*, 131(11), 1423–1428.

Smola, A. and Schölkopf, B. (2004). “A tutorial on support vector regression.” *Statistics and Computing*, 14(3), 199–222.

Suchomel, R. et al. (2010). “Comparison of different probabilistic methods for predicting stability of a slope in spatially variable< i> c</i>–φ soil.” *Computers and Geotechnics*, 37(1), 132–140.

Suman, S., Khan, S., Das, S., and Chand, S. (2016). “Slope stability analysis using artificial intelligence techniques.” *Natural Hazards*, 84(2), 727–748.

Team, R. (2009). “R: A language and environment for statistical computing. R Foundation for Statistical Computing, Viena, Austria. Web site: http://www.r-project.org/.

Tinoco, J., Gomes Correia, A., and Cortez, P. (2014a). “A novel approach to predicting young’s modulus of jet grouting laboratory formulations over time using data mining techniques.” *Engineering Geology*, 169(Feb), 50–60 http://dx.doi.org/10.1016/j.enggeo.2013.11.015.

Tinoco, J., Gomes Correia, A., and Cortez, P. (2014b). “Support vector machines applied to uniaxial compressive strength prediction of jet grouting columns.” *Computers and Geotechnics*, 55(Jan), 132–140 http://dx.doi.org/10.1016/j.compgeo.2013.08.010.

Torgo, L., Branco, P., Ribeiro, R., and Pfahringer, B. (2015). “Resampling strategies for regression.” *Expert Systems*, 32(3), 465–476.

Venables, W. and Ripley, B. (2003). *Modern Applied Statistics with S*. Springer Heidelberg, second edition edition.

Wang, H. and Sassa, K. (2005). “Comparative evaluation of landslide susceptibility in minamata area, japan.” *Environmental Geology*, 47(7), 956–966.

Wang, H., Xu, W., and Xu, R. (2005). “Slope stability evaluation using back propagation neural networks.” *Engineering Geology*, 80(3), 302–315.

Weiss, G. M. and Provost, F. (2003). “Learning when training data are costly: The effect of class distribution on tree induction.” *Journal of Artificial Intelligence Research*, 19, 315–354.

Yao, X., Tham, L., and Dai, F. (2008). “Landslide susceptibility mapping based on support vector machine: a case study on natural slopes of hong kong, china.” *Geomorphology*, 101(4), 572–582.