DOI: 10.4408/IJEGE.2021-01.S-19

NEW EXPLANATORY VARIABLES TO IMPROVE LANDSLIDE SUSCEPTIBILITY MAPPING: TESTING THE EFFECTIVENESS OF SOIL SEALING INFORMATION AND MULTI-CRITERIA GEOLOGICAL PARAMETERIZATION

Samuele SEGONI(*), Nicola NOCENTINI(*), Ascanio ROSI(*), Tania LUTI(*), Giulio PAPPAFICO(**), Michele MUNAFÒ(***), Nicola CASAGLI(*) & Filippo CATANI(****)

(*) University of Firenze - Department of Earth Sciences - Florence, Italy
(**) University of Urbino - Department of Pure and Applied Sciences - Urbino, Italy
(***) ISPRA - Italian Institute for Environmental Protection and Research - Rome, Italy
(****)University of Padova - Department of Geosciences Padova, Italy
Corresponding author: samuele.segoni@unifi.it

EXTENDED ABSTRACT

Le mappe di suscettibilità da frana sono strumenti che quantificano la probabilità spaziale di innesco di frane tramite lo studio statistico della combinazione di vari fattori predisponenti. Negli ultimi anni, la ricerca in tema di analisi della suscettibilità da frana si è focalizzata più su aspetti di tipo matematico-informatico (realizzazione di algoritmi sempre più complessi e performanti) che su aspetti di tipo geomorfologico e geologico. In questo studio, una mappa di suscettibilità viene prodotta per la Toscana settentrionale (3100 km^2) ricorrendo ad un algoritmo di machine learning detto Random Forest e, elemento di maggiore novità, affiancando ad un set di parametri predisponenti consolidati in letteratura un set di parametri innovativi derivati da carte tematiche ad alta risoluzione disponibili a scala nazionale: mappe di consumo di suolo e carte geologiche digitali in scala 1:10.000.

Da questi dati di base è stata definita una serie di parametri originali, che sono stati utilizzati come variabili indipendenti per la modellazione della suscettibilità.

Utilizzando i dati nazionali di monitoraggio del consumo di suolo, è stato derivato un indicatore del grado di antropizzazione del territorio. Il consumo di suolo può essere definito come la rimozione del suolo o la sua occlusione mediante materiali artificiali (parzialmente) impermeabili. È un processo antropico con impatti ambientali notevoli e l'indice derivato è stato testato come possibile dato di input nella modellazione della suscettibilità per tenere conto dell'interferenza delle azioni antropiche nell'assetto idrologico e geomorfologico dei versanti.

È stato inoltre introdotto un approccio multi-parametrico per introdurre nella modellazione l'informazione di tipo geologico in modo più completo. Di solito, negli studi di suscettibilità la litologia è l'unica variabile geologica considerata; ciò non è sbagliato ma costituisce un'informazione parziale: la litologia è solo una delle tante informazioni che le mappe geologiche possono fornire e dunque il potenziale di tali mappe rimane in gran parte sottoutilizzato. Nel presente studio, la cartografia geologica regionale in scala 1:10.000 è stata riclassificata secondo cinque criteri diversi (litologico, genetico, strutturale, paleogeografico e cronologico), dando luogo ad altrettanti parametri che sono stati utilizzati nella modellazione della suscettibilità per definire in modo più compiuto le caratteristiche geologiche dell'area.

Per quanto riguarda i parametri di letteratura, sono stati utilizzati quota, rugosità del rilievo, pendenza, energia del rilievo, orientamento del versante, tre tipologie di curvatura diverse, indice TWI e TPI, area drenata, uso/copertura del suolo e due variabili meteoclimatiche basate sui tempi di ritorno delle piogge. Inoltre, sono state introdotte anche due variabili casuali con lo scopo di individuare (e, successivamente, scartare) eventuali variabili che mostrano capacità predittiva simile a quella fornita da una classificazione o da una serie di misure casuali (circostanza non verificatasi nel nostro caso di studio).

L'efficacia dei nuovi parametri presentati in questa ricerca è stata valutata in modo oggettivo tramite due indicatori di performance comunemente usati in letteratura. L'AUC (area under receiver-operator characteristic curve) è stata utilizzata per misurare l'affidabilità della mappa di suscettibilità rispetto ad un dataset di verifica indipendente da quello utilizzato per addestrare il modello. L'OOBE (out of bag error - errore che verrebbe commesso dal modello se venisse ignorata una variabile) ha consentito di avere una stima dell'importanza relativa e della capacità predittiva di ogni parametro utilizzato.

È stato possibile concludere che l'impego del parametro derivato dal consumo di suolo fornisce un contributo positivo alla modellazione e non è ridondante rispetto alle variabili derivate dal Corine Land Cover, perché i due tematismi di base hanno risoluzioni spaziali, accuratezze tematiche e obiettivi di mappatura completamente diversi. Inoltre le mappe di consumo di suolo vengono aggiornate annualmente. Per quanto riguarda i parametri derivati dalla carta geologica, i risultati mostrano che l'uso combinato di tutti i parametri è migliorativo e dà risultati più accurati, aprendo nuove prospettive per l'impiego del dato geologico negli studi sulla suscettibilità da frana.

ABSTRACT

Landslide susceptibility maps (LSM) define the spatial probability of landslide occurrence based on the spatial distribution of predisposing factors. In this work, a LSM is produced for Norther Tuscany (3100 km^2) with a Random Forest algorithm. The element of novelty is the use, besides 15 state-of-the-art parameters, of some newly proposed parameters.

Starting from the national soil sealing map updated yearly by ISPRA, we derived a parameter accounting for the degree of human interference on hillslope systems. Soil sealing is the most intense form of land take, and it can be defined as the destruction (or covering) of soil by completely or partly impermeable artificial material.

A multi-criteria approach was introduced to get a more complex and complete geological information into LSM. Usually, lithology is the only geological variable used, leaving the potential of geological maps largely unexploited. We used a 1:10,000 geological map to define a set of parameters based on lithological, genetic, structural, paleogeographic and chronological criteria, and found that the joint use of all the geology-derived parameters improved the susceptibility assessment.

The outcomes of this study could be easily reproduced elsewhere in Italy, since the newly proposed parameters were generated from easily accessible datasets.

***Keywords**: landslide susceptibility, random forest, geology, soil sealing*

INTRODUCTION

Landslide susceptibility maps (LSM) can be defined as the representation of the spatial probability of landslide occurrence over appropriate spatial units (Brabb, 1984). LSMs have been extensively used for land planning (Cascini 2008; Frattini *et alii* 2010) and hazard assessment (Corominas *et alii*, 2003); more recently they have been successfully integrated also in quantitative risk assessment (Chen *et alii*, 2016) and early warning systems (Segoni *et alii*, 2018: Tiranti *et alii*, 2019).

