Jul 1st, 12:00 AM

Use of Two Nonparametric Methods (CART and MARS) to Model the Potential Distribution of Gullies in Spanish Rangelands

A. Gómez Gutiérrez

S. Schnabel

F. J. Lavado Contador

iEMSs 2008: International Congress on Environmental Modelling and Software
Integrating Sciences and Information Technology for Environmental Assessment and Decision Making
4th Biennial Meeting of iEMSs, http://www.iemss.org/iemss2008/index.php?n=Main.Proceedings
M. Sànchez-Marrè, J. Béjar, J. Comas, A. Rizzoli and G. Guariso (Eds.)
International Environmental Modelling and Software Society (iEMSs), 2008

Use of Two Nonparametric Methods (CART and MARS) to Model the Potential Distribution of Gullies in Spanish Rangelands

Á. Gómez Gutiérrez, S. Schnabel & F. J. Lavado Contador.

GeoEnvironmental Research Group (GIGA), University of Extremadura, Avda. de la Universidad, s/n, 10071, Cáceres, Spain (alvgo@unex.es)

Abstract: Gully erosion represents an important soil degradation process in rangelands. In order to take preventive or control measures and to reduce its environmental damages and economical costs it is useful to localize the points in the landscape where gullying takes place and to determine the importance of the different factors involved. The study is carried out in Extremadura, southwest Spain. The main objectives of this work are: 1) to model the spatial distribution of gullies, 2) comparison of two nonparametric schemes to construct predictive models 3) to analyze the role of prevalence and of scale for susceptibility models in geomorphology, 4) evaluating the importance of the different factors involved and 5) implementing and mapping the results with the help of a Geographical Information System (GIS). Two methods were used to model the response of a dependent variable (gullying) from a set of independent variables: Classification And Regression Trees (CART) and Multivariate Adaptive Regression Splines (MARS). Three different datasets were used; the first one for constructing the model (training dataset) and the others for validating the model (external datasets). These datasets are formed by a target variable (presence or absence of gullies) and a set of independent variables. The dependent variable was obtained mapping the locations of gullies with the help of a GPS and high resolution aerial ortophotographs. We used 32 independent variables reflecting topography, lithology, soil type, climate and land use and vegetation cover of each area. To evaluate the performance of the models we used a non-dependent threshold method: the Receiver Operating Characteristic (ROC) curve. The results show a better performance of MARS for predicting gullying with areas under the ROC curve of 0.98 and 0.97 for the validation datasets, while CART presents values of 0.96 and 0.66.

Keywords: Gully erosion; nonparametric modeling; CART; MARS; ROC Curve.

1. INTRODUCTION

A gully is a channel originated by concentrated and intermittent water flow. Gully erosion is one of the most important soil degradation phenomena in several environments around the world and has long been neglected because it is difficult to study and to predict [Valentin *et al.* 2005]. Nevertheless, the negative on-site and off-site effects of gullying are well known: soil loss, reduction of soil water retention, landscape dissection (hampering the movement of vehicles, farm machinery and animals), water contamination and, sedimentation of channels and reservoirs. Several studies have tried to model the location of gullies in the landscape based on the concept of topographical thresholds [Patton & Schumm 1975]. However, these models are difficult to apply at the regional scale with success. In this line, Gómez Gutiérrez *et al*. [2007] developed a statistical multivariable model using MARS to predict the potential distribution of gullies in rangelands of southwest Spain. In the present paper, we compare the results of this previous work with another statistical model,

Classification And Regression Trees (CART). A further objective is to analyze the role of prevalence (different proportions of presence/absence data in the dataset) and of scale for susceptibility models in geomorphology. The importance of the different factors involved is evaluated and the results are implemented and mapped with the help of a Geographical Information System (GIS).

