Available online at www.sciencedirect.com
ScienceDirect
Geomorphology 94 (2008) 379–400
GEOMORPHOLOGY
www.elsevier.com/locate/geomorph

Artificial neural networks and cluster analysis in landslide susceptibility zonation

C. Melchiorre [a,*], M. Matteucci [b,1], A. Azzoni [c,2], A. Zanchi [a,3]

[a] D.I.S.A.T., Università degli Studi di Milano–Bicocca, 20126 Milan, Italy
[b] D.E.I., Politecnico di Milano, 20133 Milan, Italy
[c] Geologist, Via Nullo, 24100 Bergamo, Italy

Received 15 April 2005; received in revised form 28 February 2006; accepted 8 October 2006
Available online 14 June 2007

Abstract

A landslide susceptibility analysis is performed by means of Artificial Neural Network (ANN) and Cluster Analysis (CA). This kind of analysis is aimed at using ANNs to model the complex non linear relationships between mass movements and conditioning factors for susceptibility zonation, in order to identify unstable areas. The proposed method adopts CA to improve the selection of training, validation, and test records from data, managed within a Geographic Information System (GIS). In particular, we introduce a domain-specific distance measure in cluster formation. Clustering is used in data pre-processing to select non landslide records and is performed on the whole dataset, excluding the test set landslides. Susceptibility analysis is carried out by means of ANNs on the so-generated data and compared with the common strategy to select random non-landslide samples from pixels without landslides. The proposed method has been applied in the Brembilla Municipality, a landslide-prone area in the Southern Alps, Italy. The results show significant differences between the two sampling methods: the classification of the test set, previously separated and excluded from the training data, is always better when the non-landslide patterns are obtained using the proposed cluster sampling. The case study validates that, by means of a domain-specific distance measure in cluster formation, it is possible to introduce expert knowledge into the black-box modelling method, implemented by ANNs, to improve the predictive capability and the robustness of the models obtained.

Keywords: Susceptibility analysis; Landslides; Cluster analysis; Artificial neural networks; Lombardy Southern Alps; Italy

1. Introduction

Landslide susceptibility maps are important tools for territorial planning. However, there are many open issues in their construction due to the complexity of natural processes in connection with their conditioning factors and with the difficulties for the cartographic generalization of such variables. The need for predicting the location of future landslides at basin scale requires a quantitative methodology to model these complex

* Corresponding author. Tel.: +39 0264482882.
E-mail addresses: caterina.melchiorre@unimib.it (C. Melchiorre), matteucc@elet.polimi.it (M. Matteucci), augusto.azzoni@poste.it (A. Azzoni), andrea.zanchi@unimib.it (A. Zanchi).
[1] Tel.: +39 0223993470.
[2] Tel.: +39 035231115.
[3] Tel.: +39 0264482859.

0169-555X/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.geomorph.2006.10.035

phenomena from past events using data gathered in the field or in laboratory. In this paper we propose a quantitative methodology to map the landslide susceptibility level given information from past mass movements and conditioning factors focusing on an automatic procedure to choose the dataset. The procedure, which involves the use of Artificial Neural Networks (ANNs) and Cluster Analysis (CA), demonstrates that an accurate sampling strategy improves the model results and increases the landslide occurrence prediction.

ANNs have been widely used for landslide susceptibility zonation (Lee et al., 2003a,b; Lu and Rosenbaum, 2003; Lee et al., 2004; Ermini et al., 2005; Gomez and Kavzoglu, 2005). In fact, in *indirect hazard mapping* the landslide prediction should be based on complex, unknown, and non-linear relationships between mass movement distribution and conditioning factors. ANNs are data-driven models and universal non-linear function approximators. The ability to learn non-linear functions from the data is an important feature in the problem of classifying landslide-prone areas. Moreover, a neural network does not require assumptions about the input variable distribution or absence of correlations between such variables. The use of the ANNs can be a valid alternative in the *indirect hazard mapping*, when the conditioning factors are not approximable by a normal distribution and are strongly correlated.

Spatial data can have inherent uncertainty. Identification and mapping of landslides are subjective: such an uncertainty is proved by large mismatches among inventory maps produced by different research teams (Ardizzone et al., 2001). Other digital maps contain errors caused by measurement and interpolation errors (e.g., slope and aspect maps) (Heuvelink, 1993). ANNs are able to give a good prediction even though trained with noisy and uncertain data.

However, ANNs and other statistical methods need two kinds of samples to estimate the probability of landslide: the first one must be representative of sliding

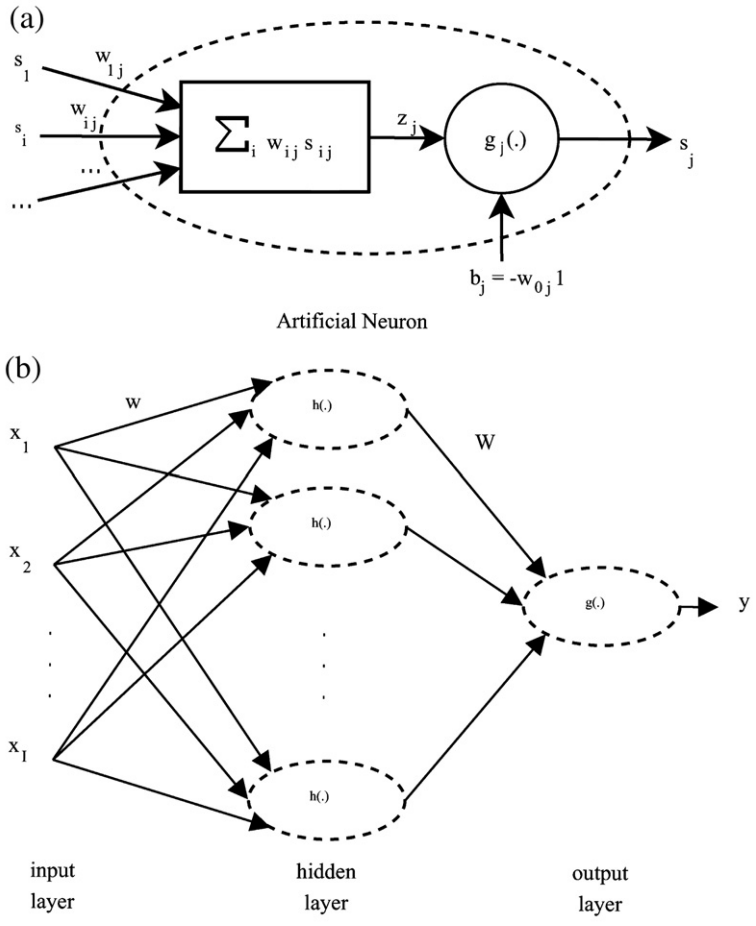

Fig. 1. Schematic representation of an artificial neuron (a) and of a simple feed-forward topology (b).

conditions (presence of landslides), the second one must be representative of stability conditions (absence of landslides). Whereas the dataset representative of sliding conditions is known (landslide inventory map), the non-landslide set is usually obtained by random sampling from the unlabeled data (pixels without landslides).

In the present work we propose a strategy to sample from the unlabeled data by using an unsupervised technique (CA), introduced in data pre-processing to analyse the data structure before sampling the non-landslide cases. A domain-specific distance is introduced to improve clustering and to introduce expert knowledge in non-landslide selection.

2. State of the art

In recent years, the development of Geographic Information Systems (GISs) and spatial analysis techniques has increased and improved *indirect hazard mapping*. The main reasons for such a widespread use of GIS technology are the efficiency in collection, manipulation, and analysis of data given by these software. Even if there are many open questions in landslide hazard mapping, as widely discussed by Carrara et al. (1999), the models obtained by means of spatial analysis show encouraging outcomes.

The common idea of the statistical models used in susceptibility mapping is to estimate:

$$P(L_{\mathrm{ua}}|\mathrm{cf}_1(\mathrm{ua}), \mathrm{cf}_2(\mathrm{ua}), \ldots\ldots, \mathrm{cf}_m(\mathrm{ua})) \quad (1)$$

the probability that the unit analysis ua (i.e, pixel, unit condition area) will be affected by landslide given the m values, one for each conditioning factor (cf). An area is classified as susceptible when its terrain conditions are comparable to landslide areas. Several frameworks have been introduced in order to analyse these terrain variables in relation to the mass movement distribution. Joint conditional probability function (Chung and Fabbri, 1999), certainty factor function (Binaghi et al., 1998) and weights of evidence (Bonham-Carter, 1994) have been used to estimate the probability P in Eq. (1). Carrara (1983), and Carrara et al. (2003) presented multivariate statistical techniques to classify study areas according to conditioning factors.

In recent years, after the developing and the spreading of soft computing, fuzzy set theory, ANNs and genetic algorithm were applied in different regression and classification problems with GIS data, mainly for their capability of analysing heterogeneous and uncertain data (Lee et al., 1998).

Focusing on landslide hazard zonation, due to their capability of being universal approximators, ANNs were used to study mass movements and to map landslide hazard. In the work of Lu and Rosenbaum (2003), ANNs were introduced as a tool for the analysis and the prediction of future ground movements based on geotechnical properties. The networks were implemented as a tool to predict the Factor of Safety (FS) and the state of stability (failed or stable); the capability of the networks to predict the FS as well as the stability of the slope was demonstrated. Lee et al. (2003b) developed landslide susceptibility analysis techniques using a multi-layered perception (MLP) network. The results were verified by ranking the susceptibility index in classes of equal area and showed satisfactory agreement between the susceptibility map and the landslide location data. Lee et al. (2003a) obtained landslide susceptibility by using neural network models and compared neural models with probabilistic and statistical ones. Lee et al., (2004) developed a method to integrate ANNs in calculating the Landslide Susceptibility Index (LSI). The network was built and trained in order to find the weights of the relative importance of different factors for landslide occurrence. Such weights were used successively for calculating the LSI. Ermini et al. (2005) used ANNs to classify terrain units considering hillslope factors and applying two neural

Fig. 2. Location of the study area.

architectures: a PNN (Probabilistic Neural Network) and a MLP (Multi-Layered Perceptron) network. A good prediction was achieved by the use of neural models with a slight preference for the MLP network. Finally, Gomez and Kavzoglu (2005) described an approach for assessing landslide risk by using parameters derived from Digital Elevation Models (DEM), remote sensing imagery, and documentary data in a MLP.