In recent years, many advances have been proposed to increase the reliability of LSMs. As instance, new sophisticated machine learning algorithms are continuously proposed and applied to susceptibility mapping (Kavzoglu *et alii*, 2019), or existing models are hybridized to increase the predictive effectiveness of the susceptibility assessment (Thai Pham *et alii*, 2019; Li & Chen, 2020). Newly proposed and well-established models are continuously compared (Yilmaz 2009; Zizioli *et alii*, 2013; Kalantar *et alii*, 2018) and many attempts have been made to produce ensemble predictions by blending the different outputs coming from different models (Tien Bui *et alii,* 2019; Di Napoli *et alii*, 2020 and 2021). Another series of works, instead than on the models, focuses on their sensitivity to different methods used to apply them to real case studies, investigating e.g. model settings (Catani *et alii,* 2013), sampling techniques (Kalantar *et alii*, 2018), mapping units (Erener & Uzgun 2012; Canavesi *et alii* 2020) and spatial resolution (Chen *et alii*, 2020). Other works introduce robust validation procedures to measure the reliability of the LSM and to explain their outcomes (Frattini *et alii*, 2010; Xiao *et alii*, 2020). Of course, beside model algorithms and model settings, the selection of the explanatory variables used in the susceptibility assessment has a strong influence on the final results, and new parameters are continuously proposed for LSM studies. The present work originates in this framework, with the objective of using for landslide susceptibility mapping two newly proposed sets of input parameters. The first one is the use of soil sealing maps, a monitoring product delivered every year at national scale for the whole Italy by ISPRA (Italian Institute of Environmental Protection and Research), to account for the ongoing anthropogenic process of urbanization and impermeabilization of the territory. The second is a multi-parametrical approach to transfer geological information from detailed geological maps to landslide susceptibility models, accounting simultaneously for the lithology, paleoenvironment, tectonic history, genetic characteristics, and age of the mapped units. Some preliminary tests on the potentiality of these two sets of parameters have been recently published (Luti *et alii*, 2020; Segoni *et alii*, 2020), but to our knowledge this is the first time they are integrated in a full model configuration to map landslide susceptibility.

MATERIAL AND METHODS

Test site description

The test site (Fig. 1) is located in the Northern part of Tuscany and it is composed by the provinces of Lucca, Prato and Pistoia. This area has been selected for its heterogeneity, both from a morphological and a geological point of view and for the availability of a detailed landslide inventory. The area is also relatively small, about 3100 km^2 wide, so it has been considered a good test-site to verify the hypotheses of work, since it allows low time-consuming simulations.

From the morphological point of view, the northern part is characterized by mountains (with altitudes up to 2000 m asl) and intermontane plains, while the central and southern parts are mainly hilly and plain areas, respectively. The geological setting of the area is the effect of the Apennines orogenesis, which resulted in the juxtaposition of several tectonic units, piled during the Tertiary under a compressive regime that was followed by extensional tectonics from the Upper Tortonian (Vai & Martini 2001). This process produced a sequence of horst-graben structures with a NW-SE alignment, that resulted in the emplacement of Neogene sedimentary basins, mainly of marine (to the West) and fluvio-lacustrine (to the East) origin.

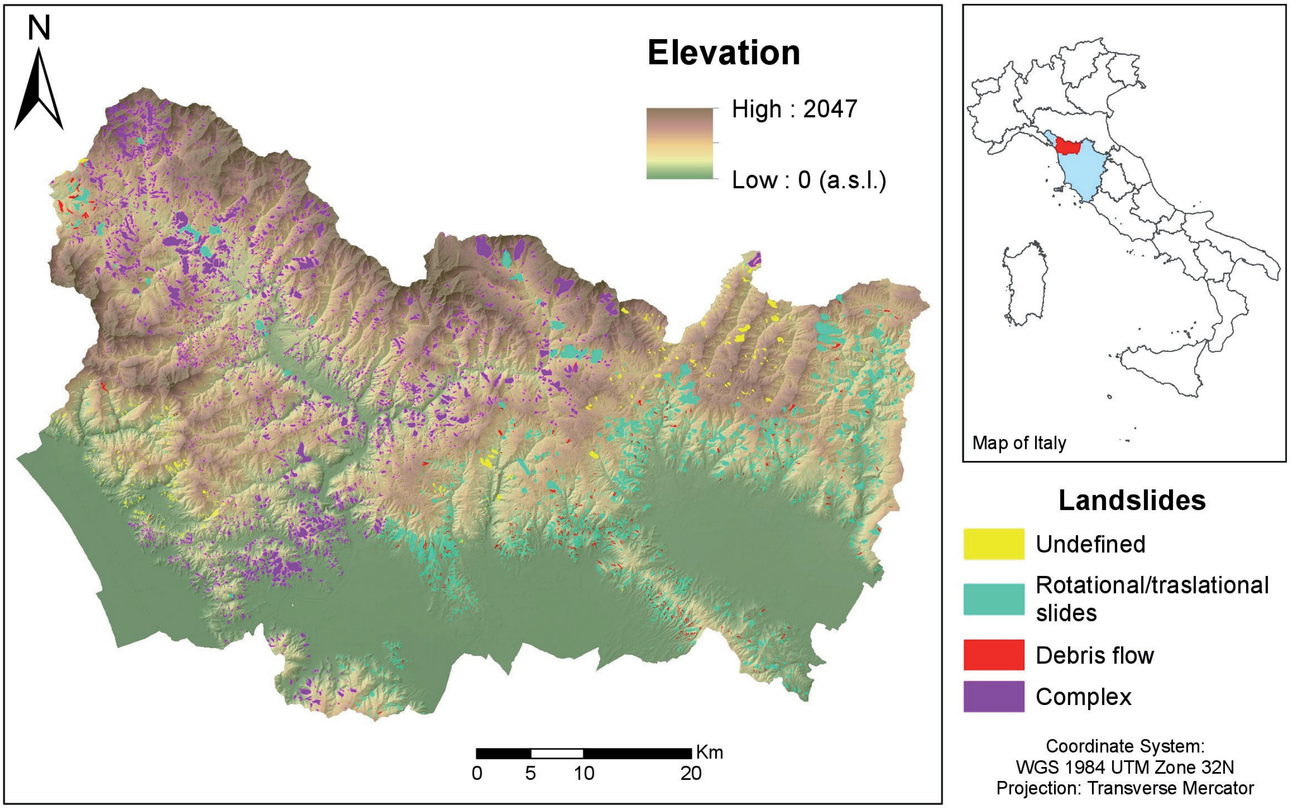

Fig. 1 - Elevation map and landslide distribution of the test site

In the test site the bedrock is mainly composed of metamorphic rocks (marls, phyllite and schists) in the western part and of layered pelitic flysch formations in the northern part, while in the central and southern parts the bedrock consists mainly of massive arenaceous flysch, covered by alluvial deposits. Further details on the geological characteristics of the study area are reported in the subsequent section describing the geological parameterization of the test site. The mountainside is mantled by a 1.5–5m thick layer of colluvial soil overlying the bedrock (Segoni *et alii* 2018), which exhibits a marked contrast in geotechnical properties with respect to the bedrock (Tofani *et alii*, 2017).

The test site is widely affected by landslides (Battistini *et alii,* 2017), where the main types can be classified as slides (rotational, translational, and compound), slow earth flows, complex movements (mainly slides evolving into flows), and, to a lesser extent, debris flows (Fig. 1) (Rosi *et alii*, 2018; Luti *et alii,* 2020).

From the climatic point of view, the region is characterized by a Mediterranean climate, with warm and dry summers and mild and wet winters. Rainfall distribution is characterized by two peaks over the year, the main one is in Autumn (November is the rainiest month) and the secondary one is in Springtime, while summer is the driest period of the year (Rapetti &Vittorini, 1994; Rosi *et alii* 2012), with sporadic and isolated rainstorm.

