COMPARING MODEL PERFORMANCE METRICS FOR LANDSLIDE SUSCEPTIBILITY MAPPING

V.Yordanov [1,2], M.A. Brovelli [1*]

[1] Department of Civil and Environmental Engineering (DICA)
Politecnico di Milano, Piazza Leonardo da Vinci 32, Milan, Italy - (vasil.yordanov, maria.brovelli)@polimi.it
[2] Vasil Levski National Military University, Veliko Tarnovo, Bulgaria

Commission IV, WG III/IVa

KEYWORDS: Landslide, Susceptibility map, Validation, Logistic Regression, Random Forest

ABSTRACT:

Landslides are one of the most diffused hazard events in the world, they can occur in different locations under different triggering factors. As such, they are also one of the most studied hazards, while the mechanism of an event is known to the scholars, more difficulties are found in forecasting the location and time of the following event. However, scholars are putting great effort into modelling the phenomena through various tools, as such susceptibility mapping is one of the initial and key steps in the hazard assessment. While effort is put on producing such maps, less is put on the evaluation of those outcomes. The current work aims to analyse the behaviour of two validation metrics – Receiver Operating Characteristics (ROC) and Precision Recall Curve (PRC). The former is widely used in susceptibility modelling, while the latter not so much utilized. However, scholars are highlighting a drawback of the ROC – it is not able to discriminate imbalanced datasets and is providing unreliable outcomes, and as an alternative is proposed the PRC which does not exhibit such flaws. In order to test the performance of both metrics, they were applied to three susceptibility models produced using Statistical Index, Logistic Regression and Random Forest for the area of Val Tartano, Northern Italy. As a result, it was determined that when the metrics are applied to balanced datasets they exhibit similar behaviour; on the contrary when imbalanced classes are introduced PRC is depicting the model performance in a more precise manner.

1. INTRODUCTION

Landslide phenomena are widely spread hazards all over the world, they are occurring due to various factors, on various geomorphological conditions and spatial extent, as in different magnitudes (Guzzetti et al., 2006; Van Den Eeckhaut and Hervás, 2012). As a huge geohazard issue, scholars are putting enormous efforts in developing new and reliable methodologies for hazard mapping and mitigation approaches. As such, landslide inventories are considered as crucial for any further hazard assessment, followed by determining the susceptibility levels. Landslide susceptibility is heavily related to the knowledge of past events (Guzzetti et al., 2012), since it is determining the probability of a hazardous event, based only on the conditions and properties of the locale. In recent years, susceptibility modelling undergoes constant progress; researchers are implementing new modelling approaches that have not been used until now for hazard assessment. As a result, numerous methodologies with various input parameters are proposed to the public with outcomes that not always can be considered as reliable. A recent study (Reichenbach et al., 2018) shows that there are more than 15 modelling strategies that are used to determine landslide susceptibility levels. The majority of studies are relying on statistically based models such as logistic regression, neural networks, data overlay and index-based methods. On the other hand, less attention is paid on the procedures related to validating the outputs of susceptibility maps. In the literature, if any validation metrics are implemented, authors are mostly relying on Receiver Operating Characteristics (ROC) curve to evaluate model fitting and prediction performance, in most cases without going further in details.

This paper has the task to discuss and evaluate the implementation of the ROC curve as a validation metric for susceptibility maps and to compare it to Precision-Recall Curve (PRC), another metric that is considered more sensitive to imbalanced datasets and can represent a more accurate evaluation. For that purpose, a case study in Northern Italy was chosen - Val Tartano, a relatively small valley with a catchment area of 51 km^2. Despite its small area, the landslide phenomena are quite abundant: an inventory lastly updated in 2017, exhibits more than 750 records of mass movements, which makes the area very suitable for the current task. For creating the landslide susceptibility maps, three statistical models – statistical index-based (SI), logistic regression (LR) and random forest (RF), were used. The before mentioned landslide inventory was used as training and test datasets divided in ratios as such 50/50, 70/30 and 90/10. In total 12 thematic variables were used into the susceptibility modelling, each of them divided into relevant classes. As a result, 79 susceptibility maps were produced incorporating different model implementations as well as a combination of variables and training/test ratios.

For validating susceptibility maps, a highly adopted metric is the Receiver Operating Characteristics (Fawcett, 2006; Reichenbach et al., 2018), which relies on the *sensitivity* and *specificity* derived from a confusion matrix. Even though the models were not considered as randomly generated by the performed validation, some of the produced maps could not be accepted as plausible from a geomorphological point of view.

Another validation metric for classifiers that is not so popular among landslide hazard studies are the PRC plots. They have

* Corresponding author

similarities with the ROC plots and can be constructed in a similar manner. In fact, some studies (Saito and Rehmsmeier, 2015) are suggesting that PRC can be more sensitive and accurate for validation purposes (compared to ROC) in cases when natural imbalance can occur between classes.

The current work aims to compare the before mentioned evaluation metrics for the purposes of landslide susceptibility mapping to investigate the suitability and reliability of ROC and PRC when implemented for modelling such phenomena.

As it was mentioned before, overall 79 landslide susceptibility maps were produced and as for the current paper, only three maps produced from three different models will be presented and discussed. However, the same trends were depicted in the rest of the models.

