ORIGINAL PAPER

Application of various strategies and methodologies for landslide susceptibility maps on a basin scale: the case study of Val Tartano, Italy

Vasil Yordanov[1,2] · Maria Antonia Brovelli[1]

Abstract

Landslide susceptibility mapping is a crucial initial step in risk mitigation strategies. Landslide hazards are widely spread all over the world and, as such, mapping the relevant susceptibility levels is in constant research and development. As a result, numerous modelling techniques and approaches have been adopted by scholars, implementing these models at different scales and with different terrains, in search of the best-performing strategy. Nevertheless, a direct comparison is not possible unless the strategies are implemented under the same environmental conditions and scenarios. The aim of this work is to implement three statistical-based models (Statistical Index, Logistic Regression, and Random Forest) at the basin scale, using various scenarios for the input datasets (terrain variables), training samples and ratios, and validation metrics. A reassessment of the original input data was carried out to improve the model performance. In total, 79 maps were obtained using different combinations with some highly satisfactory outcomes and others that are barely acceptable. Random Forest achieved the highest scores in most of the cases, proving to be a reliable modelling approach. While Statistical Index passes the evaluation tests, most of the resulting maps were considered unreliable. This research highlighted the importance of a complete and up-to-date landslide inventory, the knowledge of local conditions, as well as the pre- and post-analysis evaluation of the input and output combinations.

Keywords Landslide · Susceptibility · Statistical index · Logistic regression · Random forest · Model comparison

Introduction

Landslides are considered to be one of the most diffused and devastating geohazards in the world (Guzzetti et al. 1999; Iadanza et al. 2009; Van Den Eeckhaut and Hervás 2012; Trigila et al. 2015; Reichenbach et al. 2018), as they are movements of rock mass, debris, or earth material (Varnes 1984) rushing downslope causing enormous damage (physical and economical). Landslides are therefore the subject of past and ongoing research. Due to the fact that they occur in different locations, with a variety of type movements and magnitude (Cruden and Varnes 1996), it is not a straightforward task to deploy suitable risk mitigation strategies without an in-depth investigation of the current location and phenomena. For this purpose, the first step of hazard analysis is considered to be the preparation of a detailed landslide inventory (Guzzetti et al. 2012). The knowledge of past events is crucial for understanding the landslide phenomena with their distinctive characteristics in their current location. This is followed by the creation of a susceptibility map, which depicts the spatial probability levels for a certain area to be prone to a mass movement, based on environmental conditions. The foundation of the susceptibility maps is that areas with certain conditions similar to the ones that have been displaced have a high probability to move as well.

Since landslide hazards are spread all over the world and many scholars are involved in the study of this phenomenon, a vast number of modelling approaches can be found in the literature, as well as implementation at various scales—starting from a basin (Catani et al. 2005; Guzzetti et al.

✉ Vasil Yordanov
vasil.yordanov@polimi.it

[1] Department of Civil and Environmental Engineering (DICA), Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milan, Italy

[2] Vasil Levski National Military University, Veliko Tarnovo, Bulgaria

2005; Remondo et al. 2005), regional (Chalkias et al. 2014; Díaz et al. 2019), national (Van Den Eeckhaut and Hervás 2012; Trigila et al. 2013), and even continental (Van Den Eeckhaut et al. 2012; Günther et al. 2013). Two distinct modelling approaches are used by scholars—qualitative like geomorphologic analysis (Van Westen et al. 2003; Castellanos Abella and Van Westen 2008) and quantitative methods, such as statistical bi- and multivariate (He and Beighley 2008; Bai et al. 2010; Pourghasemi et al. 2013; Aditian et al. 2018). Additionally, recent studies directly compare the implementation of different machine learning techniques and artificial intelligence for modelling landslide susceptibility (Arabameri et al. 2019; Park and Kim 2019).

In fact, a recent review (Reichenbach et al. 2018) had collected and analysed a significant number of research on the topic. Their work had pinpointed that there are around 19 classes of statistically based models used in susceptibility mapping. However, there are big discrepancies in the types of used environmental variables; moreover, a significant number of analyses were carried out without a proper output evaluation. In fact, scholars have been trying to standardize in some manner the landslide susceptibility processing (Guzzetti et al. 2006; Fell et al. 2008; van Westen et al. 2008). On the other hand, different types of landslides located in various conditions cannot be approached in the same manner; therefore, such standardizations are hard to apply. As conditions are different, a direct comparison between models and various approaches is hard to be made (Reichenbach et al. 2018), especially when they are carried out by different experts in different locations. Yet, authors agree that some principles should be determined in susceptibility modelling.

With regard to some of the inconsistencies discussed above and to the mentioned literature, the current work aims at directly comparing three of the most widely used statistical models, namely Statistical Index, Logistic Regression, and Random Forest for landslide susceptibility on a basin scale. In addition to the comparison of approaches, the sensitivity of each algorithm is analysed through various input approaches and modelling scenarios—starting from terrain sampling, variable combinations, training ratios from a landslide inventory, validation techniques, and implementation of expert evaluation on the input datasets. The motivation for this multi-aspect comparison was that apart from the main algorithms implemented for susceptibility mapping, there are further elements to consider that receive little attention in the academic literature, despite the fact that they can significantly influence the credibility of susceptibility output. From the large number of studies, it can be deducted that the main highlights when modelling landslide susceptibility are the model and the environmental variable. While actually in some cases, e.g. machine learning, the implemented models are highly sensitive to the inputs and small changes can affect the outputs. To understand the effect of input variations, three training terrain variable sample sets with different sample counts are tested to determine their effect on the output models. In the same relation, the landslide inventory is separated in different training/ test partitions. Even though the terrain variables are predefined according to the case study and the respective landslides predisposing factors, different combinations are implemented with a focus on the effects of annual precipitation on landslide susceptibility. Different precipitation inputs are applied—two interpolations of the ground meteo-stations data (Kriging and Inverse distance weighting) or the omission of the effect of rainfall. Through these comparisons, the intent was to highlight some practices that yield better results than others, while comparing not only the modelling methods, but also other important aspects during the analyses. The results from improved and reliable studies can provide useful insights to the decision makers and local authorities, which can be followed by suitable risk mitigation measures reducing the disaster effects on socio-economic development of the affected region. To our knowledge, the presented study cannot cover all the relevant aspects of the susceptibility analysis but aims to encourage discussion and similar work.

Case study and landslide phenomena

As a suitable case study, Val Tartano, Northern Italy, was chosen due to the abundant landslide events and high concentration of different landslide types. Val Tartano is located on the southern side of the Valtellina Valley in Lombardy region, Northern Italy (46.1075° N, 9.6791° E) and has a coverage of about 51 km^2. The altitude of the Tartano Valley ranges from 250 to 2250 m a.s.l. with an average of 1861 m a.s.l. The main river of the basin is called the Tartano River, which is a tributary of the Adda River. This valley was studied by several authors (Mandelli et al. 2009; Ballio et al. 2010; Brambilla et al. 2011; Colombera and Bersezio 2011; Longoni et al. 2016) due to the geological particularities of the entire catchment. Several flooding events and landslides have been documented in this valley. The most catastrophic one happened in 1987 and caused 20 fatalities. During this event, the strong rainfall combined with snowmelt triggered hundreds of debris flows and soil slips in the span of a few days.

The strong presence of faults in the area is associated with a high number of landslide phenomena. The entire catchment is characterized by the presence of several shear zones along these faults. These areas are considered weak zones, prone to instability. The faults have clearly influenced the geometry of the stream network and gave rise to a variety of landslide phenomena. The most well-known of those are the "Pruna" landslide (Fig. 1) which is a deep-seat gravitational slope deformation (DSGSD) with a surface area of around 1 km^2 and depth that reaches up to 100 m and the "Foppa dell'Orso"—a shallow mass movement located on the left bank at the end of
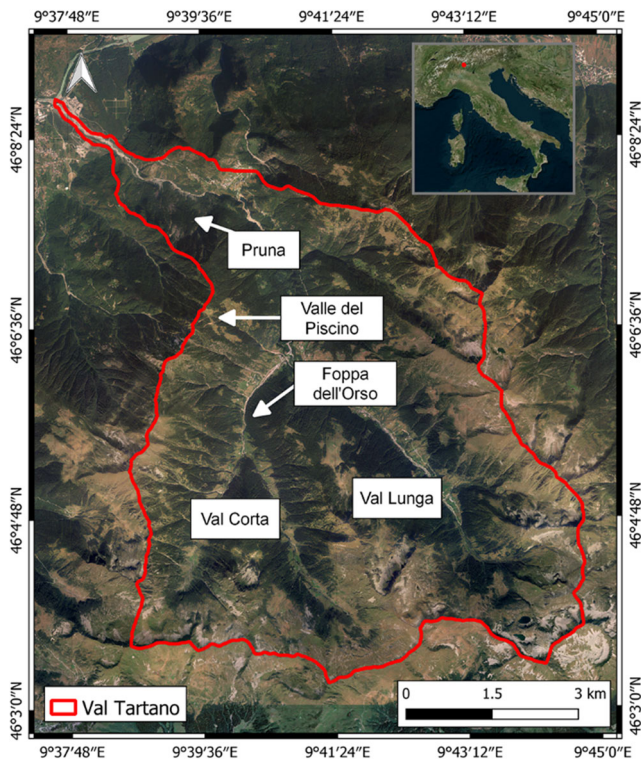
**Fig. 1** The boundaries of Val Tartano and some of the well-known landslides

Val Lunga. There are other sparse translational land instabilities that are present within the basin. Channelled debris flows appear in the entire catchment (linear landslides). The most distinguished one is the Piscino Valley (Fig. 1). The narrow trench lays on a fractured zone and starting from the top of the Piscino peak runs straight downstream featuring high slope gradient. The area is approximately 300 m long and 20 m wide and completely covered with talus and boulders originating from the rock walls surrounding the higher part of the valley (Longoni et al. 2016). The variety of landslide phenomena in this catchment renders it a rather interesting case for the application of landslide susceptibility models.

Data

Landslide inventory

In order to prepare a susceptibility map, it is important to have exhaustive data records of the location of past events that occurred in the same area, including their magnitude, movement type, and time of occurrence. Those historical records are managed in landslide inventory maps (Guzzetti et al. 2012), and through them, the geo-conditions related to terrain variables, such as slope, elevation, and lithology, can be studied (Guzzetti et al. 2006). In addition to the terrain conditions, the time of occurrence can add valuable information from a historical point

of view (Temme et al. 2020). An occurred landslide can interfere with the terrain condition and play the role of a predisposing factor for followed-up events (Samia et al. 2017). In fact, their study has reached to a conclusion that a landslide event is increasing the susceptibility level for around a period of 10 years. For the area of Val Tartano, a relatively recently updated (2017) landslide inventory at the scale of 1:10,000 is available thanks to the Italian Landslide Inventory (Inventario dei Fenomeni Franosi in Italia IFFI) project (Scienze et al. 2007; Trigila and Iadanza 2008), created for the Lombardy region. The recorded mass movements only in this case study are 759 (Fig. 2). According to the commonly adopted landslide classification (Cruden and Varnes 1996), the dominant movement types are debris flow, rockfall, rotational/translational slide, and complex. While for the current study, only the debris flows and slides were modelled. Geometrically speaking, the landslides were grouped as polygonal and linear. In the inventory database, the polygonal landslides (PLS) are represented accordingly to their affected area, while the linear landslides (LLS) are represented just as linear features due to the fact that their size in width is much smaller than their length. It was estimated that areas affected by PLS are covering 5.98 km^2 (11.75%) of the AOI area and LLS—2.80 km^2 (5.48%). Landslide distributions according to the elevation and slope angle can found in the Online Resource 1. On the other hand, multi-temporal analyses of the landslide inventory were not possible, due to the lack of such information for most of the landslide entries. For the current susceptibility mapping were
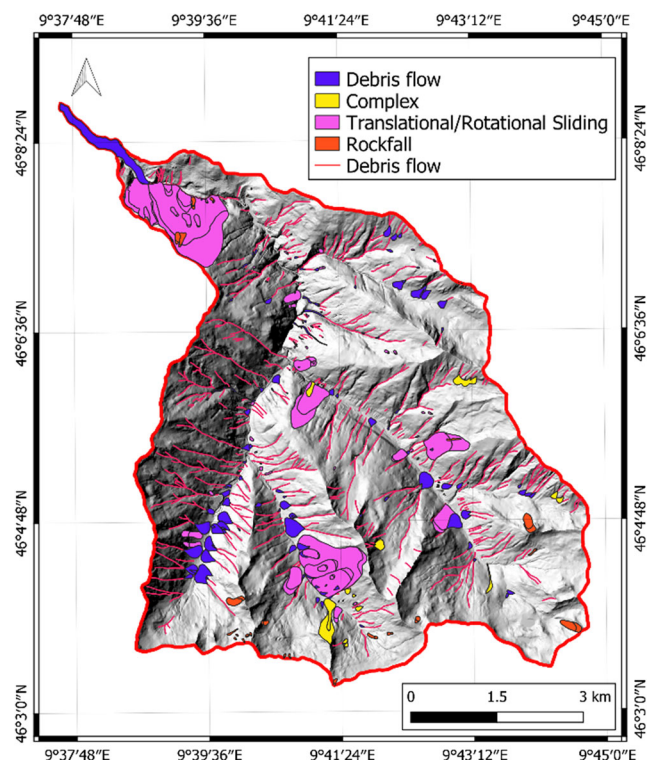
Fig. 2 IFFI landslide database for Val Tartano

omitted complex and rockfall of type, and initially the models were created for debris flow, slides, and complex. As it will be explained below after the first iterations, the landslide input was re-evaluated.

