LANDSLIDE SUSCEPTIBILITY MAPPING: A COMPARISON OF LOGISTIC REGRESSION AND NEURAL NETWORKS METHODS IN A SMALL SICILIAN CATCHMENT

ELISA ARNONE (1), ANTONIO FRANCIPANE (1), LEONARDO V. NOTO (1)
(1): Dipartimento di Ingegneria Civile, Ambientale ed Aerospaziale,Università degli Studi di Palermo,Viale delle Scienze, Palermo,90128, Italy

Susceptibility assessment concerning the estimation of areas prone to landslide remains one of the most useful approach in the analysis of landslide hazard. The use of statistical methods together with the GIS technologies is currently the most efficient tool. The correlation between the physical phenomenon and its triggering factors based on past observations is the key point of such analysis. Many methods exist in scientific literature to capture and modeling this correlation. Among these, the logistic regression and the neural networks methods have provided successful results in many applications. A comparison between both the methodologies is given in the present study, by discussing the results of the susceptibility analysis carried out on the same study area by applying the two different approaches to a small basin located in the eastern Sicily, where a number of historical events have been documented over the years.

INTRODUCTION

Every year numerous landslide events hit various areas through the world, often causing severe economic and social damages and making the field of landslide prevention an extremely current issue in territorial management.

Susceptibility mapping, based on recognition of landslide-prone terrain (Hansen [1]), is traditionally considered one of the most useful approach dealing with landslide hazard analysis. In such an analysis, the term 'susceptibility' commonly refers to the probability of a landslide occurrence over a region, following the assumption that "the past and present are keys to the future" and based on an empirical or modeled relationship between historical events and surface characteristics (Varnes and IAEG, [2]). These are identified in the so called *landslide-inducing or triggering factors*, and characterize the landslide potential of an area. It follows that estimation of susceptibility results into a typical spatial correlation analysis between the triggering factors and the occurrence of landslides and the production of thematic maps is the ultimate target.

Over the last twenty years, the availability of reliable tools to assess area prone to landslide has deeply increased, due to the development of several GIS-based methodologies to evaluate the spatial correlation. Statistical methods are particularly successful in such applications. Particularly, the generalized linear models and in particular *logistic regression*, is well suited to analyze a presence-absence dependent variable (Carrara et al., [3]; Lee et al, [4]; Lee and Pradhan, [5]; Mirabella et al., [6]), representing one of the most applied methods. Recently, a number of studies proposed the use of *Artificial Neural*

Network (ANN), belonging to the data driven methods, whose structure is suitable to analyze spatial correlation data as well (Lee et al.,[4]; Ermini et al.,[7]; Caniani et al.,[8]). An interesting combined use of the two methods have been proposed by Lee et al, [4], who used a probability method for calculating the rating of the relative importance of each triggering factor class to landslide occurrence and a ANN method for calculating the weight of the relative importance of each triggering factor.

In this study we investigate the use of both the methodologies for landslide susceptibility mapping, by applying separately the two models on a small Sicilian catchment, where a number of historical events have been documented over the years. Goodness of models and their comparison are assessed by means of the area under the ROC (*Receiving Operating Characteristic*) curves method, whose value is a measure of model fitting. Results from comparison will provide an important indications in choosing the proper method for future analysis.

METHODOLOGIES

Logistic regression - LR

Among the multivariate approaches, the logistic regression (LR) analysis is the one that best fits the case in which the dependent variable is a dichotomous variable. Moreover, it allows one to correlate the dichotomous variable with variables that may be both continuous (slope, distance from street, etc.) and polychotomous or categorical (land use, soil type, geology, etc.).

In susceptibility analysis the dependent variable (*Y*) depends on landslides occurrence and is coded as 0 (*no landslide*) or 1 (*landslide*). The conditional probability that a landslide occurs, which is given by $P[Y=1|X_i]=E[Y|X_i]$ where *X* is the vector of all the landslide-inducing factors, is expressed inside the logistic regression analysis model as:

$$P[Y=1|X_i]=\frac{1}{1+e^{-\left(\beta_0+\beta_1 X_1+....+\beta_p X_p\right)}}=\frac{1}{1+e^{-z}} \qquad (1)$$

where β_1, β_2, … β_p, are the coefficients of variables X_1, X_2, … X_p, and represent the different weight of each landslide inducing factor.

Logit function of Eq. (1) provides a simple linear relationship which allows one to make considerations about the weight that each factor has on the probability that a landslide occurs. A positive value of parameter β_i means that an increasing of variable X_i leads to an increasing of the probability that the dependent variable *Y* assumes the value 1.

