A Comparison Between Heuristic, Statistical, And Data-Driven Methods In Landslide Susceptibility Assessment: An Application To The Briga And Giampilieri Catchments

Antonio Francipane
Elisa Arnone
Francesco Lo Conti
Claudio Puglisi
Leonardo Valerio Noto

Follow this and additional works at: http://academicworks.cuny.edu/cc_conf_hic

Part of the Water Resource Management Commons

Recommended Citation
Francipane, Antonio; Arnone, Elisa; Lo Conti, Francesco; Puglisi, Claudio; and Noto, Leonardo Valerio, "A Comparison Between Heuristic, Statistical, And Data-Driven Methods In Landslide Susceptibility Assessment: An Application To The Briga And Giampilieri Catchments" (2014). CUNY Academic Works.
http://academicworks.cuny.edu/cc_conf_hic/150

This Presentation is brought to you for free and open access by CUNY Academic Works. It has been accepted for inclusion in International Conference on Hydroinformatics by an authorized administrator of CUNY Academic Works. For more information, please contact AcademicWorks@cuny.edu.
A COMPARISON BETWEEN HEURISTIC, STATISTICAL AND DATA-DRIVEN METHODS IN LANDSLIDE SUSCEPTIBILITY ASSESSMENT: AN APPLICATION TO THE BRIGA AND GIAMPIELIERI CATCHMENTS

ANTONIO FRANCIPANE(1), ELISA ARNONE(1), FRANCESCO LO CONTI(1), CLAUDIO PUGLISI(2), LEONARDO V. NOTO(1), ANTONINO SCARBACI(1)

(1) Dipartimento di Ingegneria Civile, Ambientale, Aerospaziale, dei Materiali, Università degli Studi di Palermo, Viale delle Scienze, Palermo, Italy.
(2) Dipartimento di Prevenzione Rischi Naturali e Mitigazione Effetti, ENEA, Roma, Italy.

Susceptibility assessment concerning the estimation of areas prone to landslide is one of the most useful approach in the analysis of landslide hazard. Over the last years, in an attempt to find the best approach to evaluate landslide susceptibility, many methods have been developed. Among these, the heuristic, the statistical, and the data-driven approaches are very widespread, and they all are based on the concept that the conditions which led to landslide movements in the past will control the probability of movement occurrence in the future. This study presents an assessment of landslide susceptibility in which models of the three different methodologies, such as the heuristic approach, the logistic regression, which belongs to the generalized linear models, and the artificial neural networks are used along with GIS spatial analysis techniques. We compare the results by applying the three different approaches to evaluate the debris-mud flows susceptibility to Briga and Giampilieri basins, two catchments of the city area of Messina (Sicily) where a considerable number of historical events were documented. The evaluation is carried out by comparing the AUC curves resulting from the application of the three approaches.

INTRODUCTION

The different combination of geological, morphological, climatic, and anthropic factors leads to a wide variety of hydrogeological instabilities, which differ by typology, evolution, and size of the involved area. Among these phenomena, landslides can have dimensions more considerable than others and be the cause of severe economic and social damages. It is for this reason that nowadays landslides study and prevention are among the most important problems to be dealt with in matter of territorial management.

In order to have a more efficient forecast of landslide events and, consequentially, a territorial management able to mitigate the effects of these phenomena, the risk assessment has become a fundamental tool to support the decision making process. One of the most reliable methods to identify landslide-prone areas consists in assessing their susceptibility (Hansen [1]), i.e. the probability of landslide occurrence.
Over the years, many methods have been developed as tools to assess landslide susceptibility, which can be classified as heuristic, statistic and deterministic. The deterministic methods are based on equations which simulate the physical processes of cause-effect and are generally used for small scale applications. Heuristic and statistical methods are instead based on the concept that “the past and the present are keys to the future” (Varnes and IAEG [2]) and that future landslides will be due to the same factors that caused landslides in the past. For Heuristic and statistical methods, estimation of landslide susceptibility therefore results into a typical spatial correlation analysis between the inducing factors and the occurrence or not of landslides and leads to the production of thematic maps as ultimate target.

