Modeling soil properties at a regional scale using GIS and Multivariate Adaptive Regression Splines

Álvaro Gómez Gutiérrez, Francisco Lavado Contador and Susanne Schnabel

Geoenvironmental Reseach Group, University of Extremadura

Cáceres, Spain

alvgo@unex.es

Abstract—**In this paper, five topsoil properties (clay, silt, sand, organic matter and bulk density) have been modeled using accessible environmental information in order to improve pedological cartography in the Extremadura region (Spain). Independent variables related to topography, climate, lithology and vegetation cover were used to generate the models. The statistical approach was based on using Multivariate Adaptive Regression Splines (MARS) to produce the models. The performance of the models was tested using Root Mean Square Error (RMSE), Generalized Cross Validation (GCV) and the r coefficient (classical regression analysis). The models presented moderate power prediction; with r ranging from 0.23 for organic matter to 0.47 for silt. The most important independent variables supporting the models were location (X and Y UTM coordinates), lithology and altitude. Finally the models were compiled and implemented into a Geographical Information System (GIS) to obtain continuous maps of target variables.**

I. INTRODUCTION

Digital soil mapping uses quantitative relationships between soil properties and other spatially distributed environmental variables. At the regional scale, the availability of pedological information and cartography is crucial for the development of projects related to soil science (geomorphology, hydrology, ecology, climatology, etc.). In the Extremadura region (Fig. 1: SW Spain) the existence of soil cartography and pedological information is very limited. Besides, the combined use of Geographical Information Systems (GIS) and data mining techniques has facilitated the elaboration of statistical models with a spatial component. The main benefits of these models are to estimate the distribution of a target variable, to predict future changes on its allocation and to establish the importance of each predictor in determining the distribution of the dependent variable. In this paper, GIS and data mining techniques are used to construct a model capable of predicting five different topsoil properties (clay, silt, sand, organic matter and bulk density), to analyze the import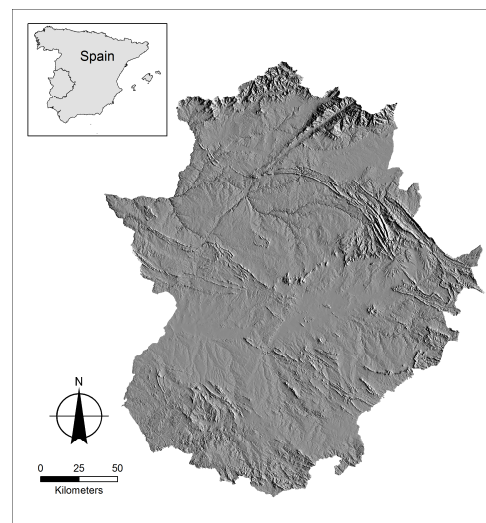ance of the factors involved and to improve the existing pedological knowledge and cartography in the Extremadura region.

II. STUDY AREA

The study was carried out in the Extremadura region located in SW Spain (41,633 km^2; Fig. 1). The landforms in the study area are quite diverse, mostly undulating peneplains and low mountain ridges. Elevations range from 100 to 2400 masl, with an average of 410 masl. The mean slope angle amounts to 5.4º. Climate is Mediterranean semi-arid to sub-humid with an annual average rainfall of 641 mm. Lithology is quite variable and acid materials are the most representative, with schist and granites as the most frequent rock types.

Figure 1. Location of the study area.

III. MATERIAL AND METHODS

Summary of the Procedure

Five different models were trained with 5 dependent topsoil variables (the content of clay, silt, sand, organic matter and bulk density) and 17 independent variables (Table 1). The training data set included 1688 records while the validation data set included 137 cases. Dependent variables used here were extracted from the National Inventory of Soil Erosion database [10] gathered all over the study area. On the other hand, a set of 17 independent variables related to climate, vegetation, topography, lithology and geographical location were collected from different sources (Table 1).