The mean annual rainfall ranges from 800 mm to 1500 mm (Maracchi *et alii*, 2005; Segoni *et alii*, 2014). Lower rainfall amounts are recorded in the plain areas of the southern part of the study area, while the mountainous northern part is usually characterized by a higher amount of rainfall. Being close to the sea and due to the presence of high relieves, the western part of the study area can be affected by orographic rainfalls, caused by cyclones moving westward from the sea, which can result in the triggering of landslides.

Random forest treebagger

The landslide susceptibility assessment was performed using the Random Forest algorithm (RF), a machine learning algorithm for non-parametric multivariate classification (Breiman, 2001), implemented in a Matlab software code (MathWorks, version R2020b) (Lagomarsino *et alii*, 2017). Random Forest relies on the construction of binary decision trees (or Bayesian trees) using the bootstrapping technique. Bootstrapping consists in a random sampling of a part of the variables, then those not included in the sampling, called "out-of-bag", will be added into the process iteratively, to obtain a set of randomly generated decision trees. This technique allows to reduce predictive errors. In addition, a subset of data is excluded from sampling and is used in the validation phase as an independent dataset (Breiman, 2001).

Random Forest is an effective and widely used algorithm in landslide susceptibility studies (Trigila *et alii,* 2013; Xiao *et alii*, 2019). Among the advantages of using the RF algorithm, there is the possibility of using numerical and categorical variables at the same time, without assumption on the statistical distribution of their values. Furthermore, Random Forest is acknowledged to be capable of handling implicitly the multicollinearity of variables, identifying the uninfluential (or the pejorative) ones (Breiman, 2001; Brenning, 2005).

The Random Forest implementation used in this study also automatically performs a validation by building a Receiver Operating Characteristic Curve (ROC Curve) and calculating the relative area under the curve (AUC). AUC is widely used as a quantitative indicator for the predictive effectiveness of susceptibility models: it can range from 0.5 (completely random predictions) to 1.0. The implementation also calculates the OOBE (out of bag error) for each variable. This parameter measures the relative error that would be committed if a given variable is excluded from the random forest classifier. OOBE can be used to assess the relative importance of each independent variable, thus representing a powerful tool to interpret the results and to rank the variables according to their importance (Catani *et alii*, 2013).

INPUT DATA

Landslide data

The spatial distribution of landslides was derived from IFFI (Inventory of Landslides in Italy), an open access national scale inventory that has been credited to have a high degree of completeness and homogeneity (Trigila *et alii,* 2010; Herrera *et alii*, 2018). In the study area, 7799 landslides are mapped as polygons at the 1:10,000 scale. They have an areal extension ranging from 102 to 106 m2 and they are classified mainly as complex movements (55%) and rotational/translational slides (35%) (Fig. 1). Since the complex movements of the area consist of rotational slides that evolve into earth flows, the triggering mechanism and the predisposing factors can be considered similar and both classes of landslides can be included in the same susceptibility assessment (Segoni *et alii*, 2016, 2020). The remaining 10% of landslides is composed by debris flows (4%) and unknown movement typology (6%) and was excluded from the susceptibility assessment as in these cases the relationship linking triggering process – predisposing factors would be too different with respect to the abovementioned landslide types, preventing a proper statistical calibration of the susceptibility model.

State-of-the-art explanatory variables

In the literature there is no consensus on the optimal number of variables to be used: this choice may be influenced by the landslide type, the physical features of the study area and the characteristics of the model. Since Random Forest is acknowledged to be capable of handling a large number of parameters and their multicollinearity (Brenning *et alii*, 2005; Catani *et alii*, 2013; Xiao *et alii*, 2020), 21 parameters were selected and prepared as raster with 100 m pixel size, since this mesoscale has been recognized as a good compromise in studies on landslide susceptibility on a regional scale (Catani *et alii,* 2013; Chacón *et alii*, 2006).

Fifteen "basic parameters" were selected among those most used in literature (Reichenbach *et alii,* 2018; Catani *et alii*, 2013) and among those that generated results with high predictive power in previous susceptibility assessments performed at this test site (Segoni *et alii,* 2016; 2018; 2020; Luti *et alii*, 2020). Some morphological and hydrological parameters were derived from a digital elevation model with 10 m cell size: elevation, roughness (defined as the standard deviation of elevation), slope gradient, standard deviation of slope gradient, aspect, total curvature, planar curvature, profile curvature, topographic wetness index (TWI), topographic position index (TPI), flow accumulation, stream power index (SPI). Land use/land cover was used as a variable as well, and it was derived by the most updated Corine Land Cover map (1:50.000 scale) available, referred to the year 2018. Another set of predisposing factors accounts for the climatological characteristics of the area and shows the attitude of the territory (expressed in terms of return time) to be affected by prolonged rainfalls or by exceptionally intense rainfalls. While the former are typically associated to landslide with a complex hydrological response (e.g. deep seated landslide in low permeability terrain), the latter are usually related to shallow landslides in terrains with a relatively higher permeability. Following the approach of Catani *et alii,* 2013, these characteristics were estimated by the parameters k_100_72 (representing the return time of a rainfall event with 72 hours duration and 100 mm total rainfall depth) and k_30_1 (representing the return time of a rainfall event of 30 mm in one hour). These parameters were already defined by Catani *et alii* (2013) with a procedure composed by a statistical analysis carried out in the 332 rain gauges of the regional network and by a kriging interpolation to the whole Tuscany territory. In addition, two artificial control variables were created by generating two rasters of random values ranging from 0 to 1 (thus simulating a variable expressed as a percentage) and from 0 to 5 (thus simulating a variable classified into 5 classes). The purpose of these control variables is to check the effectiveness and the significance of all the others, by identifying if some "weak" explanatory variable provides a contribution close to the one provided by a random field.

Newly proposed explanatory variables

Two innovative sets of variables were used in this study. The first one is the "soil sealing aggregation" (SSA), as it was named by Luti *et alii* (2020), who introduced it in a preliminary test for landslide susceptibility. Soil sealing is an extreme form of land take and is an anthropogenic process consisting in the destruction, removal or covering of soils by completely or partly impermeable artificial material (e.g. asphalt or concrete) (Prokop *et alii*, 2011). Italy is one of the European countries where the soil sealing process is more widespread, mainly because of the expansion of infrastructures and buildings (Munafò, 2019). ISPRA (Italian Institute for Environmental Protection and Research) monitors the spatio-temporal evolution of the soil sealing process and releases every year an updated national scale map. ISPRA soil sealing map consists of a high resolution (10m x 10m) raster conveying a binary information: pixels are classified either as "sealed soil" or as "non-sealed soil" (Munafò, 2019). To include this information in the susceptibility assessment, which is based on broader spatial units (100m x 100m pixels), we calculated the percentage of soil classified as sealed by the original ISPRA soil sealing map. Therefore, "soil sealing aggregation" values range from 0 (pixel composed by completely natural or semi-natural soil that fully retains his hydraulic and environmental properties) to 1 (pixel where the soil is completely artificialized). Intermediate values express different degrees of disturbance brought by artificial intervention to the soil, altering his hydrological features (Fig. 2a).

