VI

Congresso Nacional de Geomorfologia

Geomorfologia: novos e velhos desafios

Atas

.:.

Proceedings

Editores:
Adélia Nunes
Lúcio Cunha
João Santos
Anabela Ramos
Rui Ferreira
Isabel Paiva
Luca Dimuccio

21 a 23 de fevereiro de 2013
Universidade de Coimbra

Associação Portuguesa de Geomorfólogos

Departamento de Geografia
Faculdade de Letras da Universidade do Porto
Via Panorâmica,s/n
4150-564 Porto

apgeom.dir@apgeom.pt

Editores
Adélia Nunes
Lúcio Cunha
João Santos
Anabela Ramos
Rui Ferreira
Isabel Paiva
Luca Dimuccio

Design e Formatação:
Anabela Ramos
Isabel Paiva
Rui Ferreira

VI Congresso Nacional de Geomorfologia
Departamento de Geografia
Faculdade de Letras
Universidade de Coimbra
Col. S. Jerónimo
3004-530 Coimbra

21 a 23 de fevereiro de 2013

ISBN: 978-989-96462-4-7

Apoios:

Associaçao Portuguesa de Geomorfólogos

International Association of Geomorfologists

Departamento de Geografia (FLUC)

CEGOT
Centro de Estudos de Geografia e Ordenamento do Território

SOIL LANDSCAPE MODELLING – PLACING PLACE IN ITS PLACE

MODELAÇÃO DO SOLO-PAISAGEM – A IMPORTÂNCIA DA LOCALIZAÇÃO

Fonseca, I. L.,*Centro de Estudos Geográficos – IGOT, Lisboa, Portugal, i.fonseca@campus.ul.pt*
Freire, S., *Centro de Estudos de Geografia e Planeamento Regional – FCSH-UNL, Lisboa, Portugal*
Brasil, R., *Centro de Estudos Geográficos – IGOT, Lisboa, Portugal*
Rocha, J., *Centro de Estudos Geográficos – IGOT, Lisboa, Portugal*
Tenedório, J. A. *Centro de Estudos de Geografia e Planeamento Regional – FCSH-UNL, Lisboa, Portugal*

ABSTRACT

Landscape variables, which are also factors of soil formation, can be combined with existing soil map data to train Artificial Neural Networks (ANNs) in order to predict soil types in unmapped areas. In this study, the impact of location data and proximity of the training data on the performance of ANN models, for two catchments in northern Portugal, is evaluated. Results are largely concurrent between catchments, indicating that using latitude and longitude data produces more accurate models, whilst taking into account the spatial autocorrelative properties of input data makes ANN models converge for a better "local" rather than "global" solution. The conclusion is that hillslopes show some degree of connectivity which is passed onto soils, and conforms to the principles of the catena concept.

RESUMO

Os tipos de solos existentes em áreas sem cartografia de solos podem ser inferidos combinando variáveis de paisagem, igualmente factores de formação de solo, com mapas de solos, através de Redes Neuronais Artificiais. Neste estudo, aplicado a duas bacias do norte de Portugal, foi avaliado o impacto da utilização de dados de localização e da proximidade das áreas de treino na qualidade da predição dos modelos de RNA. A utilização da latitude e longitude produz modelos mais precisos, enquanto a consideração das propriedades de autocorrelação dos dados espaciais produz modelos de RNA melhor adaptados a soluções locais, do que ao total da área da bacia de drenagem. Conclui-se, portanto, que as vertentes apresentam algum grau de conectividade que se transmite aos solos, e que está de acordo com os princípios da catena.

1. INTRODUCTION

There is a growing need for soil maps at scales suitable for land management and regional planning. Soils modulate hydrological fluxes, regulate ecosystems and can play a very important role in mediating the impact of climate change. However, most European countries still lack complete soil map coverage at medium to large scales because soils surveys are expensive and time consuming. Digital Soil Mapping (DSM) is an advanced technique for mapping soil classes (Dobos *et al*., 2006) which has been developed to produce soil maps in a quicker, cheaper, more flexible and consistent fashion than the maps obtained through traditional soil survey. DSM can combine computer-based technologies, such as Geographical Information Systems (GIS), with advanced techniques, such as Artificial Neural Networks (ANNs). ANNs are sophisticated computer programs which are able to model complex functional relationships. As such, ANNs can maximize the information content of existing soil maps by learning rules that have, more or less explicitly, led to the mapping of the spatial distribution of soil classes across the landscape, as long as those rules are based on factors known to be responsible for the spatial variation of soil (McBratney *et al*., 2003). Thus, a set of variables related to soil forming factors and the respective soil class are used as training data for the ANNs, which construct rules (Tso and Mather, 2001) that can be used to predict the spatial distribution of soil classes in unmapped areas.