3. Methodology

Traditional classification problems present a set of data to be separated into two or more different groups (i.e., the classes) through the use of a discriminative rule (i.e., the classifier) and a training algorithm used to learn this rule. In the literature, different approaches to classification have been proposed ranging from the early

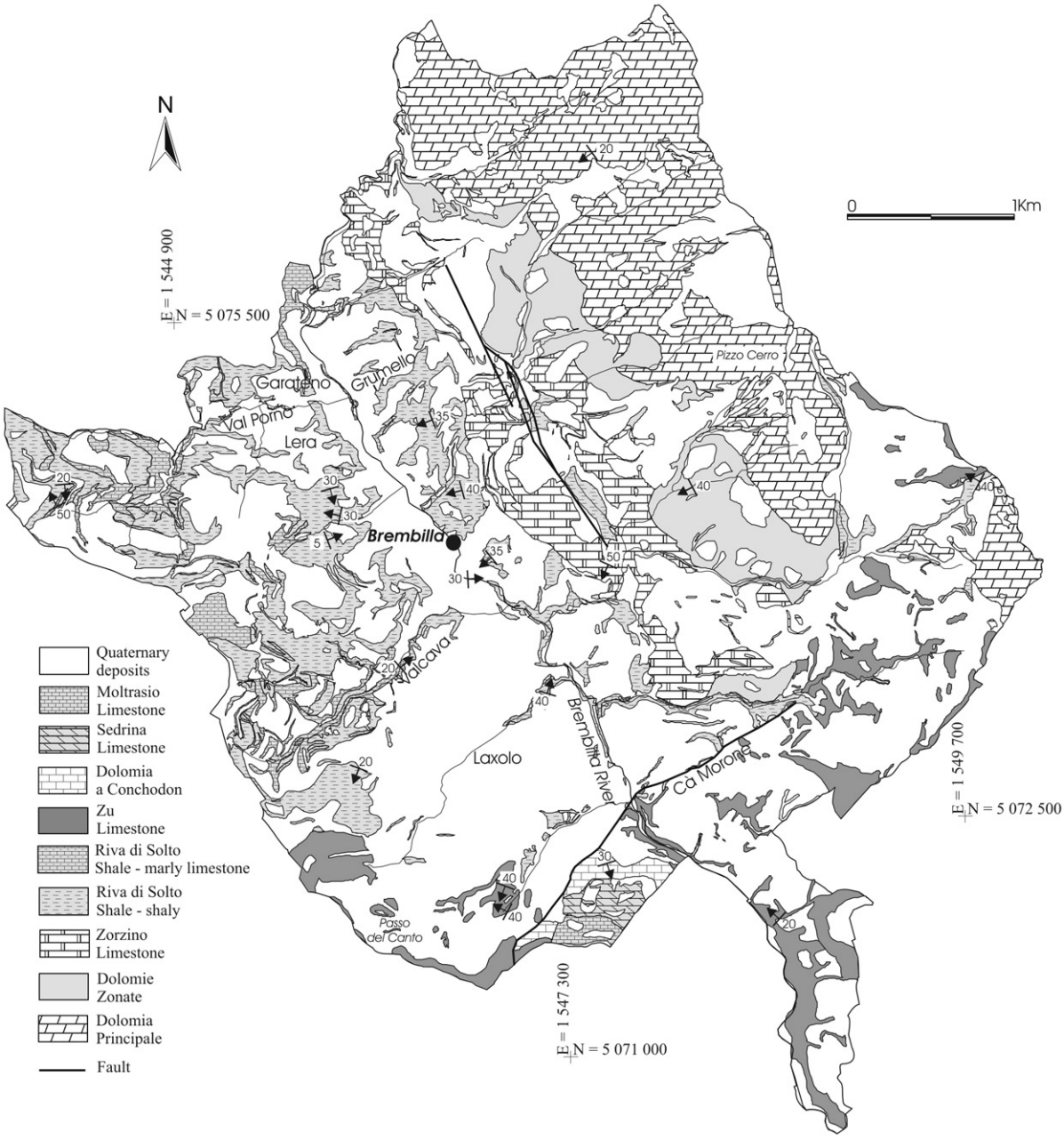

Fig. 3. Geological map of the Brembilla Municipality with bedrock units shown. This map was obtained through an original filed survey carried out at a 1:2500 scale (Azzoni and Agliardi, 2004).

Decision Trees presented by Quinlan (1986) to recent developments in support vector machines (Burges, 1998) and artificial neural networks (Bishop, 1995). However, all these methods require the use of a set of labelled data for each class.

With binary problems, in particular, we distinguish between positive and negative data, i.e., whether they belong to the interest class, or not. When we have only a set of positive data and a set of unlabelled data (data we can not yet classify as positive or negative), traditional algorithms cannot be used to effectively classify the data. In fact, the lack of negative examples makes it more difficult to build a classifier to partition the unlabelled examples into positive and not-positive cases without having a clear idea of the latter. Landslide susceptibility zonation is one of the typical examples having this type of problem. In fact, it is rather easy to identify positive examples (i.e., areas where landslides

Fig. 4. As is Fig. 3, but just showing the Quaternary deposits of the area.

have already occurred), but difficult to identify statistically meaningful examples of stable areas.

In the past, several techniques were proposed to solve this problem. The most common technique tries to identify a set of negative examples by random sampling of the unlabelled data (i.e., areas where landslides have not yet occurred), based on the assumption that the unlabelled data set contains a small number of positive examples and a large number of negative examples. The drawback of this technique is that false non-landslide cases could be selected worsening the discrimination capabilities of the trained classifier. The main idea of this research is to use an unsupervised technique to find out pattern distribution in the dataset, in order to capture aspects (presence/absence of landslides) in the data structure and devise a sampling procedure able to improve the performance of the final classifier.

3.1. Cluster analysis: the k-means algorithm

Clustering can be considered as the most important unsupervised learning technique. As for any other problem in unsupervised learning, it deals with finding an unknown structure in a collection of unlabelled data. A loose definition of clustering could be "the process of collecting objects into groups whose members are similar in some way" (Kaufman and Rousseeuw, 1990). The goal of clustering is thus to determine the intrinsic grouping in a set of unlabelled data and our intent is to exploit this result in order to reduce the number of erroneous samples introduced in the training set.

k-means (MacQueen, 1967; Hartigan and Wong, 1978) is one of the simplest unsupervised learning algorithms to solve a clustering problem. This procedure follows a simple and easy way of classifying a given data set through a certain number k of clusters (i.e., prototypical samples) fixed *a priori*. Starting from a random initialization the algorithm iterates two simple phases: takes each point belonging to a given data set and associates it with the nearest centroid and then, when no point is pending, recomputes the k new centroids as the barycenters of groups resulting from the previous step. After we have these k new centroids, new association phases are performed iteratively between the same data set points and the nearest new centroid. As a result of this loop we may notice that the k centroids change their location step by step until no more changes occur, i.e., the centroids do not move anymore.

This simple procedure can be viewed as a greedy algorithm for partitioning the dataset into k clusters so as to maximize the similarity between objects in a cluster and its prototype, while minimizing the similarity between cluster barycenters. In most cases, the similarity criterion is distance, but how to decide what constitutes a good distance metric is not trivial. It can be shown that there is no absolute "best" metric which would be independent of the final aim of the clustering (Romesburg, 1984). Consequently, it is the final user who must supply this criterion, in such a way that the result of the clustering will suit the needs.

If the components of the data instance vectors are all in the same physical units then simple Euclidean distance is sufficient to successfully group similar data instances. However, in landslide susceptibility zonation the variables used in classification are not immediately comparable. In this case, domain knowledge must be used to formulate an appropriate similarity measure. In the Section 4.3 we describe the metric used for the study area.

3.2. Artificial neural networks

ANNs are generic non-linear function approximators extensively used for pattern recognition and classification (Bishop, 1995; Haykin, 1999). A neural network is a collection of basic units, called neurons, computing a non-linear function of their input. Every input has an assigned weight that determines the impact this input has on the overall output of the node. In Fig. 1 (a) it is possible to see a schematic representation of such an artificial neuron, where w_{ji} is the weight of the connection from neuron i to neuron j, and s_j is the activation

Table 1
Frequency of active, stabilized, dormant, and inactive landslides in the Brembilla Municipality

Typology	n°	Area (m^2)	% Brembilla area
Active slide depletion areas	42	28,321	0.13
Active slide accumulation areas	42	145,316	0.68
Stabilised slide depletion areas	14	9606	0.05
Stabilised slide accumulation areas	15	47,548	0.22
Dormant slide depletion areas	51	16,941	0.08
Dormant slide accumulation areas	41	72,511	0.34
Inactive slide depletion areas	42	252,380	1.19
Inactive slide accumulation areas	26	610,118	2.87
Total		1,182,742	5.57
Total area Brembilla Municipality		21,246,386	100.00
Active slides	42	173,537	0.82
Stabilised slides	15	57,154	0.27
Dormant slides	53	89,553	0.42
Inactive slides	26	862,498	4.06
Total	137	1,182,742	5.57

or output of neuron j. Unit j output is obtained by ideally following a two step-procedure. First the total weighted input z_j is computed using the formula $z_j = \sum_i w_{ji} s_i$ where s_i is the activity level of the i-th unit in the previous layer and w_{ji} is the weight of the connection between the i-th and the j-th unit. Then, the neuron output is obtained as a non-linear function (e.g., sigmoid or hyperbolic tangent) of the total weighted input z_j minus a bias term.

By interconnecting a proper number of nodes in a suitable way and by setting the weights to appropriate values, a neural network can approximate any non-linear function with arbitrary precision (Hornik et al., 1989). This structure of nodes and connections, known as network topology, together with the weights of the connections, determines the final behaviour of the network. Fig. 1b describes a simple feed-forward topology (i.e., no