The second set of variables involves a multiparametric characterization of geological information. In the study area, a digital geological map produced by the Tuscany Region is available at the 1:10,000 scale (https://www.regione.toscana.it/-/banche-dati-cartografia-geologica), which maps 194 lithostratigraphic units. This very high thematic accuracy is impossible to be directly used for a statistical landslide susceptibility assessment at a regional scale, therefore the lithostratigraphic units should be grouped into broader classes with a specific geological meaning. The most used approach in the landslide susceptibility literature would be to group them into lithological units, as different lithologies usually have specific geotechnical properties and thus lithology could be considered a key feature in controlling the spatial distribution of landslides. Besides lithology (Fig. 2b), other parameterizations were created from the original lithostratigraphic map. The genetic parametrization considers the genetic process originating each formation and thus subdivides the areas into the following classes: clastic rocks, organogenic rocks, magmatic rocks, metamorphic rocks, soils (Fig. 2c). The structural parametrization accounts for the structural units in which the Apennine is traditionally subdivided according to its tectonic evolution (Fig. 2d). The paleogeographic parametrization is conceptually similar to the previous one, as it groups lithostratigraphic units according to the paleogeographic units they belong, thus accounting for different environment of deposition of sedimentary rocks and soils (Fig. 2e). The chronologic parameterization classifies the lithostratigraphic units according to their geological age (Fig. 2f).

MODEL APPLICATION

All explanatory variables were imported in a GIS system. For each landslide polygon, IFFI identifies a point located in the proximity of the scarp as the point of initiation of the instability. Those points were imported in the GIS system as well and were used to sample the environmental conditions associated to landslide triggering. To have a balanced prediction, an equal number of sample points was randomly generated outside the landslide polygons, to characterize the conditions associated with stable areas. The sample dataset contains 15598 points, evenly subdivided between landslide and no-landslide points. This dataset was randomly split in a model training dataset (70% of the points) and a model testing dataset (30% of the points), which both maintain the 50%-50% balance among landslide and no-landslide points.

The sampled data were used as input for the Random Forest algorithm, configured with 500 trees and 100 iterations to select the classification tree with the highest predictive effectiveness according to the internal validation procedure performed in terms of AUC. The best prediction tree was then applied to the entire set of data to have a full susceptibility map of the area. At the same time, the out of bag error of each variable was calculated to identify the variables with little or no importance in the model, in order to remove them and run the model with a reduced configuration.

In addition, for comparison, a model run was also performed with a configuration that does not encompass the newly proposed variables (soil sealing and multiparametric geological variables).

RESULTS

A preliminary test of the model was performed using all the variables defined: 15 variables coming from the "core set", 6 newly proposed variables and the two random variables. This test was not used to map the susceptibility and was useful to compare the variables with respect to the randomly defined ones. An analysis of the OOBE values revealed that the random variables can be considered pejorative as they have values close to zero, and that all the other variables have a much greater and positive contribution to the modeling effectiveness (from 0.4 to 2.6). As a consequence, to map landslide susceptibility, a "full configuration" run was performed by simply removing the random variables and using all of the "core" and "new" variables (totaling 21 variables). The resulting map is shown in Fig. 3 and the validation statistics provided an AUC value of 0.77. A subsequent procedure of pruning (progressive iteration of the classification

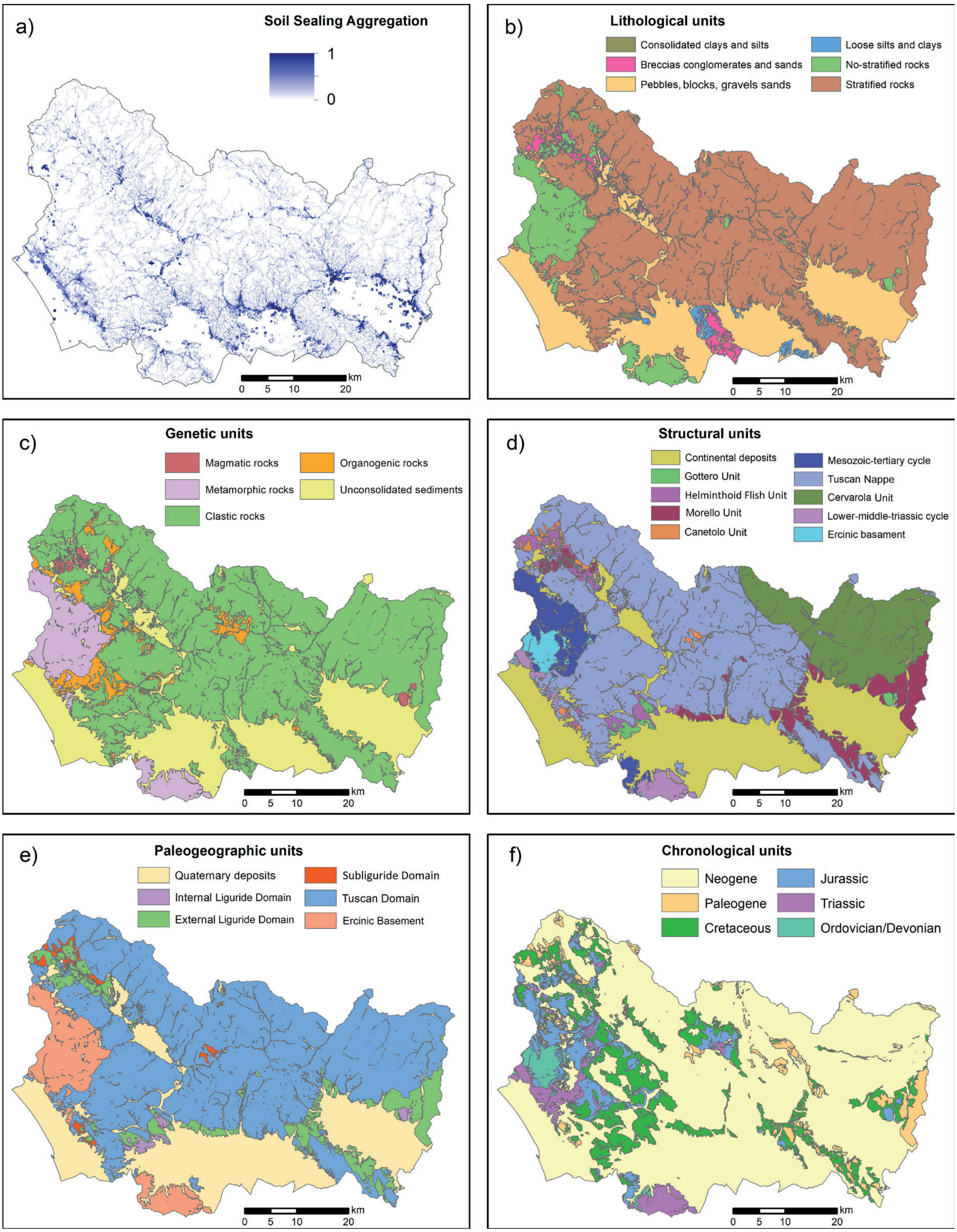

Fig. 2 – Newly proposed variables tested in this work: a) soil sealing aggregation; b) lithological units; c) genetic units; d) structural units; e) paleogeographic units; f) chronological units.