Heuristic approach is based on opinion of geomorphologic experts. Generally this approach is divided into two phases: a direct mapping analysis, in which the geomorphologists determine the susceptibility in the field directly on the base of their experience, and a qualitative map combination, in which the experts use their knowledge to determine the weighting value for each class parameter in each parameter (Bartolomei et al. [3], Puglisi et al. [4], Falconi et al. [5]).

Among the statistical methods, the generalized linear models are well suited to analyze a presence-absence dependent variable (Lee et al. [6], Lee and Pradhan [7], Arnone et al. [8]) thus representing one of the most applied methods in the field of landslide susceptibility, with particular regard to the Logistic Regression (LR) model. Recently, a number of studies have proposed the use of Artificial Neural Networks (ANNs) models, as possible tool to assess the landslide susceptibility (Lee et al. [6], Ermini et al. [9], Arnone et al. [10]), given their suitability in analyzing spatial correlation. ANNs belong to the data driven methods although sometimes they are classified under the statistical methods.

In this study we assess the performances of three different landslide susceptibility methods: the logistic regression (statistical), the ANN (data driven), and a heuristic method developed by the Natural Risks Prevention and Effect Mitigation (UTPRA-PREV) department of the Italian National Agency for new technologies, Energy, and sustainable Economic development (ENEA). The models are separately applied on two Sicilian basins, where a number of historical landslide events, more than 2000, have been documented from 2000 to 2009. Suitability of models and their comparison are assessed by means of the ROC (Receiving Operating Characteristic) curve and the area under the ROC curve (AUC), whose value is a measure of goodness of model fitting. Results from comparison provide an important indication in choosing the proper method for future analyses.

BASINS DESCRIPTION

The Briga and Giampilieri basins

Climate, hydrology, digital elevation model (DEM), and landuse data were collected in the Briga and Giampilieri catchments, which are located within the Messina district in northeastern Sicily, Italy (38° 11’ N, 15° 34’ E). Both the catchments are approximately 10 km² in size and present a rugged morphology with mountains up to about 1,000 meters high above the sea level, narrow valleys, and very steep hillslopes (Figure 1). The vegetation is diversified and mainly dominated by crops and forests.

The climate of the two catchments is typical of Mediterranean area. Mean annual precipitation ranges between 882 mm in the coastal regions and 1,149 mm in the mountain region. The mean annual temperature is 18 °C with a monthly mean maximum and minimum temperature equal to 30 °C in July and 4.5 °C in February, respectively. Runoff regime of catchments rivers is ephemeral, as many other rivers in the northeastern part of Sicily, with low-flow or null discharges during the dry season and high-flow discharges during the fall and the winter.
METHODOLOGIES AND MODELS APPLICATION

Identification of landslide inventory and landslide inducing factors
Historical landslide events and landslide inducing factors represent the main required data for a landslide susceptibility analysis.

The landslide inventory map for the Giampilieri and Briga basins was realized by the UTPRA-PREV through a detailed geomorphological and morphometric field survey and an aerial photos analysis; this study led to identify and record in a GIS database more than 1000 debris flows. In particular, all the censused phenomena were classified as debris and mud flows and each event was characterized with specific morphological elements such as the Landslides Identification Point (PIFF), trigger areas, transport areas and the Landslide Foot Identification Point (PIP). Most of the landslides were located in the eastern part of the study area, including specifically the lower-middle portions of the Giampilieri and Briga catchments (Figure 1). Landslides were categorized, according to the morphological characteristics of trigger areas, in curved, rectangular, lobed (lobed-curved, rectangular-lobed, lobed-mixed), and punctual. The source areas were divided into "channeled", when bundled into a pre-existing drainage line, and "not channeled". In order to go deep into landslides details and validate the preliminary inventory of the phenomena, a field survey was necessary. A survey form, specifically developed for this type of phenomena, was used to gather information about a total of 124 landslides. Apart from morphological and morphometric elements, the form contained information about all of the discriminating parameters and the inducing factors. Moreover, some parameters were detected from aerial photographs and a very detailed 2 m resolution DEM.