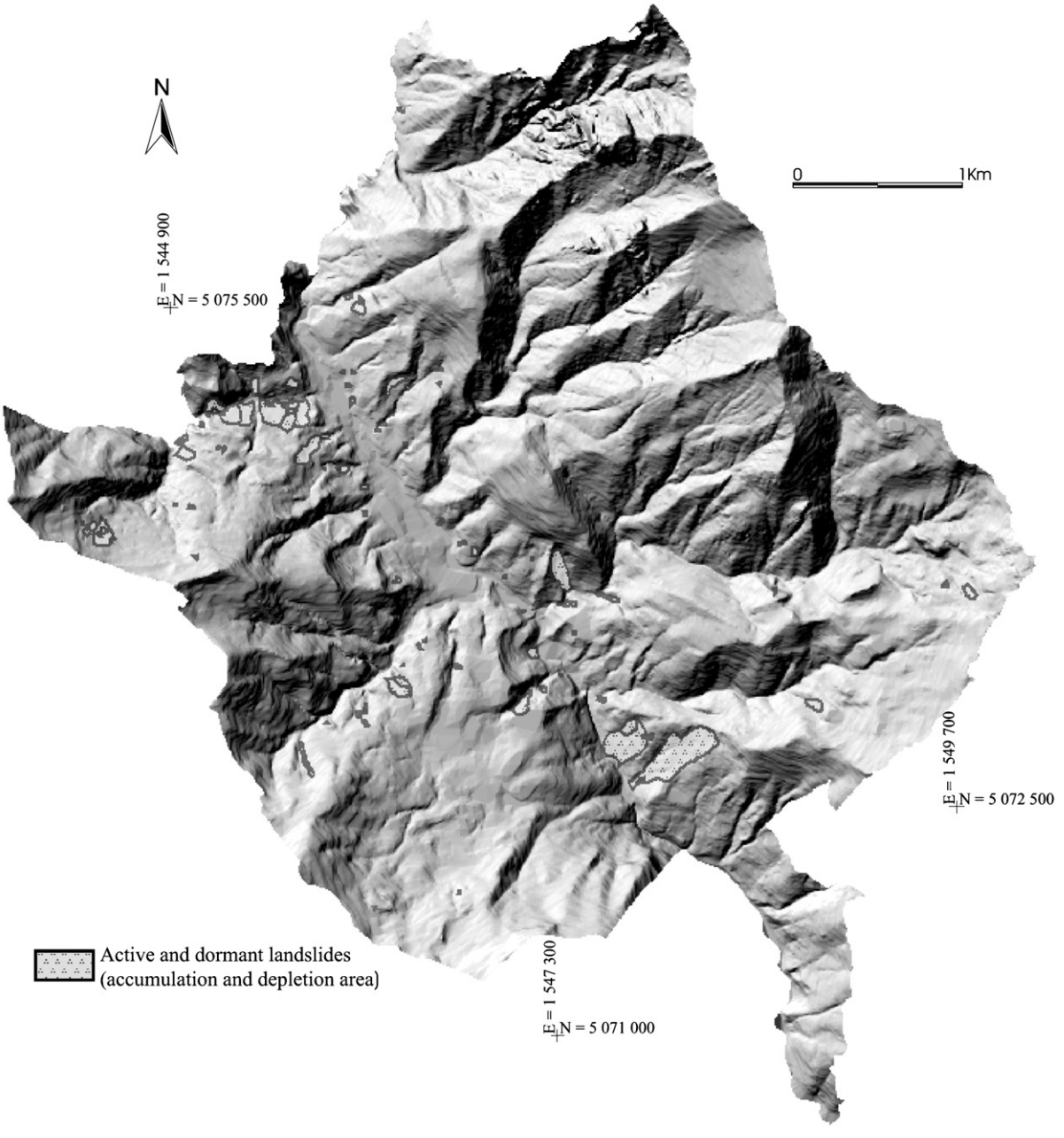

Fig. 5. Landslide map overlaid on the hillshade DTM obtained from the 1:5000 topographic map. The map shows active and dormant rotational landslides used in the analysis.

loops are present) with a single hidden layer (i.e., a layer of neurons neither connected to the input nor the output). Given a neural network topology and a training set, it is possible to optimise the values of the weights in order to minimise an error function by means of any back-propagation algorithm (Rumelhart et al., 1986), standard optimisation techniques (Press et al., 1992), or randomised algorithms (Montana and Davis, 1989).

The topology of a neural network plays a critical role in whether or not the network can be trained to learn a particular data set. A simple topology will result in a network that cannot learn to approximate a complex function, whereas a complex topology is likely to result in a network losing its generalisation capability. This loss of generalisation is the result of overfitting the training data, i.e., instead of approximating a function present in the data, the neural network memorises the training set resulting in inaccurate predictions on future samples. In this paper to improve generalisation we use the *early stopping* technique (Caruana et al., 2000), consisting of using a validation set to stop the training algorithm before the network starts learning noise in the data as part of the model. The error on the validation set can be used also as an estimate of generalisation error and thus can be used to select a proper number of hidden neurons.

4. Case study

4.1. Area description

The Municipality (Fig. 2) extends for about 20 km^2 across the lower part of the Brembilla valley, which is part of the Brembo river catchment. The area, along the southern foothills of the Alps (Prealpi Orobiche), was recently stricken by heavy rainfalls (November, 2002), triggering several unexpected landslides. The middle-age village of Ca' Morone was partially destroyed during this event and the only main road was interrupted by a landslide for one month. Other landslides occurred all over the study area, which has been recently built up and industrialized. Considering the morphological and geological conditions of the Brembilla Valley, it is evident that slope stability processes are the most relevant problem for public safety and land use.

4.2. Geological setting

The study area belongs to the frontal sector of the Southern Alps, a Late Cretaceous to Miocene south-vergent fold and thrust belt (Forcella and Jadoul, 2000). The Brembilla valley consists of a thick carbonate and shaly succession, ranging in age from Late Triassic to Early Jurassic, forming an open NW-SE trending syncline. The lowermost unit, named Dolomia Principale, consists of thick massive carbonates (100 m). It is covered by well bedded dark limestones of the Zorzino Limestone, about 100 m thick. Most of the study area consists of the Riva di Solto Shale (RRS). Its lower member, 150 m thick, shows black shales with minor intercalations of marly limestones which grade into shales, marly limestones, and calcilutites forming the 250 m thick upper member (Jadoul et al., 1994). The RRS passes upward to the Zu Limestone with 500 m of marls, bioclastic limestones and thick massive

Table 2
Erosion/weathering rating and permeability rating assigned to the geological classes

Formations and deposits	Rating	Formations and deposits	Rating
	Erosion/ weathering		Permeability
Riva di Solto Shale — shaly	1.00	Riva di Solto Shale — shaly	1.00
Eluvium on shaly bedrock	1.00	Eluvium on shaly bedrock	0.90
Riva di Solto Shale — marly limestone	0.90	Landslide deposits	0.90
Alluvial deposits	0.75	Colluvium	0.90
Artificial embankments	0.75	Riva di Solto Shale — marly limestone	0.75
Slope deposits	0.75	Moltrasio Limestone	0.75
Eluvium on calcareous bedrock	0.75	Sedrina Limestone	0.60
Landslide deposits	0.75	Zorzino Limestone	0.60
Colluvium	0.75	Zu Limestone	0.60
Coarse grained slope deposits	0.70	Dolomie Zonate	0.60
Breccias deposits	0.70	Eluvium on calcareous bedrock	0.55
Moltrasio Limestone	0.65	Dolomia a Conchodon	0.40
Sedrina Limestone	0.40	Dolomia Principale	0.40
Zorzino Limestone	0.40	Artificial landfills	0.35
Zu Limestone	0.40	Slope deposits	0.35
Dolomie Zonate	0.30	Breccias slope deposits	0.35
Dolomia a Conchodon	0.20	Alluvial deposits	0.25
Dolomia Principale	0.20	Coarse grained slope deposits	0.25
(a)		(b)	

Table 3
Rating assigned to the land-use classes

Landuse	Rating
Urban	1.00
Bedrock	1.00
Grass-pasture	0.75
Uncultivate land	0.75
Natural vegetation	0.75
Orchard	0.60
Newly reafforested land	0.60
Forest	0.40
Riverbed	0.00
Lakes	0.00

boundstones. Jurassic units form the highest part of the exposed succession and consist of the massive oolitic limestones of the Dolomia a Conchodon (100 m) covered by well bedded cherty marly limestones of the Sedrina Limestone.

The oldest units outcrop along the uppermost part of the eastern side of the Brembilla valley forming the northern limb of the fold with dips up to 45°, whereas the western side and the central part of the valley shows the youngest units. Vertical N-S and NNE-SSW strike-slip and normal faults cross the fold and bound it laterally. The Jurassic units are exposed to the east.

The structural setting strongly controls landslide occurrence. Most of the observed phenomena developed within the RRS, which is one of the most landslide-prone units in the region. This is due to the association of intensively cleaved shales and marly limestones with low geomechanical properties and the structural setting which favours the formation of bed-parallel slip causing rotational and complex slides especially along the eastern limb of the syncline. Earth-flow slides are also frequent due to deep weathering of this unit. It is worth noting that the major slide of the area, the "Ca' Morone" slide, which caused major economic damages to the whole valley developed in close association with the NE-SW normal and strike-slip fault which separates the RRS from the Zu Limestone. Rock falls are less common and are related to the large rock walls formed by the massive carbonates which surround the valley.

4.3. Available data and conditioning factors

The cartographic database used in the analysis includes several maps (topographic, land use, geological and landslides inventory maps) stored in a GIS and elaborated to obtain input and output variables suitable to perform the susceptibility analysis by means of ANNs. Six different data layers consisting of *erosion/weathering* rating, *permeability* rating, *landuse* rating, cosine *aspect*, *slope* and *contributing area* were obtained from the original data and used as input variables. The presence/absence of landslides layer was derived from the landslide inventory map and used as the output variable.

The geological map was obtained through field survey (Azzoni and Agliardi, 2004) and was used to individuate geological and geomorphological features relevant to slope stability problems. Field mapping was carried out at a 1:2500 scale. The geological units were

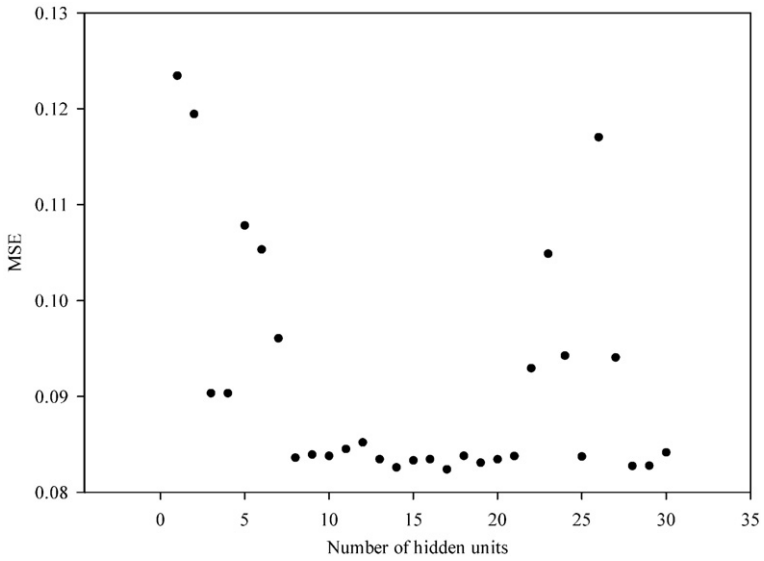

Fig. 6. MSE for the validation set of several neural architectures with different number of hidden units.

Table 4a
Weights used to perform the cluster analysis

	Aspect	Contributing area	Erosion/weathering rating	Permeability rating	Slope	Land use rating
Weight	0.2	1	0.3	0.3	0.5	0.1

classified according to their lithology, composition and origin, on which their geomechanical and hydrogeological properties depend. The bedrock (Fig. 3) has been mapped according to its litostratigraphic subdivisions, ranging from strong limestones and dolomites, to very weak shales. Single rock outcrops have been distinguished from partially covered discontinuous outcrops (maximum: 0.5 m soil cover thickness). Quaternary deposits (Fig. 4), which are mainly related to slope processes, cover most of the area. Due to the medium-high slope angle (28°) of the Brembilla valley, superficial deposits are usually quite thin; the thickest accumulations occur at the bottom of the slopes, where they are generally rich in clay.