Valley delineation and determining the width of LLS

While the area of mass movements of the polygonal type is predetermined by the catalogue record and the terrain conditions underlying the PLS are possible to be explored and extracted for each separate areal event, it is not the same case for the linear landslides. Because of the scale of the map at which the inventory is produced (1:10,000) and the fact that some mass movements have a greater length than width, high number of cases (441) are represented as linear features in the inventory, which restricts the gathering of terrain information. Since all of the LLS actually are representing debris flow which are confined in narrow valleys and channels, it was decided to approximate the width of those features which will eventually allow us to obtain the width of the LLSs and to gather the related terrain information. To obtain their width, a classification of the terrain was performed and the morphological terrain type on which LLS occur was determined. Such as classification was obtained using the Geomorphon method (Jasiewicz and Stepinski 2013) implemented in GRASS-GIS (GRASS Development Team 2017). The method is based on geomorphologic phenotypes (geomorphon), where geomorphon is 'a simple ternary pattern that serves as an archetype of a particular terrain morphology' (Jasiewicz and Stepinski 2013). The method detects various landform elements using a DEM and classifies them into 10 groups such as flat, peak, ridges, valley, and slope. The catchment area of Val Tartano was classified using this methodology (Fig. 3) with the aim of delineating the landforms that can host LLS. Therefore, for the current approximation, the classes valley and hollow were used, which are predefined by their local ternary pattern based on raster elevation values (Liao 2010; Jasiewicz and Stepinski 2013). In the case of geomorphons, the neighbourhood is not fixed, rather it is adapting to the topography. Accordingly, to define a landform as a valley, it should has only two neighbours with the same elevation and the rest should be exhibiting higher elevation. Similarly, the hollow types are having neighbours with both high and low elevations. For more information regarding the geomorphons, the reader is referred to the work done by Jasiewicz and Stepinski (2013). After overlying the landslide inventory, it was estimated that LLSs can have a width depending on the channel they are present in the range from 5 to 15 m, i.e., 1 to 3 pixels. For the final approximation of the LLS widths, the values for the approximated valley widths were used and an average value of 10 m (2 pixels) was determined for all the inventory entries from LLS class.
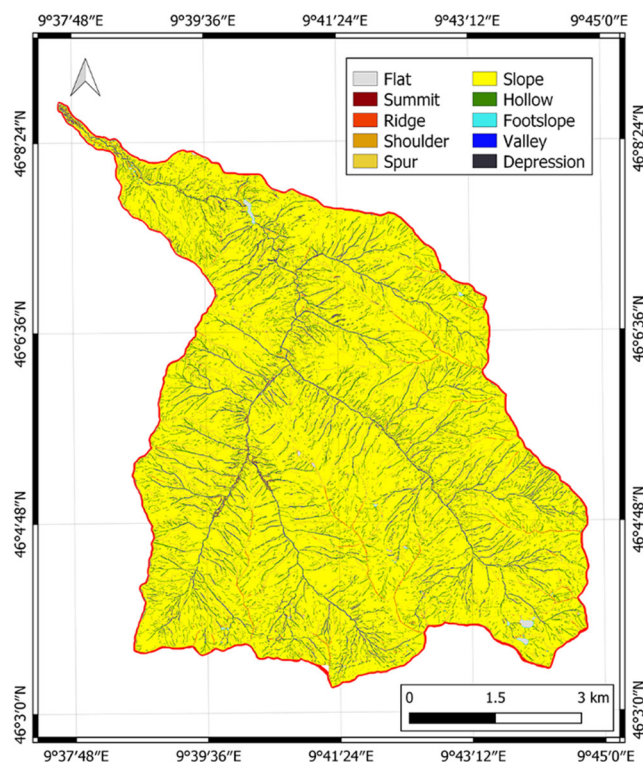

Fig. 3 Geomorphon classification

Terrain variables

A Digital Terrain Model (DTM) produced in 2015 with a spatial resolution of 5 × 5 m was downloaded from the GeoPortal (2019) of Lombardy region. Five morphological variables were used, namely elevation, slope, aspect, plan, and profile curvatures. The terrain variables are visualized in Online Resource 1.

The characteristics of the terrain variables are as follows:

- Elevation—the elevation of Val Tartano ranges between 250 and 2250 m a.s.l. The variable is subdivided into nine classes;
- Slope—the angle of slope is in a strong relationship with the occurrence of landslides. With the increase of the angle the shear stress is increasing and, depending on the material, the possibility of a movement is higher. From the histogram (Online Resource 1), it can be noted that the PLS and LLS occurrence concentration is between 30° and 40°, while for PLS, the distribution leans to the range of 20° and 30°, and the LLS is to the 40°–50° range. The considered slope classes are eight, starting from 0° to 80°.
- Aspect depicts the direction of the slope. In our case, it was reclassified in 8 classes. The orientation of the slope is related to the soil moisture content and vegetation density. For example, the north faces can host more snow and water quantity in the soil for longer periods.

- Plan and profile curvature—the curvatures highlight the profiles of the slope in both directions, where important classes are the concave which can block a water runoff and augment higher water saturation on the surface and the convex where the accumulated mass could lead to slope failure.
- Lithology—lithology is important since it describes the physical characteristics of the rocks and specifically the geo-mechanical properties. In Val Tartano, 8 classes are present, where the dominant ones are gneiss, mica schists, and the alluvial deposits.
- Land use and land cover map—a regional product (DUSAF 5.) is used. It is based on aerial photogrammetry and represents eight classes. The LULC is particularly important for the predisposal of a landslide event.
- Road and river network—road and river networks were obtained from OpenStreetMap (OpenStreetMap contributors 2017); for each network, buffer zones were produced at 50, 100, 200, 500, and more than 500 m. They have different importance related to landslide susceptibility. Rivers can be responsible for oversaturation and erosion of landslide toes, while the roads and particularly the passage of cars and trucks on them can generate loads that slopes cannot withstand.
- Normalized difference vegetation index (NDVI)—for deriving NDVI, multi-spectral optical images over the area of Val Tartano from ESA Copernicus Sentinel 2 missions (Copernicus 2019) were used. Sentinel 2 A and B are operational since 06/2015 and 03/2017, respectively. The twin satellites are producing multi-spectral data with 13 bands in the visible, near-infrared, and short-wave infrared part of the spectrum with a spatial resolution of 10, 20, and 60 m, while temporal resolution is 5 days with the same viewing angle and less than 5 days with a different angle. The images were selected with the same relative orbit and the goal was to have at least one image per month for the period 01/2016–06/2019. Since the area of interest is in a mountain region, the cloud cover can be a significant problem. The images were selected with cloud cover less than 40%, but this resulted in an insufficient number of images for the period. Therefore, the cloud cover was increased to 70% and images were manually chosen where the cloud cover was not affecting Val Tartano. The total number of used images is 64, where 29 were Level 1C and the rest 35 Level 2A. Level 1C datasets are only orthorectified, radiometrically and geometrically corrected, while Level 2A are also atmospherically corrected and ready to use. For that purpose, all of the granules at Level 1C had to be processed to L2A and the processing was done with the ESA SNAP toolbox (2019). The bands needed for the production of the NDVI are Red and Near Infrared. These bands have a spatial resolution of 10 m; therefore, after the production

of the average NDVI map, it was resampled to correspond to the primary resolution of 5 m/pix. The NDVI was computed using Eq. (1). The NDVI values are in the range [−1:1] (Pettorelli 2013), where negative values tending to −1 represent water; values in the range [−0.1:0.1] bare soils, rocks or snow; and from 0.2 (grass or shrubs) to 1 dense vegetation.

$$NDVI = \frac{(\mathbf{NIR} - \text{Red})}{(\mathbf{NIR} + \text{Red})}, \tag{1}$$

- Rainfall data—daily rainfall data were downloaded from ARPA Lombardia (ARPA Lombardia 2019) for the period 01/2016 to 12/2018 and from 37 meteo-stations around Val Tartano.

The final precipitation per station was computed by averaging the data of these 3 years. Then, the precipitation map of the whole area was computed using both Kriging and Inverse Distance Weighting (IDW) interpolation, the former method providing more accurate estimates (cross-validation was applied).

Methods

Mapping units

The importance of choosing an appropriate mapping unit for susceptibility map has been widely discussed in literature (Meijerink 1988; Guzzetti et al. 1999, 2006; Reichenbach et al. 2018). For the following analyses and processing, a grid-cell unit was chosen as a mapping unit. The grid-cell size was chosen accordingly to the DTM native resolution of 5 m × 5 m, since the DTM is the source of most of the morphological variables, which were expected to strongly correlate to landslide occurrence. Another aspect is that the adopted resolution is sufficient enough to be used for the local factors, even for the road and river networks. On the other hand, the landslide areal dimensions are big enough to be represented with the chosen grid-cell size. This resolution is suitable for both the PLSs and, after the approximation of the width, for the LLSs. As the terrain variables are at scale of 1:10,000, the vector features could be converted into raster formats with a resolution of 5 × 5 m.

Landslide inventory, training, and test sampling

An extensive landslide inventory is the main key for landslide susceptibility mapping (Guzzetti et al. 2012). Its particular

role is to provide knowledge for the terrain variables for past events. It can serve in two primary roles: (1) to be the basis on which classification models are trained on and (2) to be used as a test dataset to evaluate the performance of the produced models. For the purpose, the landslide inventory from 2017 was separated into training and test sets for the analyses. For the first iteration, a ratio of 70/30 training/test sampling was considered. Further trials using 50/50 and 90/10 were performed to assess the role of the training test sampling. As final combination input, 100% of the landslide records were also used to produce susceptibility models.

Terrain variable sampling

For the multivariate approaches, the terrain conditions had to be sampled for accurate training and testing. After dividing the inventory into the before mentioned partitions (section *4.2*), randomly generated points were created to sample the terrain variables according to their classes and the presence of mass movement or no. The bivariate Statistical Index (SI) analysis is based on the pixel count through the full extent of the basin under analysis; therefore, no further sampling was needed. Three approaches were undertaken with the purpose to improve the classification results. The initial analyses were performed using 20,000 sample training points (TP), equally divided into two presence classes. After obtaining satisfactory results, the analyses were done again using increased sample datasets consisting of 100,000 and 200,000 TP, again equally separated in both classes. For each set of training points, an exploratory analysis was carried out to determine their suitability to be used as training samples. It was determined that, even in the case of the smallest training datasets (20,000 TP), the *TP* was covering the same range of values with the same distributions and therefore they were good representatives of all terrain conditions.

Statistical Index (SI)

Many methods exist to produce susceptibility maps based on qualitative weight-values. The approach used as a first iteration is implemented in various studies (van Westen et al. 1997) and is based on a statistical index of the density classes for the terrain variable. The weight values of a factor class are computed as the natural logarithm of the landslide density in each class (N_j- pixel count with landslide for the j-th terrain variable class; M_j – pixel count of the j-th terrain variable class) over the landslide density of the entire area (N – the total pixel count with landslide; M – the total pixel count of the area of interest). The weights (SI_j) are derived on a pixel basis and determine the level of importance of each terrain variable in predisposing a possible mass movement. After obtaining the SI for each class, the final map is the sum of each terrain factor class (FC_j) multiplied by its weight

coefficient. It should be noted that for categorical variables, the method could be applied directly, while for the continuous factors, they had to be separated into relevant classes.

$$SIj = \ln\left[\left(\frac{Nj}{Mj}\right) / \left(\frac{N}{M}\right)\right], \tag{2}$$

$$\text{Susceptibility map} = \sum FC \times SI \tag{3}$$

Logistic regression (LR)

Logistic regression is a widely preferred method for determining landslide susceptibility levels (Bai et al. 2010; Mancini et al. 2010; Trigila et al. 2015). When dealing with a dependent variable and multiple independent ones, logistic regression is particularly useful to create a regression model between them, allowing the usage of continuous or categorical variable and even their combination. In the case of landslide susceptibility, the dependent variable is the presence or the absence of landslides, expressed as a probability between 0 and 1. The relation between the response variable and the explanatory ones (P_r) is defined as (4), where $\beta_1 \dots \beta_n$ are the coefficients of the independent variables (Hosmer and Lemeshow 2000).

$$P_r = \frac{e^z}{1 = e^z} = \frac{1}{1 + e^{-z}} \tag{4}$$

$$Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n; P_r = [0, 1] \tag{5}$$

Random Forest (RF)

Random Forest is a machine learning algorithm for classification and regression based on multiple decision trees working at the same time with a set of binary rules and giving as a result the predicted variable (Breiman 2001). A single tree is not sufficient for classification since it can have high variance or high bias (Taalab et al. 2018); an ensemble of n regression trees (Breiman et al. 2017) can balance the errors and produce an accurate classification.

For implementing the method, the ModelMap (Freeman et al. 2009) package, implemented in R (R Development Core Team 2011), was a used. It is based on the randomForest package (Liaw and Wiener 2002) and gives the possibility to create models, validate the results, and produce maps over geographic areas (Freeman et al. 2009). As computational parameters, the number of trees should be determined according to the estimated accuracy. The number of variables which the algorithm can iterate on a single node of the random forest tree was left by default because ModelMap has a built-in function to optimize it. Moreover, the algorithm works simultaneously with both classified and continuous variables. Therefore, the factor datasets were implemented as in

their normal form, on the contrary of beforementioned classification for continuous variable for the case of SI.

Validation

When constructing a classifying model, it is necessary to evaluate its performances—fitting and prediction. From a statistical point of view, it is fundamental to use and implement appropriate assessment metrics when dealing with classifiers and especially in the domain of risk mitigation. This is the reason why, besides computing the maps, we paid attention also on this quality aspect.

For the current work, three main groups of models, based on the classifier approach, were produced. For each one of them, the fitting and prediction performance is evaluated by means of standard metrics, such as Cohen's Kappa, overall accuracy, receiver operator characteristic curve (ROC), and precision recall plot (PRC). The last of them is not so standard in susceptibility modelling but have been discussed by authors (Saito and Rehmsmeier 2015; Yordanov and Brovelli 2020) as equally important and, in particular cases, even more reliable. In addition to the before mentioned metrics for logistic regression and random forest, two ad hoc internal evaluations were carried out. In the case of LR, the Akaike Information Criterion (AIC) (Akaike 1974), which measures the prediction quality of the models in a relative manner, was computed. Similarly, the out-of-bar error (OOB) (Breiman 1996) was estimated for the models created using Random Forest.