2. STUDY AREA

The study was carried out in 54 farms, representative of *dehesas* and pasturelands in the Iberian Peninsula (Figure 1). Dehesas commonly have a savannah-like vegetation and agrosylvopastoral land use, with farm sizes in excess of 100 ha. The relief is diverse with an average elevation of 413 m (ranging from 115 m to 902 m) and a mean slope angle of 5º. Lithology is quite diverse, though acid rocks dominate, being schists and granites the most frequent. Soils are mainly Cambisols (80%), Luvisols (10%) and Acrisols (5%). Climate is Mediterranean with continental and Atlantic influences. Mean annual temperature is approximately 16º. Precipitation presents a high spatial and temporal variability. Median annual rainfall of the study farms is 607 mm, ranging from 465 to 926 mm. More information on physical conditions of the 54 farms can be found in Tables 2 and 3.

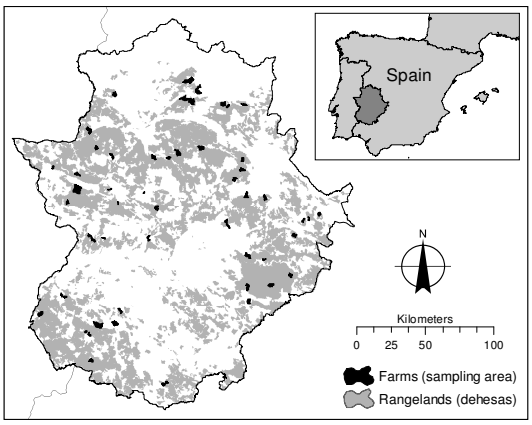

Figure 1. Location of the farms and the extension of dehesas in Extremadura region.

3. MATERIAL AND METHODS

3.1 CART and MARS

Classification And Regression Trees [CART; Breiman *et al.* 1984] is a popular data mining technique based on recursive binary partitioning. The result of CART (1) is a hierarchical binary tree which subdivides the prediction space into regions (R_m) where the values of the response variable are similar ($\cong a_m$):

$$f(x) = a_m ; \forall x \in R_m \quad (1)$$

Basically, CART predicts the pertinence of cases to a category of the dependent variable from a set of values of the independent variables. CART presents a root node with all the cases and analyzes if these belong to a single class. If the answer is yes, the algorithm finishes, but if it is no, CART divides the original dataset based on a split criterion (the Gini index of impurity) into two classes (the one which fulfills the split criterion and the rest of the dataset). The performance of the algorithm continues until a stopping rule is completed. The objective of the split criterion is to minimize the variability within each of the resulting subsets. The principal inconveniences of CART are the possible complexity of the resulting model, the hierarchical dependence between nodes at different levels and the difficulties that CART presents to reproduce smooth variations on the response. The model based on CART theory was computed using the Classification and Regression Tree tool implemented in *Statistica 7*© [Stasoft-Inc. 2004]. In this software, the variable importance is computed through the sum of the reductions in node impurity over all nodes in all computed

trees when a variable is eliminated. Then, these values are referenced as percentages of the most important variable.

On the other hand, MARS [Friedman 1991] is a relatively new technique which combines the classical linear regression, the mathematical construction of splines, the binary recursive partitioning and brute search intelligent algorithms to produce a model capable to predict the value of a target variable (categorical or continuous) from a set of independent variables. To do this, MARS approaches the underlying function through a set of piecewise functions called basic function (BF). The BF represent the information included in one or more independent variables and are selected through a step by step process. MARS general expression can be written as follows:

$$y = f(x) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(x) \qquad (2)$$

where y is the value predicted by the model by means of a function $f(x)$, which can be decomposed into an initial constant β_0 and a sum of M terms, each of them are formed by a coefficient β_m and a BF $h_m(x)$. The first step of MARS is to approximate its response with a constant (y=a). Then, MARS starts the search of the couple node-variable that produces the best fit of the model. This search continues until all possible BF have been added to the model (or a stopping rule is completed), obtaining as a result a very complex and overfitted model. Based on the overfitted model, MARS identifies which is the less important BF which is eliminated. This process is repeated until all BF are suppressed. The outcome is a sequence of candidate models with different BF numbers, from which MARS selects the optimal model via Generalized Cross Validation [GCV; Craven & Wahba 1979]:

$$GCV(M) = \frac{1}{N} \frac{\sum_{i=1}^{N} (y_i - f_M(x_i))^2}{\left(1 - \frac{C(M)}{N}\right)^2} \qquad (3)$$

where N is the number of cases, y is the dependent variable, $f_M(x_i)$ represents the model and $C(M)$ is a measure of cost-complexity for a model with M terms. The software programme MARS 2.0© [Salford-Systems 2000] was used for modeling. It enables the estimation of the importance of the independent variables. In order to determine the importance of an independent variable, MARS drops all terms including this variable and calculates the reduction of the goodness of fit. Finally, MARS references these values to the most important variables as a percentage.