The landslide inventory map has been obtained from field survey combined with aerial photo interpretation (1:25,000 scale) and includes phenomena which are distinguished according to their degree of activity and typology. Active slides include all the areas where slope instabilities occurred in recent decades and are still active at present. The most important slides mainly correspond to rotational slides and earth flows. In most cases they affect the soil cover and the uppermost weathered part of the shaly bedrock. They have been also observed in unweathered shale and limestone. The major slides are located around Cà Morone (rotational earth slump), Val Porno, at Grumello village. Minor slides have been also recognised along the Brembilla river banks, in Valcava valley. Most of the major active slides have been artificially stabilised in the last five years. The most important ones are located near Laxolo (rotational earth slump), along the Valcava valley, near Lera (rotational earth slump), and near Garateno village (rotational earth slump and earth flow).

Dormant slides include all the areas where slope instability phenomena have been clearly identified, but without any particular evidence of recent activity, such as slope deformation, cracks in houses and roads, etc. No dormant slide showed any relevant movement during the last major meteorological events (Autumn 2000, November 2002).

Inactive slides show overall good stability conditions, with vegetation usually covering old slide features. No reactivation is known from past historical records. All these slides evolved within the Riva di Solto Shale outcrop area, except for the Passo del Canto rock slide, which evolved in the Zu Limestone and produced a wide accumulation of limestone blocks and can be classified as a rock avalanche. Table 1 shows the frequency of active, stabilised, dormant, and inactive landslides in the study area.

The 1:5000 topographic map (contour interval of 5 m) has been digitized in order to prepare a DTM of the area. The 1:10,000 scale land-use map (Regione Lombardia, 1991) identifies areas with presence or absence of vegetation. The areas covered by vegetation are distinguished according to dominant species and naturalness of vegetation. Before elaborating the landslide inventory map to derive the output layer (presence/absence of mass movements), it was necessary to consider the type of landslides. Since we decided to test the susceptibility analysis on a homogeneous data population, we have selected only the rotational landslides that have occurred in the recent past from the landslide inventory map, including active and dormant slides (Fig. 5). This is justified by the fact that the rotational landslides are the most common phenomena in the study area and that the inactive landslides may not be representative of the

Table 4b
Example of subdivision of the study area in 7 clusters and example of the verified sampling condition

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Total in the seven clusters	Test set	Total in the study area
Number of landslide pixels	239	0	321	495	27	1845	3568	6495	1350	7845
Number of absence landslide pixels	63,123	78,065	144,424	153,311	114,114	111,968	174,431	839,436		839,436
Number of pixels	63,362	78,065	144,745	153,806	114,141	113,813	177,999	845,931		847,281
Sampling condition (pc/pt)	0.04	0.00	0.05	0.08	0.00	0.28	0.55			
Verified sampling condition	Yes	Yes	Yes	Yes	Yes	No	No			

present-day sliding conditions. The rotational landslides analysed are 118: their medium slope angle is 24° and their medium size is 440 m^2. Most of the landslides are smaller than 200 m^2. In order to give more emphasis to the failure conditions, we considered the highest part (50%) of each landslide unit (accumulation and depletion area). The output layer was obtained by assigning value 1 to the highest part of the landslide part and value 0 to areas with absence of mass movement.

Slope, *aspect* and *contributing area* layers have been calculated from the DTM. We chose the d-infinitive algorithm to calculate contributing area (Tarboton, 1997). We scaled the raw continuous data (slope, aspect and contributing area) into a range of 0–1. The aspect data were scaled by means of the coseno operator, since its value distribution range is between 0° and 360° and values close to minimum of the range (0°) have the same physical meaning as values close to the maximum of the

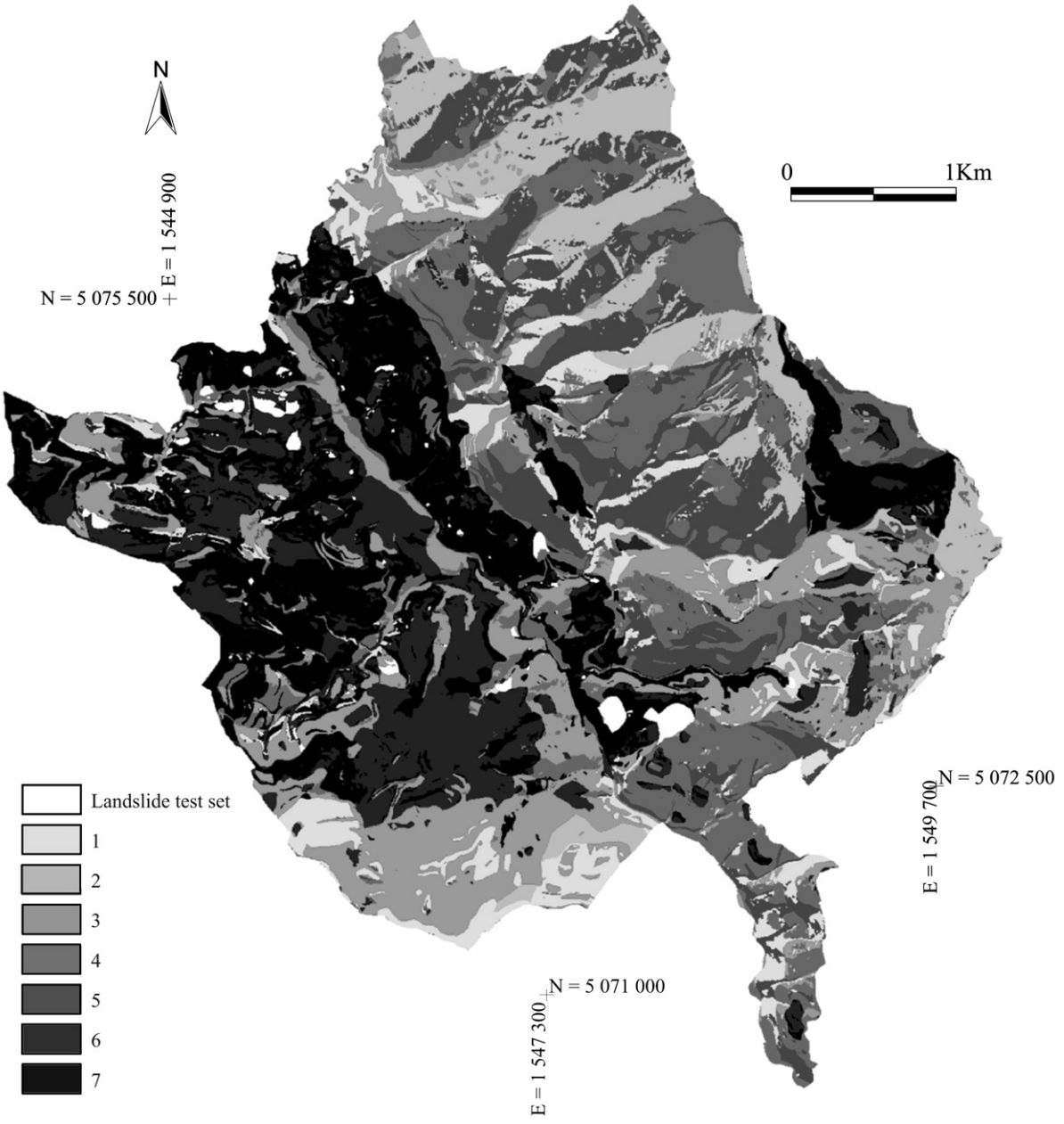

Fig. 7. An example of subdivision of the study area in 7 clusters. The landslides in the test set were excluded from the clustering.

range (360°). Taking into consideration the landslide distribution, the coseno operator is enough to discriminate between north-facing and south-facing slopes.

Categorical variables (i.e., geology, land-use) have been converted into numerical values by assigning a rating to each class between 0 and 1. We chose this approach to reduce network complexity and improve classification performance.

Geological classes have been used to define landslides susceptibility with their erosion/weathering attitude and permeability using two different ratings, according to the approach suggested by Anbalagan (1992a,b). This author presents a methodology based on an empirical approach which combines past experiences of the conditioning factors and their impact on landslide occurrence in a study area. The main idea of that rating scheme is to assign index values to each conditioning factors taking into consideration its influence on mass movements. We have converted the categorical variables into *erosion-weathering* rating and *permeability* rating by following that numerical scheme.

The *erosion-weathering* rating was derived by considering the response of rocks and deposits to these processes. As several units with similar composition (limestones, marlstones, slates) form the substratum of the area, the rating was based on their composition. Limestone is generally hard and massive, whereas the terrigenous rocks, as the RSS, are weak and often deeply weathered and rotational landslides more easily occur in that kind of rock. A rough estimate of the spacing of the beds has also been taken into account, as a wide range of conditions occur, from massive bedded (Dolomia Principale, Dolomia a Conchodon) to finely stratified units (RRS). Loose deposits have been rated based on the

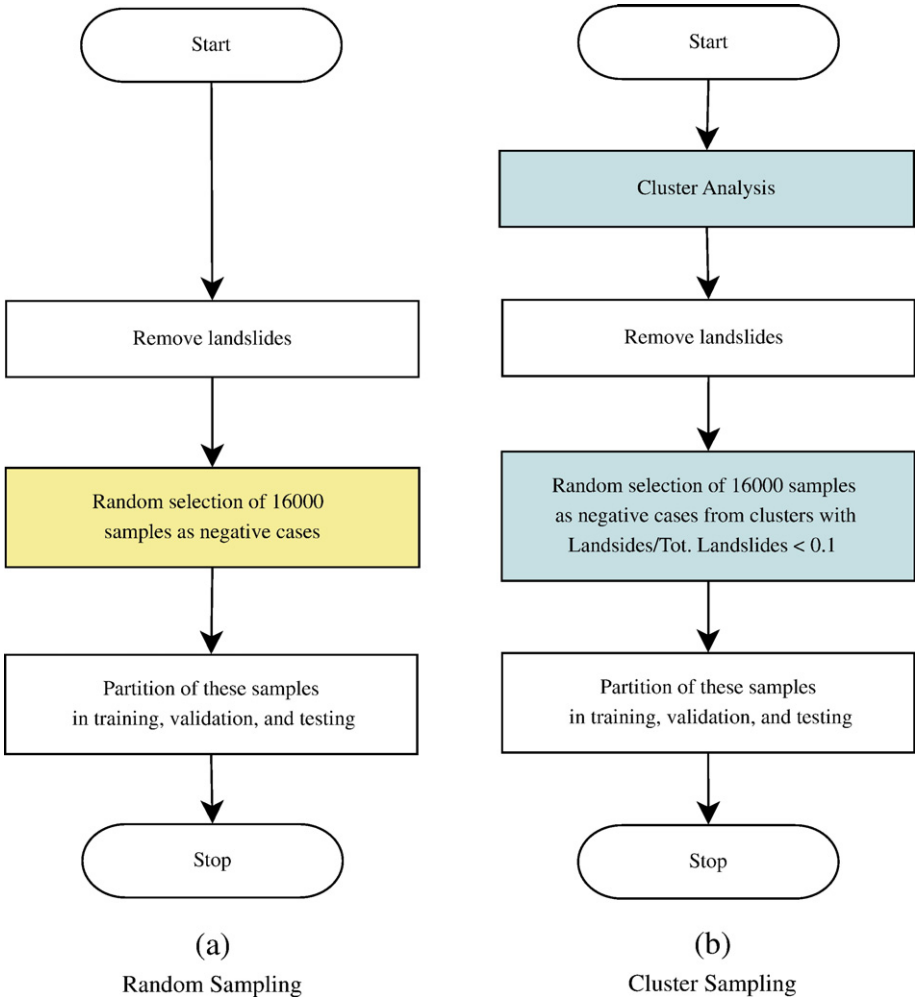

Fig. 8. The two sampling procedures compared. (a) describes the classical random sampling procedure, whereas (b) shows the algorithm proposed in this contribution. Coloured boxes highlight the differences in the two algorithms.