The area under the curve of the ROC (AUCROC) is a common method to evaluate the performance of the classifiers and it is highly adopted in the landslide susceptibility modelling. It is based on the true positive rate which is the 'sensitivity' of a model and on the false positive rate, '1- specificity'. The AUCROC ranges between 0 and 1, and the perfect performance corresponds to a value equal to 1. On the other hand, any model with a result lower than 0.5 is considered as a random classification; on the opposite results, higher than 0.6 can be considered as acceptable.

Similar to ROC, the Precision Recall curve is easily derivable from a confusion matrix and relies again on one side on the 'recall', which is the true positive rate, and on the other hand on the positive predictive value 'precision'. The AUCROC when computing the area under the PRC (AUCPRC) ranges from 0 to 1, where the latter is the perfect score. On the other hand, there is not a defined threshold as 0.5 to distinguish the classifier between random chance and model approach. In this case, the threshold is computed as the ratio between the positives *(P)* and the negatives *(N)*.

The overall schema of the performed processing for the current research is represented in Fig. 4, where two main processing flows can be distinguished: primary and refined. In the former, the iterations were carried out applying all of the before mentioned combinations related to the model training and

variable input. The latter was processed once conclusions were made based on the previously intermediate outputs and the refined inputs were deducted based on geological analyses of the previous outputs.

Moreover, due to the high number of combinations for the model creation, an abbreviation system is adopted to denominate the model combinations which will be commented below. The model and map combinations are stated as ML.TT.TP.PV(.TH), where ML will denote the modelling approach (possible values: SI, LR or RF); TT is the train/test ratio; TP is the training points used; PV is the precipitation variable. In the cases when it is referred to a map, not a model, TH is added, denoting the applied classification threshold. As an example and with the help of Table 1, RF.II.1.C.a means that the random forest is applied with a 70/30 training ratio, 20,000 training points; the used precipitation map is that obtained with IDW and the adopted classification threshold is 0.5.

Using the previously explained methods and approaches, 79 susceptibility maps were produced. Except for the application of the three different methods, the main differences were the ratios used for the training and validation sets. Regarding the input variables, additional three combinations were implemented:

PV1. the precipitation was omitted as a factor;
PV2. the precipitation was computed with the Kriging interpolation and
PV3. the precipitation was computed with IDW.

For the cases when SI has been implemented, a fixed value as a threshold (e.g. 0.5 or 0.6) was not possible to be determined since the obtained susceptibility maps had a range of values according to the statistical indices. For that reason, two approaches were used to define a threshold value: natural and quantile breaks were used for the classification of the susceptibility values and the use of four classification groups allowed a better discrimination of the susceptibility levels. On the other hand, when using LR and RF, the output is a raster with defined probability levels between [0:1]; therefore, a threshold level definition can be straightforward. The susceptibility classes were kept four (low, medium, high, and very high). The examined threshold values for high susceptibility level were 0.50 and 0.60, and what concerns the related very high critical levels—0.75 and 0.8. A summary of all model combinations and output maps is shown in Online Resource 2.

As it was discussed before, for computing the susceptibility map based on the SI method, the weight of each variable class was determined taking into account the presence of previous landslide events. By computing the class weights, the significance of a single class parameter over the whole model can be highlighted. Since the model relies on the class weights computed according to the landslide density, the resulting maps do not exhibit a probability value ranging from 0 to 1 rather it can be a range of values from different sizes, even starting from
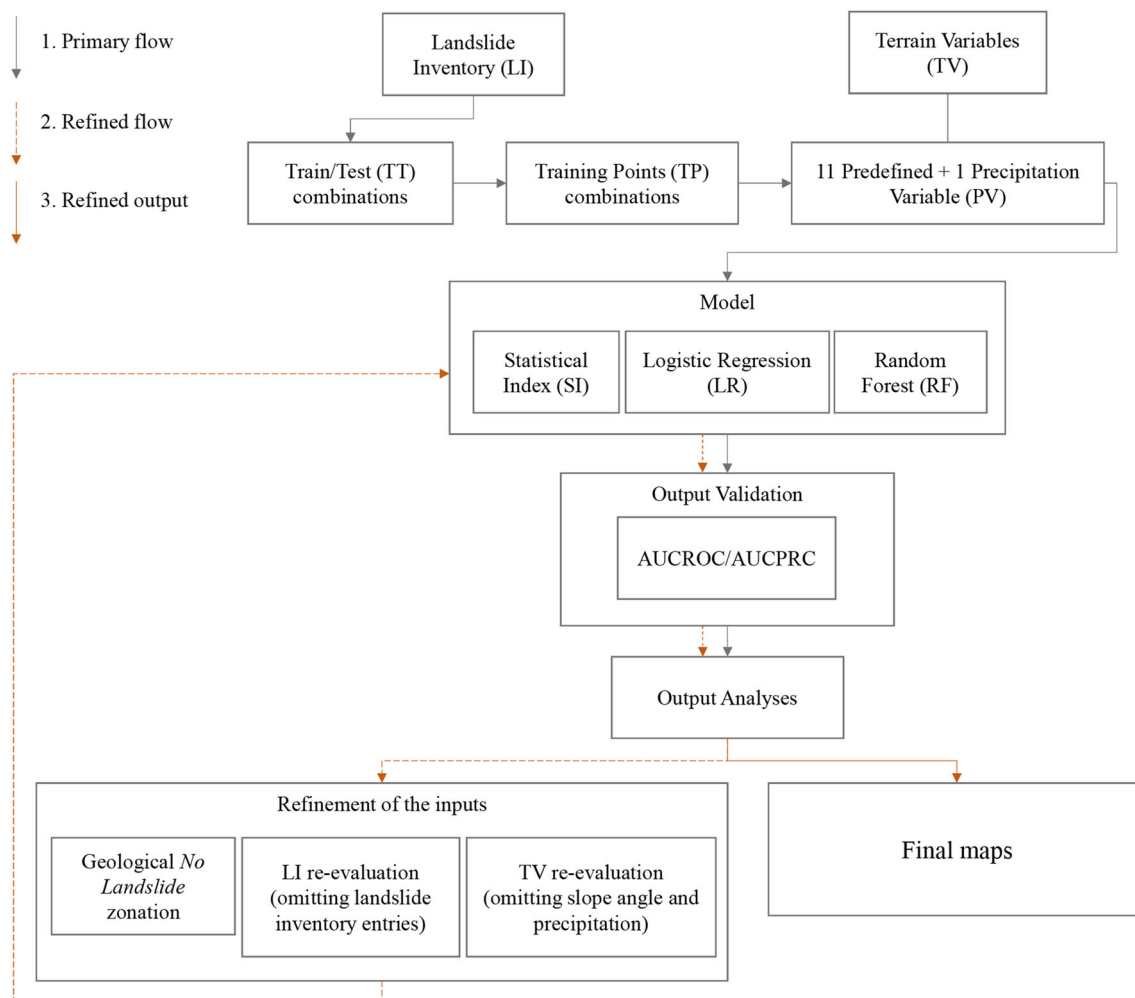
Fig. 4 Processing flow diagram

the negative domain. As mentioned before, two thresholding approaches brought to satisfactory results—natural and quantile breaks. Yet to be consistent in further comparisons with the *LR* and *RF* models, all SI maps were initially normalized in the range 0 to 1 and threshold baselines corresponding to 0.5 and 0.6 were tested. The model predictive performances of these maps (Fig. 6) are on the border to be considered as a random modelling. Therefore, no further processing and analyses were applied to these results, rather the focus was on the maps obtained with natural and quantile breaks.

Results from the Primary Flow

The overall improvement with the various modelling strategies and combinations is in favour of the Random Forest approach, where the least satisfactory is the results of the Statistical Index. The RF performance can overpower between 20 and 40% over the SI, and 10 to 30% over the LR, depending on the metric (Fig. 5).

The maps produced with the highest performing models (summary in Table 2) are represented in Fig. 6.

Table 1 Reference system for denoting the produced models/maps

Model (ML)	Train/test ratio (TT)	Training points (TP)	Precipitation variable (PV)	Threshold (TH)
Statistical Index (SI)	50/50 (I)	20,000 (1)	No precipitation (A)	Natural breaks[1] (a)
Logistic Regression (LR)	70/30 (II)	100,000 (2)	Kriging Interpolation (B)	Quantile[1] (b)
Random Forest (RF)	90/10 (III)	200,000 (3)	IDW Interpolation (C)	0.5^2 (a)
	100% (IV)			0.6^2 (b)

a Model Performance Comparison
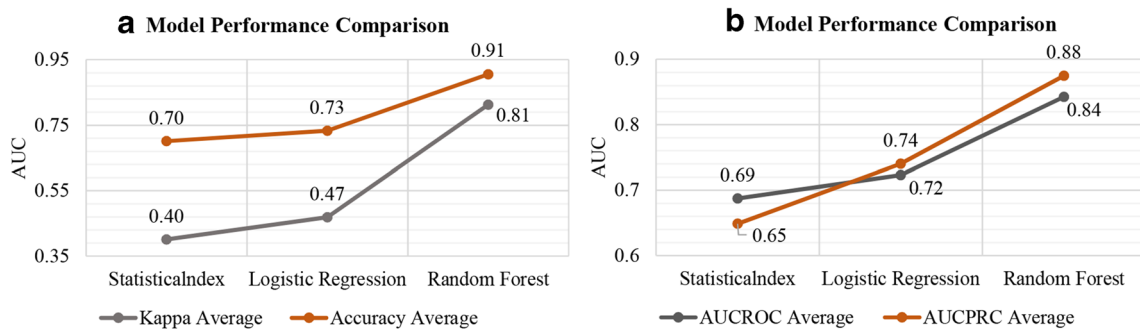
b Model Performance Comparison

Fig. 5 Overall model performance comparison

Results from Statistical Index

As it was discussed before, for computing the susceptibility map based on the SI method, the weight of each variable class was determined taking into account the presence of previous landslide events. By computing the class weights, the significance of a single class parameter over the whole model can be highlighted. Since the model relies on the class weights computed according to the landslide density, the resulting maps do not exhibit a probability value ranging from 0 to 1 rather it can be a range of values from different sizes, even starting from the negative domain. As mentioned before, two thresholding approaches led to satisfactory results—natural and quantile breaks. Yet to be consistent in further comparisons with the LR and RF models, all *SI* maps were initially normalized in the range 0 to 1 and threshold baselines corresponding to 0.5 and 0.6 were tested. The model predictive performances of these maps were on the border to be considered as a random modelling. Therefore, no further processing and analyses were applied to these results, rather the focus was on the maps obtained with natural and quantile breaks.

The analysis of the fitting performance (Fig. 7a) through Cohen's Kappa and Overall Accuracy (OA) depicts slight to moderate reliability between the evaluation metrics (between 0.17 and 0.57), even though the OA percentage is higher than 60%. It was observed that the threshold and the presence of precipitation in the terrain variables do not make a significant change in the results. On the opposite, the adoption of different training/test ratios does affect both the Kappa coefficient and the OA. In particular, the use of 70/30 *TT* ratio stands out with a

better performance; these considerations are also confirmed by the scores assigned from AUCROC and AUCPRC (Fig. 7b). The produced models are passing the test procedures and are not considered as randomly classified. Nevertheless, the results are not enough satisfactory considering that, on average, the AUCROC values are 0.69 and the AUCPRC ones 0.65. Both sets of results show similar trends and it can be noted that the optimal results are obtained in the case of classification in 4 classes through natural rather than quantile breaks.

The prediction map with the highest performance can be considered as the maps with AUCROC = 0.78 and AUCPRC = 0.75. In fact, even if the *SI.II.FE.B.a* has similar results (AUCROC = 0.77, AUCPRC = 0.74), yet the map exhibits the effect of high absolute weight from a single terrain variable (precipitation) and its classes significance. Clear differentiation between the susceptibility levels is notable; this is due to the high impact of the weights derived from the precipitation classes Fig. 8.

An example of the weights is reported in Table 3. Analysing the values, some considerations can be highlighted.

(1) The variable classes with the highest positive impact are as follows:

a NDVI with values ranging from − 0.036 to 0, which represent barren areas with a weight of 2.346;

b the land use class of abandoned areas with 2.213 and elevation class between 250 and 500 m a.s.l with 2.167.

2 The variable class with the highest negative impact is the precipitation between 1420 and 1570 mm/year with a weight of − 4.168.

These values are in correspondence with the landslide inventory; the high impact weights are due to concentration of the factor in relatively small areas and in the meantime high density pixels with of landslides. For example, in the case of NDVI − 0.036 to 0, the total area in terms of pixels is 457 while the pixels affected by previous mass movements are 422. The explanation for the high concentration bare areas affected by landslides could be related to the linear landslides which mostly contain loose material and no vegetation.