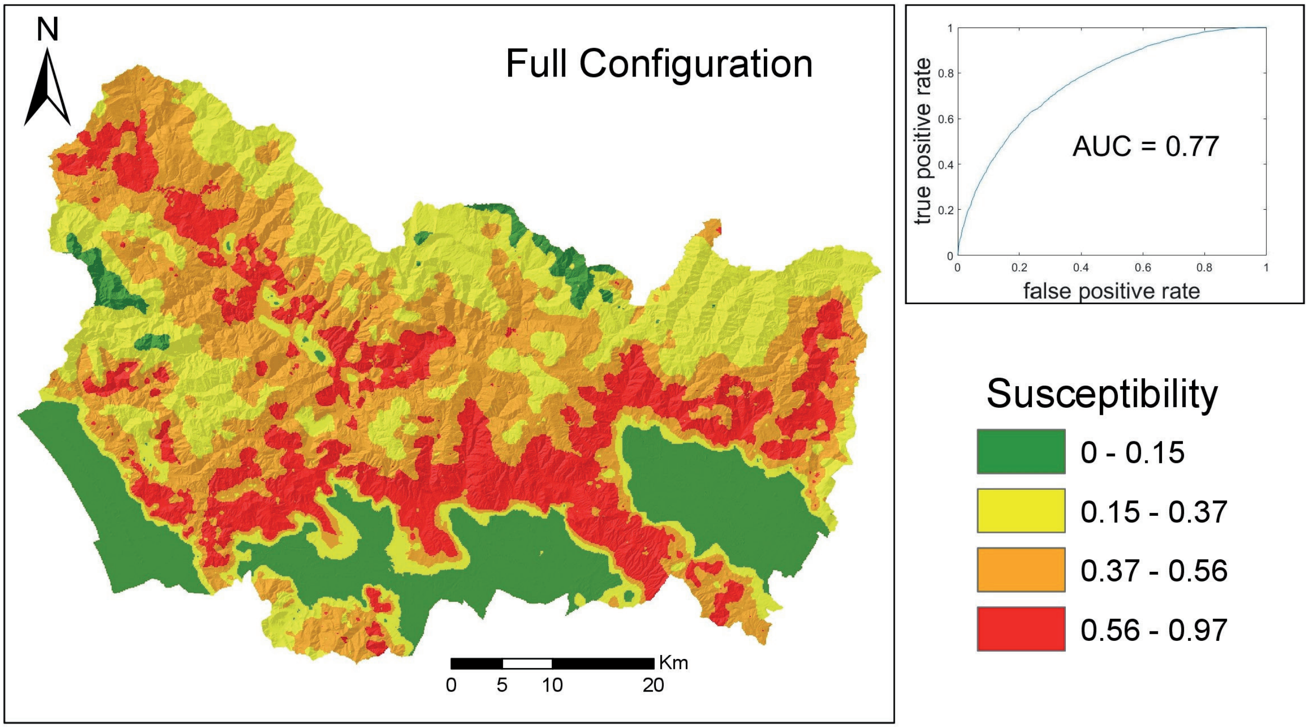

Fig. 3 – Landslide susceptibility map obtained with the full configuration of the model, encompassing soil sealing information and multiparametric characterization of geology.

removing the parameter identified as least important during the previous run) revealed that the full configuration is the one that produces the highest validation statistics. Indeed, the AUC value slowly decreases as the least important parameter is pruned out of the model configuration: e.g. AUC was 0.76 removing aspect and 0.75 removing aspect and roughness. In addition, a run was performed also with a "base" configuration, which does not include the newly proposed parameters (soil sealing aggregation and all the geological parameters other than lithology). The base configuration reported a lower effectiveness than the full configuration, with a 0.75 AUC: further insights on this difference will be discussed in the next section.

DISCUSSION

The out of bag error (Fig. 4) can be used as a proxy for the importance of each variable in the model, providing some insights on the landslide processes of the area. The most influential variables are elevation, return period of abundant rainfalls (K_100_72) and mean slope gradient. Concerning elevation and slope, this outcome is not surprising, as it is quite common to see in the literature these statistics ranked as the most important. Concerning rainfall, the use of a meteo-climatic parameter as a predisposing factor is not so common in landslide susceptibility studies and the second rank reported in this work stresses the importance of this modeling approach. Furthermore, the outcome that K_100_72 has a higher importance than K_30_1 allows to infer that the landslides typically affecting the study area are more sensitive to long and abundant rainfalls than to short and intense cloudbursts. This can be probably explained with the low-permeable residual soils mantling the bedrock (Tofani et al., 2017) and is especially true for relatively deep rotational slides. It should be stressed that debris flows (accounting for 4% of the landslides mapped in the area), which are more sensitive to short and intense rainfalls, were excluded from the susceptibility analysis. Two other very important variables (4th and 7th, but with very similar OOBE values) are land cover and soil sealing. Both of them account for the anthropogenic or natural processes that shape the territory and cover the soil. These two variables have complimentary pros and cons: Corine Land Cover has a very high thematic accuracy but a coarse spatial resolution (the minimum mapping unit is 25 hectares), while the soil sealing map produced by ISPRA has a high spatial resolution (10 m pixels) but a low thematic accuracy, reporting a dichotomous subdivision between sealed soil and (semi)natural soil. The results of this application provide a strong evidence of the potential of using parameters derived from soil sealing in landslide susceptibility assessments: while Corine land cover may be used to account for different land uses and soil covers, soil sealing provides a