TABLE I.   SUMMARY OF THE VARIABLES USED IN THE MODELS.
D=DEPENDENT, I=INDEPENDENT

Variable	Units-description
Clay (D)	Topsoil clay content in %
Silt (D)	Topsoil silt content in %
Sand (D)	Topsoil sand content in %
Organic matter (D)	Topsoil organic matter content in %
Bulk density (D)	Topsoil bulk density in kg·m^{-3}
Coordinates X (I) and Y (I): 2 variables	UTM coordinates in m
Elevation (I)	Altitude extracted from SRTM DEM in m (version 4.1 with 90 m pixel size)
Slope (I)	Steepness in degrees obtained from SRTM
Curvature (I): 4 variables	General, profile, plan and tangential curvature
Specific catchment area (I; 2 variables)	Upslope area per unit width of contour obtained from SRTM with a) D-Infinity algorithm and b) D-8 algorithm
Topographic wetness index (I)	Beven and Kirkby [1]
Relative stream power index (I)	Moore et al. [12]
Sediment transport capacity index (I)	Moore and Wilson [13]
Annual rainfall (I)	Obtained from the Digital Climatic Atlas of the Iberian Peninsula [15] in mm
Total vegetation cover (I)	Obtained from the INES [10] in %
Tree cover (I)	Obtained from the INES [10] in %
Lithology (I)	Obtained from the INES [10]

Statistical Modelling

Several methods exist for constructing predictive models. However, for solving complex multivariable problems in geomorphology, non-parametric approaches (such as Artificial Neural Networks and Classification and Regression Trees) usually produce the best results [2,6,7,9]. Amongst non-parametric techniques MARS [5] is frequently used and was selected for being, in most of the cases, more accurate, faster and easier to interpret than others [3,14]. MARS combines the classical linear regression, the mathematical construction of splines, the binary recursive partitioning and brute search and intelligent algorithms to produce a model capable to predict the value of a target variable from a set of independent variables. It approaches the underlying function through a set of piecewise functions called basics functions (BF). The BFs 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:

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

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 $h_m(x)$.

STATISTICA®, ArcGIS® and WhiteBOX Geospatial Analysis Tools® software were used to integrate, process and analyze the data.

IV. RESULTS

A summary of the observed and predicted values for the five target variables is presented in Table II. The best performance of MARS was obtained for silt and sand content models with values for r of 0.47 and 0.46 respectively. The silt and sand content models presented also RMSE values of 9.59% and 12.11% for validation datasets with mean values of 25.43% (standard deviation=10.68) and 50.39% (standard deviation=13.15). Moreover, the clay content and bulk density model showed a similar accuracy with values for r of 0.42 and 0.41 respectively. However, the model generated for soil organic matter content was the worst with a RMSE of 3.89% over an average value for the validation dataset of 3.56% and a standard deviation of 3.95%. In addition, the r coefficient between predicted and observed values showed a poor relationship (r=0.23). Probably, additional variables need to be included in this model (such as land use and vegetation descriptors or higher resolution topographical variables). Fig. 2 shows the relationship between the observed and predicted clay content.

In general terms, the models were complex and included a large number of BFs and terms, ranging from 67 BFs and 37 terms in the clay content model to 20 BFs and 11 terms in the soil bulk density model.

TABLE II.   SUMMARY OF THE MODELS.

Clay (%)			
		Observed	**Predicted**
Mean		24.14	23.89
Standard Deviation		6.77	5.08
RMSE=6.52	GCV=41.67	r (O-P)[a]=0.42	
Silt (%)			
		Observed	**Predicted**
Mean		25.43	24.99
Standard Deviation		10.68	6.95
RMSE=9.59	GCV=66.70	r (O-P)[a]=0.47	
Sand (%)			
		Observed	**Predicted**
Mean		50.39	50.51
Standard deviation		13.15	9.36
RMSE=12.11	GCV=11.72	r (O-P)[a]=0.46	
Organic matter content (%)			
		Observed	**Predicted**
Mean		3.56	3.48
Standard deviation		3.95	1.58
RMSE=3.89	GCV=11.72	r (O-P)[a]=0.23	
Bulk density (kg·m^{-3})			
		Observed	**Predicted**
Mean		1233.4	1222.1
Standard deviation		169.7	82.1
RMSE=154.20	GCV=20,805	r (O.P)[a]=0.41	

a. Root Mean Square Error, Generalized Cross Validation error and r coefficient between predicted and observed values.