The current work is structured as follows: in the following Section 2 the susceptibility mapping strategies will be discussed, including the case study, data used, modelling approaches and briefly the outcomes. In Section 3 we will deepen into details of the validation metrics and the results from their implementation will be discussed in Section 4, concluding on the findings in Section 5.

2. LANDSLIDE SUSCEPTIBILITY MAPPING

2.1 Case Study

Located in the Northern of Italy, Val Tartano (Figure 1) covers an area of 51km^2 and it is characterized by steep slopes and an elevation ranging from 250 to 2250 m a.s.l.

Geologically speaking, the valley represents a huge interest to scholars from different backgrounds (Colombera and Bersezio, 2011; Longoni et al., 2016); of particular interest is the presence of numerous faults accompanied by shear zones, prone to instabilities. Coupling the zones with the river network turns the area into suitable terrain for hosting landslides of different types. In fact, a local landslide inventory exhibits numerous entries of different types.

2.2 Data Used

During the landslide susceptibility modelling various combinations and approaches were applied to obtain the most suitable and acceptable map from both modelling and geomorphologic points of view. For this reason, three classification models, three sampling strategies and twelve terrain variables were implemented. The different combinations of the inputs highlighted the sensitivity of the model outcomes towards the inputs. In the following sections, the used susceptibility inputs will be briefly discussed, while the focus will be on the combinations implemented for the results discussed in Section 2.5.

2.2.1 Landslide Inventory: An exhaustive landslide inventory was obtained from Lombardy region. The database was created by the IFFI (Inventario dei Fenomeni Franosi in Italia) project (Scienze et al., 2007; Trigila and Iadanza, 2008), where mass movements were recorded at a scale of 1:10,000 and categorized according to the widely accepted landslide classification proposed by Varnes (1978) and then revised by Cruden and Varnes (1996).

The most distributed types in the current case study are: debris flow, rockfall, toppling and translational/rotational sliding. As mentioned before, the area of Val Tartano is abundant of slope failures, more than 750 landslides entries, with a landslide density of approximately 14.70 landslides/km^2. Mostly the area is affected by channelled debris flows (Colombera and Bersezio, 2011) that can reach 300 m in length. In addition, shallow landslides are also present in almost all parts of the valley.

One of the most well-known and studied landslide in the Val Tartano is 'Pruna' (Ballio et al., 2010; Longoni et al., 2016), a deep-seated gravitational slope deformations (DSGSD) with a surface area of around 1 km^2 and a depth that can reach 100m.

The landslide susceptibility analyses for the current study were focused only on debris flows and slidings, while rock falls and DSGSDs were omitted (Figure 2).

2.2.2 Terrain variables: For producing landslide susceptibility maps it is crucial to define a set of suitable, for each case study, factors that can be considered as predisposing or controlling (Trigila et al., 2015), which then will be used as input variables in the models. Twelve factors were chosen to be investigated and included as inputs. A digital terrain model (DTM) from 2015 with a spatial resolution of 5 meters is freely available via a dedicated geoportal of the Lombardy region (GeoPortale Lombardia, 2019). From the DTM five of terrain variables were derived, namely: aspect, elevation, slope angle, plan and profile curvature. Road and river networks were obtained from OpenStreetMap (2017). For deriving the normalized difference vegetation index (NDVI), three years of Sentinel 2 A/B data (Copernicus, 2019) was downloaded and processed. Rainfall map was obtained using the data of 37 stations around the area of Val Tartano for the same period and interpolated using two methods: Kriging and Inverse Distance Weighting. Land use, lithology and the location of geological faults were obtained from the local Italian catalogues (GeoPortale Lombardia, 2019).

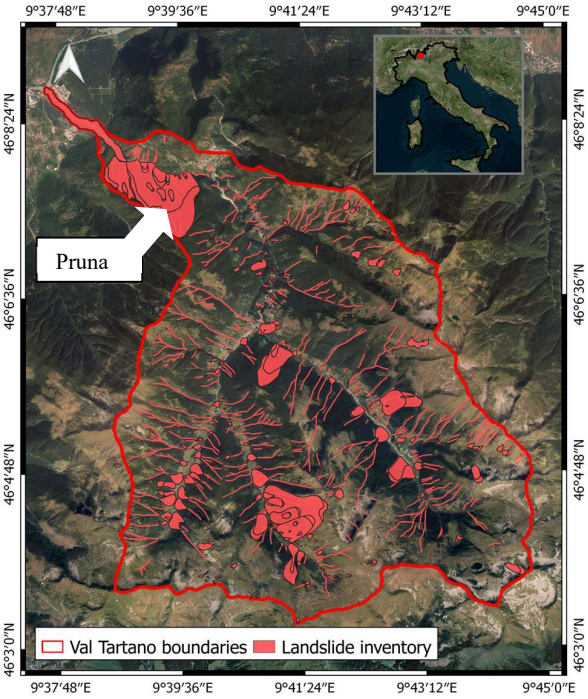

Figure 1. Val Tartano, Northern Italy

Upon geological expertise, for the area of Val Tartano, were included *no landslide* zones into the model training process. The zones represent areas that are unlikely to host a landslide due to some terrain conditions. The *no landslide* zonation included areas

where the slope angle is less than 20° and more than 70°. In addition, from lithological point of view bare intact rocks were also considered as stable enough to be included in the *no landslide* delimitation. However, in the current case study, they are overlapping with the highly inclined areas (>70°). To verify the reliability of such *no landslide* zonation, it was compared to the existing landslide inventory and less than 2% of error was measured, which can be explained with the prolonged dimensions of some mass movements and their material deposition especially in the flat areas. Since the discussed outputs in Section 2.5 included the *no landslide* zonation in the training process, the slope angle was omitted as an input variable for those models because its contribution was considered as already biased. Moreover, the experience of previous modelling combinations showed a low contribution of the precipitation to the output maps, due to low value variability. Therefore, precipitation was also excluded from the current modelling.

