A Supervised Machine Learning Spatial Tool for detecting terrain deformation induced by landslide phenomena.

Paraskevas Tsangaratos[1], Ioanna Ilia[2]

[1] *Ph.D, Mining Engineering, National University of Athens, School of Mining and Metallurgical Engineering, Department of Geology, Laboratory of Engineering Geology and Hydrogeology, Iroon Polytechneiou 9, 15780, Zografou, Greece, ptsag@metal.ntua.gr*

[2] *Ph.D, M.Sc, Geologist, National University of Athens, School of Mining and Metallurgical Engineering, Department of Geology, Laboratory of Engineering Geology and Hydrogeology, Iroon Polytechneiou 9, 15780, Zografou, Greece, gilia@metal.ntua.gr*

Abstract

The present study uses Google Earth's imagery, a library of high resolution satellite imagery and aerial photography of the entire Earth's surface to extract spectral and morphometric information in order to detect large terrain deformation and produce a landslide inventory map. The developed supervised machine learning spatial tool for landslide detection was verified by experimentation on a Google Earth Images sequence (16 images) that covered a watershed area with numerous recorded landslide incidences in the municipality of Kimi, Euboea Island, Greece. The evaluation of the outcomes of the performed classification showed satisfactory results, with the overall accuracy presented to exceed 88%.

Keywords: Supervised machine learning spatial tool, Naïve Bayes algorithm, Landslides.

1. Introduction

One of the most critical issues that arise while trying to couple with the problem of landslide manifestation is the identification and classification of landslide prone areas. Specifically, the detection of landslide prone areas and the ability to provide landslide specific information to policy makers and emergency managers is the most essential element in urban development scenarios (Tralli et al. 2005). Visual interpretation of aerial photographs, combined with field investigations, is the major source for landslide inventory map preparation. The approach for recognition and classification of landslides is mainly derived from the knowledge developed by experts for detection of landslides during image interpretation of aerial photographs. However, aerial photographs are often not available in a timely manner for the majority of landslide prone areas. On the other hand, satellite imagery has become an alternative data source since it allows a more economic assessment of larger landslide affected areas.

In the last decade, there has been an increase in the number of applications that use pixel-based and object-based image classification techniques along with Remote Sensing technology for fast retrieval of information that concern natural hazards (Martha et al. 2010). Real time integration of remotely sensed imagery and Geographical Information System (GIS) data has been carried out with expert knowledge in hazard mapping, wildfire monitoring, and crop disease surveillance since the 1990 (Kimes et al. 1991; Zerger 2002; Wan and Lei 2009; Pradhan and Lee 2010; Wan et al. 2010). Recent studies have made use of very high resolution imagery (QuickBird, Ikonos, WorldView-1, Cartosat-1-2, SPOT-5 and ALOS-PRISM) and other Remote Sensing techniques used in landslide analysis that include shaded relief images produced from LiDAR and Synthetic Aperture Radar (SAR) interferometry based DEMs (Moine et al. 2009; Martha et al. 2012).

In this direction, the present study uses Google Earth's imagery, a library of high resolution satellite imagery and aerial photography of the entire Earth's surface to extract spectral and morphometric information such as hue, saturation, value (HSV color space), slope inclination, flow direction, terrain curvature, plan and profile, in order to detect large terrain deformation and produce a landslide inventory map. The phase of detection was achieved by the

Ψηφιακή Βιβλιοθήκη Θεόφραστος - Τμήμα Γεωλογίας. Α.Π.Θ.

implementation of a supervised machine learning spatial tool, which utilized a pixel – based technique that embodies a Naive Bayes classifier (NB). The main idea behind the developed tool is to exploit the expert knowledge during the training phase in order to determine the prior probabilities of an area to be classified into a specified category. The categories that can be discriminated by the developed classifier include water-bodies, streams, road network, urban settlements, grassland, landslide area and bare land (rock formations). The seven (7) parameters that are related to spectral and morphometric information of a target pixel and its eight (8) surrounding pixels were used for classification. Having gained this knowledge the algorithm evaluates each pixel of the image and produces an outcome based on the comparison between the calculated probabilities for each possible category that the area could be classified. The pixel would be classified to the class that corresponds to the higher probability.

2. Study area and data used

2.1 Study area

Geographically, Euboea Island is located eastern of Attica – Athens and extends along the eastern coast of Greece covering an area of 167.6 km^2 (Fig. 1). The watershed cover an area of about 14.68km^2 bounded by the hills of Koutsiko (515m), Mesiani Rachi (640m) at southwest and Mesovouni at the east. The morphological relief of the entire basin is characterized as intense with slopes ranging between 40^o to 50^o. Milder slopes, about 20^o, are present at the main section of the basin, between the town of Kimi and the site Patero. The stream network has been formed by the tectonic activity of the wider area. Two secondary networks with East-West direction contribute to the main stream that has a North-South direction. The Apoulistis River has a significant flow during the rain session, which remains constant for a longer time as it is recharged by a number of springs that are located in the intense fragmented zones of the flysch formations (Ilia 2013).