The variables coefficients represent the model unknown quantities and are estimated maximizing a log-likelihood function. Once the parameters are estimated, it is possible to evaluate the probability of landslides occurrence through Eq. (1).

Artificial Neural Network - ANN

Among the Artificial neural networks (ANNs), the feed-forward Multilayer Perceptron (MLP) network has been chosen in this application. A MLP consists of a number of units

(perceptrons or neurons) which are connected by weighted links. The units are organized in layers: an *input layer*, with a number of neurons equal to the number of input independent variables of the problem, an *output layer*, with a number of neurons equal to the number of output dependent variables, and ultimately one or more so called *hidden layers*. In a generic MLP, an arbitrary input vector is propagated forward through the network. The hidden layers of neurons make a linear combination of input signals and convert it through a generally nonlinear function (activation function). The hidden layer output becomes then the input to the following layers. Moreover in a MLP there are no feedbacks, i.e. the neurons of each layer are linked only to the neurons of the following layer.

The key instrument that allows the network to learn the dynamics of a particular phenomenon is called *training phase*. During the training phase a set of known input-output couples are presented to the network and the weights are updated by following some pre-determined learning rule, so that the resulting output vector of the net is almost equal to the target vector. Weights are updated on the base of distance between the target and the actual output vector, measured by a cost function *E*, through a minimization of the same function *E*. The most widely used learning rule to perform a gradient descent along the cost surface of the network is the "error backpropagation rule" (EBP).

Model performance of ANN and LR can be evaluated by means of the AUC method, i.e. Area Under the Curve *ROC* (*Receiving Operating Characteristic*). 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. The ROC curve is built plotting the sensitivity versus 1-specificity over all possible cutoff.

CASE STUDY

The Timeto cathment

The Timeto catchment was used in the evaluation of landslide susceptibility. The basin is located in northeastern Sicily, within Messina district (Figure 1) characterized by the highest number of landslides and the largest area of soil removed by landslides in Sicily.

The cathment is approximately 95 km^2 in size and its elevation range between 0 and 1350 m a.s.l.. The morphology of the basin is typical of the Peloritani mountains with narrow valleys and very steep hillslopes. Hydrology is torrential with low-flow discharges during the dry season and high-flow discharges during the fall and the winter.

Information about landslides come from the *Carta Inventario delle Frane*, made by *Regione Sicilia - Assessorato Territorio Ambiente* ([9]); this map gives 4 different types of landslides: flows, falls/topless, slides and complex landslides (Figure 1). The map was transformed in a raster layer at 100 m resolution and represents the dependent dichotomous variable (1 if landslide occurs and 0 in otherwise) used by the two methods. Landslide-inducing factors identified for the analysis are listed in Table 1.

Slope, curvature, and aspect were derived from a 100 m resolution DEM (*Digital Elevation Model*). Mean annual rainfall data were obtained by means of an interpolation of data coming from all the pluviometric stations located inside the basin. The hydrological parameters *a* ed *n* are the parameters of the rainfall *depth-duration curve*, expressed as

$h=ad^n$, and were derived from Lo Conti *et al*., [10]. These parameters allows one to take into account the very intense rainfall effect. Information about land use come from the Corine Land Cover map (*APAT*, [11]), while the pedological layer was assessed by using the Sicilian soil map edited by Fierotti and Ballatore, [12]. The basin is characterized by the dominance of agricultural crops and trees with some areas for grazing and arable land (wheat and grass). Geology information were derived from the Geological Map 1:50000 scale of the Province of Messina. Finally, distances from stream network, faults, and roads were estimated using the Euclidean Distance tool of ArcGis 9.1 starting from the corresponding vector data.

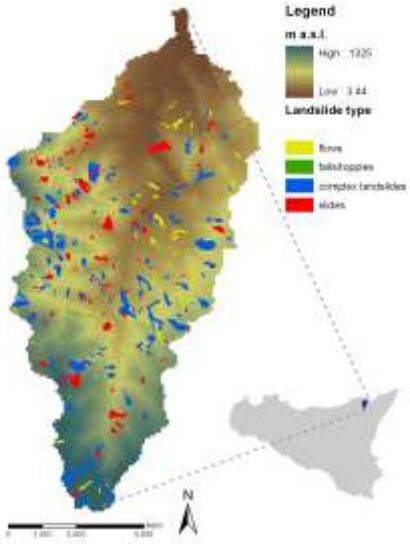

Figure 1. Timeto catchment and landslide locations.