3.2 Evaluation and validation of the models

In this paper, gullying was treated as a binary variable (1: presence and 0: absence). Therefore, the results of CART are easily comparable with reality. However the output of MARS is a continuous value, normally between -0.5 and 1.5 for a target variable encoded as 0 or 1. To evaluate the results of predictive models for categorical or binary target variables traditionally a confusion matrix was used. This matrix consists in crossing the predicted and the observed values. However, the confusion matrix and its derived statistics (misclassification rate, model efficiency, odd ratio, etc.) depend on a predefined threshold which divides the output into categories (1,0). However, these are influenced by prevalence of the dataset [Beguería 2006], typical in geomorphology studies. Therefore the use of a method is recommended which is not based on a predefined threshold and with an external dataset different from the training dataset [Muñoz & Felicísimo 2004; Beguería 2006].

The results of the models were analysed using the Receiver Operating Characteristic curve [ROC; Deleo 1993] and the Area Under the Curve (AUC). The ROC curve shows the skill of a model to discriminate presences from absences (representing the complementary of specificity versus sensitivity). A model with a good performance should present a ROC curve fitted to the upper-left side of the graph and the AUC should be a value close to one.

Two different external datasets were used to validate the model: *Guadalperalón* and *Monroy* (Table 1). These datasets were obtained in areas having similar environmental characteristics than the areas of the training dataset.

Table 1. Characteristics of the validation areas.

Dataset	Area	Elevation	Slope	Lithology	Soil types	Tree Cover	Annual Rainfall
Guadalperalón	131 ha	353 m	5.47 º	Schists	Distric Cambisol	30.2 %	595 mm
Monroy	114 ha	397 m	2.01 º	Schists	Distri-endolep. Cambisol	30.0 %	601 mm

3.3 Elaboration of the dataset

The general scheme of this part of the work was to generate a map showing the presence or absence of gullies in the study area and to obtain maps of different factors related with the gully formation. The presences of the target variable were obtained mapping existing gullies in the field with a GPS and the absences were obtained with a random point generator tool implemented using the Hawth's Analysis Tools [Beyer 2002] of ArcGis 9.0 [ESRI 1999]. A total of 662 Presences were mapped, as compared to 7001 Absences.

Different types of independent variables were used (Tables 1 and 2). To represent the topography a Digital Elevation Model (DEM) was used, with a pixel size of 5 m, extracted from contour lines and elevation points of the topographical map of Spain (scale 1:10,000) using the *Topo to Raster tool* [Hutchinson 1993] implemented in ArcGis 9.0 [ESRI 1999]. With this DEM we generated derived Digital Terrain Models to represent slope, curvature, ruggedness and some hydrological features of the surface (Table 1). Soil type was extracted from maps of 1:300,000. This small scale represents a handicap, but it is the only soil map available for the whole study area, and we think that it can represent regional trends in some soil types for the development of gullies. Lithology is based on the digital version of the geological cartography of Spain (scale 1:50,000). Maps of mean annual, seasonal and monthly rainfall were elaborated using a dataset (1960-1990) from 222 meteorological stations in the study area, [using as interpolation method thin plate splines; Hutchinson 1991]. Finally we used five variables from the forest map of Extremadura [scale 1:50,000, EGMASA 2002]: land use and management of the farms, dominant tree species, tree cover, total vegetation cover and structure of the vegetation cover.

Table 2. Quantitative independent variables used and their values for the study area (training dataset).

