MODELLING THE OCCURENCE OF GULLIES IN SEMI-ARID AREAS OF SOUTHWEST SPAIN

Gómez, Á.*, Schnabel, S., Felicísimo, Á.

GeoEnvironmental Study Group, Área de Geografía Física, Universidad de Extremadura, Av. de la Universidad, Cáceres, Spain. *alvgo@unex.es

1. Introduction

Modern methods and techniques of predictive modelling have rarely been applied to predict the location of geomorphic phenomena. In this paper we use a relatively new method, *Multivariate Adaptive Regression Splines* (Friedman 1991), in order to determine the potential distribution of gullying in silvopastoral systems of southwest Spain. We selected MARS instead of Neural Networks because the results are easily interpretable and they do not suffer from black box limitations.

In these areas gully erosion can be the most important process of soil degradation together with sheet erosion (Schnabel 1997).

2. Methodology

Multivariate Adaptive Regression Splines (MARS) is a non-parametric method for modelling the response of a dependent variable (gullying) from a set of independent variables. Gullies were located in the field within a selection of 54 farms distributed in Extremadura, Spain. With the help of a GPS and aerial orthophotographs of high resolution gullies were mapped. A set of 36 different maps of the study areas were used to define the independent variables reflecting topography, climate, soils, lithology, land use and vegetation cover.

2.1. Topography

To represent the topography of the study areas Digital Elevation Models (DEM) were elaborated with a resolution of 5 m, which is the best available. Furthermore the model is developed for being applied in large areas such as Extremadura with a surface area of 41,634 km^2. The resolution of the *DEM* can be considered as the error of prediction, i.e. whether a gully exists or not. This error is considered not significant for localizing a gully in the field. Starting from this *DEM* we generated maps of slope gradient, curvature (general, profile and plan), Roughness (using three different windows of three, seven and eleven pixels), catchment area, total upslope length and longest upslope length of flow path terminating at each grid cell.

2.2. Rainfall

From a large dataset of precipitation (1960-1991) for 222 meteorological stations in the study area, we generated maps of mean annual, seasonal and monthly rainfall.

2.3. Lithology and soils

Rock type was obtained from the geological maps of the study area (scale 1:50,000). Two soil maps using the FAO-WRB classification system were digitalized. (1:300,000).

2.4. Land use and vegetation cover

Five variables were used to represent land use and vegetation cover, each of them were obtained from the Forest Map of Extremadura (scale 1:50.000): land use and management of the farm, dominant tree species, tree cover, total vegetation cover and structure of vegetation cover.

The performance of the model was analyzed using the ROC curve (*Receiver Operating Characteristic*) that represents the ability of a predictive model to differentiate areas with and without gullies. The ROC curve represents the values of sensibility (true positives) and the complementary of specificity (false positives) across all possible thresholds. The threshold is defined as the value used for classifying a quantitative result of the model like gullied or ungullied. For an optimal model, the ROC curve must be fitted to the upper-left side of the graphic, and the value of the AUC (*Area Under the Curve*) must be approximately one. AUC is a measure of model accuracy, but it does not provide a rule (threshold) for the classification of areas with or without gullying. The optimal threshold was selected using the sum of the percentage of correct cases of presence and absences of gullies.

In addition to the ROC curve and AUC we carried out a validation of the model with an external dataset from five different areas. For these datasets a confusion matrix was obtained representing a comparison between reality and model results.

The contribution of each variable to explain the model was estimated from the Generalized Cross Validation algorithm (1) (Craven and Wahba 1979) which estimates the reduction of the goodness of fit when each variable is eliminated from the model (Table 1):

$$GCV(M) = \frac{1}{N}\frac{\sum_{i=1}^{N}\left(y_i - f_M\left(x_i\right)\right)^2}{\left(1-\frac{C(M)}{N}\right)^2} \qquad (1)$$

IV International Symposium on Gully Erosion. J. Casalí and R. Giménez (Eds.)
© 2007 Universidad Pública de Navarra, Spain. ISBN 978-84–9769-198-7

where *N* is the number of cases of the initial dataset, *y* is the dependent variable, is the model and *C(M)* is a measure of cost-complexity of the model with *M* terms.