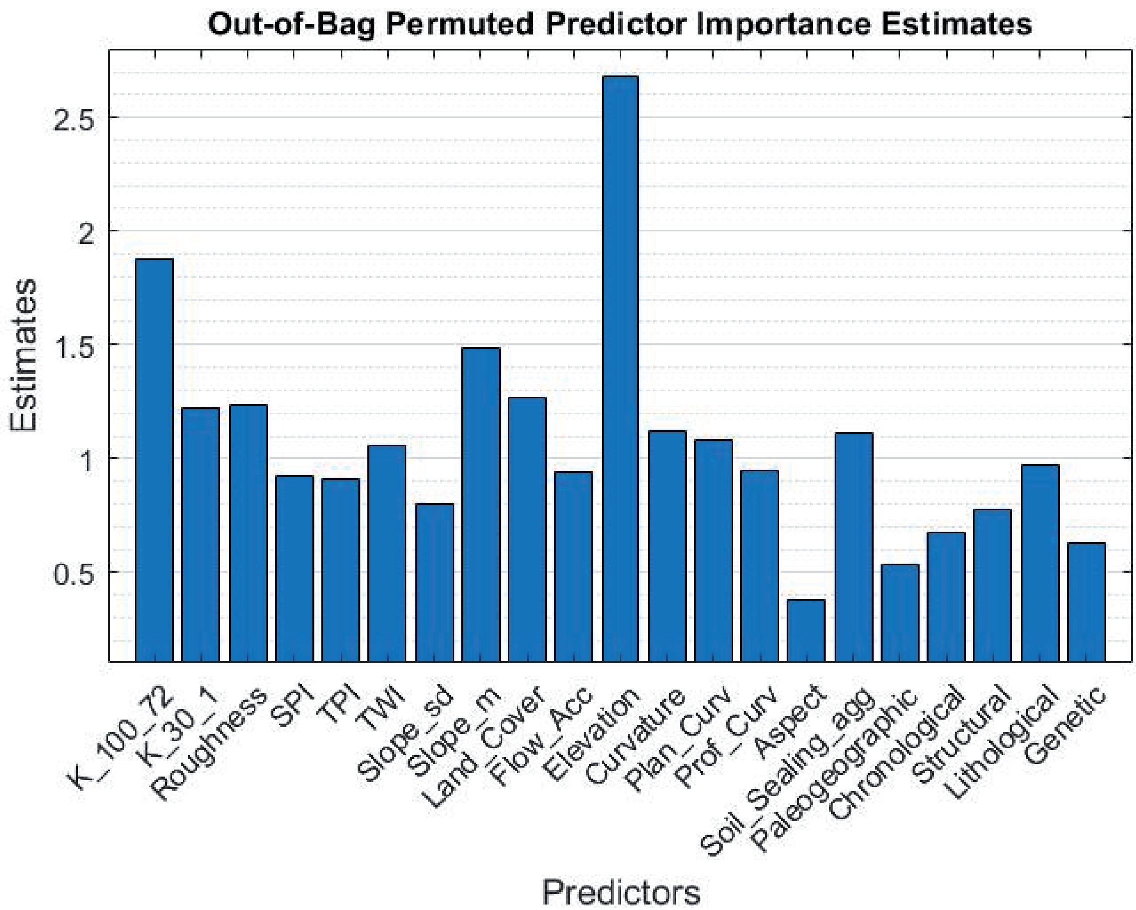

Fig. 4 – Explanatory variables used in the susceptibility assessment and their "out of bag error" (OOBE), which can be considered an indicator of the relative importance of each variable in contributing to the model results.

high resolution information on the human disturbance on the hillslope system, thus the two parameters can be used together and complete each other. This especially true when the urban texture is sparse and scattered across the territory, like in the study area, because small villages, small infrastructures and isolated buildings are usually neglected by CLC. This brings about the risk of underestimating the role played by human activity in landsliding, while on the contrary the influence of land use changes and urbanization in predisposing slope instability is widely acknowledged (Persichillo *et alii*, 2017; Mendes *et alii*, 2018; Martino *et alii*, 2019; Dikshit *et alii,* 2020): soil sealing maps can thus be regarded as an effective method to include the anthropic processes in landslide susceptibility studies.

In landslide susceptibility studies, geology is of course a very important element (Reichenbach *et alii*, 2018; Luti *et alii,* 2020). This is not in disagreement with the results reported in Fig.4, where the geological variables only apparently seem to have a relatively low importance. Indeed, it should be stressed that the geological information has been included in this susceptibility model with five different parameters and all of them share the weight of training the machine learning algorithm to understand how geology is connected with landslide distribution. The use of five different geological parameters allows the model to provide a more complete description of the complex interactions between geological features and predisposition to landsliding. Traditional studies only account for lithology, which is obviously connected with landsliding, but it is only a partial information provided by geological maps and accounts only for one of the geological features that influence the landslide distribution. As instance, the same lithology could be found in two different structural units, which may have undergone completely different tectonic histories and stresses, thus leading to two different settings and resistance to failure. Paleogeographic units, accounting for the different environment of sedimentation, could highlight differences in the chemical composition, in the texture of the parent material or in the stratigraphy, which may reflect in different predisposition to instability. Similarly, the chronological criterium may be useful to differentiate otherwise similar units that may differ for the age of deposition, in turn reflecting a different degree of weathering.

As a last test of the effectiveness of the parameterization used in this study, the result obtained with the full configuration were compared with those obtained by a base configuration that does not include the newly proposed parameters (soil sealing aggregation and all the geological parameters other than lithology). According to an approach recently introduced by Xiao *et alii* (2020), the difference is not expressed only in terms of AUC scores (the full configuration slightly outperforms the base configuration, as their AUC values are 0.77 and 0.75, respectively), but the susceptibility values are compared on a pixel-by-pixel basis. Using a GIS software, the susceptibility values of the base configuration were subtracted by the ones of the full configuration. The resulting map shows the spatial distribution of the difference of the results produced by the two parameterizations (Fig. 5). Negative values are systematically found in the North-East sector and in the South sector. Both circumstances find a possible explanation with the multiparametric characterization of geological information. If the lithology alone is considered as in the basic configuration, the North-East sector is not differentiated by the rest of the stratified rocks dominating the largest part of the test site, while e.g. the structural parameterization allows differentiating the Tuscan Nappe structural unit from the Cervarola Unit and the Morello Unit (Fig. 2). This finer detail, if inserted into the susceptibility model with the full configuration, leads to refine the susceptibility map resulting in different susceptibility values in large sectors of the Cervarola and Morello units (Fig. 5).

A similar explanation can be found for the hotspots of negative differences in the South sector (Fig. 5): they are in correspondence of metamorphic rocks that outcrop more

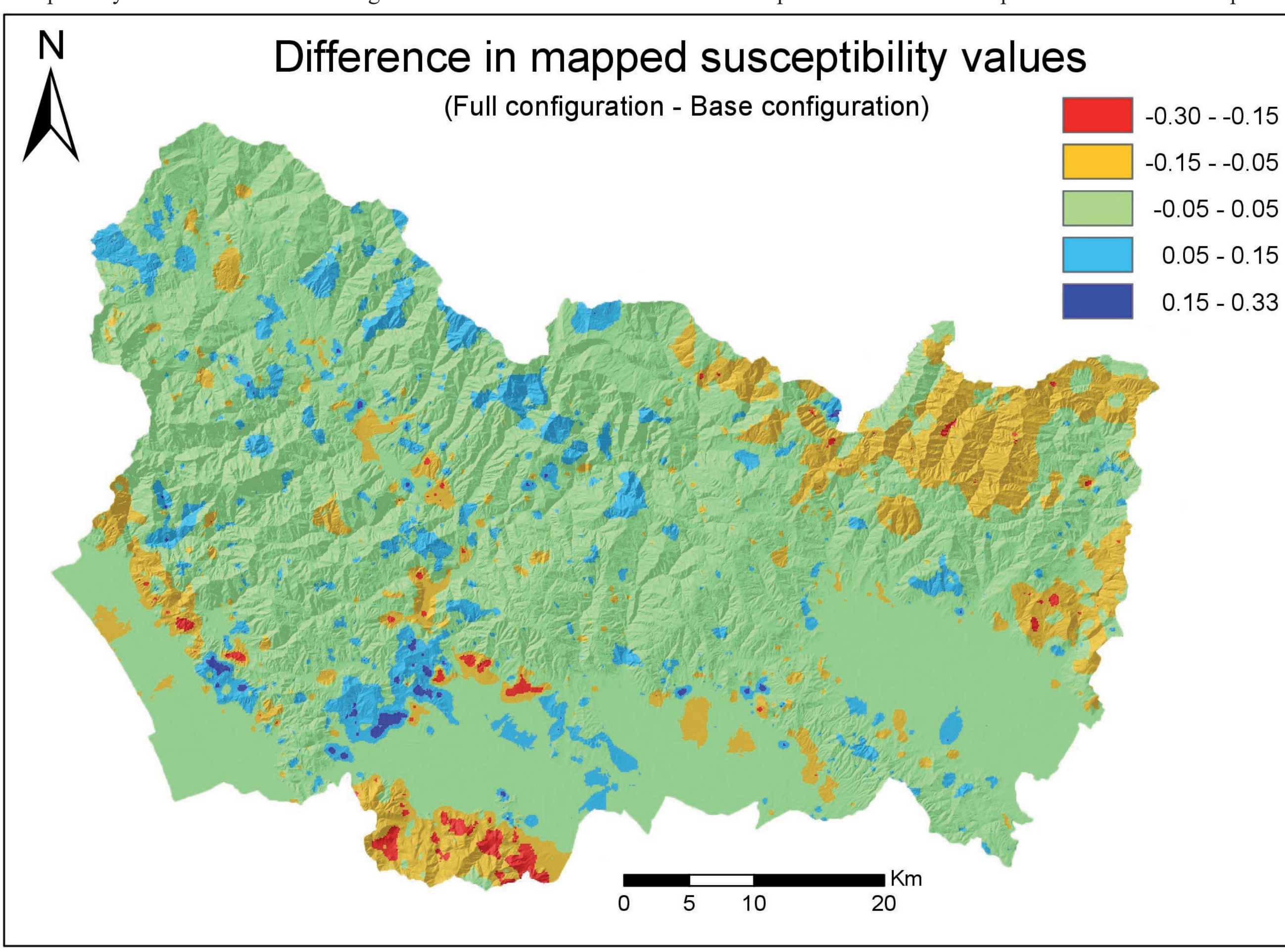

Fig. 5 – Map showing the difference in landslide susceptibility values calculated by the full configuration (21 parameters) and the basic configuration (15 parameters)