Whilst the literature provides a number of examples where ANNs has been developed successfully and the spatial variation of soil shown to be induced by a limited number of soil forming factors (Mora-Vallejo *et al*., 2008), still little is known about the impact that the high levels of spatial autocorrelation found on the landscape variables most commonly used to predict soils can have on the performance and accuracy of ANN models. High levels of spatial autocorrelation mean that close neighbours are more likely to have similar soil types. Thus, the objective of this study is to evaluate (1) if taking into account the spatial location of training samples through the addition of latitude and longitude to the set of input variables improves the accuracy of ANN models; (2) if adopting a sampling strategy that simultaneously takes into account the range of variation of the landscape variables and minimizes the distance between training sites allows ANNs to learn better; (3) if the results of these two hypotheses are consistent across the landscape by comparing the results obtained for two catchments; (4) if the pattern holds when comparing results obtained for two sets of data with morphometric variables derived at two different resolutions, i.e. if results change with scale; and, (5) if impact of location and sampling strategy varies with the type of ANN architecture used.

2. MATERIAL AND METHODS

2.1. Study Areas

Two catchments in northern Portugal were selected for this study (Fig. 1): Mondim de Basto (MB) in the NW region (911km^2) and Vila Real (VR) in the NE region (468km^2). These catchments were chosen because they are located in areas covered by published 1:100 000 soil maps. Additionally, they were chosen because they present diverse geomorphological and ecological characteristics and include soil classes representative of those found in the NW and NE soil mapping regions.

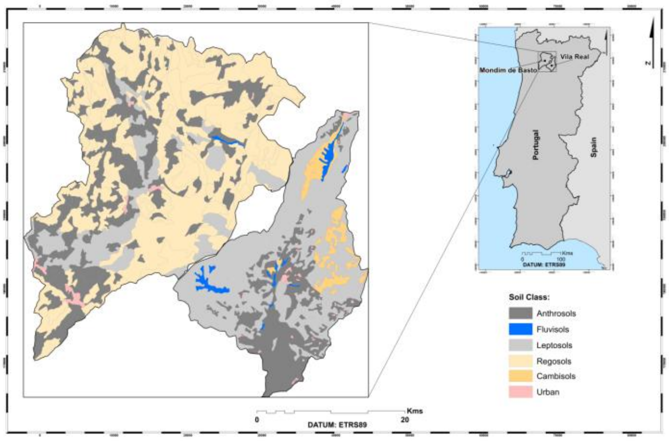

Figure 1 – Soil Class Distribution in the Studied Catchments: Mondim de Basto and Vila Real.

2.2. Data and Methods

Independent variables used for training the ANNs included continuous terrain data and categorical maps. One terrain dataset is the SRTM digital elevation data with a 90m resolution (2001), freely available for download (courtesy of the USGS), and the other dataset was derived from a TIN model supplied by IGeoE at 1:25 000, from which a DEM with a 25m resolution was produced. Morphometric variables were derived from each of the DEMs using ArcGIS software, and seven variables were selected after multicollinearity tests showed little data redundancy: slope, plan and profile curvatures, upslope catchment area, dispersal area, wetness index and potential solar radiation. All the continuous variables were re-scaled to a 0-255 value range. Categorical data used in the

models included land use and lithology. Land use data were extracted from the Corine Land Cover 2006 dataset (CLC2006) and lithology classes were obtained from digitised geological maps of northern Portugal at 1:200 000 (1989 and 2000).
In addition to the independent variable dataset, digital soil data were extracted from 1:100 000 maps to train the ANNs. One dataset was provided by DRAEDM (the NW regional department of agriculture) and the other freely available for download, courtesy of the University of Trás-os-Montes and Alto Douro (UTAD). All datasets were clipped to the catchment areas and converted to a raster structure with 90m and 25m cell sizes, using the ETRS 1989-TM06. Spatial autocorrelation was evaluated through Moran´s I test, which indicated that in both catchments autocorrelation is significantly high for slope (0.76/0.82) and very high for potential solar radiation (0.88/0.88) and altitude (0.99/0.98). Thus, in order to account for the possible effects of spatial autocorrelation, the x, y coordinates (longitude and latitude) were also included in the input set to indicate location.