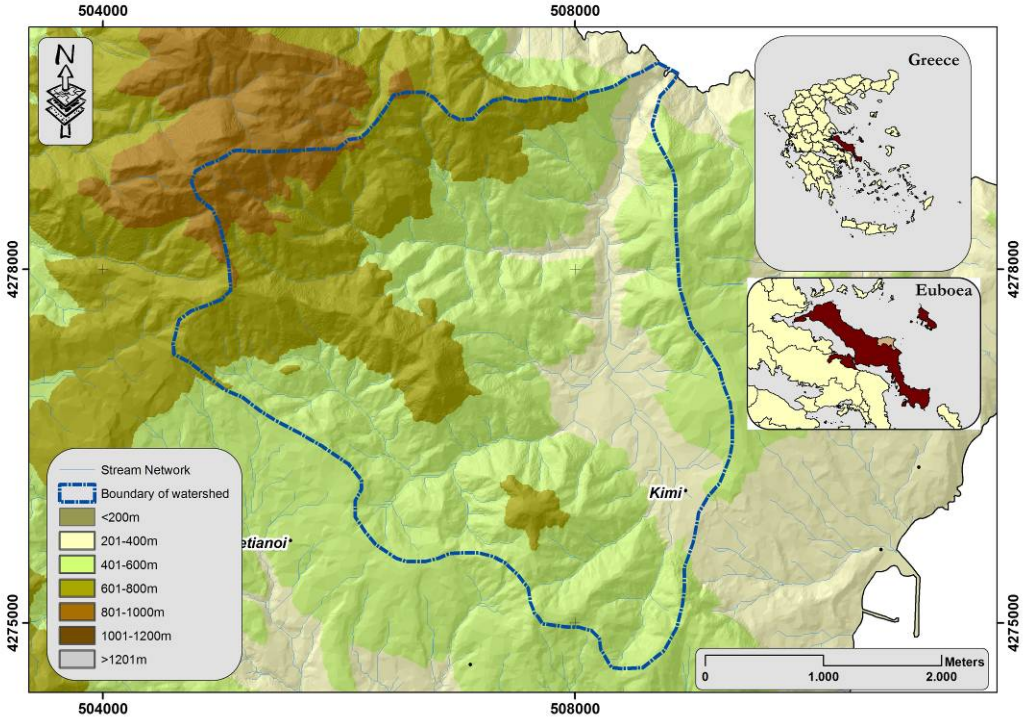

Figure 1 Study area

In the study area, quaternary deposits, upper horizons of neogene sediments, base conglomerate and flysch formation are present. Also, peridotites, dounits and serpentine peridotites outcrop in the eastern part of the examined area, while carbonate rocks are presented with Mesozoic and Paleozoic limestones. Granites Carboniferous in age covers the

Ψηφιακή Βιβλιοθήκη Θεόφραστος - Τμήμα Γεωλογίας. Α.Π.Θ.

Western part of Kimi basin. Concerning the climatic features the average annual amount of rainfall is 1071mm, whereby in highland zones of the region, the annual amount of rainfall is expected to be higher. Concerning the seasonal distribution of rainfall, a rate of 48.5% of the average total annual amount is attributed to the winter, followed by 27.4% in autumn, 20.3% during the spring and 6.8% during the summer.
Landslides in the area of Kimi, Euboea, are distributed in several locations mainly due to the general geotechnical behaviour of the geological formations and also in most cases due to human activities (Ilia et al. 2010; Tsangaratos et al. 2013; Ilia 2013). Several landslides resulted in the destruction of the main coastal road network as well as in the collapsing of several structures, causing disastrous socio-economic implications. With regard to the landslide type, the study area is affected by creep, rotational and translational landslides, lateral spreads and rock falls. Most of them occurred on steep slopes after heavy rainfall or tectonic activity. The spatial distribution of landslides was captured in a landslide inventory map using information of field surveys during the period 2006 – 2012 along with the combined knowledge obtained from the evaluation of aerial photos, satellite images and previous studies. The field surveys resulted in identifying a total amount of 21 landslide event locations within the boundaries of the watershed (Ilia 2013).