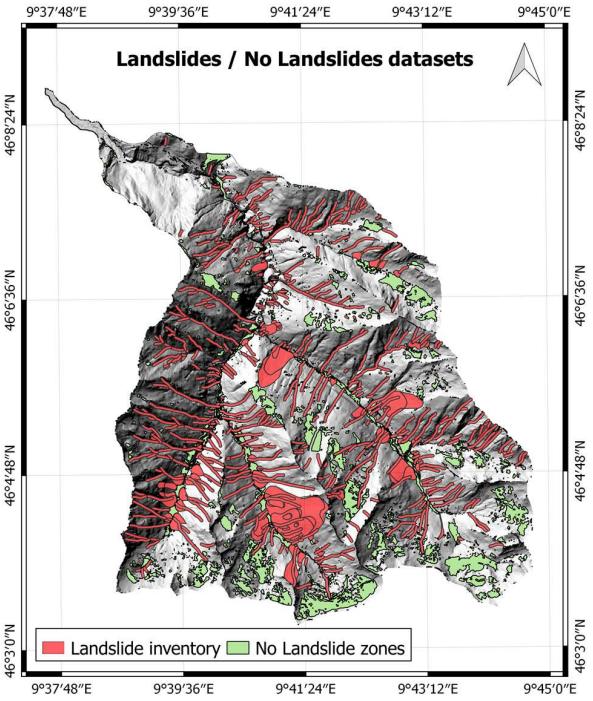

Figure 2. Input datasets for the training/testing the models

2.3 Susceptibility Models

The landslide susceptibility represents the probability for a region to be affected by landslides (Brabb, 1985; He and Beighley, 2008). For determining the probability of a landslide event, in the current study, three methods – statistical index, logistic regression and random forest- had been implemented.

2.3.1 Statistical Index (SI): The method exploits the relationship between the spatial distribution of landslides and the terrain conditions (Lee and Talib, 2005; He and Beighley, 2008; Chalkias et al., 2014; Aditian et al., 2018;). The statistical weight for each class is computed as a natural logarithm (Equation 1) of the ratio of the landslide density for the particular class over the landslide density for the whole area:

$$SI = \ln\left(\frac{N_i}{M_i} / \frac{N_T}{M_T}\right) \quad (1)$$

$$Susceptibility\ map = \sum FC_j\ x\ SI \quad (2)$$

where N_i = the landslide area of the i-th variable class
M_i = the total area of the i-th variable class
N_T = the total landslide area of the AOI
M_T = the total area of the AOI
FC_j = the i-th variable class
AOI = Area of Interest

The output weights are clearly highlighting positive and negative correlation of a terrain variable to landslides. Therefore, high positive values highlight high landslide density for the particular class and vice versa.

2.3.2 Logistic Regression (LR): The approach has been already discussed and implemented in the domain of susceptibility mapping (Bai et al., 2010; Mancini et al., 2010; Trigila et al., 2015). It is a great tool to create a regression model when dealing with dependent variables (in the current case presence or absence of a landslide event) and independent ones (terrain variables). The relationship between the variables is done through contribution coefficients (Equation 4) and the final output is an event probability between 0 and 1 (Equation 3).

$$P_R = \frac{e^Z}{1+e^Z} = \frac{1}{1+e^{-Z}};\ P_R = [0,1] \quad (3)$$

$$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n \quad (4)$$

2.3.3 Random Forest (RF): The machine learning algorithm is used for classification and regression, based on single decision trees that are working in a group (Breiman, 2001). The idea of using multiple decision trees is based on the fact that a single decision tree can produce high variance or bias, while in a group the decision on the final classification will be based on the most voted class among the forest. In order words, n-number of trees will balance the error and output higher precision classification through uncorrelated models (Breiman, 1996).

2.4 Train/test datasets and terrain variable sampling

Except for the modelling techniques, it is an important aspect to determine suitable sampling approach and training/test partitions. The landslide inventory has a crucial role (Guzzetti et al., 2012) since it provides knowledge of the phenomena to the model (training) and can be used for testing the performance of a classification model. Along with the combinations of the modelling approaches and input variables, three approaches for creating the training and test partitions – 50/50, 70/30 and 90/10 were tested.

For sampling the terrain variables another three approaches using 10,000, 100,000 and 200,000 training points to evaluate how their increase will affect the output susceptibility maps were applied.

The maps presented in this paper were created through the 70/30 ratio since it provides a sufficient amount of data for good training and testing purposes. The variables were sampled with 100,000 training points, due to the better performance compared to the 10,000 cases and the less computational demand compared to 200,000. Moreover, no significant improvement was noticed in the latter cases.