Table 2 Summary of the map evaluation metrics

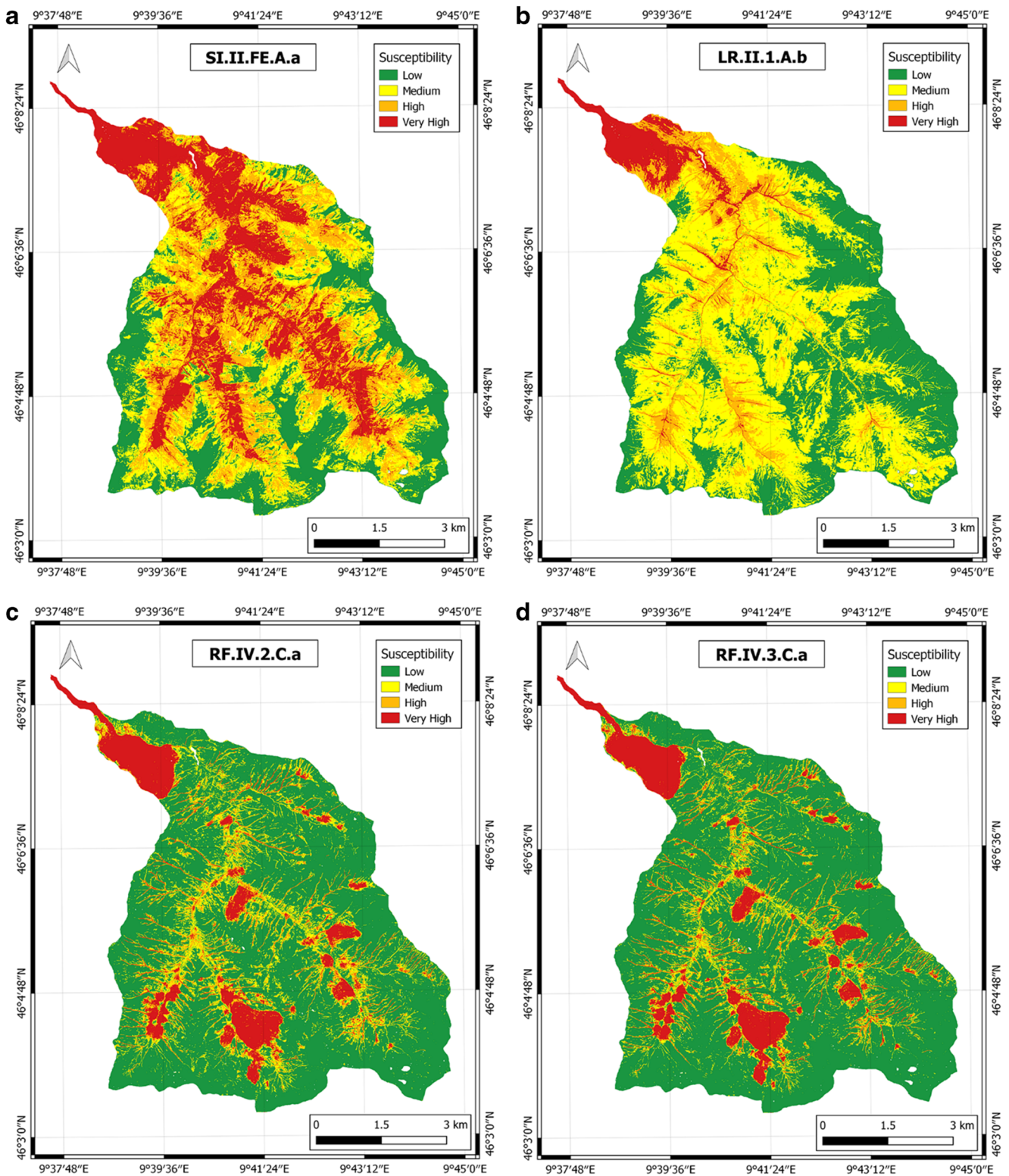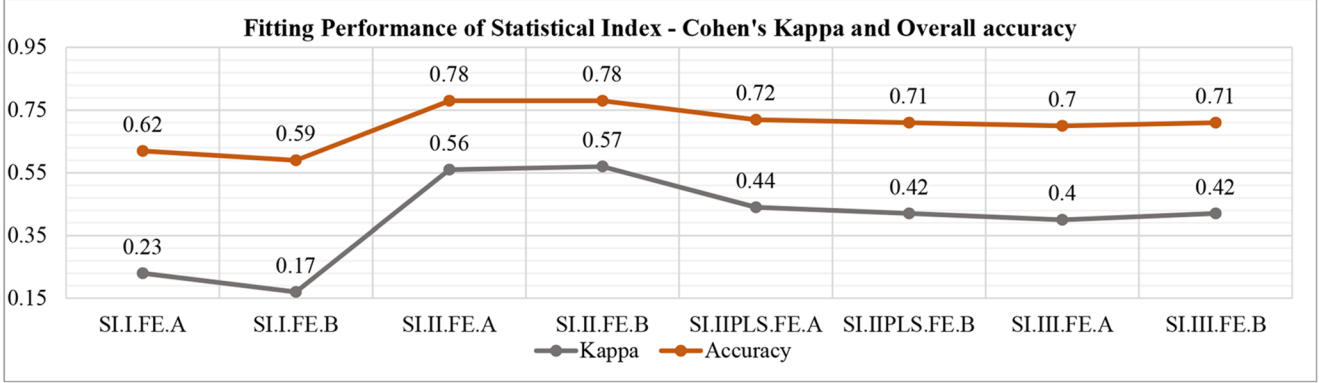| Model	Fitting performance			
	Kappa	OA	AUCROC	AUCPRC
SI.II.FE.A.a	0.39	0.69	0.65	0.61
LR.II.1.A.b	0.44	0.72	0.69	0.79
RF.IV.2.C.a	0.85	0.93	0.83	0.77
RF.IV.3.C.a	0.88	0.94	0.83	0.77

Fig. 6 The highest performing maps from the applied approaches

The map with the effect of precipitation is biased by a large number of landslide points in the south of the basin. Since there we have a particular range of rainfall values, the model assigns a high negative weight to the rainfall, which would bring to the incorrect result that more rainfall corresponds to fewer landslides. The opposite effect can be noted in the upper part of the basin where high precipitation is associated with low susceptibility. Therefore, the model appears to be wrong

a

b
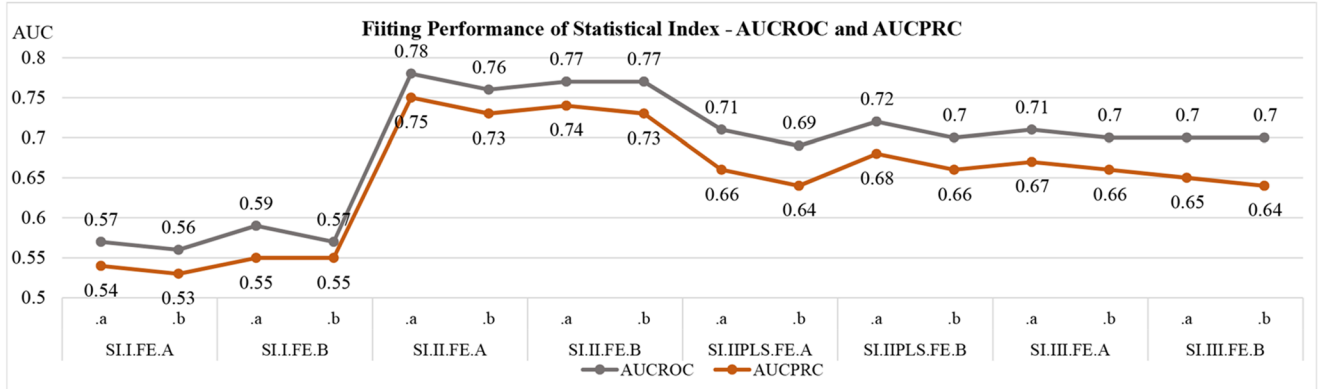
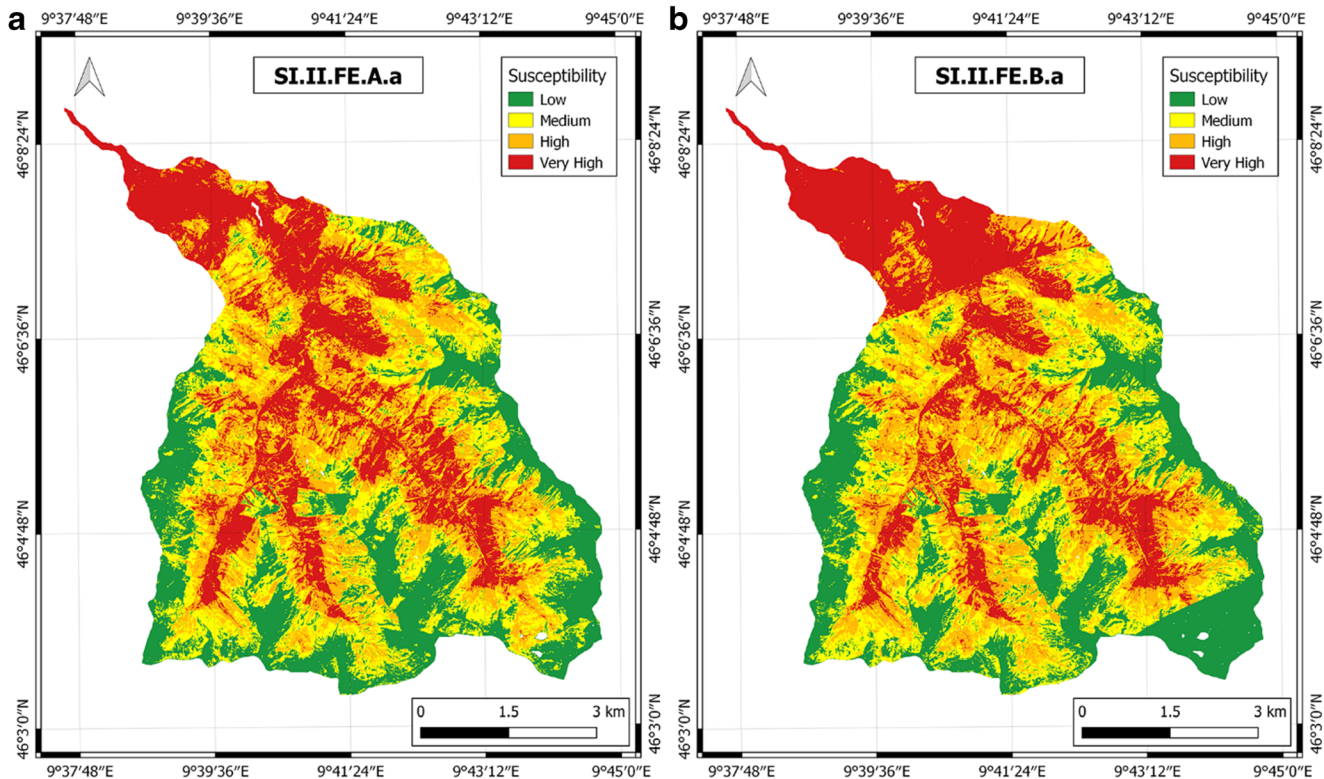**Fig. 7** Fitting performance of SI—Cohen's Kappa and overall accuracy, AUCROC and AUCPRC

Fig. 8 Two susceptibility maps produced through SI: **a** SI.II.FE.A.a—without precipitation and **b** SI.II.FE.B.a—with precipitation included as a variable

Table 3 Variable-assigned weights

Variable	NDVI	DUSAF	Elevation	Elevation	Elevation	Elevation	Profile Curvature	Aspect	River Distance	Faults Distance
Class	− 0.036-0	Abandonded areas	250–500	750–1000	500–750	1000–1250	− 0.4-− 0.2	315–360	50 m	50 m
Weight	2.35	2.21	2.17	1.46	1.29	0.57	0.54	0.47	0.43	0.43

and rainfall cannot be a reliable predictor (at least not at this (coarse) precipitation resolution). The presence of concentrated landslide points in the bottom part of the basin conditions the model performance. For this reason, the precipitation, as a factor, was removed in further simulations.

The map related to the results is reported in Table 3 and can be found in Fig. 8a.

Results from Logistic Regression

For the cases of LR, the training/test ratios were kept the same as in the previous SI case. Only a threshold value of 0.5 was applied, due to its better performance in the previous models. The number of maps obtained through the LR models compared to the 16 produced using the SI is a bit higher due to the fact that the inverse distance weighting (IDW) interpolation was introduced to determine what is the influence of this interpolation method of the rainfall data in the susceptibility modelling and the fact that for the last set of maps (3), an increased number of sampled training points (100,000 in total), equally distributed between landslide/no landslide, was introduced. The latter step was made with the purpose of assessing the sensitivity of the model using larger training samples and can be therefore considered as a further improvement of the previous results.

An internal model validation was performed with the AIC (Akaike 1974) (Fig. 9). Since the AIC is related to the sample and response data, it is not possible to compare the model using different TP sets; therefore, in the following graphs, the results are reported separately for models with 20,000 and 100,000 TP. Two conclusions can be made from AIC analyses—(1) the use of different TT rations can improve

the model (TT = 90/10) and (2) the implementation of IDW precipitation yields better results than the Kriging or no precipitation at all.

Similar to the SI approach contribution coefficients (β_1 … β_n) for the dependent variables, in the case of LR, the coefficients are related to the terrain variables, although they were not subdivided into classes. As it can be noted from Table 4 where the regression coefficients and intercept for models produced using the three combinations for the precipitation variable are reported, when present, the precipitation has a relatively high impact on the result.

The Kappa coefficient and the OA (Fig. 10a) were again computed for measuring the fitting performance of the various models. As in the case of SI, the effect of the TT ratio is highlighted and when TT = 70/30 the model exhibits higher results with Kappa = 0.78 and OA = 0.55. The model performance, evaluated through the AUCROC and AUCPRC (Fig. 10b), proved that the results are significantly better than those of SI. Two models obtained the highest results with AUCROC = 0.75 and AUCPRC = 0.80 for *LR.III.1.B.a* and *LR.III.1.C.a.* These results highlight the significance of the included precipitation interpolations. In the meantime, Fig. 11b highlights the model performance difference when applying two different thresholds.