2.2 Data

The high resolution satellite data and digital elevation models are the main requirements in order to sufficiently model the process of landsliding. The images that were used for training were captured from the developed supervised machine learning spatial tool that embodies the Google Earth API. The Google Earth Plug-in and its JavaScript API is a free programmable Interface that provides a library of satellite imagery and aerial photography of the entire Earth's surface, thus supplying integrated coverage and monitoring images. Google Earth uses digital elevation model (DEM) data collected by NASA's Shuttle Radar Topography Mission (SRTM) where most land areas are covered in satellite imagery with a resolution of about 15 m per pixel (Farr et al. 2007). The base imagery of the research area is a 30 m multispectral Landsat which is pan-sharpened with the 15 m (panchromatic) Landsat imagery. The tool is also capable of extracting altitude information and calculate the slope inclination, flow direction, curvature, plan and profile of each area in focus (Tsangaratos 2011).

3. Detecting terrain deformations

The developed methodology for the detection of terrain deformations and particularly landslides is mainly derived from expert knowledge. The approach for landslide recognition is described in the following phases (Fig. 2).

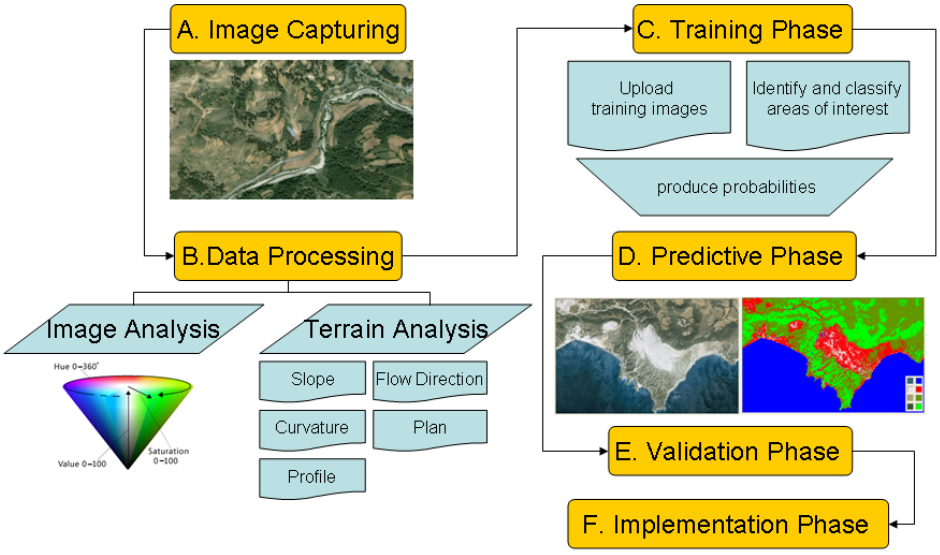

Figure 2 Flow chart of the developed methodology

3.1 Image Capturing

In the first phase, the researcher captures the research area, which automatically is separated by the tool into smaller images, 1200x1200 pixels and 200 dpi. Each training image should contain a sufficient number of spatial features that could be identified. In the present study the entire image set of sixteen (16) images. Four (4) images where used for training the algorithm and to calculate the spectral characteristics of each category and one (1) image for validating. The rest eleven (11) images where used to implement the algorithm.

3.2 Data Processing

During the phase of data processing and specifically the stage of image analysis, all the selected regions are converted from RGB (Red-Green-Blue) color space to HSV (Hue-Saturation-Value) space pixel by pixel. The tool automatically obtains the elevation data and stores them in the database. The elevation data are used to create the appropriate morphological variables that are needed, such as slope inclination, flow direction, curvature, plan and profile.

3.3 Training Phase

The next phase is for the expert to identify the spatial features, such as the water-bodies, streams, road network, urban settlements, grassland, bare land and landslide areas, using four (4) training images while the tool automatically applies to the spatial features the HSV and the topographical aspect values. When this process of identification is finished the tool creates 2-dimensional charts of Hue-Saturation in order to evaluate the discrimination power of the tool.

3.4 Predictive Phase

The categories discriminated by the developed classifier include water-bodies, streams, road network, urban settlements, grassland, landslide area and bare land (rock formations). To classify each pixel into one of the categories, seven parameters related to the values from HSV color space and the topographic aspects of a target pixel and its eight surrounding pixels were used. The developed spatial tool uses a naive Bayes classifier to determine if the given specifications (a set of independent variables) results in a specific category (dependent variables). A naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions (Soria et al. 2011). A naive Bayes classifier considers each of these features to contribute independently to the probability regardless of the presence or absence of the other features. Given an observation consisting of k attributes x_i, (i = 1,2,..,k), xi is the conditioning factor (hue, saturation, value, slope inclination, flow direction, terrain curvature and stream network data), and y_j, (j = 1 – water bodies, 2 - streams, 3 - road network, 4 - urban settlements, 5 - grassland, 6 - landslide area and 7 - bare land) the output class. The prediction is made for the class with the highest posterior probability as indicated in the following equation:

$$y_{NB} = \arg\max P(y_j) \prod_{i=1}^{k} P(\frac{x_i}{y_j}) \qquad (1)$$

$$y_j \in \{1,2,..k\}$$

The prior probability $P(y_j)$ can be estimated using the proportion of the observations with output class y_i in the training dataset. The conditional probability is calculated in the present study assuming that the values associated with each class are distributed according to a Gaussian distribution:

$$P\left(\frac{x_i}{y_j}\right) = \frac{1}{\sqrt{2\pi std}} e^{\frac{-(x_i-\mu)^2}{2std^2}} \qquad (2)$$