Through a statistical analysis of the landslides inventory the more significant landslide inducing factors (i.e., geological, morphological, morphometric, and anthropic conditions that contribute to determine the landslide susceptibility of a given area) were identified (Table 1).

ENEAG Heuristic Method - EHM
The ENEA Heuristic Method (EHM) allows one to make a heuristic-statistical elaboration on landslide risk with the aim of obtaining reliable results as a function of potential landslides
areas, possible areas of transit and/or accumulation of material moved by the landslide, and modelling of energy dissipation (Puglisi et al. [4], Falconi et al. [11], Puglisi et al. [12]). For the sake of brevity, just the part of the methodology pertinent to the evaluation of landslide susceptibility is here described.

Once the landslide inventory and the inducing factors are identified, the susceptibility evaluation with the EHM requires the identification of discriminating parameters (i.e., geological and morphometric parameters defined as necessary conditions, but not sufficient, so that a portion of territory is susceptible to failure). Through a statistical analysis of the landslides inventory, an index and a weight are assigned to each landslide inducing factor on the base of its contribution to the instability. An opportune function of susceptibility implements the indexes and the weights of all the factors and extracts a map of landslide susceptibility, $S$, through the following relationship:

$$ S = (I_{cop} \times I_{pend}) \times \sum_{n} \left( i_{n} \times P_{n} \right) / \sum_{n} P_{n}, $$

(1)

where $I_{cop}$ and $I_{pend}$ are the indexes of the discriminating parameters of coverage and slope, respectively, $i_{n}$ and $P_{n}$ are the index and the weight of the $n$-th inducing parameter, respectively.

In order to quantify the influence on the susceptibility assessment with respect to the others, a weight from 0 to 5 and an index from 0 to 9 were assigned to each discriminating parameter and predisposing factor, respectively. Discriminating parameters and predisposing factors, implemented within a GIS framework, were used to produce Homogeneous Territorial Units (HTU) and then draw a susceptibility map through a map algebra analysis for the considered basins by means of Eq. (1).

**Logistic Regression Model- LRM**

The Logistic Regression Model (LRM) is a multivariate method that allows one to correlate the occurrence, or the no-occurrence, of an event (e.g., a landslide) with some continuous (e.g., slope, distance from the street, etc.), polychotomous or categorical (e.g., land use, soil type, geology, etc.) variables (Hosmer et al. [13]). Among the multivariate approaches, the LRM is the one that best fits the case in which the dependent variable is a dichotomous variable. As in linear regression, given a sample of $(X, Y)$ pairs, the goal is to estimate the regression coefficients in a model. In susceptibility analysis the dependent variable ($Y$) depends on landslides occurrence, coded as 0 (no landslide) or 1 (landslide), while $X$ is the vector of all the landslide-inducing factors, which can be numerical or categorical. The conditional probability that a landslide occurs, i.e. $P[Y = 1 | X]$ is given by the following:

$$ P[Y = 1 | X_{j}] = \frac{1}{1 + e^{-(\beta_{0} + \beta_{1}X_{1} + ... + \beta_{p}X_{p})}} = \frac{1}{1 + e^{-z}}, $$

(2)

where $\beta_{1}, \beta_{2}, ..., \beta_{p}$ are the coefficients of variables $X_{1}, X_{2}, ..., X_{p}$, and represent the different weight of each landslide inducing factor.

Among the selected landslide inducing factors, the choice of the most significant variables to take into account in the LRM was made with the stepwise method, which either includes or excludes a variable on the basis of the increase in goodness of fit introduced by different variables. For the choice of the most successful parsimonious model the Akaike Information...
Criterion (AIC) (Akaike [14]) was used. The lower value of AIC indicates the best model. In this analysis, 13 steps, corresponding to the total number of inducing factors (Table 1) were performed. Free software R, here used, determines a coefficient for each continuous variable and a number of coefficients equal to the number of the classes minus one (class assumed as class of reference) for each categorical variable. Following the AIC, the optimal model was obtained at step 6 and contains, in order, the following variables: land use (coeff\textsubscript{landuse}), mean annual precipitation (MAP), slope (slope), pedology (coeff\textsubscript{pedology}), parameter a of ddf curve (a), and distance from river network (net\_dist). In this case, the variable \(z\) of the chosen LRM is:

\[
z = -166.80 + \text{coeff}_{\text{landuse}} + (-0.00664 \cdot \text{MAP}) + (0.053370 \cdot \text{slope}) + ... \\
+ \text{coeff}_{\text{pedology}} \cdot y + (4.217 \cdot a) + (0.002634 \cdot \text{net\_dist}). \tag{3}
\]