Table 1. Landslide-inducing factors

LANDSLIDE-INDUCING FACTORS	
Continuous variables	Slope
	Curvature
	Aspect
	Mean annual rainfall
	Parameter *a* of *ddf* curve
	Parameter *n* of *ddf* curve
	Distance from river networks
	Distance from faults
	Topographic index
	Distance from roads
Polichotomous variables	Land use
	Pedology
	Lithology

Application of logistic regression model

Estimating regression coefficients implies to fit the model to the dataset and assessing the coefficient significance. The most successful parsimonious model is pursued selecting the variables that result in a best model within the constraints of the available data following a stepwise procedure. Such a procedure involves the estimate of coefficients at different model configurations which either includes or excludes a variable on the basis of the increase in goodness of fit introduced by different variables. Many software for statistical computing include algorithms capable to provides the estimate of regression coefficients together with values of statistical indexes to assess the significance of the coefficients.

In our analysis, 13 steps have been performed, corresponding to the total number of factors. Free *software R* has been used, which determines a coefficient for each continuous variable, while in case of categorical variables, for each factor, it determines a number of coefficients as the classes minus one class (assumed as class of reference).

The best fitted model was obtained at step 11, corresponding to a value AUC equal to 0.645. The most parsimonious results in the model with 3 variables, with a value AUC equal to 0.633. Here, the more onerous but still acceptable model has been chosen, with the following 11 variables: pedology (*pedo*), litology (*lito*), land use, distance from river network (*dist_riv*), curvature (*curvat*), slope, parameter n, aspect, distance from faults (*dist_faults*), distance from roads (*dist_roads*) and mean annual rainfall (*rain*). The resulting coefficients and their significance were here omitted for sake of brevity, referring the reader to Mirabella et al. (2010) for more details.

Given the estimated coefficients, either as vector or scalar, the variable *z* of the logistic regression model results as follows:

$$z = -0.156 + Coeff_{pedo} + Coeff_{lito} + Coeff_{landuse} + (0.0009 \cdot dist_{riv}) + (0.1077 \cdot curvat) + (0.0191 \cdot slope) + (-12.05 \cdot n) + (0.0006 \cdot aspect) + + (0.0003 \cdot dist_{faults}) + (0.0005 \cdot dist_{road}) + (0.0015 \cdot rain) \quad (3)$$

The probability of landslide occurrence is then given by introducing *z* into Eq. (1).

Application of artificial neural network model

Phases involved in the development of an MLP network are various and need to be carefully defined by the operator in order to build up the most successful model which best describes the modeled phenomenon. They can be identified in the following: *data selection for training phase*, *design definition* (input, hidden and output layer), *training phase* (choosing activation and transfer functions), *classification phase*. Analysis has been carried out within the *Neural Network Toolbox* implemented into the software for numerical computing Matlab (*MathWorks*).

Selection of proper dataset for training phase is far from being obvious and a clear criterion has not been provided yet. It is common use selecting randomly a subset corresponding to 1/2 (Lee et al.,[4]) or 1/3 (Ermini et al., [7]) of the entire database. In our application, two different subsets have been defined in order to pursuit the best configuration, according to the following criterion: the entire dataset was first divided between cells experiencing landslides (*landslides*) and those not experiencing landslides (*non-landslides*); the 50% of landslides cells were randomly selected (464 cells) while the non-landslides cells were randomly selected in a percentage obtained by varying the ratio between landslides and non-landslides cells, respectively equal to 1:2 (*subset 1*) and 1:3 (*subset 2*). The two dataset have dimensions respectively of 1392 and 1856 cells.

Structure of input vector depends on number and type (continuous or categorical) of triggering factors and on the methodology used in representing data. In order to minimize the subjectivity connected to a classification of data, we adopted the methodology used by Ermini et al., [7]), which organizes the factors as binary variables and the input vector (relative to each cell) as a binary string composed by a number of positions equal to the total number of classes of all variables (*73*); each position results from 1/0 binary switches as a function of the presence/absence of a class of the variables in a given cell. Such a method, although capable to provide an efficient objective approach, increases considerably the number of computational nodes. A single hidden layer is commonly used in landslide analysis applications, whose number of nodes can be defined using various empirical criteria available from scientific literature, which relate that number to those of nodes in

input and of training cells. Here, *140* nodes have been imposed. Lastly, one node is designed in the output layer, corresponding to the output value of susceptibility at each cell. The overall network structure is denoted as *73 x 140 x 1*. Among all the *backpropagation* algorithms available in literature, one of the most suitable to treat a large amount of data and here used is the GDM (*Gradient Discendent with Momentum*) algorithm. The chosen transfer function is a sigmoid function (*sgm*) which returns values ranging from 0 to 1.