μ= mean, std = standard deviation of x_i

3.5 Validation Phase

The validation phase involves the estimation of the predictive performance of the tool, using as input values the areas that were classified by the expert during the training phase. Specifically, for supervised learning with two (2) possible classes, all measures of performance are based on four (4) numbers obtained from applying the classifier to the test set (Tsangaratos and Benardos 2014). These numbers are called true positives (tp), false positives (fp), true negatives (tn), false negatives (fn). A sample is characterized as true positive when it is estimated to belong to the i^{th} class and it truly belongs. A sample is characterized as false positive when the sample is estimated to belong to the i^{th} class but it truly does not belong. A sample is characterized as true negative when the sample is estimated not belonging to the i^{th} class and it truly does not belong to it. Finally a sample is characterized as false negative when the sample is estimated not belonging to the i^{th} class but it truly does belongs to it. Depending on the application, many different summary statistics are computed from these entries. In particular: accuracy, precision, recall, and F-m index (Tab. 1).

Table 1 – Validation Index

Index	Equation	Description
Accuracy	$accuracy = \frac{t_p + t_n}{t_p + f_p + t_n + f_n}$	Defined as the proportion of true results (both true positives and true negatives) in the population.
precision	$precision = \frac{t_p}{t_p + f_p}$	Defined as the number of positive instances retrieved over the total number of instances declared positive by the classifier
recall	$recall = \frac{t_p}{t_p + f_n}$	Defined as the number of true positive instances retrieved over the total number of instances that are positive in the set
F-measure	$F - m = \frac{2t_p}{2t_p + f_p + f_n}$	F-measure, a single measure combining precision and recall, can be used as an efficient single-valued metric

3.6 Implementation Phase

The final phase is the implementation phase where each image of the database is uploaded and classified according to the generated rules. The final product is a geo-referenced classified map. The algorithm performs well, showing robustness against confusion with other zones that are close in colour. Because of its ability to learn from samples, it's capable of improving its estimation percentage with the increase of the number of the samples. Figure 3a illustrates a sample training image. The following figures, 3b, 4a, 4b, 5a, and 5b are the slope inclination, flow direction, curvature, plan and profile layers calculated by the Terrain Analysis procedure. Figure 6 shows the outcome of the combined analysis (Image Analysis and Terrain Analysis) with only four (4) features of interest, landslide areas, grassland, road network and bare land.

Ψηφιακή Βιβλιοθήκη Θεόφραστος - Τμήμα Γεωλογίας. Α.Π.Θ.