2.5 Susceptibility Outputs

In Figures 3, 4 and 5 the three outcomes using SI, LR and RF as classification models are reported. The input variables, training/test ratios and sampling approaches were kept the same for all of them, as described in the previous sections.

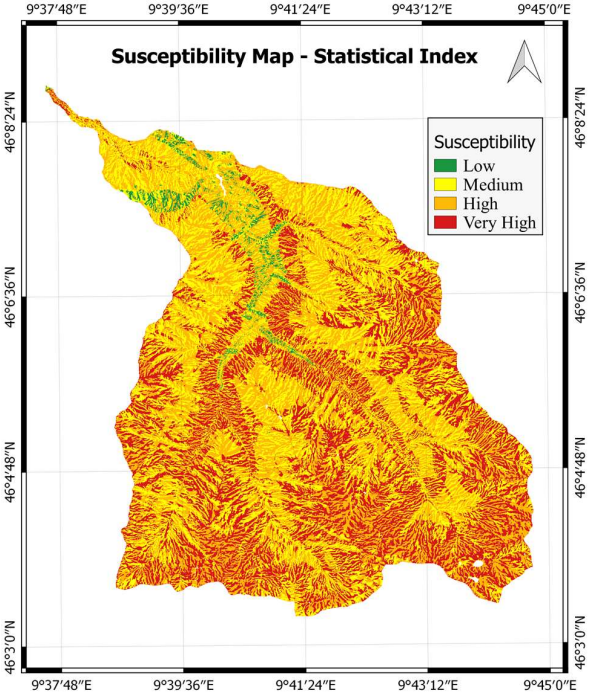

Figure 3. Susceptibility map produced with Statistical Index

As it can be seen from a visual inspection of the susceptibility maps, there are some fundamental differences among them. It is more obvious the huge contrast between the map produced via the statistical index and the other two (LR, RF). For the SI the class weights are directly related to the landslide density and the analysis of the indices depict a high significance for the *plan curvature* variable, which can explain the pattern visible on the map (Figure 3). From a geomorphological point of view, the current map cannot be considered as an acceptable and true representation of the reality, as it depicts more than 65% of the area as highly and very highly susceptible.

On the other hand, logistic regression and random forest maps exhibit more similarities. The *no landslide* zonation, especially in the valleys' bottoms, is easily depicted – more visible is in the *low* susceptibility levels in the RF case, while in LR they are classified as *medium*. Both maps have an 'agreement' on *low* susceptibility to what concerns the *no landslide* areas on high slope gradients. It is an interesting fact, that even though the models were not trained for DSGSD, LR and RF models produced maps that are recognizing the area of 'Pruna' as highly susceptible to landslides.

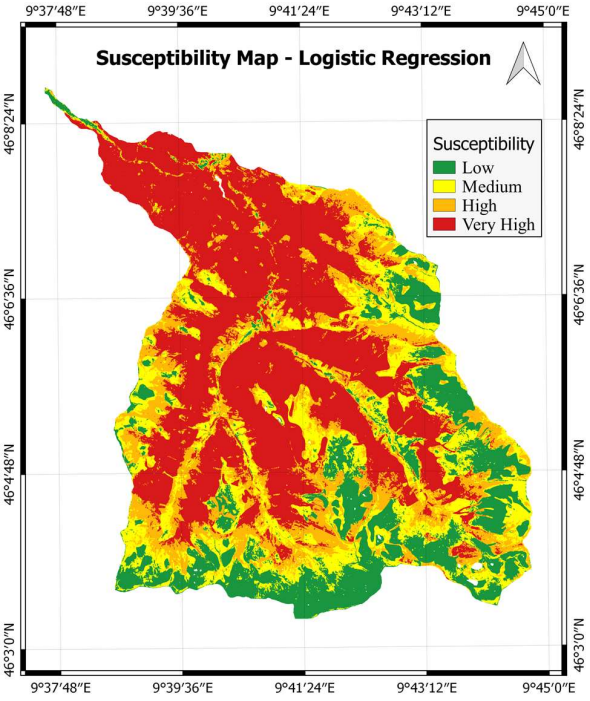

Figure 4. Susceptibility map produced with Logistic Regression

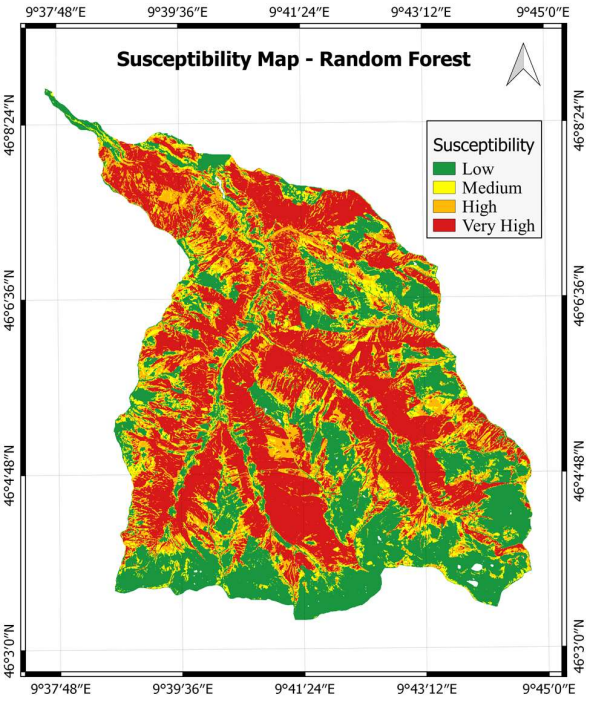

Figure 5. Susceptibility map produced with Random Forest

3. VALIDATION METRICS

In recent years, a great attention is paid on building and implementing more complicated models to obtain more accurate results (Xie et al., 2011; Reichenbach et al., 2018). However, less attention is paid on evaluating model performance, whether fitting or predictive. In a review carried out by Reichenbach et al.