Table 5
Results of the t-test

Sensitivity results			*T*-test result	
	Sensitivity		Degrees of freedom	179.0264
	Cluster sampling	Random sampling	*t* value (one-tailed)	26.06814
Minimum	54.92	10.16	$P(x>t)$	<0.0001
Maximum	98.44	67.63		
Mean	78.64	35.82		
Standard deviation	9.43	13.45		
Number of observations	100	100		
(a)			(b)	

The *t*-test was calculated to evaluate the statistical difference in performance measures after random and cluster sampling.

relative abundance in fine-grained sediments (silt–clay/ versus sand–gravel) and according to their consolidation. The maximum *erosion-weathering* rating was assigned to the weakest degradable terrains (RSS-shaly and colluvium on shaly bedrock), as shown in Table 2a. The *permeability* was scored by attributing the greatest rating to terrains prone to the development of high pore pressure conditions (RRS formation-shaly) and the lowest rating to porous and coarse-grained deposits (Table 2b).

The land-use rating layer was finally obtained from the land-use map again by using the method of Anbalagan (1992a,b), which considers the type and structure of vegetation, its stability or its absence. Since ground cover affects erosion, weathering, fluctuation in the water table, etc., and thus the stability of the slope, we assigned the maximum rating to areas devoid of vegetation (Table 3). The minimum value was given to areas where landslides do not occur (i.e., lake and riverbed) and 0.4 value to forested areas with shade-tree vegetation ensuring the maximum protection with respect to superficial erosional processes. Intermediate values have been given to other classes.

5. Analysis and results

The ANN was trained with six network inputs (erosion/weathering rating, permeability rating, land-use rating, coseno aspect, slope, and contributing area) scaled in the range 0–1 and the network output was also defined in the range 0–1 by setting the output value to 1 for landslide presence and to 0 for landslide absence.

The analysis was performed using an MLP network with the Levenberg–Marquardt training algorithm (Marquardt, 1963; Hagan and Menhaj, 1994). We used the *early stopping* technique (Caruana et al., 2000) to improve the generalization of the network. Several structures with different numbers of hidden units have been tested to find the best one. Results are shown in Fig. 6. In the left part of the plot, the Mean Square Error (MSE) for the validation set decreases if the number of hidden units increases, i.e., the higher the network complexity, the better the performance on the validation set. In the right part of the plot the error increases, since the network is too complex and overfits the training data. The curve in Fig. 6 has two minimums, the first one at 14 hidden units, the second one at 16. The structure with 14 hidden units has been chosen, as it ensures the best generalization without excessively increasing the network complexity.

Landslides from the output layer were subdivided into three subsets: training, validation, and test set. By means of random permutation, 100 different subdivisions of the landslide dataset were found. Although the analysis was performed on a pixel-by-pixel basis, each landslide was considered as a unit during the dataset subdivision.

To find clusters meaningful for the instability analysis, a measure based on variable domains and the importance of each factor to landslide occurrence was used in the cluster formation. In order to identify similar objects, the different variables have been differently weighted. In particular, a weighted sum of the absolute value of the difference has been used as a distance measure:

$$D = \sum_{k} |x_{ik} - x_{jk}| w_k$$

where D is the distance between the two objects i and j, and k is the number of factors (variables). The weights (Table 4a) have been chosen considering the distribution of each variable in its domain with the aim of avoiding the formation of clusters over-influenced by the distribution of a single variable. In this particular case, weight values have been assigned in order to identify clusters showing the separation of classes based on presence and absence of

Table 6
Mean Sen and mean (1-Spe) values calculated using different cut points

Cut point	Random sampling		Cluster sampling	
	Sensitivity	1 − Specificity	Sensitivity	1 − Specificity
0.1	0.87	0.61	0.86	0.51
0.2	0.76	0.43	0.81	0.42
0.3	0.65	0.31	0.80	0.39
0.4	0.53	0.22	0.80	0.38
0.5	0.36	0.15	0.79	0.38
0.6	0.26	0.09	0.79	0.38
0.7	0.14	0.05	0.78	0.37
0.8	0.07	0.03	0.77	0.37
0.9	0.03	0.01	0.73	0.35

landslides. Landslides in the test sets have been excluded from the dataset and from the CA. They have been used to validate both the trained neural models and the effectiveness of the sampling procedure.

We want to underline that the weights are not assigned to the input of the model, so they do not represent the importance (i.e., weight) of each factor for landslide occurrence. Weights are considered only in the clustering procedure to define a proper similarity criterion.

Introducing a proper metric into CA allows us to identify clusters in which the presence of landslide pixels is dominant. In fact, two of the seven clusters found after performing CA contain the most of the landslide pixels, as shown in Table 4b. Those clusters have to be excluded from the selection cases, since they are representative of instability conditions. An example of the subdivision of the study area into the 7 clusters is shown in Fig. 7.

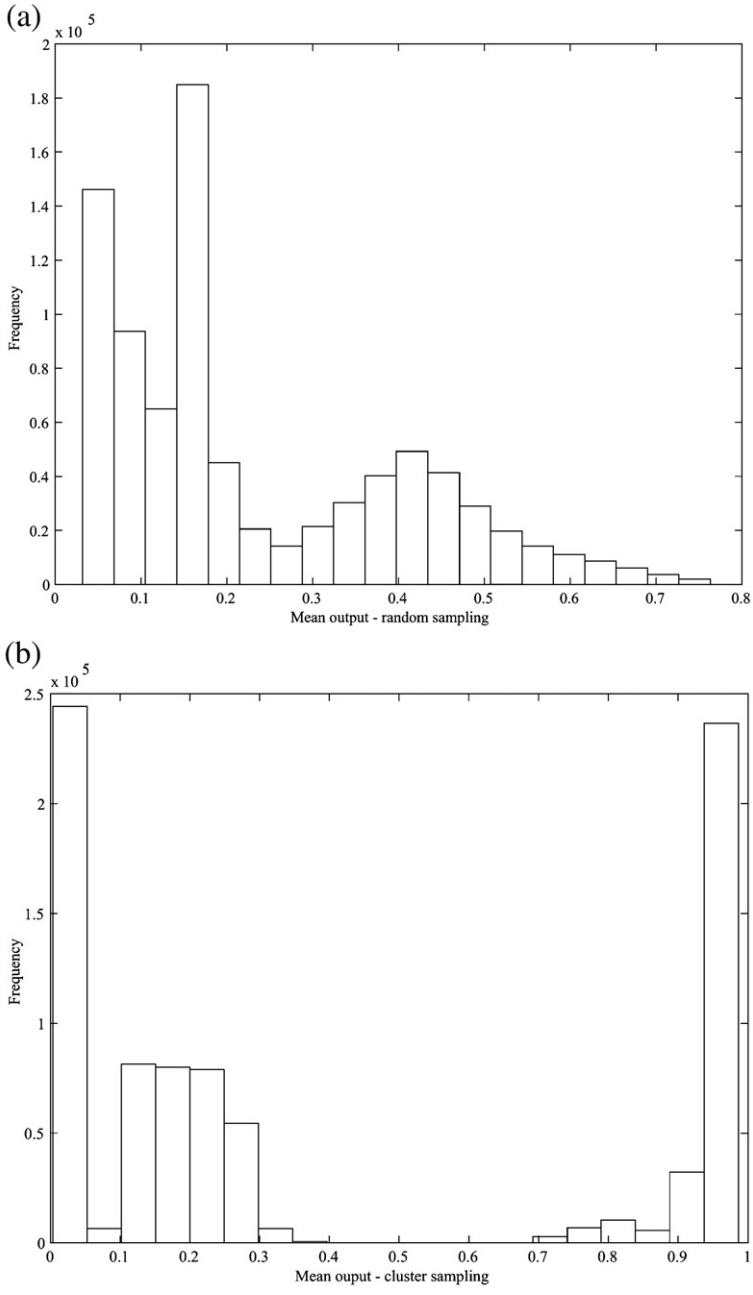

Fig. 9. Plot of the mean values of the trained networks against the frequency after random sampling (a) and after cluster sampling (b).

For each of the 100 different landslide subdivisions, the procedures for random sampling and cluster sampling were carried out as follows. In the random sampling, the negative cases (stable pixels) were sampled from unlabelled data (pixels without landslides) by randomly selecting 16,000 pixels and dividing those pixels into three subsets (training, validation, and test set). In the cluster sampling, after excluding the landslide test set from the dataset, the k-means clustering was performed and the 16,000 negative cases were uniformly sampled only from clusters in which the following ratio was verified:

$$\frac{pc}{pt} < 0.1 \quad (2)$$

where pc is the number of unstable pixels in the cluster and pt is the number of total unstable pixels in the training and validation sets.