Variable	Average	Deviation	Units
Elevation	413	125	Meters
Slope	5.06	5.60	Degrees
Curvature	0.00	1.15	10^{-2} meters
Vertical curvature	0.01	0.56	10^{-2} meters
Horizontal curvature	0.01	0.81	10^{-2} meters
Total upslope length	2677	23,046	Pixels
Longest upslope length	80.0	218.7	Pixels
Drainage area	1,057.9	9,093.0	Pixels
Total vegetation cover	76.3	25.4	%
Tree cover	27.1	17.2	%
Monthly rainfall	(12 variables)	-	Millimeters
Seasonal rainfall	(4 variables)	-	Millimeters
Annual rainfall	695.6	145.3	Millimeters

The effect of prevalence was analysed varying the number of absences on the training dataset and maintaining the same number of presences and calculating the success of the resulting models using the AUC.

Table 3. Qualitative independent variables used in the model.

Variable	Number of categories	Dominant categories
Lithology	139	Schists and granites
Soil type	47	Distri-endoleptic Cambisols and distric Cambisols
Soil family	8	Cambisols and Luvisols
Vegetation structure	63	Mixed woodland and low shrubs
Land use	8	Pasturelands
Dominant tree specie	15	*Quercus rotundifolia* Lam.

4. RESULTS AND DISCUSSION

4.1 CART

The result of CART is a tree with 34 non-terminal nodes and 35 terminal nodes. Table 3 shows part of the model classification rules. This model presents an AUC of 0.97 for the training dataset. The values of the AUC for the validation datasets are quite different, while *Guadalperalón* presents a high value (0.96), the AUC of the *Monroy* dataset is only 0.66.

Table 4. CART model classification rules. These expressions must be compiled and implemented on a GIS to obtain susceptibility maps for gullying.

Terminal node	Classification rule
4	If [(lithology=9 or 11 or 40 or 80 or 99 or 109) and (Soil type=0 or 11 or 16) then the pixel is classified as ungullied area (0).
…	…
54	If [(lithology ≠ 9 or 11 or 40 or 80 or 99 or 109) and (vegetation structure = 22) and (lithology = 84 or 90 or 127) and (planimetric curvature ≤ 0) and (longest upslope length grid ≤ 225.13) and (elevation >251) and (rainfall in 4th trimester ≤ 343) and (december rainfall >84.5) and (elevation ≤ 474.5) and (elevation ≤ 380.5)] then the pixel is classified as gullied area (1).
…	…
123	If [(lithology ≠ 9 or 11 or 40 or 80 or 99 or 109) and (vegetation structure ≠ 22) and (rainfall in 3rd trimester >35.5) and (soil type ≠ 18) and (lithology ≠ 20 or 85) then the pixel is classified as ungullied area (0).

Regarding the influence of independent variables, elevation (DEM) and annual rainfall seem to be the most important variables. In addition, lithology and soil type are observed as root node or nodes in the upper part of the tree structure. Furthermore, variables related with topography are situated in the lower part of the tree. This trend seems to be related with the scale of the independent variables (Figure 2). Since the variables were obtained from maps with only three different scales (1:10,000, 1:50,000 and 1:300,000), we can deduct from the tree structure that the quality and resolution of the model is supported by large scale variables (elevation, slope, curvature, drainage area, etc.) due to their position in the lower part of the tree. This argument could be an important key in the future, because the tree structure shows us which variables must be elaborated carefully (with high resolution and precision), and this will probably improve the success of the model for predicting the location of gullies in similar environments. In the case of our study the variables in the upper part of the tree indicate areas with a risk of gully formation and the variables in the lower part of the tree which have higher spatial resolution enable the location of the gully.

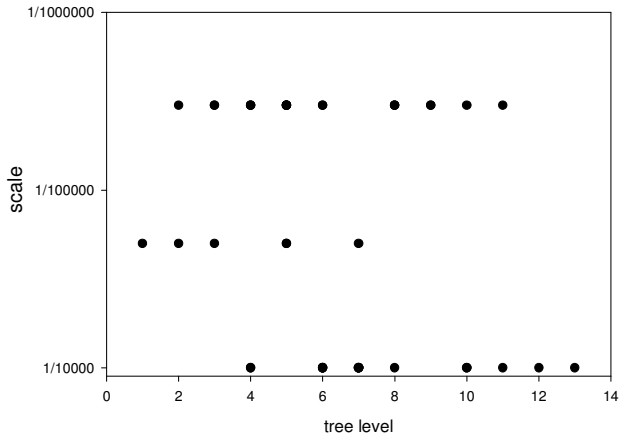

Figure 2. Relation between the scale of each cartographic source and the position where its derived variables are located on the tree.