(2018), high percentages (16.3% and 20.0%) of the analysed articles (565) did not implement fitting or performance evaluation metrics. In the meantime, they have noted an increase after the year of 2000 of the implemented metrics used for validation purposes of susceptibility maps. As most diffused among scholars, they have highlighted the use of success /prediction rates (Chung and Fabbri, 1999, 2003) and the ROC curves (Ayalew and Yamagishi, 2005). This could be explained with the fact that ROC curves are easy to build and to interpret (Fawcett, 2006).

On the other hand, Saito and Rehmsmeier (2015) discussed the use and suitability of Area Under The Curve ROC (AUCROC) and the (PRC) plots when validating imbalanced binary datasets. An imbalanced dataset is considered when there is a difference between the positive and negative cases in a binary classification (Saito and Rehmsmeier, 2015). The focus of their work is on the performance evaluation of a model after the classification is carried out. The reason for that is that, even though the testing set is created according to the class distribution as in the training set, a natural inequality in the classes can arise due to the phenomena under consideration. They suggest that class inequality (imbalance) is not often taken into account and evaluation outcomes are usually misinterpreted. The work of Saito and Rehmsmeier (2015) highlights the use of PRC plots as much more sensitive over ROC when datasets under consideration exhibit high inequality between the positive and negative classes.

3.1 Receiver Operating Characteristics (ROC) curves

The ROC plots have been used for a long time in various applications for visualising and determining suitable classifier based on their performance (Fawcett, 2006). The evaluation metric is widely used for classification problems due to its simplicity and straightforward interpretation.

When dealing with binary classification, the instances are distributed into two main classes – *positives* and *negatives*. The applied classifier is then putting the outcomes into the possible classes of *positives* and *negatives.* All the outcomes can be summed into a confusion matrix, which is visually representing the performance of the algorithm (Figure 6). The construction of a ROC plot is based on the information contained in a confusion matrix.

		Actual values: Positive	Actual values: Negative
Predicted values	**P**ositive	**T**rue **P**ositive	**F**alse **P**ositive
	Negative	**F**alse **N**egative	**T**rue **N**egative

Figure 6. Confusion matrix

To build the ROC plot two parameters are needed, which are easily computed from the confusion matrix: *sensitivity* (true positive rate) and *specificity* (true negative rate), described in Equations 5:

$$sensitivity = \frac{TP}{TP + FN},\ specificity = \frac{TN}{TN + FP} \tag{5}$$

Visually analysing a ROC plot is not sufficient enough. To measure a classifier performance is more meaningful to assign a qualitative value. As such, computation of the Area Under the Curve of ROC plot acts as a valuable interpretation of the ROC plot and model behaviour. An area larger than 0.5 is accepted as a classification produced on a model basis, while less than 0.5 is considered as a random classification. A perfect classification is considered when the area is equal to 1.

3.2 Precision Recall Curve (PRC)

Similarly to ROC curves, the PRC relies on the confusion matrix and the sensitivity (which is equal to the recall). On the other hand, the PRC is a quantifier of the positives classes. The precision is computed through Equation 6:

$$precision = \frac{TP}{TP + FP} \tag{6}$$

As it can be seen from the way it is constructed, it is clear that PRC is not using the *TrueNegative*, which in most of the classification models are the majority of the classification outputs. Therefore, even when introduced imbalanced datasets, the metric is not affected by them.

To use PRC as an evaluation, again the area under the curve can be computed and area of 1 is considered as a perfect classifier. In the case of AUCROC, the threshold for a classifier is 0.5, however, in the case of AUCPRC the threshold is computed based on the ratio between the predicted *positives* and *negatives* (Equation 7). Only in the cases where the input data is balanced, 0.5 is actually the threshold. Therefore, the interpretation of the AUCPRC is not as straightforward as in the case of ROC.

$$PRC_{THRESHOLD} = \frac{P}{P + N} \tag{7}$$

4. VALIDATION RESULTS

From the modelling point of view, the visual interpretation of susceptibility maps is not a sufficient approach to determine the performance of the classification models. The metrics discussed in Section 3 were applied for all the susceptibility maps (Figures 3-5). In addition, to AUCROC and AUCPRC, the *Cohen's kappa coefficient* and the *overall accuracy* were computed. A summary of the obtained results is represented in Figure 7, where the fitting and predictive performances are reported.

An initial comparison of the AUCROC and AUCPRC model fit evaluations can note that all of the metrics are in overall agreement among them. Moreover, they are highlighting the poorer performance of the model produced through SI (AUCROC=0.59; AUCPRC=0.52), while the visual uncertainty between the LR (AUCROC=0.67; AUCPRC=0.80) and RF (AUCROC=1.00; AUCPRC=1.00) is disproved, and RF is exhibiting excellent results. Those findings are confirmed also by the *kappa* and *overall accuracy*, where the *kappa's* extremums are at SI (-0.013) and RF (1.00).