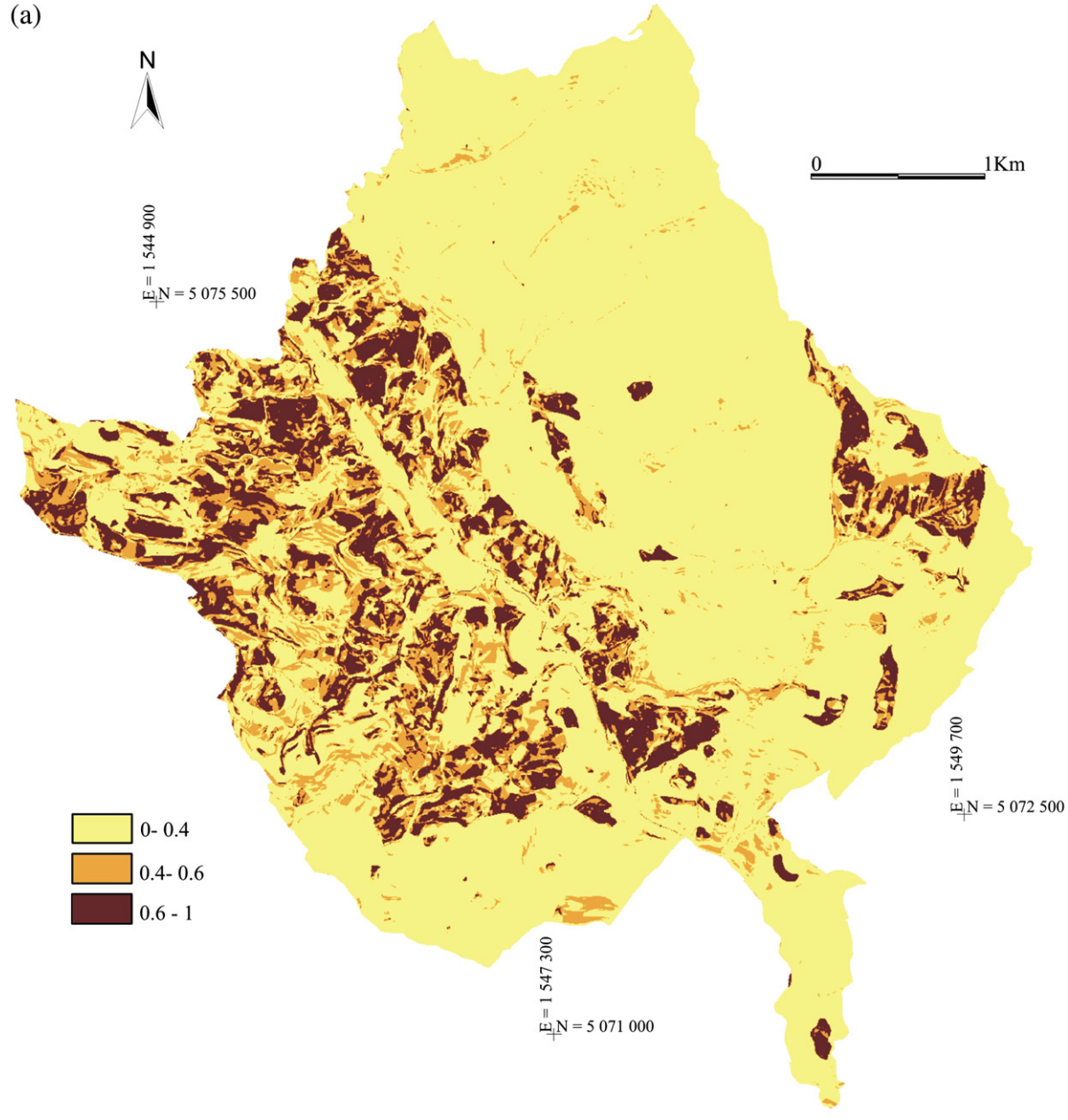

Fig. 10. Output of the network classified into 3 ranges for the model after random sampling (a) and for the model after cluster sampling (b).

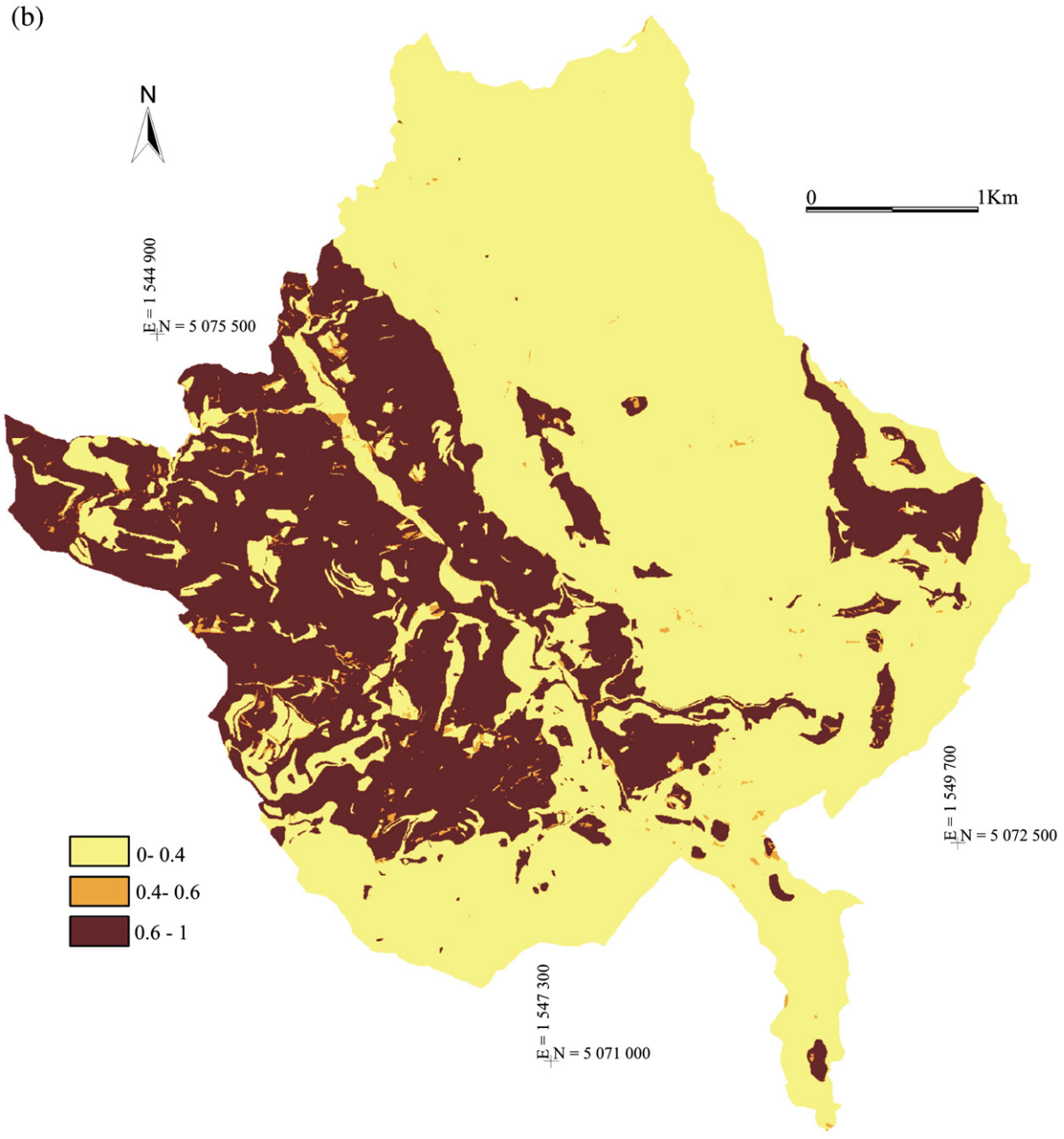

Fig. 10 (*continued*).

In order to compare classification results on many datasets, we found two non-landslide sample sets for each of the 100 subdivisions of all landslides: the first one was obtained by random sampling, the second one by cluster sampling (Fig. 8). The performance of the models after random and cluster sampling was evaluated through the sensitivity, because the main aim of the proposed approach is to improve landslide-prone area classification. This means that sensitivity, i.e., the percentage of correctly classified landslide cases, is the most useful measure to compare random and cluster sampling strategies. The sensitivity Sen (i.e., true positive rate) is defined as:

$$\mathrm{Sen} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \times 100 \tag{3}$$

in which TP represents the number of true positives (pixels with the presence of landslides classified as

unstable) and FN represents the number of false negatives (pixels with the presence of landslides classified as stable).

The single network output was used to separate the two classes. The first classification was obtained using 0.5 as threshold. Output values in the range 0.5–1 represent the landslide class and the values in the range 0–0.5 the non-landslide class. Table 5 illustrates the difference in performance measures after random or cluster sampling. A Student's *t*-Test shows that there is a significant difference between the averages of the means of the two samples with a *p*-level <0.0001. The minimum and the maximum in sensitivity show that the network is able to separate stable and unstable classes only if it is trained after pre-processing data with CA. The mean value of the sensitivity after the random sampling is 35%. Random sampling does not allow us to classify unstable pixels,

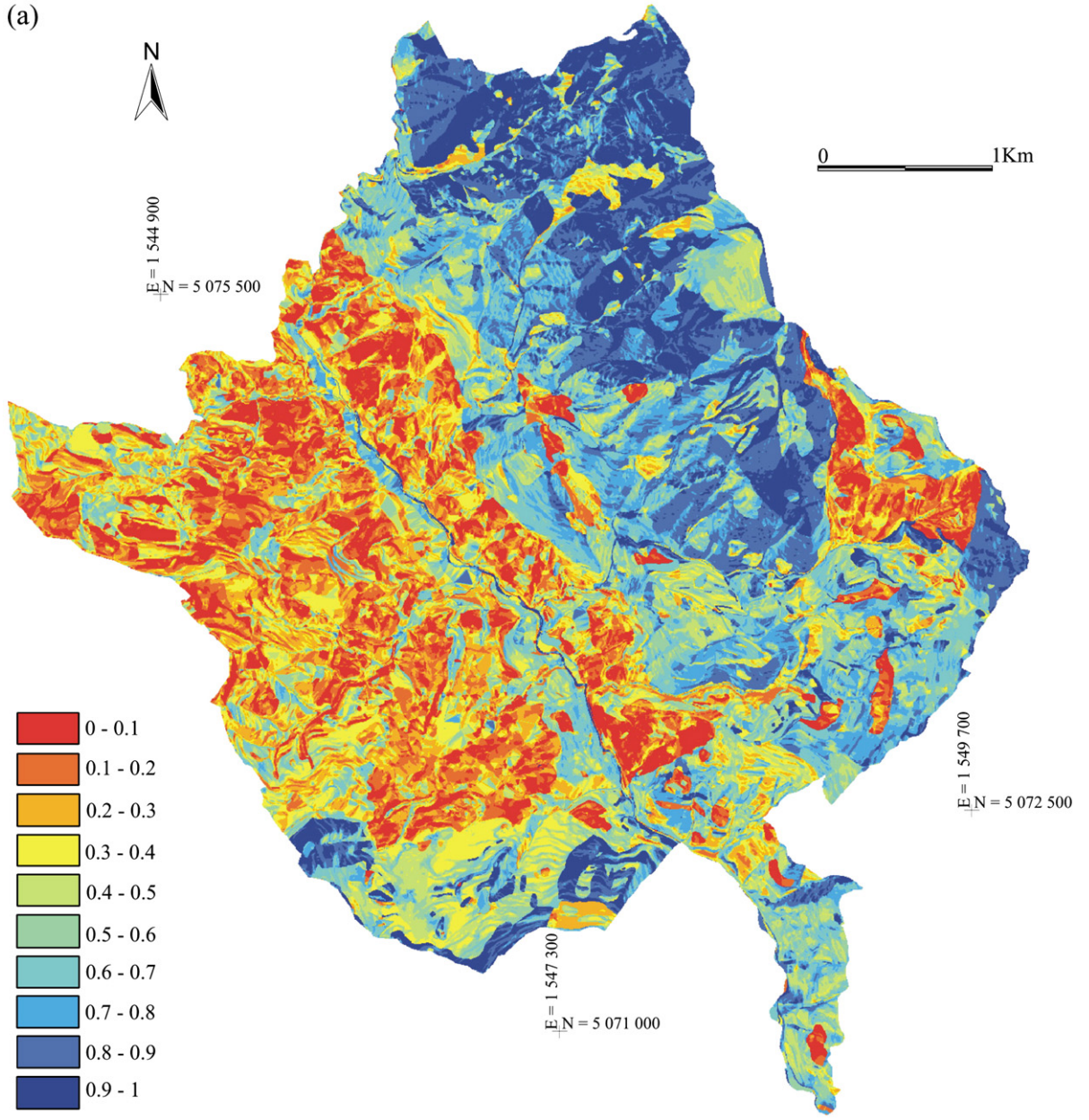

Fig. 11. Susceptibility maps after random sampling (a) and after cluster sampling (b). The area more susceptible is the 0–0.1 class, which represent the 10% more susceptible area. There are some differences between the two maps: the different sampling strategy influence the behaviour of the network and the final response of the model.