Results from Random Forest

The same training combination approaches were used in case of Random Forest. In addition, the number of trees was assigned as 500, as being sufficiently low and stable the estimated error. As mentioned before, ModelMap has the ability to test and determine the number of variables to iterate on a single node. For the presented models, this value was
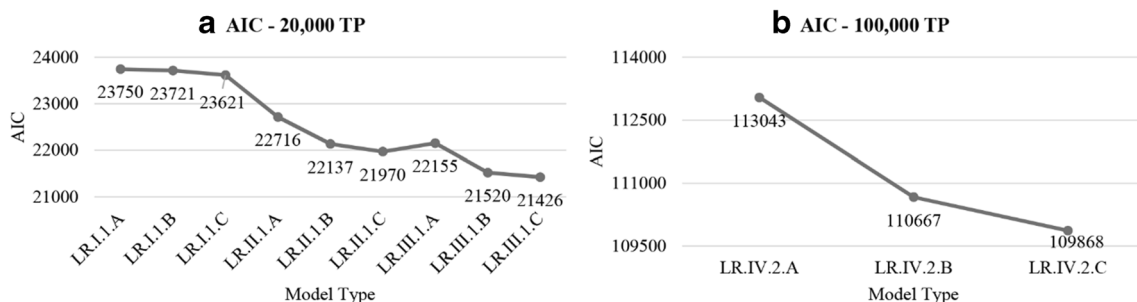
Fig. 9 AIC progress in the modelling approaches using **a** TP = 20,000 and **b** 100,000 points

**Table 4** Examples of regression coefficients

Model	Intercept	Aspect	Elevation	NDVI	Plan Curv.	Profile Curv.	Slope Angle
LR.II.1.A	5.4210	0.0019	− 0.0039	− 2.8690	− 1.2500	− 2.7160	0.0058
LR.II.1.B	3.3834	0.0020	− 0.0050	− 2.1900	− 1.3262	− 2.5976	0.0090
LR.II.1.C	3.2861	0.0021	− 0.0052	− 2.2978	− 1.3069	− 2.5369	0.0101
Model	LULC	Distance to faults	Lithology	Precipitation	Distance to rivers	Distance to roads	
LR.II.1.A	0.1249	− 0.2213	0.0229	-	0.0386	0.4004	
LR.II.1.B	0.1392	− 0.2048	− 0.0270	0.8135	0.1822	0.4516	
LR.II.1.C	0.1628	− 0.2040	− 0.0347	0.8663	0.1939	0.4694	

estimated to six variables and in just three of the cases to three variables to iterate on. A widely used metric for internal validation of the RF models is the OOB error estimate. The concept is that, for training a single tree, not all of the training datasets are used but only around the 60% of the set. Then, the tree can be tested on the sample sets not used for the training. In Online resources 3, the estimated OOB error for all of 13 models with error ranging between 5.88 and 11.74% and the relative influence of the variables for the models RF.IV.2.C.a and RF.IV.3.C.a are reported. In addition, it is quite clear the contribution of the inclusion of the precipitation as a variable. In fact, when not included, a peak in the error plot is visible with a higher introduced error. Significant improvement is noticeable when a higher number of training points is introduced (cases IV and V).

In Breiman (1996, 2001), it was concluded that OOB is enough accurate and there is no need for implementing a test dataset. Nevertheless, the results obtained from RF models were tested against the dedicated samples, to be consistent in comparing with the previous models and in evaluating their prediction capabilities. The abovementioned trends are also easily distinguished in the plots of the kappa coefficient and the overall accuracy (Fig. 11). The weight of the factors in this model is quite different from the preceding two. The elevation is outlined here as the most influencing factor. This could likely be once again the influence of the concentration of points classified as landslide zone at the downstream end of the basin.

The analyses of the area under the curve for ROC and PRC plots are revealing a significant improvement of the models compared with SI and LR. While the performance of the model using TT = 50/50 is not exhibiting any advancement, the other cases obtained higher results in the range of 20–25% compared to the rest of the models and combinations. It can be noted once again the better performance when a threshold of 0.5 is adopted. As in the previous modelling results, the combination of TT = 70/30 and TP = 20,000 (AUCROC = 0.96; AUCPRC = 0.93), and all using TP = 100,000 & 200,000 (AUCROC = 0.95–0.97; AUCPRC = 0.94–0.96) are yielding the most satisfactory scores.
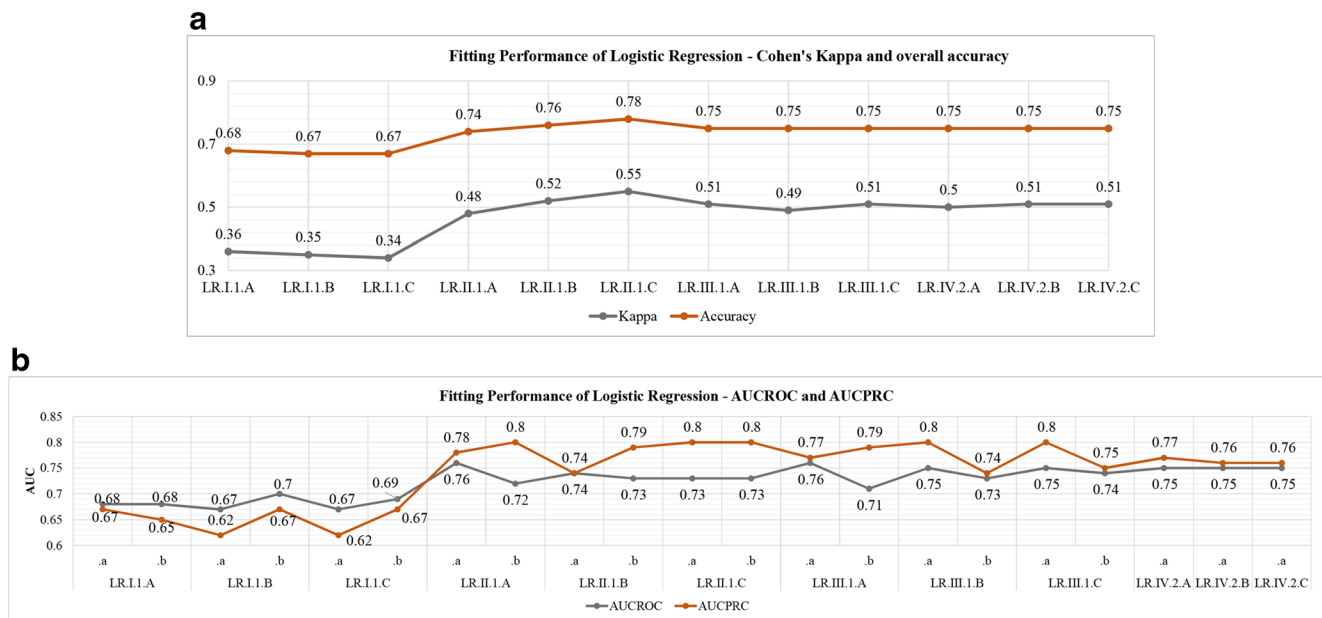
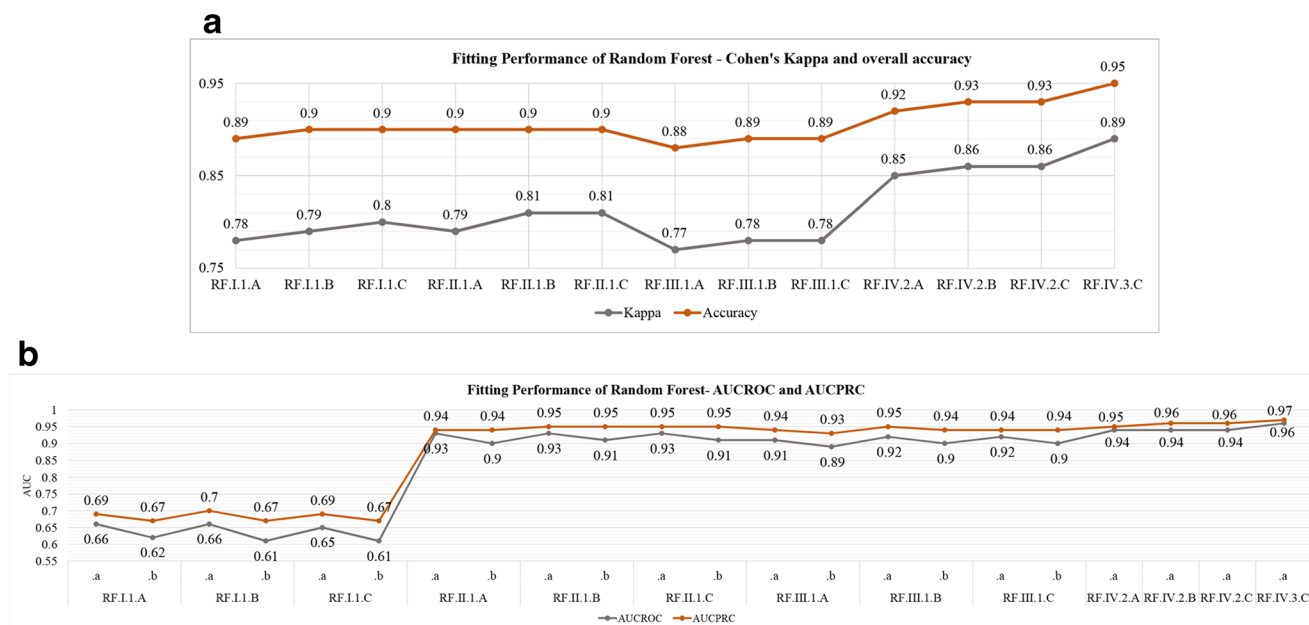**Fig. 10** Fitting performance of LR—Cohen's Kappa and overall accuracy, AUCROC and AUCPRC

a

b

Fig. 11 Fitting performance of RF—Cohen's Kappa and overall accuracy, AUCROC and AUCPRC

Comparing models predictive performances with statistical metrics

Since all modelling approaches exhibited satisfactory results (with some exceptions) in terms of fitting performance, it was an important step to evaluate also their predictive efficiency. As in the model fit assessment, the same metrics (namely Cohen's coefficient, overall accuracy, AUCROC, and AUCPRC) were implemented in this case for all three classifiers. To evaluate the performance of the models, they have to be tested against a dataset not used in the training process. The partitions from training/test ratios (e.g. 10, 30, and 50%) were used as external test sets. For the cases when 100% of the landslide inventory was used for training purposes, as an external test portion was assigned the 10% from TT = 90/10. Using the 10% partition is not exhibiting a problem in overlapping point samples as training and test due to the fact for each TP case new random points were created.

In the reported plots (Fig. 12a) of Kappa and OV, it can be noticed that the trends exhibited in model fit are similar to the performance once. In fact, the resulted values for the three modelling approaches are slightly lower than before which can be expected. The results are just confirming the remarks of the previous analyses and that the models produced with SI and TP = 50/50 and trained just with the polygonal landslides cannot be considered reliable due to the low Kappa ranging in 0.22–0.30 and OA between 0.61 and 0.64. Moreover, the AUCROC and AUCPRC are confirming with their low outcomes in the in 0.54–0.58 (Fig. 12b). Nevertheless, the models created with TT = 70/30 and 90/10 are showing satisfactory results.

As in the case of SI, the predictive capabilities (Fig. 13) of the LR are lower than the modelling. It is interesting to note that there is no clear delineation between the modelling combinations—the Kappa, OA, and AUCROC are almost steady with their values with small biases. On the other hand, the AUCPRC displays big differences between the different threshold outcomes; in some cases (LR.II.1.A), they can be up to 10% in favour of the 0.6 threshold. The resulted scores are showing an improvement in the susceptibility modelling with.

The plot (Fig. 14) containing the Kappa coefficients and OA shows that the classification done through the Random Forest algorithm yields much higher performance results and, in this case, the jump between TT = 50/50 and the rest of the combinations is missing. AUCROC and AUCPRC, on the opposite, are showing a steady improvement in the models with the increase of the number of training points. Different of LR case, the differences between the two threshold combinations are small and the best results are obtained with the 0.5 value.

Reassessment of the input data and results from the Refined flow

The results derived till now had demonstrated a high success rate (from a model point of view) with certain modelling approaches (e.g. RF and LR) and parametric combinations (e.g. *II.2.A, II.3.A, III.2.A*, etc.). A geological evaluation was performed in order to refine the inputs related to the training samples. Two main aspects were focused on: (1) the assignments of zones that can be considered with really low almost zero probability of landslide occurrence and (2) evaluation on
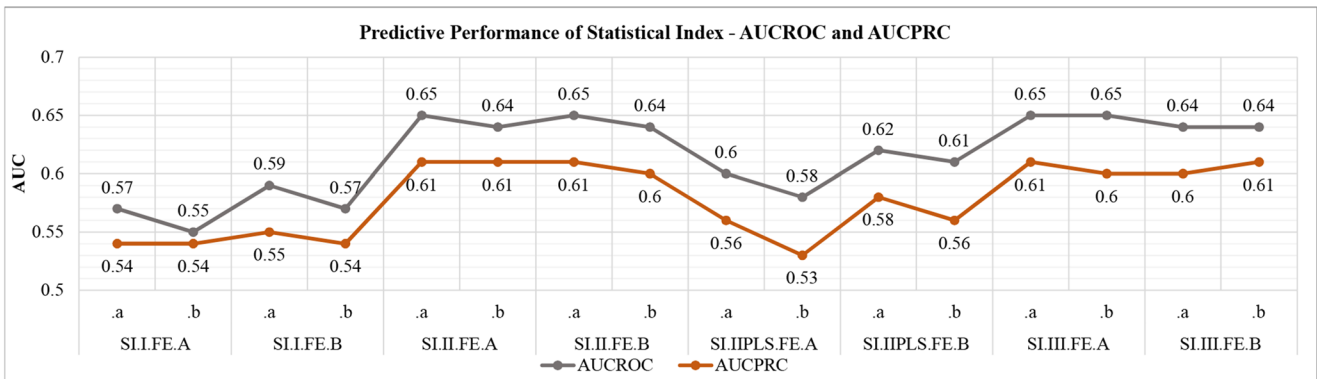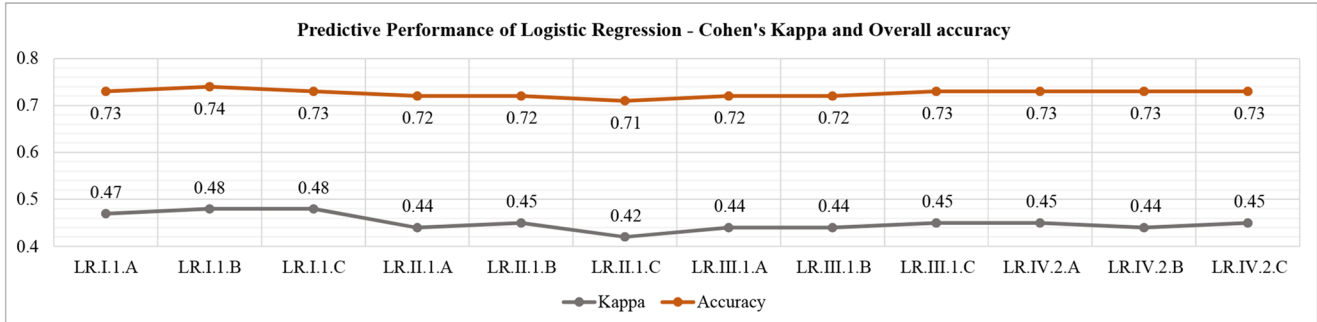
a
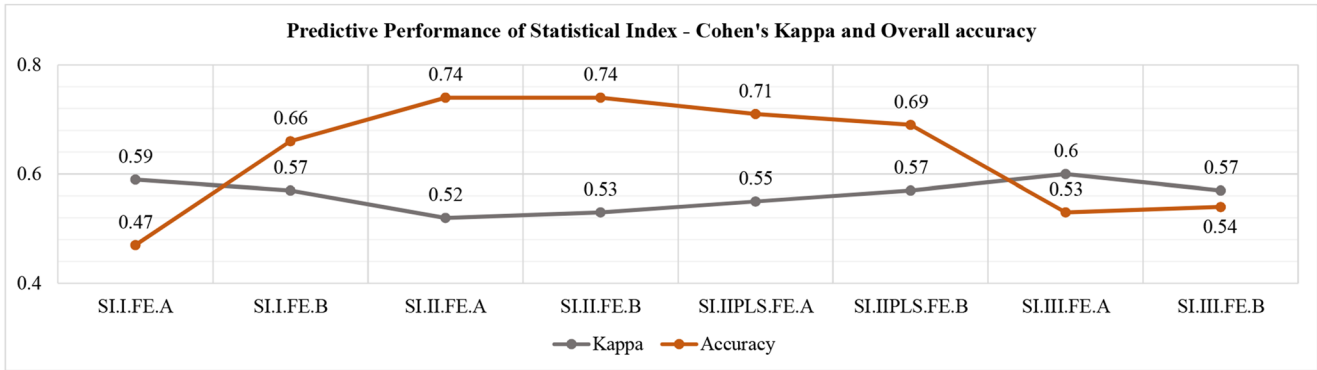
b