As expected, the predictive performance metrics exhibit lower values than the model fit. However, the overall trend in the model performances is kept – the worst case being the SI (AUCROC=0.61; AUCPRC=0.61), and the best using RF (AUCROC=0.89; AUCPRC=0.89).

To verify the model performances, the ROC and PRC metric values should be compared to their actual threshold – whether the classification is done on a model basis or at random chance. For the case of AUCROC the threshold is 0.5, while in the case of AUCPRC it should be computed according to Equation 7 for each case. The PRC thresholds are reported in Figure 9 for the relevant cases.

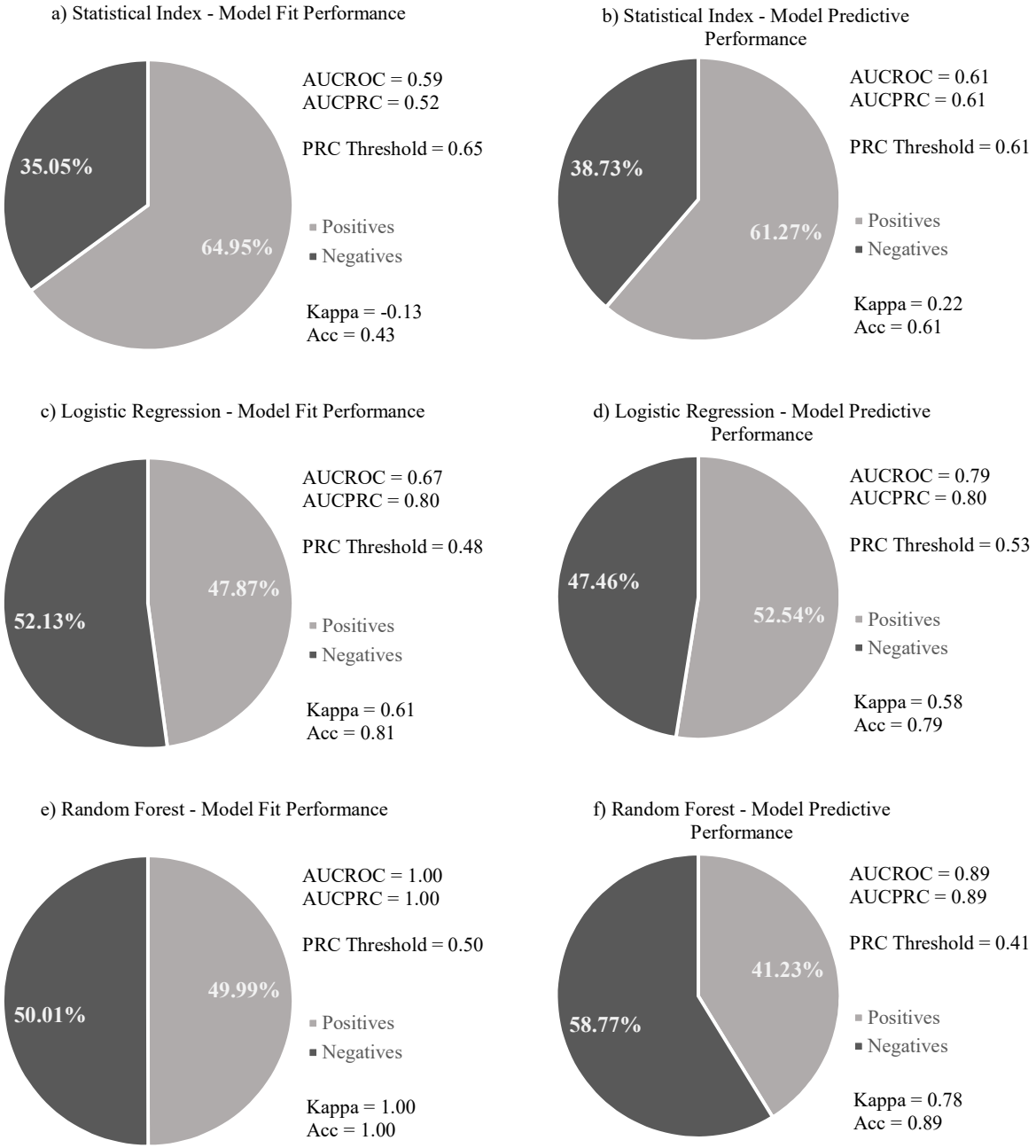

Figure 7. Summary of the model evaluations

All of the produced models exhibit AUCROC > 0.5, therefore they can be considered as successful outputs (disregarding the level of performance). Again, the worst performance corresponds to SI (AUCROC=0.59 Figure 7a) and on the opposite AUCROC=1.00 is yielded by RF.

On the other hand, the conclusions on the model performances differ when analysing the AUCPRC. The two classifications done through LR and RF are above their thresholds, which is in agreement with AUCROC. While in the case of SI, the model fit is not above the threshold of 0.65 and the predictive performance is exactly on it (0.61). Combining these results with the values of *kappa, overall accuracy* and the geomorphologic plausibility, the conclusion is that the classification done through SI cannot be accepted as a landslide susceptibility map.