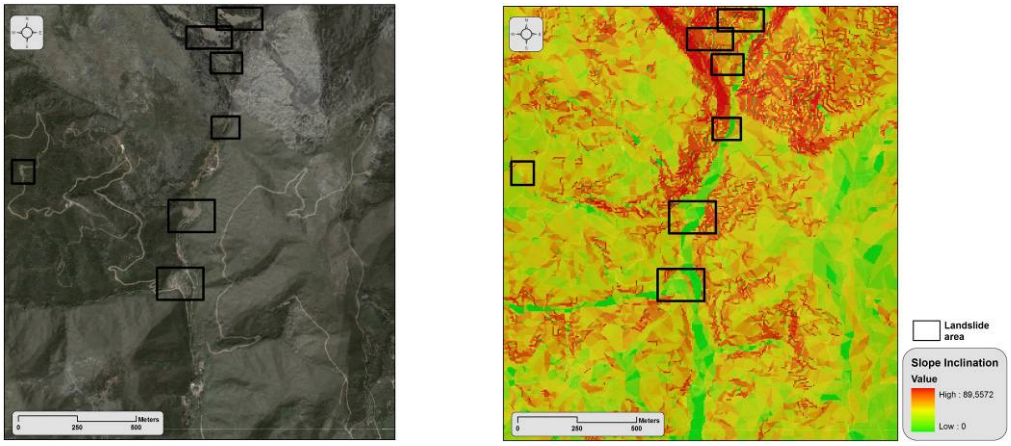

Figure 3 (a) training area (b) slope inclination

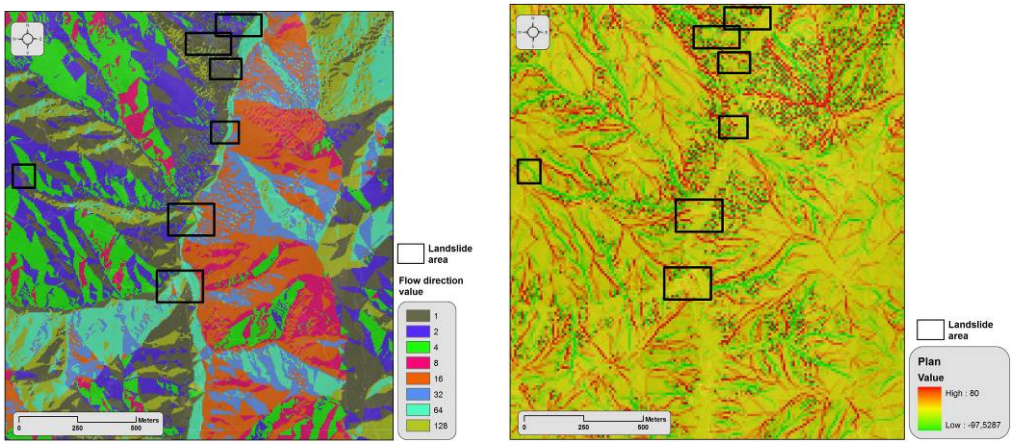

Figure 4 (a) flow direction (b) plan

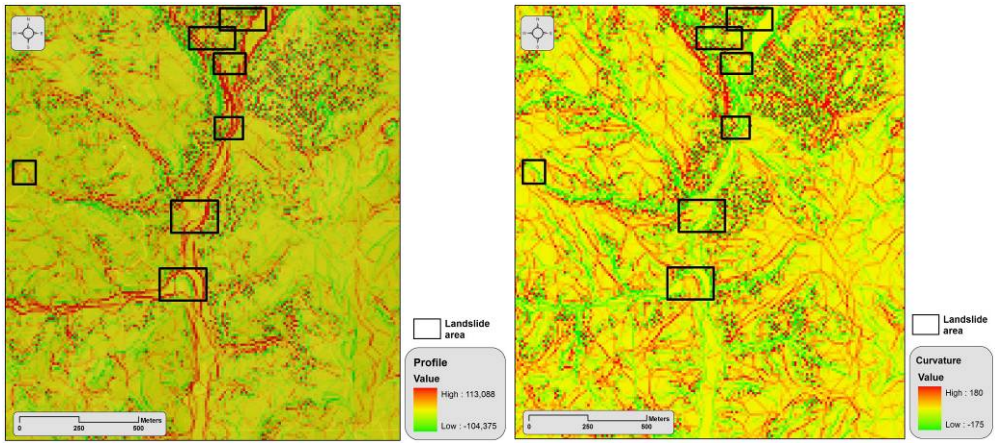

Figure 5 (a) profile (b) curvature

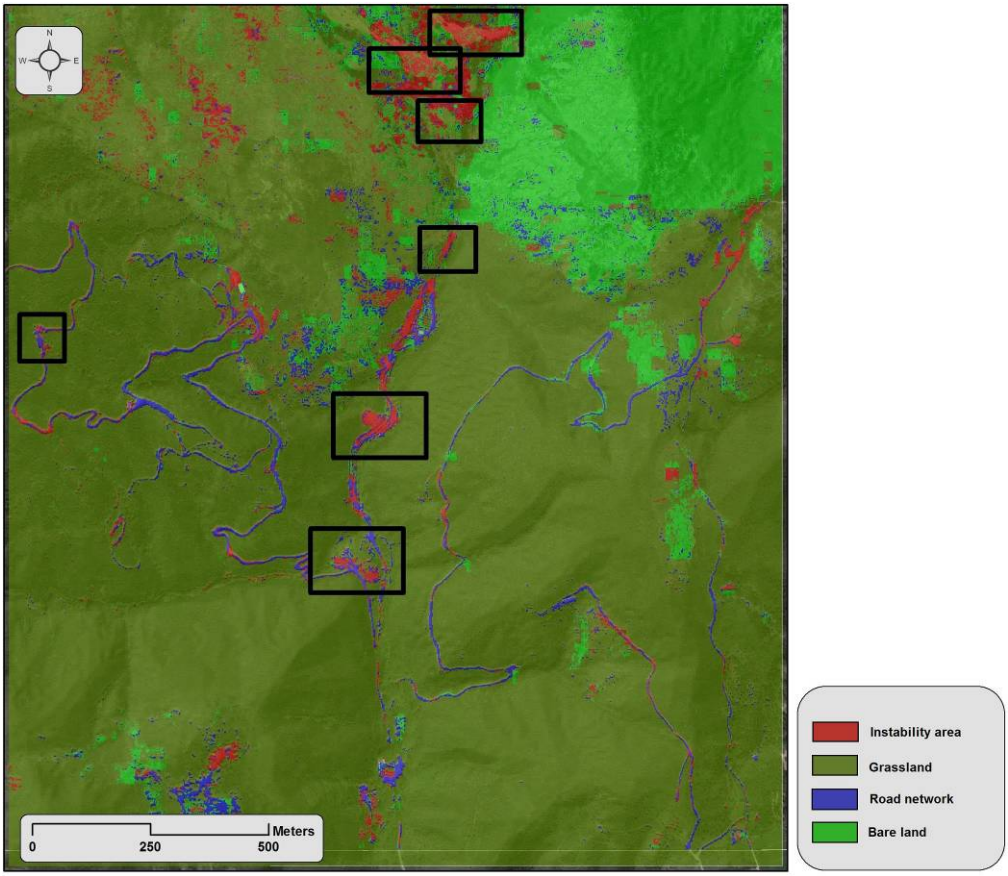

Figure 6 Final Classification of the first training image

4. Results

Table 2 shows the results obtained during the validation phase and particular the accuracy, precision, recall and f-measure indexes. According to the methodology the validation was performed on data obtained from the four (4) training images that were classified according to expert knowledge.

Table 2 – Validation Index

Categories	tp	tn	fp	fn	accuracy	precision	recall	F-measure
1 – water bodies	21	3	3	1	0.8571	0.8750	0.9545	0.9130
2 – streams	24	1	1	2	0.8929	0.9600	0.9231	0.9412
3 – road network	20	3	2	3	0.8214	0.9091	0.8696	0.8889
4 – urban settlement	24	1	3	0	0.8929	0.8889	1.0000	0.9412
5 – grassland	24	2	2	0	0.9286	0.9231	1.0000	0.9600
6 – bare land	26	1	1	0	0.9643	0.9630	1.0000	0.9811
7 – landslide area	22	2	2	2	0.8571	0.9167	0.9167	0.9167
Average					0.8878	0.9194	0.9520	0.9346

For each class twenty eight (28) points where used, totaling one hundred and ninety six (196) validation points, appointed by the expert from the four (4) training images. The tool showed an overall accuracy equal to 88.78%, precision equal to 91.94%, recall equal to 95.20% and f-measure equal to 93.46%. The highest index values were obtained in the class of bare land (rock formations) since it presented homogeneity through the entire watershed. The lowest index values were observed in the class of road network. It may appears that the road network needed to be classified into to two (2) subsets since the tool didn't distinguished paved and

unpaved road network. Concerning landslide areas the tool performed an 85.71% accuracy that is satisfactory.

The twelve (12) images that have not been used for training and also the four (4) training images that covered the watershed were introduced to the spatial tool in order to produce the final geo-referenced classified map. Figure 7 illustrates the research area, the total of the sixteen (16) images, while figure 8 the final geo-referenced classified map produced by the tool.

It can been observed that the produced map classified correctly seventeen (17) out of twenty one (21) landslide areas that have been identified in previous studies in the 12 images that have not been used for training, however it also classified several additional areas as landslide areas. A careful investigation of those sites revealed that there was a high probability that those sites may encounter instability problems, while others where misclassified (they were inclined cultivated areas). If geological information was included in the analysis the discrimination power of the classifier could be improved concerning the landslide areas classes.