Another possible output of predictive models is the generation of maps which present the risk or potential of gullying. The results were used to generate those maps for the study areas and analyze the surface area that can be potentially affected by gullies.

3. Results

The result of the ROC curve (Fig. 1) shows a model with a good performance and sufficient ability to differentiate gullied and ungullied areas, with values of the AUC of 0.98.

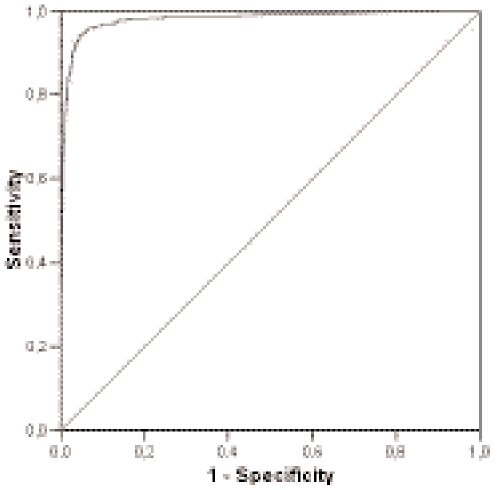

Fig. 1. ROC curve for the final model elaborated.

The threshold selected to classify the model results into gullied and ungullied areas was 0.45 which assures the highest percentage of areas being classified correctly.

The most important variables to support this model were lithology, vegetation structure and the average amount of autumn rainfall (September to November). Other important variables for explaining the model results were elevation, soil type, total upslope length of flow paths and average monthly precipitation of April, June, October and May (Table 1).

Table 1. Reduction of the goodness of fit of the model caused by eliminating each of the 10 most important variables.

Variable	*-GCV*
Lithology	0.075
Vegetation Structure	0.050
Autumn Rainfall	0.037
Elevation	0.034
Rainfall April	0.031
Soil type	0.031
Total upslope length	0.031
Rainfall June	0.031
Rainfall October	0.031
Rainfall May	0.031

Table 2. Summary of the importance of different factors in the gully model, where *N* is the number of variables, -*GCV* is the sum of *GCV* of each group variables and -$GCV_{average}$ is the average *GCV* by group of variables.

Group of Variables	*N*	*-GCV*	$-GCV_{average}$
Topographic	11	0.358	0.033
Rainfall	17	0.513	0.030
Lithology-Soil type	3	0.135	0.045
Land Use and Veg.	5	0.167	0.033

The Generalized Cross Validation (GCV) parameter shows that by groups, the most important factors in determining the distribution of gullying are lithology-soil type and topographic factors (Table 2).

Five different datasets were used for an external validation of the obtained model. The results show a model efficiency in determining the location of gullies of 80 % and in 99.65 % of the cases the model determined correctly the absence of gullies (Table 3). For three of the five validation datasets the values of efficiency in classifying gullied areas were higher than 90%.

Table 3. Confusion matrix that presents the results of validation for five external datasets.

		Reality	
		Absence of Gullies (0)	Presence of Gullies (1)
Model	Absence of Gullies (0)	4822	91
	Presence of Gullies (1)	17	364
Percent Correct (%)		99.65	80.00

4. Conclusions

Modern predictive models represent a powerful tool for predicting and analyzing geomorphological phenomena like gullying. The model obtained presents a high percentage of success in classifying gullied and ungullied areas. Nevertheless, in some areas the prediction of the occurrence of gullying was worse. Further studies need to be carried out in order to understand the reasons for its poor performance in certain areas. However, an improved model could be an important management and planning tool for silvopastoral areas of southwest Spain.

References

Craven, P. and G. Wahba. 1979. “Smoothing noisy data with spline functions.” *Numerische Mathematik* 31: 377-403.

Friedman, J. H. 1991. “Multivariate adaptive regression splines.” *Annals of Statistic* 19: 1-141.

Schnabel, S. 1997. *Soil erosion and runoff production in a small watershed under silvo-pastoral landuse (dehesas) in Extremadura, Spain*. Logroño.