Fig. 12 Predictive performance of SI—Cohen's Kappa and overall accuracy, AUCROC and AUCPRC

a

b

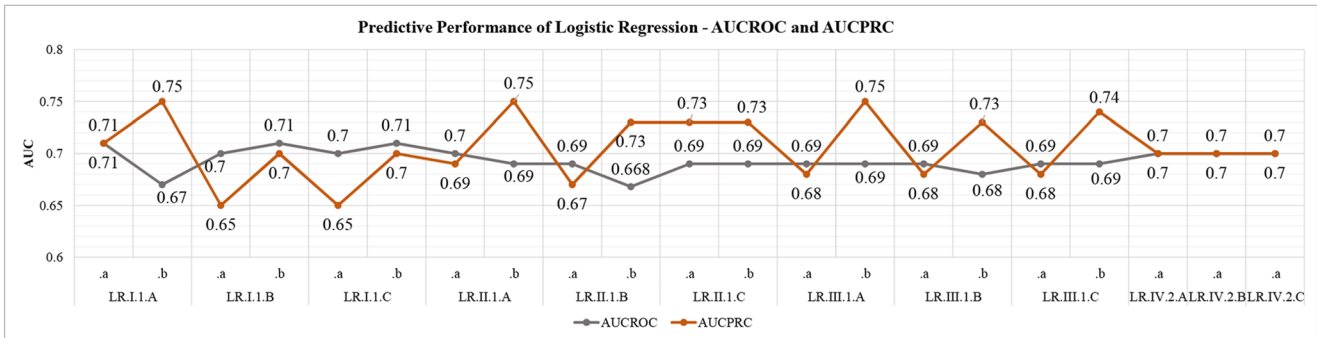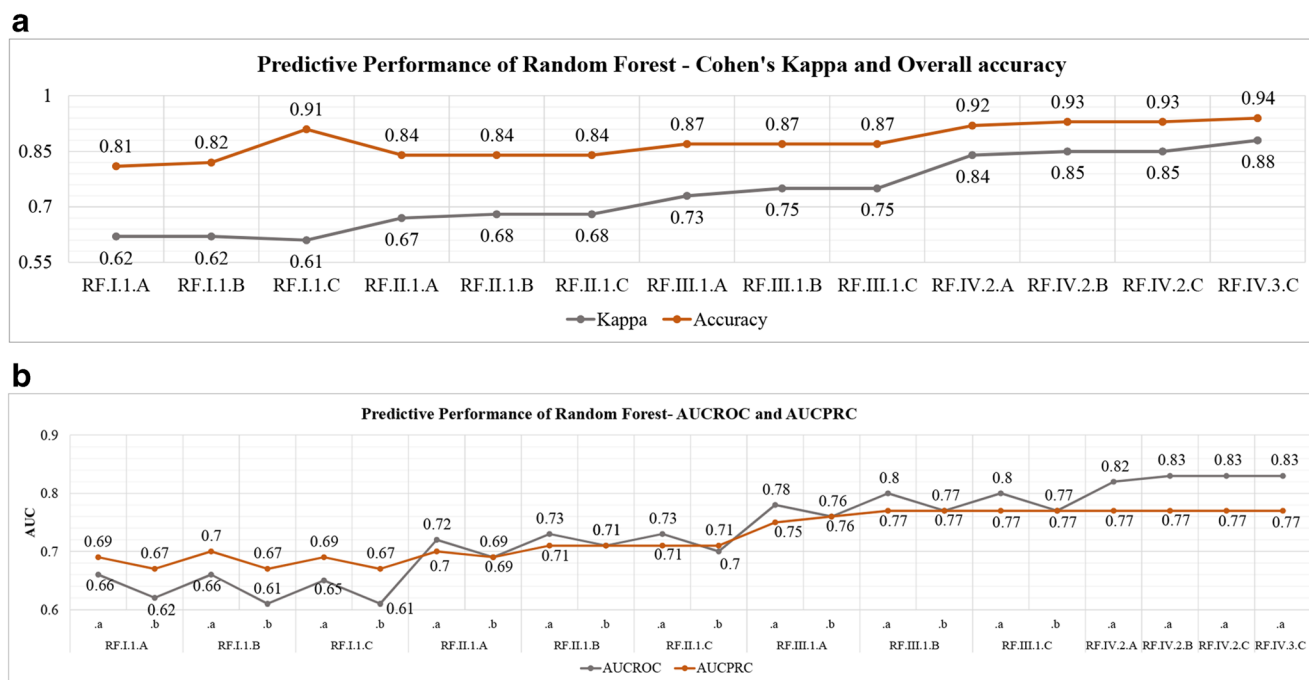Fig. 13 Predictive performance of LR—Cohen's Kappa and overall accuracy, AUCROC and AUCPRC

a

b

Fig. 14 Predictive performance of RF—Cohen's Kappa and overall accuracy, AUCROC and AUCPRC

the suitability of the landslide records in the inventory for the current modelling approaches.

Introducing 'no landslide' zone

As discussed at the beginning for defining the 'no landslide', training/testing sample sets were used areas outside the landslide inventory. While in a general case, this approach can deliver satisfactory results; in the case of Val Tartano and high landslide density, it was decided to discriminate zones suitable for 'no landslide' class training on the basis on a geological evaluation. The motivation was that due to the high event density in the valley, 'landslide' training samples are covering the same range of values compared to the variables' range; moreover, the frequency distributions are in the same order. This observation can derive two conclusions: (1) the training samples are good representatives of the terrain variable and (2) the 'landslide' cases are covering all of the variables ranges; therefore, the assumption of not having a landslide record at the current time is not sufficient enough for training.

Upon a geological evaluation were determined areas which are considered to be less likely to involve mass movements. The main assumption for their delineation was related to the inclination of the slope—very low (< 20°) hillslopes are unlikely to host landslides and very high slopes (> 70°) generally composed of rock, unlikely to retain soil deposits prone to land-sliding. Another aspect under evaluation was areas of bare intact rock. Those areas, however, tend to overlap with the areas in which the hillslopes are characterized by an

inclination higher than 70°. Therefore, the latter conditions will be considered as determinant for the 'no landslide' zones.

As a result, most of the areas correspond to low inclinations and therefore plane surfaces, covering a total area of 5.79 km². In fact, further validation of the suitability of these zones (from now on and in the model reference system will be referred as *GeoNoLS*) was carried out in order to detect a possible overlap of the *GeoNoLS* zones with the pre-existing landslide inventory. It was determined that less than 10% of *GeoNoLS* are located in the inventory database. This percentage falls into the lowest possible inclination class and can be deducted it is due to deposit accumulations. After excluding the 10% of the overlapped areas, the rest of *GeoNoLS* zones were used for sampling of the terrain variables and included into new models produced through random forest and logistic regression. The input variables and approaches were kept the same as some of the best performing models. The only main difference, except for the training samples of class 'no landslide', was the exclusion of the *slope angle* as a terrain variable. It was motivated to the fact that it is already externally biased with the focusing on some particular slope angle classes, and it will highly affect the outputs of the models.

At initial visual analyses, two aspects can be noted from the newly obtained maps (Online Resource 4): (1) as expected, the areas defined as *GeoNoLS* zones are classified as low susceptible zones and (2) the areas with very high landslide probability are increased severely compared to their previous analogue models. The frequency distribution of the 'very high' susceptibility levels and the true positives inside the landslide test areas demonstrate a significant improvement

Table 5 Summary of the evaluation metrics for the *RF.II.3.A.a.GeoNoLS*

RF.II.3.A.a.GeoNoLS			
OA	Kappa	OOB	AUCROC
0.98	0.90	1.70%	1

of the modelling approach which is demonstrated in Online Resource 4. The most evident improvement is notable in the map *RF.II.3.A.a.GeoNoLS* (summary of the results in Table 5).

The model results can be divided in more conservative or less conservative, as some delineate almost the entire area of the catchment as 'very high' susceptibility zone, while others are tending to the other extreme—most of the area is identified as 'low' susceptibility zone. A result that is a compromise between these two extrema should be the most suitable since areas prone to the phenomena are not being assigned with a critical status (e.g. LR.III.3A.a). Then again from a model point of view, the model performance is not sufficient.

As it appears, the *RF.II.3.A.a.GeoNoLS* is the best-performing model among the ones with the implemented *GeoNoLS* zone. It appears to be grasping the areas that are indeed prone to instability in a good way—high-gradient soil covered areas on both sides of the valleys as well as channels characterized by the presence of loose material. Instead, plain areas and high-altitude zones characterized by bare rock and the absence of soil cover are identified as low susceptibility zones. The bottoms of the valleys are correctly identified as low susceptibility zones as well. On one hand, the zones prone to failure are correctly identified but on the other hand, the 'very high' level of criticality could be too alarming.

Refined input of the landslides from the IFFI database

The assumption that the Validation Landslide is an outlier of the training samples; the fact that low elevation is considered as a highly important factor in the previous models urged a re-

evaluation of the IFFI landslide inventory. In fact, the information related to the type of movement is quite restricted. Upon investigation of the landslide records and their relationship with elevation as a factor, it was discovered a feature in the IFFI database (Fig. 15 denoted as 'debris flow' hosted in the lowest elevation class between 250 and 500 m a.s.l with a slope angle < 10°. Through visual analysis, it was determined that this feature resembles more to debris accumulation, rather than a debris flow. Moreover, the debris accumulation could occur due to the fact that it is located in the exact vicinity of the Pruna landslide, which can be considered as a source of the material. As mentioned in the beginning, Pruna landslide is covering a surface area of around 1 km^2, can reach depth to 100 km, and is located on an elevation range between 550 and 1200 m a.s.l. The fact that there is no similar to Pruna mass movement feature in the area of Val Tartano makes it the outlier of the inventory and combined with the other 'debris flow' they have a great weight on all of the models. Therefore, they were decided to be excluded from the training and testing samples and new ones were prepared using the reduced landslide inventory and the *GeoNoLS* zone.

Due to the previous modelling approaches, the implementation of various parametric combinations was omitted. Moreover, based on the evaluation results, training/test ratio was selected as 70/30 and the count of sampling points to TP = 100,000. Although the scenarios using 200,000 samples yielded higher performance, with the reducing of the zones it could not be any more ensured a minimum distance of 5 m (according to the pixel size) and high redundancy of the samples was anticipated. As for the modelling approach, it was decided to implement all of the three previously used methods. The maps are reported into Fig. 17, while the overall modelling and predictive performance are in Fig. 16.