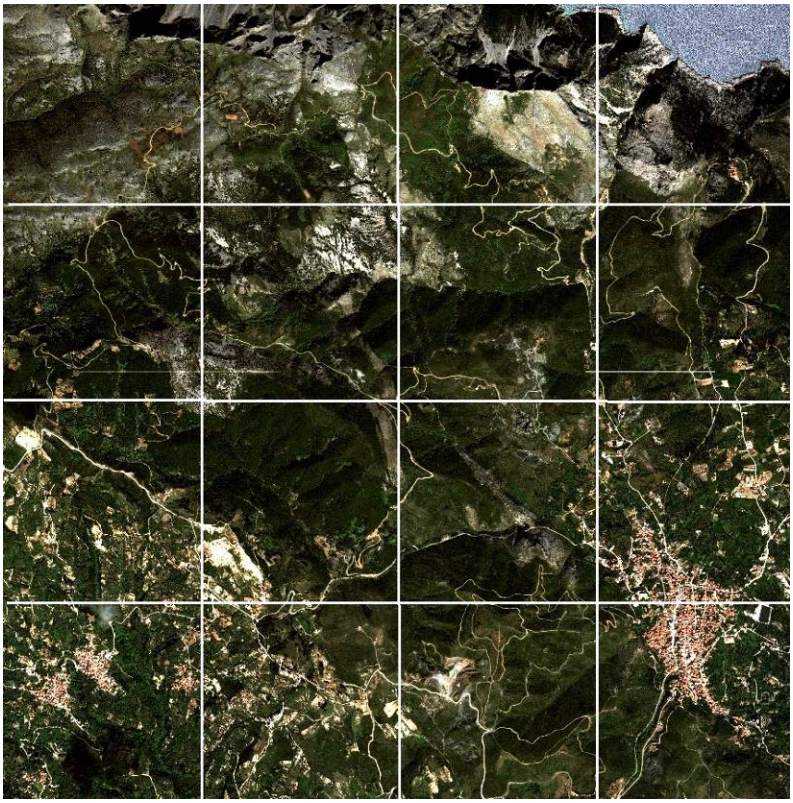

Figure 7 Research area

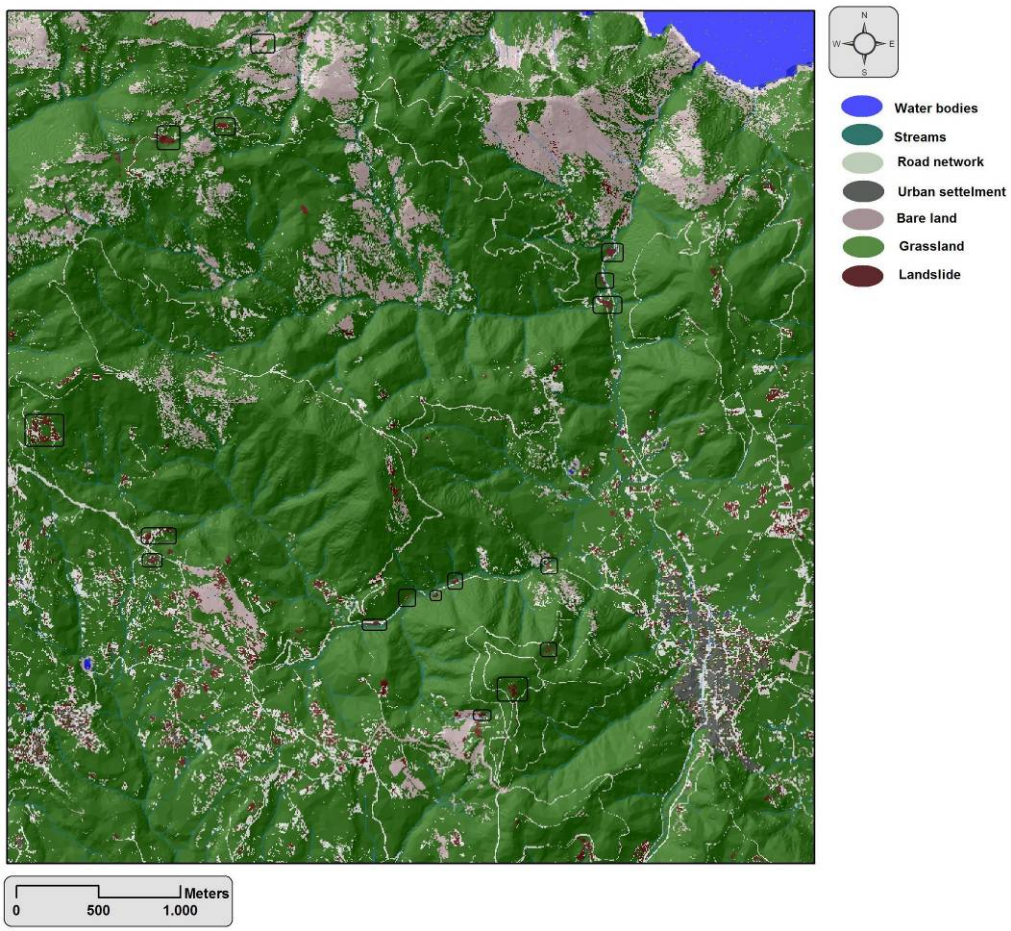

Figure 8 Final Classification Map

5. Discussion and Conclusions

In this study a supervised machine learning spatial tool which utilized a pixel – based technique that embodies a Naive Bayes classifier (NB). The tool exploits the expert knowledge during the training phase in order to determine the prior probabilities of an area to be classified into a specified category. The categories that can be discriminated by the developed classifier include water-bodies, streams, road network, urban settlements, grassland, bare land (rock formations) and landslide area. Seven (7) parameters related to spectral information and morphometric information of a target pixel and its eight (8) surrounding pixels were used for classification. Having gained this knowledge the algorithm evaluates each pixel of the image and produces an outcome based on the comparison between the calculated probabilities for each possible category that the area of interest could be classified. The pixel would be classified to the class that corresponds to the higher probability. The expert identified the required sub region from the images that were included into the database during the training phase. The method depends on collecting a large number of training images in order to classify to specific class having as an indicator color and topographic variations. An advantage of the naive Bayes classifier is that it requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. The results of validation procedure demonstrated that developed tool utilized a robust algorithm that was effective and accurate despite the poor RGB colors of Google Earth Imagery and inaccurate elevation data. The overall accuracy was estimated to be equal to 88.78%.