In order to determinate the importance of an independent variable supporting the models, all terms including a specific variable were suppressed and the reduction of the goodness of fit was calculated. Finally, all values were referenced to the most important variable as a percentage. The most significant independents variables supporting the model were spatial location (Y and X coordinates, elevation and lithology. Other important variables to explain the spatial distribution of soil properties in the study area were total tree cover and specific catchment area. However, in this kind of model the importance of the independent variables should be interpreted with caution [4]. In addition, a better resolution DEM is recommended to improve ithe performance of some models, mainly organic matter content model. However, recent works have shown that DEMs of finer pixel size and higher resolution in Spain are still unavailable [8].

Finally the equations of the model were implemented into ArcGIS and applied to the whole study area to create maps of soil properties. Fig. 3 shows a map representing spatial distribution of predicted bulk density.
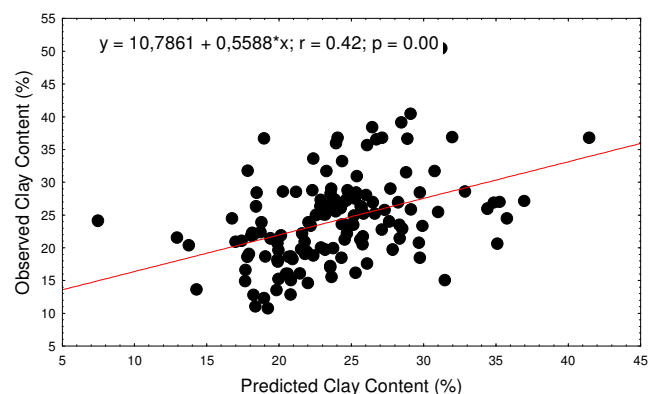

Figure 2.   Relationship between observed and predicted clay content.

CONCLUSIONS

Five spatial models representing topsoil properties were elaborated combining data mining techniques and GIS. These models presented moderate prediction power.

The equations obtained using MARS were implemented into a GIS and applied to the whole study area to produce continuous maps for five topsoil properties. These maps could be used to fill the lack of digital soil cartography in the Extremadura region.

Future challenges include incorporation of new data, mainly to improve topsoil organic matter predictions, and refining the existing dataset. A more detailed topographic dataset would be needed. Although the predictive power of the model is not very

high, the presented results are promising because the incorporation of other variables such as land use and vegetation and the use of a more detailed topographic dataset is expected to improve the model. Future challenges include the generation of a digital soil atlas for the Extremadura region.
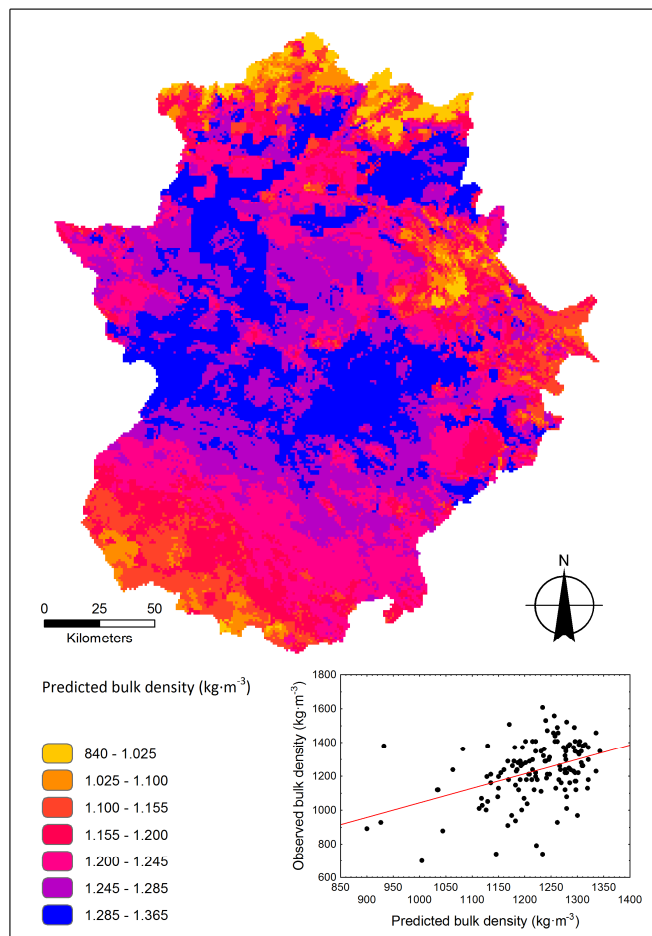
Figure 3. Predicted bulk density map for the whole study area, including the relationship between observed and predicted values for the validation dataset at the bottom. Categories in the legend were generated using the natural breaks optimization method [11].