4.2 MARS

The outcome of MARS is a model with 78 BF, 57 terms and a ROC curve that is fitted to the upper left side of the plot. The AUC for the training dataset is 0.98. Using a preliminary threshold of 0.5 we get a sensitivity of 0.953 and a specificity of 0.86. For the *Guadalperalón* validation dataset the AUC reaches a value of 0.98. For the *Monroy* validation dataset the obtained AUC is similar (0.97), showing a good performance of the model. A portion of the MARS model for predicting gullying in the study area is as follows:

$$Y = 0.065 + 11.973 * BF1 + ... + 0.317 * BF108 + \ldots + 0.000104 * BF120 \quad (4)$$

where *y* is the predicted value and BF1= (lithology = 9,11,40,60,80,99 or 109), BF108 = [(areas with rainfall in march below 73 mm – 73) * BF1] and BF120 = [(lithology = 40, 90 or 109) * BF4], and finally, BF4= [areas with a value of total upslope length bellow 27186 pixels - 27186].

MARS used 27 variables to construct the model. Individually, the most important variable is lithology followed by vegetation structure, rainfall in the 3rd trimester of the year and elevation. By groups, we can observe that climatic variables are the most important in absolute terms, principally due to their large number (17). However, if the variables are grouped (topography, soil type-lithology, climate, and land use-vegetation cover) the highest variable importance corresponds to lithology and soils. These results should be interpreted carefully because MARS and CART models can explain a portion of the spatial distribution of gullies, but cannot establish causal relationships between independent variables and gully occurrence [Donati & Turrini 2002].

4.3 The role of prevalence on the training dataset

No clear tendency was observed. The AUC values for training datasets with different proportion of presences/absences in the training dataset are unstable (Figure 3). Although some authors have attributed an important role of prevalence on the training datasets for modelling geomorphological phenomena, we have not found an influence on the success of the models. Possibly the independence of the AUC validation statistic from prevalence [Beguería 2006] and the quality of the presences and absences data can minimize this effect.

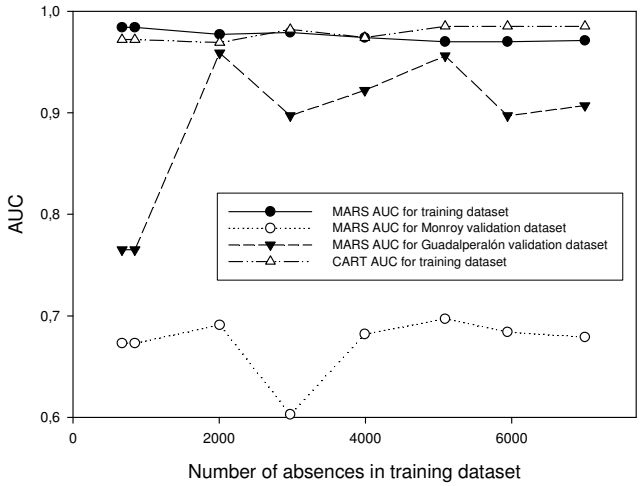

Figure 3. AUC values using different number of absences in the training dataset and with a fixed number of presences.

4.4 Implementing the models and mapping the results

After obtaining the models, they were implemented into a GIS to obtain maps of gully susceptibility. Finally, a composition map was elaborated combining the results of MARS and CART. For this the strategy consisted in using the AUC of each model as a weight for transforming the values and afterwards summing the results and generating the map.