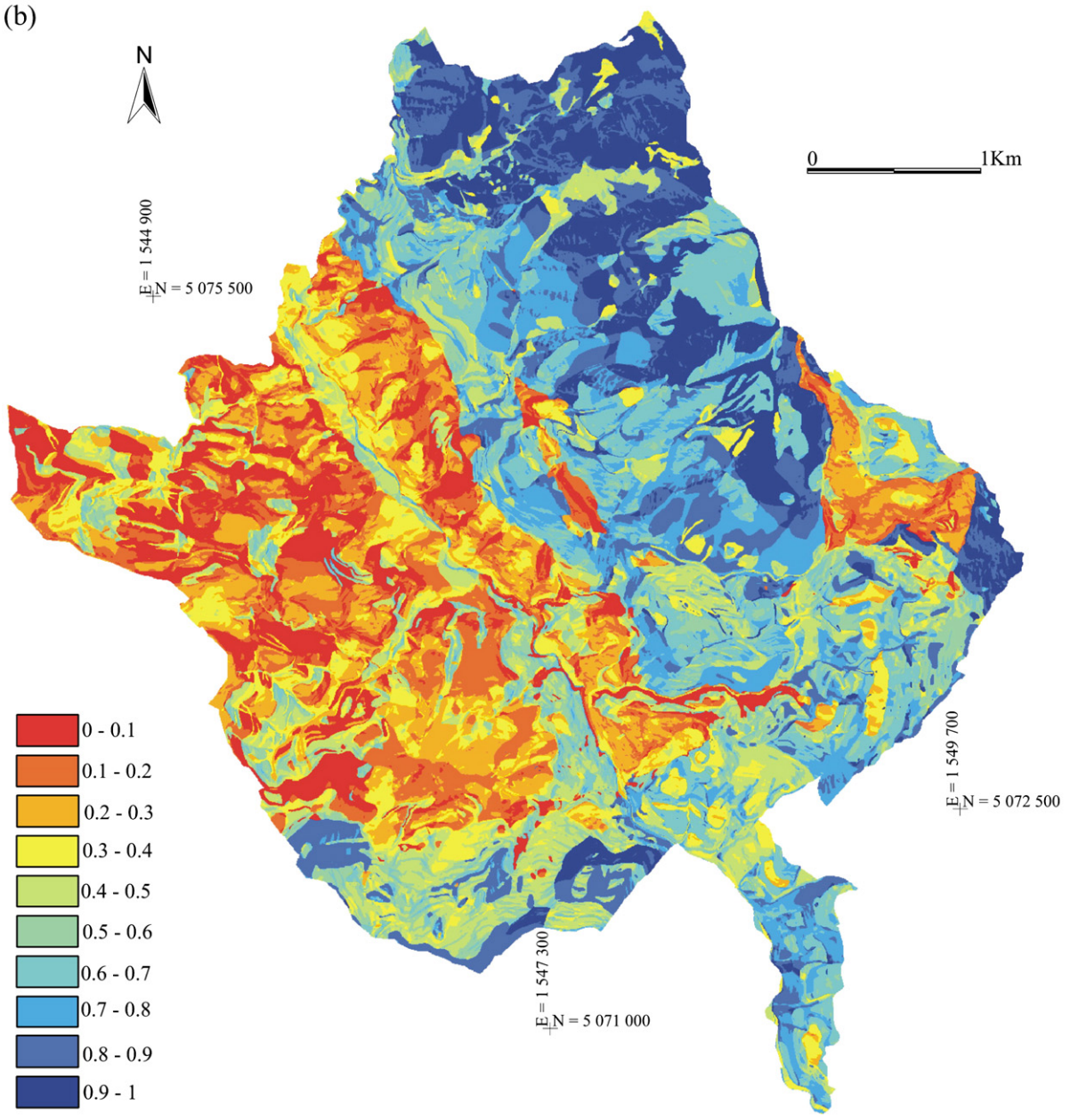

Fig. 11 (*continued*).

since the network is not able to extract the features of slope instability from the data. However, CA assists in the selection of non-landslide data and allows the ANNs to understand hidden data structures, which improves the classification of unstable areas (Table 6).

One could think that after cluster sampling, the sensitivity increases since the ANN classifies all the samples in the unstable area class. In order to disprove that, performance measures were also calculated after choosing different cut-off points. In this case we also kept in consideration also the specificity Spe (true negative rate) defined as:

$$\mathrm{Spe} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \times 100 \tag{4}$$

The value (1 − Spe) is the rate of false positives.

Table 6 shows the performance measures, calculated using different cut-off points. The mean sensitivity of the

cluster sampling-ANN model does not significantly change with the cut-off point. On the contrary, the mean sensitivity of the random sampling-ANN model is strictly related to the threshold chosen, demonstrating the robustness of the classifier trained after cluster sampling.

The trained networks were simulated for the whole study area. We have obtained two mean values for each pixel in the study area, one by averaging the 100 networks trained after the random sampling and one by averaging the 100 networks trained after the cluster sampling. The mean value against the frequency after random sampling (Fig. 9a) shows that the output of the networks is close to low values (stable areas) or close to 0.4. The value region around 0.5 is an uncertain region for the classifier. Fig. 9a

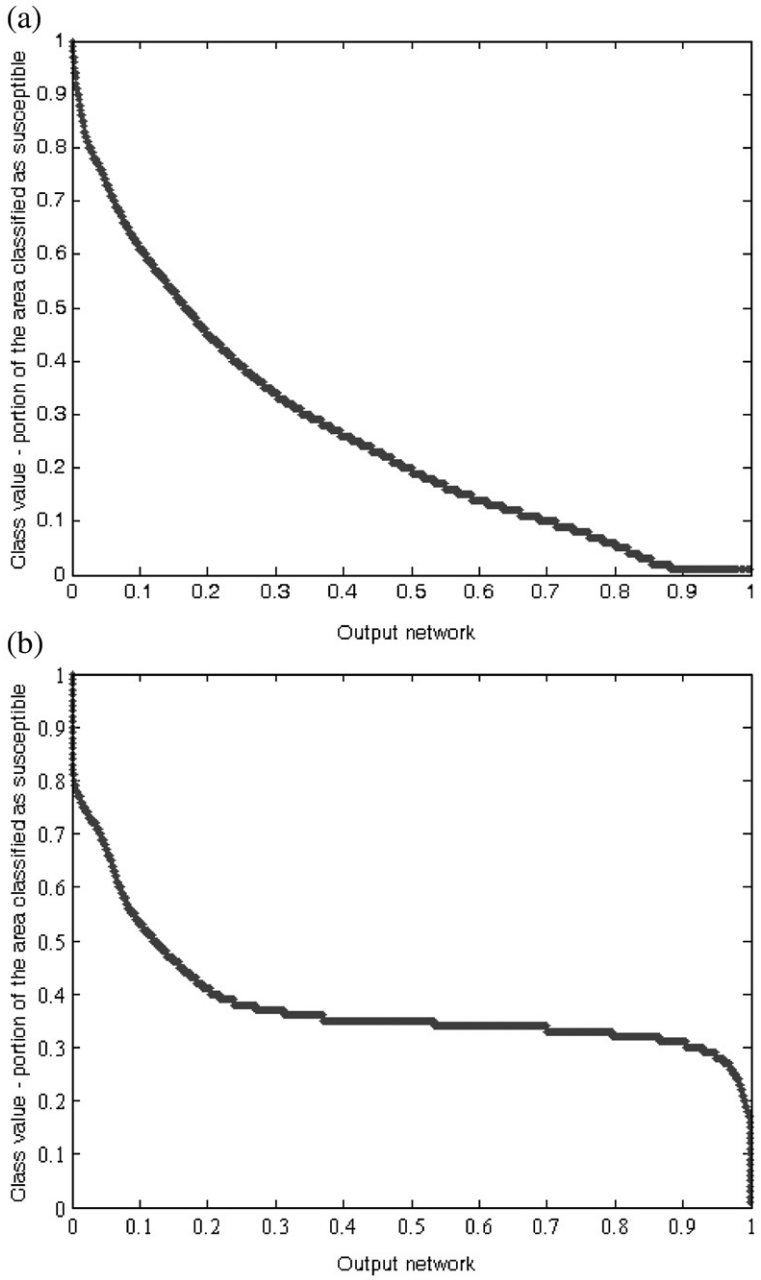

Fig. 12. The relationship between the output of the network and the susceptibility classes after random sampling (a) and after cluster sampling. In the random sampling — ANNs model some of the susceptibility classes are defined in the uncertain region of the classifier.

underlines once more that the neural classifier trained after random sampling, is able to distinguish only the stable areas. In contrast, the frequency after cluster sampling (Fig. 9b) shows values close to 0 or 1 and the uncertain regions contain only a few pixels. A preliminary classification of the two mean outputs in three classes, where the class 0.4–0.6 represents the uncertain region of the classifier, shows that the difference in the uncertain area size between the random sampling (Fig. 10a) and the cluster sampling map (Fig. 10b) is evident. Looking at the map in Fig. 10, one could think that the neural classifier trained after cluster sampling, only shifts the output to higher values. In order to verify the shifting of the output values, we have chosen two networks (one trained after cluster sampling, one after random sampling) and we have analysed the simulation (output values) of these networks for the study area. Two susceptibility maps were produced, first ordering the output values in increasing order and then ranking the values in classes with the same number of pixels. The 0–0.1 class includes the 10% of the pixels with highest mean output value. The 0.1–0.2 class includes the following 10% more susceptible, and so on. We obtained two susceptibility maps, one for the random sampling-ANN model, shown in Fig. 11a, the second one for the cluster sampling-ANN model, shown in Fig. 11b. At each pixel, we have two values: the output of the network and the class value. Fig. 12 shows the scatter plot of these pairs, each point represents one pixel in the study area. The curves in Fig. 12 express the relationship between the original output values and the defined classes for the random sampling-ANN model (a) and for the cluster sampling-ANN model (b). The differences in the susceptibility classes of the two maps (Fig. 11) prove that the neural classifier does not shift the output to higher values, but it weighs the input variables in a different way. Since the sensitivity is highly increased (Table 5), it is possible to conclude that the cluster sampling-ANN model is better to distinguish and classify landslide-prone areas.