From the results in Fig. 16, the almost perfect performances of RF according to all four implemented metrics can be noted. The LR exhibits some intermediate but satisfactory performances, while the SI is again slightly above the threshold (0.5 for AUCROC) to be considered as random classification. On the other hand, according to AURPRC, the output is not
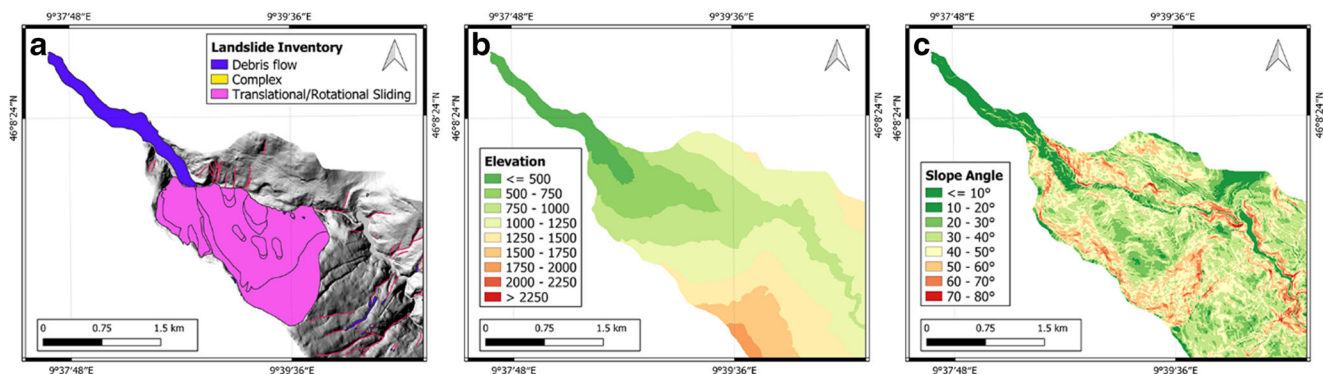
Fig. 15 Terrain variables for the areas of Pruna landslide and debris flow feature on a flat surface: **a** is representing the Pruna landslides and the debris flow; **b** the elevation; and **c** the slope gradient for the same area
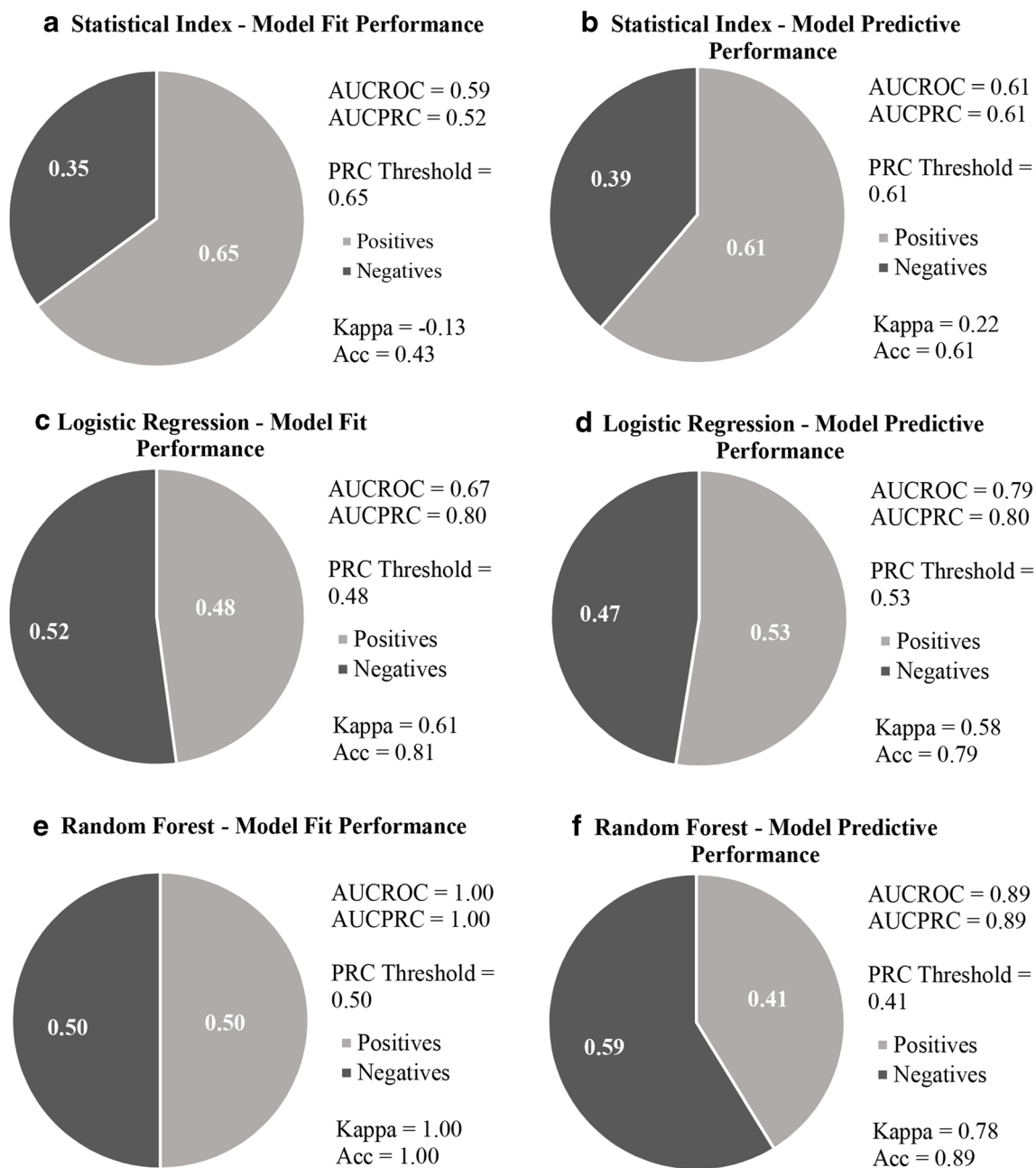
a **Statistical Index - Model Fit Performance**

AUCROC = 0.59
AUCPRC = 0.52

PRC Threshold = 0.65

■ Positives
■ Negatives

Kappa = -0.13
Acc = 0.43

b **Statistical Index - Model Predictive Performance**

AUCROC = 0.61
AUCPRC = 0.61

PRC Threshold = 0.61

■ Positives
■ Negatives

Kappa = 0.22
Acc = 0.61

c **Logistic Regression - Model Fit Performance**

AUCROC = 0.67
AUCPRC = 0.80

PRC Threshold = 0.48

■ Positives
■ Negatives

Kappa = 0.61
Acc = 0.81

d **Logistic Regression - Model Predictive Performance**

AUCROC = 0.79
AUCPRC = 0.80

PRC Threshold = 0.53

■ Positives
■ Negatives

Kappa = 0.58
Acc = 0.79

e **Random Forest - Model Fit Performance**

AUCROC = 1.00
AUCPRC = 1.00

PRC Threshold = 0.50

■ Positives
■ Negatives

Kappa = 1.00
Acc = 1.00

f **Random Forest - Model Predictive Performance**

AUCROC = 0.89
AUCPRC = 0.89

PRC Threshold = 0.41

■ Positives
■ Negatives

Kappa = 0.78
Acc = 0.89

Fig. 16 Summary of the model evaluations

classified on a model basis, rather on random chance, which is also supported by the kappa and accuracy values. Moreover, the pie charts are highlighting the discussion (in the "Validation" section) related to the evaluation metrics according to imbalanced datasets, that AUCROC is insensitive to such imbalance and can provide a false sense of success. Except for the model fit of RF, the rest are also exhibiting some imbalance, but it is almost negligible compared to SI or the imbalance is actually in favour of the negatives. Such class outnumbering is not taken into account during the PRC computations.

The resulted maps exhibit the same pattern in comparison to the previous *GeoNoLS* maps—relatively flat areas are assigned to low probability, while the critical zones are again widely spread and correctly classified also according to the test landslide set. An interesting pattern is noted into the map produced with SI, due to the fact that the highest weights are assigned to a class of the plan curvature. On the other hand, the modelling and predictive performances of *SI* cannot be considered as acceptable. A slight increase is noted into the *LR* map and the highest results from the metrics are achieved through the Random Forest model. The high influence of the

plan curvature in the SI is also confirmed in the *RF* model (Online Resource 3), next to the effects from the aspect. Areas where specific classes for both factors are reasonable to be highly prone to failure. Convex curvature can increase the instability in a slope, as well as its north orientation due to increased stored moisture.

The area covered by the Pruna landslide is classified into high and very high susceptibility level. Even though it was considered as an outlier from the whole IFFI database, the model performance correctly recognizes the area. In fact, almost all of the maps the landslide is situated in an area that is generally classified as 'high/very high' susceptibility zone and so it appears to be a sort of an outlier. One explanation could be related to the time when the DTM was produced and the occurrence of the landslide. The topography of the slope was in some way modified by the removal of material. In case the DTM was generated after the occurrence of the landslide, the estimation of geomorphometric factors has been carried out on a topography modified by the occurrence of the landslide. Unfortunately, the exact occurrence of the landslide could not be exactly determined. The assumption that it is outlier and underrepresented in the training samples is supported also due to its relatively small size in comparison to the majority of the records inside the landslide database.

In addition, the previously removed 'debris flow' record is assigned to low probability, which is in an agreement with our expectations. Based on the scores and the geological evaluation, it can be reasoned that the map with the best performances scores and parameter combinations is the one resulting from the Random Forest model (Fig. 17c).

Discussion

With the high number of produced different models (46) and the resulting maps (79) in the presented work, it was possible to highlight mapping approaches that can derive reliable results and, in the meantime, to point out some weak elements in the parametric combinations that can lead to misleading outcomes. During the iterative processes, we were able to test, verify, and improve the suitability of three statistically based models (Statistical Index, Logistic Regression, and Random Forest); the different combinations into train/test partitioning; various terrain sampling approaches; variable inclusion; validation techniques; and the importance of landslide inventory.

Our results are able to point the fact that a small modification in the input can change the output map in a different way. Moreover, it is clear that model performance is dominated by the Random Forest, and the Statistical Index in most of the cases was yielding results that are on the border to be accepted as sufficient. The approaches using Logistic Regression, highly used and reliable model, were still with very good results. Nevertheless, the outcomes of RF and LR have their

similarities and differences only in the evaluation metric but also in the maps itself. For both of them, high similarities can be found in the extrema of the probability range. While RF tends to be more conservative and 'secure', the medium probability range is almost brought to a minimum. LR on the hand is not producing such a firm baseline between the 'high' and 'low' susceptibility levels, which of course has a trade-off with the predictive capabilities.

Based on the results, it can be concluded that a training/test ratio of 50/50 is not an approach to be considered. The outcomes from the other two scenarios are more than satisfactory results. Where small difference can be noted between 70/30 and 90/10, in terms of model and predictive performances. Much higher attention should be allocated to the landslide inventory used as a reference. Pruna landslide highlighted the importance of constantly updated and thoroughly compiled landslide database. Moreover, in concern to the susceptibility mapping, it is always suggested to map a particular or a group of similar types of mass movements.

The adopted scenarios concerning the sample amount showed that a reasonable amount of training points should be selected. In all three approaches, the models managed to obtain results more than satisfactory. Of course, the higher number the closer to reality are the samples, yet in our cases, even 10% of the total pixel count was good representative. On the hand, it should be paid attention not to introduce too high amount for the area interest which can create redundancy in the data, unreliable outcome and high computational time.

At what concerns the terrain variables input, a thorough investigation of the causative factors should be carried on and for the further modelling to be used the ones that are the most relevant. Otherwise, it is risky to introduce a high number of variables that can lead to unstable models. As for this study, the behaviour of the three scenarios in relation to the precipitation input was analysed—kriging, inverse distance weighting, and no precipitation at all. When included, the better results were obtained using the IDW yet based on the variable importance, it was concluded that the overall importance of the factor is low, therefore, at a certain point, was omitted. In fact, the case study was considered as too small in relation to the precipitation differences that can occur in the basin. Meaning that in some areas ten times bigger, the rainfall amount may vary drastically from one part to the other and can have different triggering magnitude, while for the Val Tartano, the variations are not so drastically different.

The concentration of landslides in a particular area could erroneously assign high weights on factors with negligible influence on susceptibility or evidence strictly site-specific characteristics in a basin. In particular, the presence of a large 'landslide' zone in the lower end of the catchment led to the estimation of high factor weights on low elevation and low precipitation—an outcome with a little physical meaning with respect to the mechanisms of the landslide phenomena.
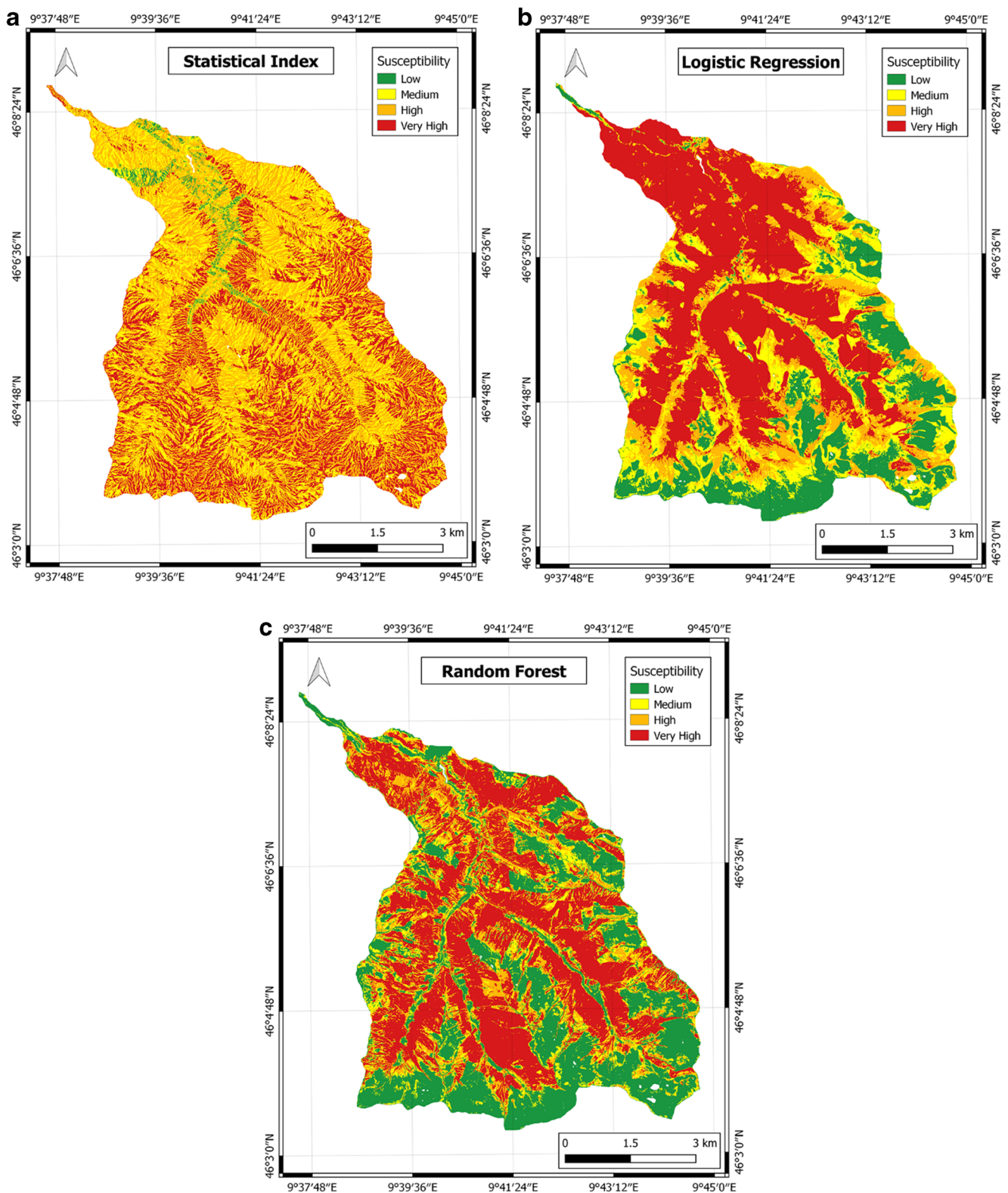
Fig. 17 The results susceptibility maps using the *GeoNoLS* and reduced inventory—**a** Statistical index, **b** Logistic Regression, and **c** Random Forest

Nevertheless, after the selective input data procedure, LR and RF models correctly identified the high susceptibility to failure of the zone that was excluded from the training dataset, which vouches for the reliability of the model performance when possible bias effects are removed from its input. Moreover, our conclusions for the performance of the LR and RF can be

compared with similar findings in the literature (Ayalew and Yamagishi 2005; Trigila et al. 2015; Aditian et al. 2018).