diffusely also in the West sector. The two sectors cannot be distinguished by lithology or genesis as they pertain to the same units (no-stratified rocks and metamorphic rocks). But if the structural units and the chronological units are taken into account, these two sectors can be considered as independent, because from a structural and chronological point of view the Ercinic basement is differentiated from the Mesozoic-Tertiary cycle and the Lower-middle-Triassic cycle. As a result, the multiparametric characterization of geology can better train the model to differentiate between the two areas. The positive differences in susceptibility value are less clustered, but a hotspot of high positive differences can be identified close to the South of the test area. Again, from a lithological point of view this site is not distinguished from the rest of the stratified rocks, but the structural and the paleogeographic parameterizations highlight some differences and specific features that are transferred to the full configuration susceptibility model. The higher AUC value resulting from the full configuration allows considering the resulting map as more reliable than the one derived from the base configuration; consequently, the multiparametric characterization of geology can be considered more reliable.

CONCLUSION

A Random Forest machine learning algorithm was applied in Northern Tuscany (provinces of Lucca, Pistoia and Prato) to map landslide susceptibility. Besides a core set of 15 state-of-the-art predisposing factors, 6 newly proposed variables were used as input for the susceptibility model. The first of the new variables (soil sealing aggregation) was derived from the soil sealing map at 10 m resolution released yearly by ISPRA and accounts for the degree of human disturbance in the hillslope system.

The other variables were derived from a detailed geological map of the Tuscany Region and serve as a multi-criteria geological characterization: the area was subdivided into lithological units, chronological units (according to their age), genetic units (according to the process that formed each unit), structural units and paleogeographic units.

Results show that the full configuration of the model (using both state-of-the-art and newly proposed parameters) produces a more accurate spatial prediction of landslide susceptibility in terms of AUC values respect to the base configuration (using state-of-the-art and lithology as explanatory variables).

In particular, the main differences in the mapped susceptibility values are located in portions of the area where the lithology is the same than in the surroundings, but other geological variables (mainly the ones derived by the structural, paleogeographic, or chronological criteria) highlight some differences in the bedrock composition.

This can be considered as a further proof that susceptibility models may benefit from a multi-criteria parameterization of geological information, rather that relying only on lithological classifications, as commonly occurs in susceptibility studies. Regarding soil sealing parameter, its effect on the resulting map is less evident at first sight because of its scattered spatial distribution, but the high out of bag error values reveals that its contribution to the modeling is highly beneficial and not redundant respect to other parameters describing land cover (like Corine Land Cover derived maps), thus highlighting its capability to link the anthropogenic processes of transformation of the territory with the landslide susceptibility.

These findings open new interesting perspectives for landslide susceptibility studies in Italy, as the new variables tested in this study can be easily derived anywhere else in the national territory: an updated soil sealing map is provided for free at the nation scale every year by ISPRA, while detailed geological maps are available almost everywhere in Italy, since many regions have a 1:10,000 mapping of their territory and at national level Italy is covered by 1:100,000 maps which are being updated to a 1:50,000 scale.

REFERENCES

Battistini A., Rosi A., Segoni S., Lagomarsino D., Catani F. & Casagli N. (2017) - *Validation of landslide hazard models using a semantic engine on online news.* Applied geography, **82**: 59-65.

Brabb E.E. (1984) - *Innovative approaches to landslide hazard mapping*. 1st edn. Proceedings 4th International Symposium on Landslides, Toronto,307–324.

Breiman L. (2001) - *Random forests*. Mach Learn, **45**: 5–32.

Brenning A. (2005) - *Spatial prediction models for landslide hazards: Review, comparison and evaluation.* Nat. Hazards Earth Syst. Sci., **5**: 853–862.

Canavesi V., Segoni S., Rosi A., Ting X. Nery T., Catani F., & Casagli N. (2020) - *Different Approaches to Use Morphometric Attributes in Landslide Susceptibility Mapping Based on Meso-Scale Spatial Units: A Case Study in Rio de Janeiro (Brazil).* Remote Sensing, **12** (11), 1826.

Cascini L. (2008) - *Applicability of landslide susceptibility and hazard zoning at different scales*. Eng. Geol., **102** (3-4):164–177.

Catani F., Lagomarsino D., Segoni S. & Tofani V. (2013) - *Landslide susceptibility estimation by random forests technique: sensitivity and scaling issues.* Natural Hazards and Earth System Sciences**,** **13** (11): 2815-2831.

Chacón J., Irigaray C., Fernández T. & El Hamdouni R. (2006) - *Engineering geology maps: Landslides and geographical information systems*. Bull. Eng. Geol. Environ., **65:** 341–411.

Chen L., Van Westen C.J., Hussin H., Ciurean R.L., Turkington T., Chavarro-Rincon D. & Shrestha, D.P. (2016) - *Integrating expert opinion with modelling for quantitative multi-hazard risk assessment in the Eastern Italian Alps.* Geomorphology, **273** (15):150–167.

Chen Z., Ye F., Fu W., Ke Y. & Hong H. (2020) - *The influence of DEM spatial resolution on landslide susceptibility mapping in the Baxie River basin, NW China*. Natural Hazards, 1-25.

Corominas J., Copons R., Vilaplana J.M., Altimir J. & Amigó J. (2003) - *Integrated landslide susceptibility analysis and hazard assessment in the principality of Andorra.* Nat. Hazards, **30** (3): 421–435.

Di Napoli M., Carotenuto F., Cevasco A., Confuorto P., Di Martire D., Firpo M., ... & Calcaterra D. (2020) - *Machine learning ensemble modelling as a tool to improve landslide susceptibility mapping reliability.* Landslides, **17** (8): 1897-1914.

Di Napoli M., Di Martire D., Bausilio G., Calcaterra D., Confuorto P., Firpo M., ... & Cevasco A. (2021) - *Rainfall-Induced Shallow Landslide Detachment, Transit and Runout Susceptibility Mapping by Integrating Machine Learning Techniques and GIS-Based Approaches.* Water, **13** (4): 488.

Dikshit A., Sarkar R., Pradhan B., Segoni S. & Alamri A. M. (2020) - *Rainfall induced landslide studies in Indian Himalayan region: A critical review.* Appl. Sci., **10**, 2466.