Once all these phases are ultimate, the network is fully designed and ready for the final simulation (i.e. the classification stage) which returns the susceptibility values on the basis of the weights found during the training phase. Networks used for the analysis are the following: *NN1*, which uses the *subset 1* as training subset, and *NN2* which uses the *subset 2*; both the networks use the GDM algorithm.

RESULTS AND COMPARISON

Both the approaches return the distribution of probability of landslides occurrences predicted over the basin with values ranging from 0 to 1. Values have been classified into five level of probability (*very low*, *low*, *medium*, *high*, *very high*) in order to obtain the final susceptibility maps in risk levels and to make their comparison easier (Figure 2).

A first observation can be done on the spatial distribution of susceptibility areas over the five classes. LR model returns a 'smoothed' distribution of values, capable to identify areas classified at various level of risk; particularly, *very low* and *low* susceptibility areas are located downstream the basin and in flat areas; large areas in the western part are instead classified as medium susceptible; within these areas, some well defined spots are classified as *high* susceptible; lastly, very few but still well defined areas are located in *very high* susceptible class, where landslide events where recorded.

Susceptibility values returned by ANN model are mostly distributed over only two of the five classes available, i.e. the basin is mostly classified either as *very low* susceptible or *very high* susceptible, with a various small spots (mostly corresponding to single cells) spread on the basin and classified as *medium* or *high* susceptible. Moreover, results show a perfect agreement with the existing landslide location data which are all classified within the *very high* risk class in both the networks NN1 and NN2. Particularly, in NN2 total areas classified within the *very high* class are larger, due to the different subset chosen for the training phase (*subset 2*) which includes a greater number of non-landslide cells. The corresponding relative frequency distributions are given in Figure 3a, which confirm the analysis above described.

A quantitative evaluation of goodness of fit of the models and their quantitative comparison are given by the ROC curves (*sensitivity* versus *1-specifity*) shown in Figure 3b and relative AUC. Results clearly denote the superiority of the ANN model in terms of goodness of fit, since the two relative curves approach very fast to the unit value. Consequently, values of AUC are greater for ANN models, resulting respectively equal to 0.645 (LR), 0.819 (NN1) and 0.824 (NN2). NN2 networks shows the best fit, suggesting that a more numerous sample in running the training phase improve, at least numerically, the goodness of model.

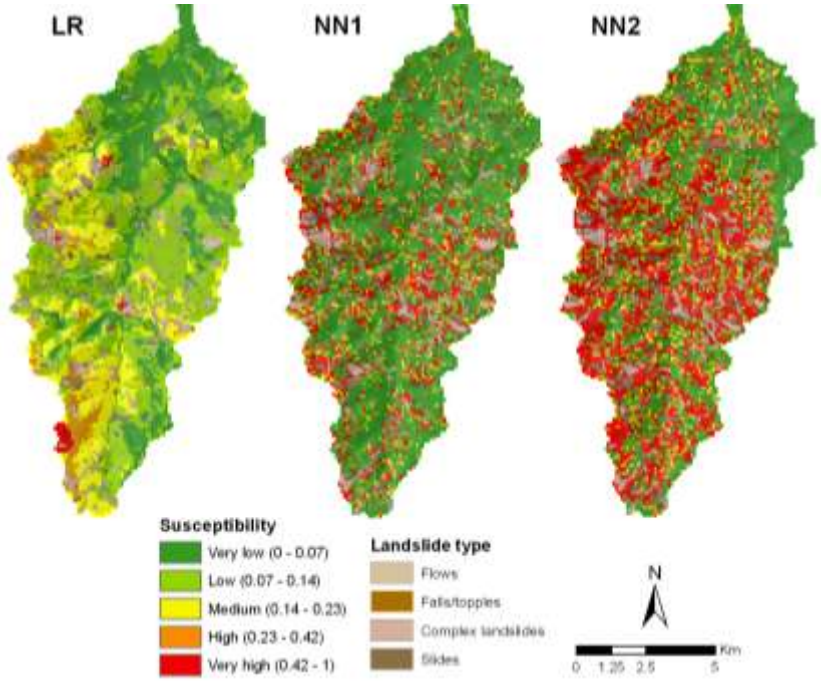

Figure 2. Maps of susceptibility, reclassified in classes of risk, obtained from models.