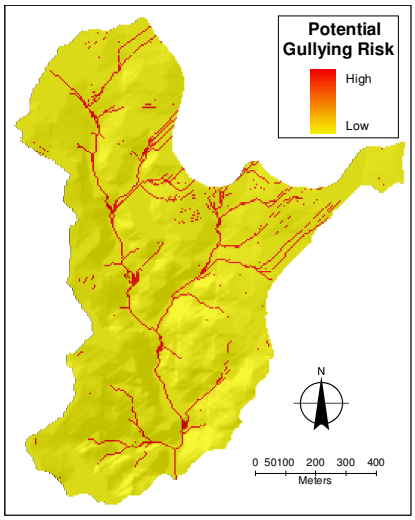

Figure 4. Composition map for the *Guadalperalón* validation area.

5. CONCLUSIONS

The two models present a good performance, with AUC from 0.66 to 0.98. MARS produced better results for the two validation datasets. Some differences between the two models regarding the importance of variables were found. Although variable importance cannot be interpreted as a causal relation, in CART the position of each variable in the tree can help us to understand the relation between variable scale, model structure and processes. Regarding the influence of prevalence on the success of the models, no clear

tendency was found. Finally we have used a simple method to create a map combining the results of both models and using the AUC values as weights. Susceptibility maps of gully erosion can be a useful tool for regional planners.

ACKNOWLEDGEMENTS

The work has been supported by the Junta de Extremadura and European Union funded project Montado/Dehesa (INTERREG IIIA, SP4.R13) and by the Spanish Ministry of Education and Science (CGL2004-04919-C02-02).

REFERENCES

Beguería, S. Validation and Evaluation of Predictive Models in Hazard Assessment and Risk Management. *Natural Hazards* **37**: 315-329, 2006.

Beyer, H. L. Hawth's Analysis Tools For ArcGis. In, 2002.

Breiman, L., Friedman, J. H., Olsen, R. A. & Stone, C. J. *Classification and regression trees*. Wadsworth, Belmont: CA. 1984.

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

Deleo, J. M. Receiver operating characteristic laboratory (ROCLAB): software for developing decision strategies that account for uncertainly. In: *Proceedings of the Second International Symposium on Uncertainlty Modelling and Analysis*, pp. 318-325. College Park: Computer Society Press, 1993.

Donati, L. & Turrini, M. C. An objective method to rank the importance of the factors predisposing to landslides with the GIS methodology: application to an area of the Apennines (Valnerina; Perugia, Italy). *Engineering Geology* **63**: 277-289, 2002.

EGMASA Mapa de vegetación y recursos forestales de Extremadura a partir del Mapa Forestal de España. In: *Mapa Forestal de Extremadura*, ed. C. d. A. y. M. Ambiente, 2002.

ESRI ArcGis 9.0. In, 1999.

Friedman, J. H. Multivariate adaptive regression splines. *Ann. Statist.* **19**: 1-141, 1991.

Gómez Gutiérrez, Á., Schnabel, S. & Felicísimo, A. Modelling the ocurrence of gullies in semiarid areas of southwest Spain. In: *Progress in gully erosion research*, eds. J. Casalí & R. Giménez, Pamplona: Public University of Navarra, 2007.

Hutchinson, M. F. Climatic analyses in data sparse regions. In: *Climatic risk in crop production. Models and Management for the semiarid tropics and subtropics*, eds. M. C. Chow & J. Bellamy, pp. 55-71. CAB International, 1991.

Hutchinson, M. F. Development of a continent-wide DEM with applications to terrain and climate analysis. In: *Environmental Modeling with GIS*, ed. G. M. F. e. al., pp. 392-399. New York: Oxford University, 1993.

Muñoz, J. & Felicísimo, Á. M. Comparison of statistical methods commonly used in predictive modelling. *Journal of Vegetation Science* **15**: 285-292, 2004.

Patton, P. C. & Schumm, S. A. Gully erosion: a threshold phenomenon. *Geology* **3**: 88-90, 1975.

Salford-Systems MARS. In, 2000.

Stasoft-Inc. Statistica: Data analysis software system. In, 2004.

Valentin, C., Poesen, J. & Yong, L. Gully erosion: Impacts, factors and control. *Catena* **63**: 132-153, 2005.