In the current work, it was decided to implement some of the most used evaluation metrics in the susceptibility modelling—AUCROC, Cohen's Kappa, and overall accuracy. Moreover, it was included AUCPRC with the intent to evaluate its suitability for the particular purposes. In most of the results, all four were exhibiting the same trends, which was a positive aspect to easier distinguish the better performing models. Yet in the case of LR, AUCPRC exhibited great fluctuations between the different scenarios, for which was difficult to explain the behaviours. It should be noted that the AUCPRC does not exhibit straightforward result reading, meaning that the baseline of 0.5 (as in AUCROC) is not present and it is computed as the ratio of the positives and negatives. Aside from the statistical metrics, map evaluation should be performed by geologists with expertise in the case study area and the occurred events in it. It is possible that upon their evaluation, a map with not the best assessment scores would be more reliable and closer to the ground truth.

Based on the obtained results from the model fit and predictive performance, it was deduced from analyses on the test landslide set and expert assessment that the most precise and accurate map (Online Resource 5) is the one obtained with Random Forest; TT = 70/30; TP = 100,000; GeoNoLS; excluded precipitation, slope, and Pruna landslide.

Conclusion

The presented work covers a wide range of susceptibility modelling scenarios through three main modelling techniques. The advantages of using Random Forest and Logistic Regression were depicted, while the reliable performance of Logistic Regression was confirmed through all the scenarios. On the other hand, Random Forest exhibited much better performance and more valuable outputs. For the case of Statistical Index, even though the results were positive, although not by all metrics, they cannot be considered to be reliable. The implementation of different sampling approaches proved that sample count should be adequately determined: on the one hand, not to under-sample the terrain conditions, and on the other hand, not to increase the computational time without significant result improvement. The partitioning of the test dataset was analysed, and 70/30 was found to represent a good ratio for training the model and to provide sufficient data for its testing. On the other hand, the widely adopted AUCROC curve demonstrated that it is not always reliable and should not be accepted as a fool proof metric. Rather, in this work, it was suggested that other evaluation metrics, such as AUCROC and Cohen's Kappa, should be implemented. While the discrepancies between AUCROC and PRC can be the topic of further work, it is strongly suggested that cross-

validation of produced susceptibility mapping and even re-evaluating the inputs used as terrain variables should be done. In fact, in the presented work, geological evaluation of the intermediate results contributed highly to improving the quality of the maps. While the choice for the most suitable terrain variables cannot be standardized and is highly dependent on each particular case study, in this work, variables that are most relevant for the particular case of Val Tartano were used, based on variables that are most utilized in literature. Therefore, it would be of great benefit to explore the contribution of others that are less used but still relevant for the case. The inclusion of time as a factor in the landslide inventory can bring out other important aspects in the modelling and the results. Lastly, but very importantly, the question of the most suitable validation metric is still open and efforts should be applied to achieve a reliable metric for such modelling. Nevertheless, it is suggested that great attention be paid to input data and combinations when building susceptibility models.

Data availability Upon a request.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Code availability Upon a request

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 318:101–111. https://doi.org/10.1016/j.geomorph.2018.06.006

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Automat Contr 19:716–723. https://doi.org/10.1109/TAC.1974.1100705

Arabameri A, Pradhan B, Rezaei K, Lee C-W (2019) Assessment of landslide susceptibility using statistical-and artificial intelligence-based FR–RF integrated model and multiresolution DEMs. Remote Sens 11:999

ARPA Lombardia (2019) Dati e Indicatori. https://www.arpalombardia.it/Pages/Ricerca-Dati-ed-Indicatori.aspx. Accessed 5 Feb 2019

Ayalew L, Yamagishi H (2005) The application of GIS-based logistic regression for landslide susceptibility mapping in the Kakuda-Yahiko Mountains, Central Japan. Geomorphology 65:15–31. https://doi.org/10.1016/j.geomorph.2004.06.010

Bai SB, Wang J, Lü GN, Zhou PG, Hou SS, Xu SN (2010) GIS-based logistic regression for landslide susceptibility mapping of the Zhongxian segment in the three gorges area, China. Geomorpholgy 115:23–31. https://doi.org/10.1016/j.geomorph.2009.09.025

Ballio F, Brambilla D, Giorgetti E, et al (2010) Evaluation of sediment yield from valley slopes: a case study. In: WIT Transactions on Engineering Sciences

Brambilla D, Longoni L, Papini M et al (2011) On analysis of sediment sources toward proper characterization of hydro-geological hazard for mountain environments. Int J Saf Secur Eng 1:423–437

Breiman L (1996) Out-of-bag estimation

Breiman L (2001) Random forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324

Breiman L, Friedman JH, Olshen RA, Stone CJ (2017) Classification and regression trees

Castellanos Abella EA, Van Westen CJ (2008) Qualitative landslide susceptibility assessment by multicriteria analysis: a case study from San Antonio del Sur, Guantánamo, Cuba. Geomorphology 94:453–466. https://doi.org/10.1016/j.geomorph.2006.10.038

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. https://doi.org/10.1007/s10346-005-0021-0

Chalkias C, Ferentinou M, Polykretis C (2014) GIS-based landslide susceptibility mapping on the Peloponnese peninsula, Greece. Geoscience 4:176–190. https://doi.org/10.3390/geosciences4030176

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 Landf 36:1632–1646. https://doi.org/10.1002/esp.2178

Copernicus (2019) Sentinel data 2019

Cruden DM, Varnes DJ (1996) Landslide types and processes. Spec Rep - Natl Res Counc Transp Res Board

Díaz S, Settele J, Brondízio E, et al (2019) Summary for policymakers of the global assessment report on biodiversity and ecosystem services – unedited advance version. Ipbes

ESA SNAP toolbox (2019) SNAP ESA Sentinel Application Platform

Fell R, Corominas J, Bonnard C, Cascini L, Leroi E, Savage WZ (2008) Guidelines for landslide susceptibility, hazard and risk zoning for land use planning. Eng Geol 102:85–98. https://doi.org/10.1016/j.enggeo.2008.03.022

Freeman E, Frescino T, Moisen G (2009) ModelMap: an R package for modeling and map production using Random Forest and Stochastic Gradient boosting. USDA For Serv Rocky Mt Res Stn 507

GeoPortal (2019) GeoPortale Lombardia. http://www.geoportale.regione.lombardia.it/

GRASS Development Team (2017) Geographic Resources Analysis Support System (GRASS GIS) Software, Version 7.2

Günther A, Reichenbach P, Malet JP, van den Eeckhaut M, Hervás J, Dashwood C, Guzzetti F (2013) Tier-based approaches for landslide susceptibility assessment in Europe. Landslides. 10:529–546. https://doi.org/10.1007/s10346-012-0349-1

Guzzetti F, Carrara A, Cardinali M, Reichenbach P (1999) Landslide hazard evaluation: a review of current techniques and their application in a multi-scale study, Central Italy. Geomorphology 31:181–216. https://doi.org/10.1016/S0169-555X(99)00078-1

Guzzetti F, Reichenbach P, Cardinali M, Galli M, Ardizzone F (2005) Probabilistic landslide hazard assessment at the basin scale. Geomorphology. 72:272–299. https://doi.org/10.1016/j.geomorph.2005.06.002

Guzzetti F, Reichenbach P, Ardizzone F, Cardinali M, Galli M (2006) Estimating the quality of landslide susceptibility models. Geomorphology. 81:166–184. https://doi.org/10.1016/j.geomorph.2006.04.007

Guzzetti F, Mondini AC, Cardinali M et al (2012) Landslide inventory maps: new tools for an old problem. Earth Sci Rev 122:42

He Y, Beighley RE (2008) GIS-based regional landslide susceptibility mapping: a case study in southern California. Earth Surf Process Landf 33:380–393. https://doi.org/10.1002/esp.1562

Hosmer DW, Lemeshow S (2000) Applied logistic regression. Wiley, New York

Iadanza C, Trigila A, Vittori E, Serva L (2009) Landslides in coastal areas of Italy. Geol Soc Spec Publ 322:121–141. https://doi.org/10.1144/SP322.5

Jasiewicz J, Stepinski TF (2013) Geomorphons-a pattern recognition approach to classification and mapping of landforms. Geomorphology. 182:147–156. https://doi.org/10.1016/j.geomorph.2012.11.005

Liao WH (2010) Region description using extended local ternary patterns. In: Proceedings - International Conference on Pattern Recognition

Liaw A, Wiener M (2002) Classification and regression by randomForest. R News

Longoni L, Papini M, Brambilla D, Barazzetti L, Roncoroni F, Scaioni M, Ivanov V (2016) Monitoring riverbank erosion in mountain catchments using terrestrial laser scanning. Remote Sens 8. 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 10:1851–1864. https://doi.org/10.5194/nhess-10-1851-2010

Mandelli M, Longoni L, Papini M, et al (2009) Modellazione del trasporto di sedimenti sul bacino del Tartano (Valtellina)

Meijerink AMJ (1988) Data acquisition and data capture through terrain mapping units. ITC J

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

Park S, Kim J (2019) Landslide susceptibility mapping based on random forest and boosted regression tree models, and a comparison of their performance. Appl Sci 9. https://doi.org/10.3390/app9050942

Pettorelli N (2013) The normalized difference vegetation index

Pourghasemi HR, Moradi HR, Fatemi Aghda SM (2013) Landslide susceptibility mapping by binary logistic regression, analytical hierarchy process, and statistical index models and assessment of their performances. Nat Hazards 69:749–779. https://doi.org/10.1007/s11069-013-0728-5

R Development Core Team R (2011) R: a language and environment for statistical computing

Reichenbach P, Rossi M, Malamud BD et al (2018) A review of statistically-based landslide susceptibility models. Earth Sci Rev 180:60

Remondo J, Bonachea J, Cendrero A (2005) A statistical approach to landslide risk modelling at basin scale: from landslide susceptibility to quantitative risk assessment. Landslides. 2:321–328. https://doi.org/10.1007/s10346-005-0016-x

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 10:e0118432. https://doi.org/10.1371/journal.pone.0118432

Samia J, Temme A, Bregt A, Wallinga J, Guzzetti F, Ardizzone F, Rossi M (2017) Do landslides follow landslides? Insights in path dependency from a multi-temporal landslide inventory. Landslides 14:547–558. https://doi.org/10.1007/s10346-016-0739-x

Scienze CLM, Prof TG, Baroni C (2007) Progetto IFFI : Inventario dei Fenomeni Franosi in Italia. http://www.progettoiffi.isprambiente.it

Taalab K, Cheng T, Zhang Y (2018) Mapping landslide susceptibility and types using Random Forest. Big Earth Data 2:159–178. https://doi.org/10.1080/20964471.2018.1472392

Temme A, Guzzetti F, Samia J, Mirus BB (2020) The future of landslides' past—a framework for assessing consecutive landsliding systems. Landslides 17:1519–1528

Trigila A, Iadanza C (2008) Landslides in Italy

Trigila A, Frattini P, Casagli N, et al (2013) Landslide susceptibility mapping at national scale: the Italian case study. In: Landslide science and practice. Springer, pp. 287–295

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. 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. 139-140:545–558. https://doi.org/10.1016/j.geomorph.2011.12.006

Van Den Eeckhaut M, Hervás J, Jaedicke C et al (2012) Statistical modelling of Europe-wide landslide susceptibility using limited landslide inventory data. Landslides. 9:357–369. https://doi.org/10.1007/s10346-011-0299-z

van Westen CJ, Rengers N, Terlien MTJ, Soeters R (1997) Prediction of the occurrence of slope instability phenomenal through GIS-based hazard zonation. Geol Rundsch 86:404–414. https://doi.org/10.1007/s005310050149

Van Westen CJ, Rengers N, Soeters R (2003) Use of geomorphological information in indirect landslide susceptibility assessment. Nat Hazards 30:399–419. https://doi.org/10.1023/B:NHAZ.0000007097.42735.9e

van Westen CJ, Castellanos E, Kuriakose SL (2008) Spatial data for landslide susceptibility, hazard, and vulnerability assessment: an overview. Eng Geol 102:112–131. https://doi.org/10.1016/j.enggeo.2008.03.010

Varnes D (1984) Landslide hazard zonation : a review of principles and practice. Nat Hazards

Yordanov V, Brovelli MA, (2020) Comparing model performance metrics for landslides susceptibility mapping. Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XLIII-B3-2020, 1277–1284, https://doi.org/10.5194/isprs-archives-XLIII-B3-2020-1277-2020