The developed supervised machine learning spatial tool could assist local authorities and government agencies as an effective tool for identifying terrain deformations and instability areas with limited resources.

Ψηφιακή Βιβλιοθήκη Θεόφραστος - Τμήμα Γεωλογίας. Α.Π.Θ.

6. References

Farr, GT., Rosen P., Caro, E., Crippen R., Duren, R., Hensley, S., Kobrick, M., Paller, M., Rodriguez, E., Roth, L., Seal, D., Shaffer, S., Shimada, J., Umland, J., Werner, M., Oskin, M., Burbank D., Alsdorf, D., 2007. The Shuttle Radar Topography Mission, Reviews of Geophysics, vol. 45 (2), doi: 1029/2005RG000183.

Ilia, I. 2013. Engineering geological features of marls in the wide area of Kimi, Euboea, their impact on construction problems and their treatment, PhD Thesis, National Technical University of Athens, School of Mining and Metallurgical Engineering, Athens, Greece, p.328.

Ilia, I., Tsangaratos, P., Koumantakis, I., Rozos, D. 2010. Application of a Bayesian approach in GIS-based model for evaluating landslide susceptibility. Case study Kimi area, Euboea, Greece. Bulletin of the Geological Society of Greece, vol.3, pp.1590-1600.

Kimes, D.S., Harrison, P.R., Ratcliffe, P.A., 1991. A knowledge-based expert system for inferring vegetation characteristics. Int. J. Remote Sens., vol.12, pp.1987-2020.

Martha, T.R., Kerle, N., Jetten, V., van Westen, C.J. and Vinod Kumar, K., 2010. Characterising spectral, spatial and morphometric properties of landslides for automatic detection using object-oriented methods. Geomorphology, vol. 116(1-2), pp.24-36.

Martha, T.R., Kerle, N., van Westen, C.J., Jetten, V. and Vinod Kumar, K., 2012. Object-oriented analysis of multi-temporal panchromatic images for creation of historical landslide inventories. ISPRS Journal of Photogrammetry & Remote Sensing, vol. 67, pp.105-119.

Moine, M., Puissant, A. and Malet, J.-P., 2009. Detection of landslides from aerial and satellite images with a semi-automatic method-Application to the Barcelonnette basin (Alpes-de-Haute-Provence, France). In: J.-P. Malet, A. Remaitre and T. Bogaard (Editors), Landslide Processes: from geomorphological mapping to dynamic modelling. CERG, Strasbourg, France, pp. 63-68.

Pradhan, B., Lee, S., 2010. Landslide susceptibility assessment and factor effect analysis: back propagation artificial neural networks and their comparison with frequency ratio and bivariate logistic regression modeling. Environmental Modeling & Software, 25(6), 747–759.

Soria D., Garibaldi M., Ambrogi E., Biganzoli M., Ellis I.O., 2011. A non parametric version of the naive classifier, Knowledge – Based Systems, vol. 24(6), pp. 775-784.

Tralli, D.M., Blom, R.G., Zlotnicki, V., Donnellan, A. and Evans, D.L., 2005. Satellite remote sensing of earthquake, volcano, flood, landslide and coastal inundation hazards. ISPRS Journal of Photogrammetry and Remote Sensing, vol.59(4): pp.185-198.

Tsangaratos P, Ilia I, Rozos D, 2013. Case Event System for landslide susceptibility analysis. Landslide Science and Practice, editors Margottini, Canuti, Sassa, Springer Berlin Heidelberg, pp.585-593.

Tsangaratos P., 2011. Virtual Globes and Geological Modeling, International Journal of Geosciences, Vol. 2 No. 4, pp. 648-656. doi: 10.4236/ijg.2011.24066.

Tsangaratos, P., Benardos, A., 2014. Estimating Landslide Susceptibility through an Artificial Neural Network classifier. Natural Hazards, May 2014 published online (http://link.springer.com/article/10.1007%2Fs11069-014-1245-x.

Wan, S., Lei, T.C., 2009. A knowledge-based decision support system to analyze the Debris-Flow problems at ChenYu-LanRiver,Taiwan.Knowledge-Based Systems, vol.22, pp.580–588.

Wan, S., Lei, T.C., Chou, T.Y, 2010. An enhanced supervised spatial decision support system of image classification: consideration on the ancillary information of paddy rice area. International Journal of Geographical Information Science. DOI:10.1080/13658810802587709.

Zerger, A., 2002. Examining GIS decision utility for natural hazard risk modelling. Environmental Modelling & Software, vol.17(3), pp.287–29.

Ψηφιακή Βιβλιοθήκη Θεόφραστος - Τμήμα Γεωλογίας. Α.Π.Θ.