The value of \(z\) estimated with Eq. (3) is used inside the Eq. (2) to determine the susceptibility map within a GIS environment.

**Artificial Neural Network - ANN**

The feed-forward MultiLayer Perceptron (MLP) network is one of the most suitable and adopted Artificial Neural Networks (ANNs) for landslide susceptibility applications.

In an MLP, the units, named perceptrons or neurons, are organized in layers and connected by weighted links. The input layer has a number of neurons equal to the number of input variables; in the output layer the number of neurons is equal to the number of output variables; between the two layers there are one or more so called hidden layers, whose number of neurons varies depending on the network complexity. The working mechanism is the following: the input signals are propagated forward through the network while neurons of the hidden layers make a linear combination of input signals and convert it through a generally nonlinear function (activation function). The network learns the dynamics of the studied phenomenon through a training procedure, in which a set of known input-output couples are fed to the network and the weights are updated with the aim to minimize the difference between the output and the target vectors, through minimization of a cost function \(E\).

In order to develop a successful MLP network, a number of phases need to be carefully defined: network design (input, hidden and output layer), data selection for training phase, training phase (choosing activation and transfer functions), classification phase.

Structure of input vector depends on the methodology used to represent the triggering factors (Arnone et al. [10]), which can considerably increase the number of computational nodes but provide an efficient objective approach, or keep a low number of nodes but introduce a rate of subjectivity. Given the considerable amount of data, in this study we adopted the latter approach, which assigns one neuron to each input variable and requires a reclassification of the categorical factors into numerical values, in order of importance for the landslide susceptibility analysis. In order to limit the subjectivity, the weights of each class were estimated based on the frequency ratio method (Carrara [15]). The used ANN algorithm will then normalize all input variables in the range 0-1. Characteristics of network design are shown in Table 2. High flexibility to the network was given by choosing an elevated number of nodes in the hidden layer. Selection of proper dataset for training phase is far from being obvious, as discussed in Arnone et al. [10]. In this study, we randomly selected the 50% of cells experiencing landslides (landslides) and a number equal to its double for those not experiencing landslides (no-landslides) (Arnone et al. [10]); these details are shown in Table 2 together with the adopted functions. Analysis was carried out within the software for numerical computing Matlab.
by using the implemented function “patternnet”. Once all these phases are ultimate, all the inducing factors are fed into the designed MLP network. The network returns the susceptibility value at each cell grid on the basis of the weights found during the training phase. For each cell, the relative position in the grid structure is recorded and used to reconstruct the susceptibility map within a GIS framework.

Table 2. ANN characteristics for network design, training phase and chosen functions.

<table>
<thead>
<tr>
<th>ANN characteristics</th>
<th># neurons input layer</th>
<th># neurons hidden layer</th>
<th># neurons output layer</th>
<th>landslide pixel for training phase</th>
<th>NO-landslide pixel for training phase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Network design</td>
<td>13</td>
<td>80</td>
<td>1</td>
<td>23872 (50%)</td>
<td>47744 (1)</td>
</tr>
<tr>
<td>Training phase</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Functions</td>
<td>Transfer function</td>
<td>'logsig'</td>
<td>Training function</td>
<td>'traingdm'</td>
<td></td>
</tr>
</tbody>
</table>

RESULTS AND COMPARISON

Results of models application and the corresponding susceptibility maps are reported in figure 2. The LRM and the ANN return a distribution of probability of landslides occurrences with values ranging from 0 to 1, while the EHM returns a classification into 10 classes of landslide susceptibility (from 0 to 9). In order to obtain the final susceptibility maps in levels of risks and to make their comparison easier, each map was classified into five nominal levels of susceptibility (very low, low, medium, high, very high) (Figure 2a, b, c).