The explanation for the different conclusions between the AUCROC and AUCPRC, can be found in the ratios between the *positives* and *negatives* values in the classifications, represented as pie charts. In the cases of LR and RF the classes are almost balanced, with the exception of the predictive performance of the RF (Figure 7f), where the higher portion is covered by the *negatives.* On the other hand, the pie charts for SI (Figures 7a, b) are highlighting imbalance but with a higher weight of the

positives. As mentioned before, the last two observations are important because the PRC is a quantifier of the positive class and therefore the *positives* will have more influence. On the other hand, the ROC is not influenced by such inequalities and it is considering the SI output as a successful classification.

The confusion matrices in Table 1 are providing clearer explanation of the imbalance in the *positives* and *negatives.*

		Model Fit		Model Predictive	
SI		Actual values		Actual values	
Predicted values		29134	35813	36257	25011
		20866	14187	13743	24989
LR		Actual values		Actual values	
Predicted values		39186	8681	40787	11752
		10814	41319	9213	38248
RF		Actual values		Actual values	
Predicted values		49981	6	40037	1190
		19	49994	9963	48810

Table 1. Confusion matrices for the three models

The imbalance of the *positives* in the SI model is due to the high count of *FalsePositives*, coupled with high *FalseNegatives* and then it is even clearer the incapability of the current model to classify correctly the area under study.

The imbalance in the model prediction of the RF can be easily explained also through the related confusion matrix and the *TrueNegatives*. The landslides are natural phenomena that even though they are widely spread and can occur in different locations, under different conditions, they are still affecting minor percentage of Earth's surface; the areas that are not affected by landslide or that will not be, are much larger. Therefore, such an imbalance where the instances are classified as *TrueNegatives*, is more natural to be expected and correct.

As mentioned in Section 1, in total 79 susceptibility maps were produced, using different modelling approaches and input combinations. While in the current paper just a fraction of them were discussed, an overall average modelling behaviour for the rest of the models is summarized in Figure 8. It should be noted that the insensibility

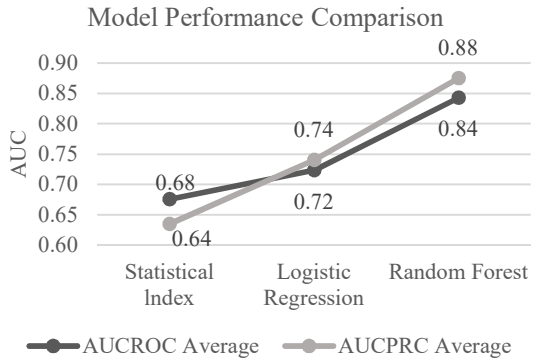

Figure 8. Average Model Fit Performance

of AUCROC towards the imbalanced datasets was more evident in the SI produced models, rather than LR and RF. The effect is notable both in the detailed validations of the current work (Figure 7a, b) and in the overall model performance comparison in Figure 8.

It should be noted that even though the SI example in this study did not exhibit positive outcomes (ref. Figure 7a, b), other maps produced with SI under different conditions and included in the overall model comparison (Figure 8), yielded more positive results. Consequently, SI should not be definitely excluded as a suitable model and can be used for susceptibility mapping, even if taking into account its lower performance compared to LR and RF.

5. CONCLUSIONS

In the current work three modelling approaches for mapping landslide susceptibility were presented. They were assessed by means of two different evaluation metrics: ROC and PRC. The ROC is widely implemented in hazard mapping, in most of the cases without further in-depth analysis of the evaluation values. It is an accepted metric and it is used mainly due its simplicity in interpreting its output. However, it was criticized as an approach due to it insensitiveness (Fawcett, 2006; Saito and Rehmsmeier, 2015) to imbalance between positive and negative instances of datasets. This inequality often arises due to the natural setting of phenomena under consideration. Therefore, the misinterpreting of the outcomes, which sometimes noticeably do not represent the truth, can lead to a false sense of secure and success, which is in the domain of risk mitigation can be considered as a huge error.

In the meantime, the work highlighted the use of PRC plots that can be considered as even more suitable for susceptibility mapping, by themselves or in combination with other indices. In the paper, the misinformation that the use of the ROC could provide in landslide susceptibility mapping was highlighted and the PRC, which yields more accurate results disregarding whether the dataset is perfectly balanced or imbalanced, was proposed as an alternative.

Far from saying that the PRC metric can be seen as the only and most accurate metric, the suggestion provided as a conclusion of our work is rather to use a set of tools and analyses to correctly evaluate susceptibility maps. In the presented study additional metrics and geomorphologic analyses of the produced susceptibility maps were included and it was highlighted the advantage of PRC over ROC for evaluating landslide hazard maps, even though its interpretation is not as straightforward as for the ROC.

ACKNOWLEDGEMENTS

This research was funded by Fondazione CARIPLO, grant number MHYCONOS.