REFERENCES

[1] Beven, K. J. and M. J. Kirkby (1979). "A physically based, variable contributing area model of basin hydrology." *Hydrological Sciences Bulletin* 24: 43-69.

[2] Bou Kheir, R., J. Wilson and Y. Deng (2007). "Use of terrain variables for mapping gully erosion susceptibility in Lebanon." *Earth Surface Processes and Landforms* 32(12): 1770-1782.

[3] De Veaux, R. D., D. C. Psichogios and L. H. Ungar (1993). "A Comparison of two nonparametric estimation schemes: MARS and neural networks." *Computers & Chemical Engineering* 17(8): 819-837.

[4] Donati, L. and M. C. Turrini (2002). "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.

[5] Friedman, J. H. (1991). "Multivariate adaptive regression splines." *Annals of Statistics* 19: 1-141.

[6] Gómez Gutiérrez, Á., S. Schnabel and A. Felicísimo (2009). "Modelling the ocurrence of gullies in rangelands of SW Spain." *Earth Surface Processes and Landforms* 34(14): 1894-1902.

[7] Gómez Gutiérrez, Á., S. Schnabel and J. F. Lavado Contador (2009). "Using and comparing two nonparametric methods (CART and MARS) to model the potential distribution of gullies." *Ecological Modelling* 220(24): 3630-3637.

[8] Gómez Gutiérrez, Á., J.F. Lavado Contador and S. Schnabel (2011). Testing the quality of open-access DEMs and their derived attirbutes in Spain: SRTM, GDEM and PNOA DEM. In Geomorphometry 2011, T. Hengl, Evans, I.S., Wilson, J.P. and Gould, M. 53-56. Redlands, CA, 2011.

[9] Hughes, A. O., I. P. Prosser, J. Stevenson, A. Scott, H. Lu, J. Gallant and C. J. Moran (2001). Gully erosion mapping for the National land water resources audit, CSIRO Land and Water Technical report: 19.

[10] INES (2007). Inventario Nacional de Erosión de Suelos. M. d. M. A. y. M. R. y. Marino. Cáceres and Badajoz spanish provinces.

[11] Jenks, G. F. (1967). "The data model concept in statistical mapping." *International Yearbook of Cartography* 7: 186-190.

[12] Moore, I. D., R. B. Grayson and A. R. Ladson (1991). "Digital terrain modelling: a review of hydrological, geomorphological and biological applications." *Hydrological Processes* 5: 3-30.

[13] Moore, I. D. and J. P. Wilson (1993). "Length-Slope factors for the Revised Universal Soil Loss Equation: Simplified method of estimation." *Journal of Soil and Water Conservation* 47: 423-428.

[14] Muñoz, J. and Á. M. Felicísimo (2004). "Comparison of statistical methods commonly used in predictive modelling." *Journal of Vegetation Science* 15: 285-292.

[15] Niyerola, M., X. Pons and J. M. Roure (2005). Atlas Climático Digital de la Península Ibérica. Metodología y Aplicaciones en Bioclimatología y Geobotánica, Universidad Autónoma de Barcelona: Bellaterr; http://opengis.uab.es/wms/iberia/pdf/acdpi.pdf.