All the models are able to classify as high susceptible the south scarp areas of the basin along the two main channels, where many of the historical events were observed. The rest of the basin is mainly classified as low and very low susceptible by the LR model, with small regions at medium susceptibility (Figure 2b). Instead, the EHM and the ANN are capable to identify areas classified at various level of risk (Figure 2a and 2c, respectively); both the maps show similar patterns which mainly reproduce the spatial distribution of the steepest areas (not shown here), even though often assign a different risk class to the same areas. Particularly, in both the maps very low and low susceptibility areas are located along the main stream paths and in the coastal area. Moreover, the ANN classifies the upper part of the basin as very low and low susceptibility areas, whereas the EHM mainly classifies this area with a higher class of risk (low and medium). Both the methods classify the central part of the basin as medium susceptibility areas with some areas classified as high and very high susceptibility. The ANN provides with the most severe scenario, with a clear defined region at very high susceptibility class in south central part of the basin. These areas mostly correspond to the abandoned shelves (not shown here). Finally, most of the historical events falls within the very high class of susceptibility and the model is able to capture the areas that most likely are prone to induce landslide based on the considered factors. To be more precise, the ANN identifies about the 10% and the 2.5% of the domain as high and very high susceptibility, respectively, against almost the 1.5% and the 0.43% for the LRM and almost the 13.5% and the 0.15% for the EHM, respectively.

A quantitative evaluation of the three models performances is made, as previously said, through the AUC method, i.e., Area Under the ROC (Receiving Operating Characteristic) Curve. This is built plotting the sensitivity (i.e., percentage of landslide cells that are correctly
identified as experiencing the event) versus 1-specificity (i.e., percentage of no-landslide cells that are correctly identified as not experiencing the event) over all possible cutoff. The AUC ranges from 0 to 1 and gives a measure of the model's ability to discriminate between the elements experiencing the outcome of interest versus those which do not.

Figure 2d shows the comparison of the three ROC curves. All the method are able to provide at least a "satisfactory" description of reality, with values higher than 0.7: in particular, the best performance is given by the LRM (0.89) followed by the ANN (0.85). The EHM application provided with the 'lowest' capability to distinguish between areas that experienced landslides and areas that did not (0.78), but still with satisfactory results. In this case the AUC obtained with the EHM and the LRM at step 1 is equal to 0.78 and 0.77, respectively, while the values of the AUC obtained with the LRM at steps 6 and 13, which are almost the same (0.889 and 0.893, respectively), and the value of the AUC obtained with the ANN (0.845) correspond to a "good" description of reality.

Figure 2. Susceptibility maps of Briga and Giampilieri basins obtained with a) EHM, b) LRM (at step 6), and c) ANN and d) their relative AUC curves.

CONCLUSIONS

In this work an assessment of three different approaches for studying the landslide susceptibility, i.e., the heuristic, the statistical, and the data-driven models is presented. We applied each method to the Briga and Giampilieri basins, two catchments of the city area of Messina (Sicily) where a considerable number of historical events were documented.

Results demonstrated that the ANN is capable to provide the best agreement with the existing landslide location data, which have been classified within the higher susceptibility classes, as compared with the other two approaches. Also in terms of AUC, the ANN is closer to the LRM, which provides the best results, than the EHM, which provides the worst results. One of the main disadvantage of the heuristic approach pointed out from this study is its high level of subjectivity; the decision rules to create the susceptibility map depend on the experience of the researchers. On the other hand, the heuristic approach presents the advantage to allow the researcher, on the basis of a regressive analysis of results, to make adjustments in the model in order to improve its performances. Statistical and data-driven methods are more
objective than the heuristic approach but they are computationally expensive. However, although the satisfactory results, the ANN models do not offer any chance to make considerations on the role of each landslide-inducing factor. This possibility is instead given by the LRM which allows one to evaluate the influence of each variable and each class in determining the susceptibility, and thus to better understand the physical relations between factors and modeled phenomenon.

REFERENCES