REFERENCES

Aditian, A., Kubota, T., Shinohara, Y., 2018. Comparison of GIS-based landslide susceptibility models using frequency ratio, logistic regression, and artificial neural network in a tertiary region of Ambon, Indonesia. Geomorphology. https://doi.org/10.1016/j.geomorph.2018.06.006

Ayalew, L., Yamagishi, H., 2005. The application of GIS-based logistic regression for landslide susceptibility mapping in the Kakuda-Yahiko Mountains, Central Japan. Geomorphology. https://doi.org/10.1016/j.geomorph.2004.06.010

Bai, S.B., Wang, J., Lü, G.N., Zhou, P.G., Hou, S.S., Xu, S.N., 2010. GIS-based logistic regression for landslide susceptibility mapping of the Zhongxian segment in the Three Gorges area, China. Geomorphology. https://doi.org/10.1016/j.geomorph.2009.09.025

Ballio, F., Brambilla, D., Giorgetti, E., Longoni, L., Papini, M., Radice, A., 2010. Evaluation of sediment yield from valley slopes: A case study, in: WIT Transactions on Engineering Sciences. https://doi.org/10.2495/DEB100131

Brabb, E.E., 1985. Innovative approaches to landslide hazard and risk mapping, in: International Landslide Symposium Proceedings, Toronto, Canada. pp. 17–22.

Breiman, L., 2001. Random forests. Mach. Learn. https://doi.org/10.1023/A:1010933404324

Breiman, L., 1996. Out-of-bag estimation.

Chalkias, C., Ferentinou, M., Polykretis, C., 2014. GIS-based landslide susceptibility mapping on the Peloponnese Peninsula, Greece. Geosci. https://doi.org/10.3390/geosciences4030176

Chung, C.J.F., Fabbri, A.G., 2003. Validation of spatial prediction models for landslide hazard mapping. Nat. Hazards. https://doi.org/10.1023/B:NHAZ.0000007172.62651.2b

Chung, C.J.F., Fabbri, A.G., 1999. Probabilistic prediction models for landslide hazard mapping. Photogramm. Eng. Remote Sensing.

Colombera, L., Bersezio, R., 2011. Impact of the magnitude and frequency of debris-flow events on the evolution of an alpine alluvial fan during the last two centuries: Responses to natural and anthropogenic controls. Earth Surf. Process. Landforms. https://doi.org/10.1002/esp.2178

Copernicus, 2019. Sentinel data 2019.

Cruden, D.M., Varnes, D.J., 1996. Landslide types and processes. Spec. Rep. - Natl. Res. Counc. Transp. Res. Board.

Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognit. Lett. https://doi.org/10.1016/j.patrec.2005.10.010

GeoPortale Lombardia, 2019. URL http://www.geoportale.regione.lombardia.it/

Guzzetti, F., Mondini, A.C., Cardinali, M., Fiorucci, F., Santangelo, M., Chang, K.T., 2012. Landslide inventory maps: New tools for an old problem. Earth-Science Rev. https://doi.org/10.1016/j.earscirev.2012.02.001

Guzzetti, F., Reichenbach, P., Ardizzone, F., Cardinali, M., Galli, M., 2006. Estimating the quality of landslide susceptibility models. Geomorphology. https://doi.org/10.1016/j.geomorph.2006.04.007

He, Y., Beighley, R.E., 2008. GIS-based regional landslide susceptibility mapping: A case study in southern California. Earth Surf. Process. Landforms. https://doi.org/10.1002/esp.1562

Lee, S., Talib, J.A., 2005. Probabilistic landslide susceptibility and factor effect analysis. Environ. Geol. 47, 982–990.

Longoni, L., Papini, M., Brambilla, D., Barazzetti, L., Roncoroni, F., Scaioni, M., Ivanov, V.I., 2016. Monitoring riverbank erosion in mountain catchments using terrestrial laser scanning. Remote Sens. https://doi.org/10.3390/rs8030241

Mancini, F., Ceppi, C., Ritrovato, G., 2010. GIS and statistical analysis for landslide susceptibility mapping in the Daunia area, Italy. Nat. Hazards Earth Syst. Sci. https://doi.org/10.5194/nhess-10-1851-2010

OpenStreetMap contributors, 2017. Planet dump retrieved from https://planet.osm.org .

Reichenbach, P., Rossi, M., Malamud, B.D., Mihir, M., Guzzetti, F., 2018. A review of statistically-based landslide susceptibility models. Earth-Science Rev. https://doi.org/10.1016/j.earscirev.2018.03.001

Saito, T., Rehmsmeier, M., 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. https://doi.org/10.1371/journal.pone.0118432

Scienze, C.L.M., Prof, T.G., Baroni, C., 2007. Progetto IFFI : Inventario dei Fenomeni Franosi in Italia. URL http://www.progettoiffi.isprambiente.it

Trigila, A., Iadanza, C., 2008. Landslides in Italy.

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. https://doi.org/10.1016/j.geomorph.2015.06.001

Van Den Eeckhaut, M., Hervás, J., 2012. State of the art of national landslide databases in Europe and their potential for assessing landslide susceptibility, hazard and risk. Geomorphology. https://doi.org/10.1016/j.geomorph.2011.12.006

Varnes, D., 1978. SLOPE MOVEMENT TYPES AND PROCESSES. Transp. Res. Board Spec. Rep.

Xie, X., Ho, J.W.K., Murphy, C., Kaiser, G., Xu, B., Chen, T.Y., 2011. Testing and validating machine learning classifiers by metamorphic testing, in: Journal of Systems and Software. https://doi.org/10.1016/j.jss.2010.11.920