Erener A.R.Z.U. & Düzgün H.S.B. (2012) - *Landslide susceptibility assessment: what are the effects of mapping unit and mapping method.* Environmental Earth Sciences, **66** (3): 859-877.

Frattini P., Crosta G. & Carrara A. (2010) - *Techniques for evaluating the performance of landslide susceptibility models*. Eng. Geol., **111** (1-4): 62–72.

Herrera G., Mateos R.M., García-Davalillo J.C. et al. (2018) - *Landslide databases in the Geological Surveys of Europe.* Landslides **15**: 359–379.

Kalantar B., Pradhan B., Naghibi S.A., Motevalli A. & Mansor S. (2018) - *Assessment of the effects of training data selection on the landslide susceptibility mapping: a comparison between support vector machine (SVM), logistic regression (LR) and artificial neural networks (ANN).* Geomatics, Natural Hazards and Risk, **9** (1): 49-69.

Kavzoglu T., Colkesen I. & Sahin E.K. (2019) - *Machine learning techniques in landslide susceptibility mapping: A survey and a case study.* Landslides: Theory, practice and modelling, 283-301.

Lagomarsino D., Tofani V., Segoni S., Catani F. & Casagli N. (2017) - *A tool for classification and regression using random forest methodology: applications to landslide susceptibility mapping and soil thickness modeling*. Environmental Modeling & Assessment, **22** (3): 201-214.

Li Y. & Chen W. (2020) - *Landslide susceptibility evaluation using hybrid integration of evidential belief function and machine learning techniques.* Water, **12** (1): 113.

Luti T., Segoni S., Catani F., Munafò M. & Casagli N. (2020) - *Integration of Remotely Sensed Soil Sealing Data in Landslide Susceptibility Mapping.* Remote Sensing, **12** (9): 1486.

Maracchi G., Genesio L., Magno R., Ferrari R., Crisci A. & Bottai L. (2005) - *The diagrams of the climate in Tuscany.* CNR-Ibimet, LAMMA-CRES, Florence (In Italian).

Martino S., Bozzano F., Caporossi P., D'angiò D., Della Seta M., Esposito C., Fantini A., Fiorucci M., Giannini L.M., Iannucci R., et al. (2019) - *Impact of landslides on transportation routes during the 2016–2017 central Italy seismic sequence.* Landslides, **16**: 1221–1241.

Mendes R.M., De Andrade M.R.M., Tomasella J., De Moares M.A.E. & Scofield G.B. (2018) - *Understanding shallow landslides in Campos do Jordão municipality-Brazil: Disentangling the anthropic effects from natural causes in the disaster of 2000.* Nat. Hazards Earth Syst. Sci., **18**: 15–30.

Munafò M. (2019) - *Consumo di Suolo, Dinamiche Territoriali e Servizi Ecosistemici*. SNPA: Rome, Italy: 224 pp.

Persichillo M.G., Bordoni M., & Meisina C. (2017) - *The role of land use changes in the distribution of shallow landslides.* Science of the total environment, **574**: 924-937.

Prokop G., Jobstmann H. & Schobauer A. (2011) - *Overview on Best Practices for Limiting Soil Sealing and Mitigating Its Effects in EU-27*. European Communities: Brussels, Belgium.

Reichenbach P., Rossi M., Malamud B.D., Mihir M. & Guzzetti F. (2018) - *A review of statistically-based landslide susceptibility models.* Earth Sci. Rev., **180:** 60–91.

Rapetti F., Vittorini S., (1994) - *Rainfall in Tuscany: Observation about extreme events*. Riv. Geogr. Ital., **101**: 47–76.

Rosi A., Segoni S., Catani F. & Casagli N. (2012) - *Statistical and environmental analyses for the definition of a regional rainfall threshold system for landslide triggering in Tuscany (Italy)*. J.Geogr.Sci., **22** (4): 617–629.

Rosi A., Tofani V., Tanteri L., Stefanelli C.T., Agostini A., Catani F. & Casagli N. (2018) - *The new landslide inventory of Tuscany (Italy) updated with PS-InSAR: geomorphological features and landslide distribution*. Landslides, **15** (1): 5-19.

Segoni S., Pappafico G., Luti T., & Catani F. (2020) - *Landslide susceptibility assessment in complex geological settings: Sensitivity to geological information and insights on its parameterization.* Landslides, **17**: 2443–2453.

Segoni S., Tofani V., Lagomarsino D. & Moretti S. (2016) - *Landslide susceptibility of the Prato-Pistoia-Lucca provinces,* Tuscany, Italy. J. Maps., **12**: 401–406.

Segoni S., Tofani V., Rosi A., Catani F. & Casagli N. (2018) - *Combination of rainfall thresholds and susceptibility maps for dynamic landslide hazard* *assessment at regional scale*. Front Earth Sci., **6**: 85

Thai Pham B., Shirzadi A., Shahabi H., Omidvar E., Singh S. K., Sahana M., ... & Lee S. (2019) - *Landslide susceptibility assessment by novel hybrid machine learning algorithms*. Sustainability, **11** (16): 4386.

Tien Bui D., Shirzadi A., Shahabi H., Geertsema M., Omidvar E., Clague J. J., ... & Lee S. (2019) - *New ensemble models for shallow landslide susceptibility modeling in a semi-arid watershed.* Forests, **10** (9): 743.

Tiranti D., Nicolò G. & Gaeta A.R. (2019) - *Shallow landslides predisposing and triggering factors in developing a regional early warning system.* Landslides **16** (2): 235–251.

Tofani V., Bicocchi G., Rossi G., Segoni S., D'ambrosio M., Casagli N. & Catani F. (2017) - *Soil characterization for shallow landslides modeling: a case study in the Northern Apennines (Central Italy)*. Landslides, **14** (2): 755-770.

Trigila A., Iadanza C. & Spizzichino D. (2010) - *Quality assessment of the Italian landslide inventory using GIS processing.* Landslides 7: 455–470.

Trigila A., Frattini P., Casagli N., Catani F., Crosta G., Esposito C., ... & Lari S. (2013) - *Landslide susceptibility mapping at national scale: the Italian case study.* In Landslide science and practice (287-295 pp). Springer, Berlin, Heidelberg.

Vai G.B. & Martini I.P. (2001) - *Anatomy of an Orogen: the Apennines and adjacent Mediterranean basins.* Kluwer Acad. Publ, Dordrecht 633 pp.

Xiao T., Segoni S., Chen L., Yin K. & Casagli N. (2020) - *A step beyond landslide susceptibility maps: A simple method to investigate and explain the different outcomes obtained by different approaches*. Landslides, **17** (3): 627-640.

Xiao T., Yin K., Yao T. & Liu S. (2019) - *Spatial prediction of landslide susceptibility using GIS-based statistical and machine learning models in Wanzhou County, Three Gorges Reservoir, China*. Acta Geochim **38**: 654–669.

Yilmaz I. (2009) - *Landslide susceptibility mapping using frequency ratio, logistic regression, artificial neural networks and their comparison: a case study from Kat landslides (Tokat—Turkey).* Computers & Geosciences, **35** (6): 1125-1138.

Zizioli D., Meisina C., Valentino R. & Montrasio L. (2013) - *Comparison between different approaches to modeling shallow landslide susceptibility: a case history in Oltrepo Pavese, Northern Italy*. Natural Hazards and Earth System Sciences, **13** (3): 559-573.

Received February 2021 - Accepted May 2021