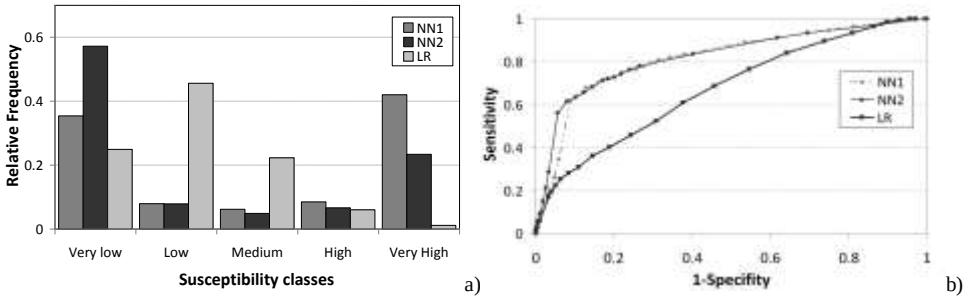

Figure 3. (a) Relative frequency distribution of susceptibility values over the three models. (b) ROC curves of LR model and two ANNs.

CONCLUSIONS

The use of probabilistic and strictly statistical methods to evaluate the susceptibility is being the preferred approach by academic and research institutions, since they allow a good comprehension of the relationships between landslides and inducing factors (Ermini et al., [7]) in term of weight of each factor on the overall analysis. The conceptual approach of multivariate statistical methods is very close to those used by ANN, which can be considered, in some way, statistical methods. In the application to landslide susceptibility both techniques can be classified as “black box models”, and furthermore, several ANNs

have been developed on a statistical basis (Patterson, [13]) even if the ANNs do not have the disadvantage to be dependent of the statistical distribution of data.

In this work the characteristics of the two approaches in a susceptibility mapping application have been analyzed by applying the models to a small Sicilian catchment. Results showed that ANNs models are capable to provide very satisfactory agreement with the existing landslide location data, which have been classified within the higher susceptibility classes. This was easily pointed out by a simple visual observation of maps and then proved by a quantitative comparison through the AUC values. Moreover, comparison between two different ANNs proved also that the use of a greater sample size in the training phase gives higher values of AUC, because it allows the network to 'better learn' the reality. 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 LR models which allow 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

[1] Hansen, A. "*Landslide hazard analysis*", Slope instability (1984).

[2] Varnes, DJ. and IAEG. "Landslide hazard zonation: A review of principles and practice", *The UNESCO Press*, Paris , (1984), pp 63.

[3] Carrara, A., M. Cardinali, R. Detti, F. Guzzetti, V. Pasqui, and P. Reichenbach. "Gis techniques and statistical models in evaluating landslide hazard*", Earth Surface Processes and Landforms* , 16(5), (1991), pp 427-445.

[4] Lee S., J. H. Ryu, J. S. Won, J. H. Park "Determination and application of the weights for landslide susceptibility mapping using an artificial neural network". *Engineering Geology*, 71, (2004), pp 289

[5] Lee, S. and B. Pradhan.. "Landslide hazard mapping at Selangor, Malaysia using frequency ratio and logistic regression models", *Landslides*, 4, (2007), pp 33-41.

[6] Mirabella C., E. Arnone, A. Francipane, L.V. Noto, G. La Loggia, "Utilizzo di modelli lineari generalizzati nella derivazione di mappe di suscettibilità per il rischio idrogeologico". *Atti del XXXII Conv. Nazionale di Idraulica e Costruzioni Idrauliche*, Palermo 14-17 Settembre (2010)

[7] Ermini L., F. Catani , N. Casagli, "Artificial Neural Networks applied to landslide susceptibility assessment", *Geomorphology* 66, (2004), pp 327-343.

[8] Caniani D., S. Pascale, F. Sdao, A. Sole, "Neural networks and landslide susceptibility: a case study of the urban area of Potenza". *Nat Hazards*, 45(1),(2007), pp 55-72

[9] Regione Sicilia - Assessorato Territorio e Ambiente. "Piano Stralcio per l'Assetto diBacino per l'Assetto Idrogeologico (P.A.I.). Bacino Idrografico del Fiume Timeto". Palermo, Italy (2006).

[10] Lo Conti, F., L.V. Noto, M. Cannarozzo, and G. La Loggia. "Regional frequency analysis of extreme precipitation in Sicily, Italy", in *2nd International Workshop on Hydrological Extremes: Variability in space and time of extreme rainfalls, floods and drought*., Cosenza (2007).

[11] APAT, "La realizzazione in Italia del progetto Corine Land Cover 2000". *APAT, Rapporti 36/2005*, Rome, Italy (2005).

[12] Fierotti, G., G. P. Ballatore, "Carta dei suoli della Sicilia, scala 1:250.000, con nota illustrativa". *Università degli Studi di Palermo, Unione delle Camere di commercio, industria, artigianato e agricoltura della Regione Sicilia*, Palermo, Italy, (1988).

[13] Patterson, D., "*Artificial Neural Networks*". Prentice Hall, Singapore (1996).