6. Conclusions and discussions

Considering the satisfactory results achieved with MLP networks, we introduced an integrated use of supervised and unsupervised techniques to improve the results of neural classifiers. Moreover, the use of a domain-specific distance introduces expert knowledge to the black-box neural models which would allow extension of the methodology to different study areas with different conditions of mass movements. In this contribution particular attention was given to the choice of the samples to use in a landslide susceptibility model. We applied a cluster sampling method before performing the analysis by means of ANNs. However, it is important to underline that such a method can be used for other statistical techniques (e.g., discriminate analysis, logistic regression, etc.), which need two sample groups to separate and to classify cases into landslide and non-landslide groups. We have described a possible way to sample from unlabelled data that would be applicable in other case studies: the proposed method requires an expert knowledge of sliding conditions and a preliminary analysis of the domains of each conditioning factor.

We have demonstrated that an accurate sampling strategy outperforms random sampling, when training a landslide classifier. In fact, sensitivity of classification without cluster sampling was not sufficient for solving the stable area classification problem and the sensitivity has been clearly increased after performing CA.

Moreover, the CA and the possibility to choose the distance measure make it possible to introduce expert knowledge to a black-box model such as the neural one. In fact, the user can condition and strongly control the selection of the data by the use of cluster sampling. Even without looking inside the black-box model, it is thus possible to achieve a better definition of the landslide conditions by weighting the relative importance of each conditioning factor.

Finally, the most robust susceptibility map was obtained after the cluster sampling, since the cluster sampling-ANN model is able to distinguish and separate the unstable areas and thus to identify more reliable susceptibility classes. Although the results are encouraging, the model output and the discrimination of unstable areas can be improved. In the analysis some variables (e.g, bedding dip domain, distance from drainage lines, geomechanical properties, etc) have not yet been considered. The introduction into the dataset of these variables could improve the classification of unstable areas, since the cluster formation and then the negative case selection would be influenced. In fact, an increased number of conditioning factors would influence the discrimination and the selection of cluster meaningfully for the instability classification problem.

Learning with only positive labelled data is a quite sensible issue in machine learning and quantitative approaches to modelling in general. This issue has been already faced in the literature especially in the text classification field. Techniques generally used in that field have been extensively tested in the case of text document classification (Nigam et al., 1998; Li and Liu, 2003; Liu et al., 2003). The main idea behind those techniques is to identify (guess) a set of reliable negative samples from the unlabelled dataset, based on the supposition that the unlabelled set contains a small

number of positive examples and a large number of negative examples (i.e., as in the case of landslide-susceptibility zonation). After an initial guess has been made, the classifier is built by iteratively applying a classification algorithm to the positive and reliable negatives and then refining this partitioning once a preliminary model has been defined. This approach, to our knowledge, has never been applied in contexts different from text classification. It is quite different from the approach we proposed in this paper since the set of positive and negative examples change during the learning process and vary both in size and composition. This refining procedure of selected data should be a future direction of investigation to obtain better results in landslide-susceptibility zonation.

Acknowledgements

This work was realized in the framework of the Dottorato di Ricerca in Scienze Ambientali with an Italian Phd scholarship. Fieldwork was supported with MIUR 40% grants. The authors thank the Local Authority of Brembilla Municipality for its kindness in making the data available. D. Alexander, G.B. Crosta, and P. Frattini are also thanked for their reviews. We are indebted to the editor in chief of the volume, A. Harvey, for his final suggestions.

References

Anbalagan, R., 1992a. Landslide hazard evaluation and zonation mapping in mountainous terrain. Engineering Geology 32, 269–277.

Anbalagan, R., 1992b. Terrain evaluation and zonation mapping in mountainous terrain. Engineering Geology 32, 269–277.

Ardizzone, F., Cardinali, M., Carrara, A., Guzzetti, F., Reichenbach, P., 2001. Impact of mapping errors on the reliability of landslide hazard map. Natural Hazard and Earth System Sciences 2, 3–14.

Azzoni, A., Agliardi, F., 2004. Note illustrative del piano geologico del Comune di Brembilla.

Binaghi, E., Luzi, L., Madella, P., Pergalani, F., Rampini, A., 1998. Slope instability zonation: a comparison between certainty factor and fuzzy Dempster-Shafer approaches. Natural Hazards 17, 77–97.

Bishop, C.M., 1995. Neural Networks for Pattern Recognition. Oxford University Press, Oxford. 482 pp.

Bonham-Carter, G., 1994. Geographic Information Systems for Geoscientists: Modelling with GIS. Pergamon, New York. 398 pp.

Burges, C.J.C. (Ed.), 1998. A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery, vol. 2. Kluwer Academic Publishers, Hingham, MA, USA, pp. 121–167.

Carrara, A., 1983. Multivariate models for landslide hazard evaluation. Mathematical Geology 15, 403–426.

Carrara, A., Guzzetti, F., Cardinali, M., Reichenbach, P., 1999. Use of GIS technology in the prediction and monitoring of landslide hazard. Natural Hazards 20, 117–135.

Carrara, A., Crosta, G.B., Frattini, P., 2003. Geomorphological and historical data in assessing landslide hazard. Earth Surface Process and Landforms 28, 1125–1142.

Caruana, R., Lawrence, S., Giles, C.L., 2000. Overfitting in neural nets: backpropagation, conjugate gradient, and early stopping. Proceedings of Neural Information Processing Systems. Denver, Colorado, USA, pp. 402–408.

Chung, C.F., Fabbri, A.G., 1999. Probabilistic prediction models for landslide hazard mapping. Photogrammetric Engineering & Remote Sensing 65, 1389–1399.

Ermini, L., Catani, F., Casagli, N., 2005. Artificial neural network to landslide susceptibility assessment. Geomorphology 66, 327–343.

Forcella, F., Jadoul, F., 2000. Carta geologica della Provincia di Bergamo a scala 1:50.000, Grafica Monti, Bergamo.

Gomez, H., Kavzoglu, T., 2005. Assessment of shallow landslide susceptibility using artificial neural networks in Jabonosa River Basin, Venezuela. Engineering Geology 78, 11–27.

Hagan, M.T., Menhaj, M., 1994. Training feed forward networks with the Marquardt algorithm. IEEE Transactions on Neural Networks 5, 989–993.

Hartigan, A., Wong, M.A., 1978. A *k*-means clustering algorithm. Applied Statistics 28, 100–108.

Haykin, S., 1999. Neural Networks, A Comprehensive Foundation. 2nd Edition. Prentice Hall, Englewood Cliffs, New Jersey, USA. 696 pp.

Heuvelink, G.B.M., 1993. Error Propagation in Quantitative Spatial Modelling: Application in Geographical Information Systems. Gedrukt door Drukkerij Elinkwijk, Utrecht. 151 pp.

Hornik, K.M., Stinchcombe, M., White, H., 1989. Multilayer feedforward networks are universal approximators. Neural Networks 2, 359–366.

Jadoul, F., Masetti, D., Cirilli, S., Berra, F., Claps, M., Frisia, S., 1994. Norian–Raethian stratigraphy and paleogeographic evolution of the Lombardy Basin (Bergamasc Alps). In: Carannante, G., Tonielli, R. (Eds.), 15th IAS Regional Meeting, April 94, Ischia, Italy. Post-meeting fieldtrip guidebook, pp. 5–38.

Kaufman, L., Rousseeuw, P.J., 1990. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York. 338 pp.

Lee, S., Cho, S., Wong, P.M., 1998. Rainfall prediction using Artificial Neural Network. Journal of Geographic Information and Decision Analysis 2, 233–242.

Lee, S., Lee, M., Yu, Y., 2003a. Quantitative analysis of landslide susceptibility: a case study of Korea. Geophysical Research Abstract 5, 04923.

Lee, S., Ryu, J., Min, K., Won, J., 2003b. Landslide susceptibility analysis using G.I.S. and artificial neural network. Earth Surface Processes and Landforms 28, 1361–1376.

Lee, S., Ryu, J., Won, J., Park, H., 2004. Determination and application of weights for landslide susceptibility mapping using an artificial neural network. Engineering Geology 71, 289–302.

Li, X., Liu, B., 2003. Learning to classify texts using positive and unlabeled data. Proceedings of International Joint Conference on Artificial Intelligence (IJCAI-03), Acapulco, Mexico, pp. 587–594.

Liu, B., Dai, Y., Li, X., Lee, W.S., Yu, P., 2003. Building text classifiers using positive and unlabeled examples. Proceedings of International Conference on Data Mining (ICDM), Melbourne, Florida, USA, pp. 179–188.

Lu, P., Rosenbaum, M.S., 2003. Artificial neural network and grey system for the prediction of slope stability. Natural Hazards 30, 383–398.

MacQueen, B.J., 1967. Some methods for classification and analysis of multivariate observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, vol. 1. University of California Press, Berkeley, California, pp. 281–297.

Marquardt, D., 1963. An algorithm for least-squares estimation of nonlinear parameters. SIAM Journal of Applied Mathematics 11, 431–441.

Montana, D.J., Davis, L., 1989. Training feed forward neural networks using genetic algorithms. Proceedings of the Third International Conference on Genetic Algorithms 3, 762–767.

Nigam, K., McCallum, A., Thrun, S., Mitchell, T., 1998. Learning to classify text from labeled and unlabeled documents. Proceedings of 15th National Conference on Artificial Intelligence (AAAI), Madison, Wisconsin, USA, pp. 792–799.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., 1992. Numerical Recipes in C: The Art of Scientific Computing. University Press, Cambridge, UK. 994 pp.

Quinlan, J.R., 1986. Induction of decision trees. Machine Learning 1, 81–106.

Regione Lombardia, 1991. Cartografia Geoambientale a scala 1:10000. Carta dell'uso del suolo ad orientamento vegetazionale.

Romesburg, H.C., 1984. Cluster Analysis for researchers. Lulu Press, Belmont, North Carolina. 334 pp.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagating errors. Nature 323, 533–536.

Tarboton, D.G., 1997. A new method for the determination of flow direction and upslope areas in grid digital elevation models. Water Resources Research 33, 309–319.