Two types of ANN architectures were used to predict soil classes using IDRISI Taiga software (Clark Labs): the highly popular supervised method, Multi-Layer Perceptron (MLP), known as error back-propagating algorithm (Haykin, 1999) and the Kohonen´s Self-Organizing Map (SOM). Notwithstanding the importance of the parameterisation of the ANNs, it is not the focus of this paper to provide a comprehensive description of all the model runs. However, for both ANN algorithms equivalent experimental settings were tested whenever possible, in terms of network topology and parameters, training parameters and stopping criteria. Several tests were performed, changing the training parameters progressively at each run, in order to find out the best combination of parameters that achieved as high accuracy as possible. Training ended when one of the stopping criteria was achieved and only the best results in terms of accuracy are analysed herein. Note however that for the same sampling method and parameter combination, five model runs were performed to average their accuracies because results may vary due to slight differences in seeding of training pixels.

Two different sampling strategies were implemented for training the MLP and the SOM networks: one random and one stratified. An even number of training sites (500 pixels) were selected, whenever possible, for each soil type (Fluvisols often did not cover an area large enough to provide those many training sites). Despite the number of training sites being proportional to the number of soil classes in both catchments, random sampling implies that sites were not chosen evenly in each soil type strata. For the stratified sampling, training pixel vectors were located by choosing the nearest coordinates for each soil type that simultaneously were representative of the distribution of the predicting variables (morphometric data, land use and lithology), i.e. coordinates were also chosen evenly in the frequency space. In addition to testing the impact of training the networks with data collected from a close neighbourhood, tests were performed for both 10 and 12 predicting variables, where the latter included latitude and longitude as input data, and for standardized and non-standardized data. The ANN modules in IDRISI evaluate accuracy on part or whole of the training sample, which does not provide an independent assessment. Thus, accuracy was assessed both on the training sample and at the catchment level, i.e. based on the accuracy to predict soil classes in the whole catchment. The latter using the Map Comparison Kit 3.2.2 software.

3. RESULTS AND DISCUSSION

Table 1 presents a synthesis of the best accuracy results obtained for Mondim de Basto and Vila Real, which varied between 21% and 87%. On the whole, adding location as input data tends to improve the level of accuracy, independently of the catchment studied, type of neural network architecture and sampling strategy used. Nevertheless, the trend is somewhat weaker if the SOM network is used. A detailed analysis shows that not only the accuracy improvement is lower for SOM, but also, in those cases that accuracy declines if using 12 variables, the drop in accuracy is substantially high, therefore concealing the average trend. Results similar to those obtained with SOM are also obtained for the Vila Real catchment. No clear pattern in accuracy change is detected with standardization of data or change in resolution of morphometric data, as in Mondim de Basto with 25m pixel resolution and in Vila Real with 90m pixel resolution there is no significant difference in the number of experiments that showed either improvement or decline in accuracy levels. Nevertheless, on the whole, adding the x, y coordinates has a higher positive impact on accuracy if morphometric data has a coarser resolution.

Table 1 – Synthesis of best ANN accuracy results obtained for the Mondim de Basto and Vila Real catchments

			25m resolution MB				90m resolution MB				25 m resolution VR				90m resolution VR			
			RD		SN		**RD**		SN		**RD**		SN		**RD**		SN	
	ANN	Validation Reference	10 Var	**12 Var**	10 Var	12 Var	10 Var	**12 Var**	10 Var	12 Var	10 Var	**12 Var**	10 Var	12 Var	10 Var	**12 Var**	10 Var	12 Var
Non-Standard	**MLP**	Sample	63.3	64.5	49.8	69.6	59.3	64.9	54.3	71.8	70.2	71.7	46	61.3	72.9	74.4	39.8	86.9
		Catchment	63.4	**66.3**	33.4	29.5	63.7	**67.9**	36.7	35.7	72.1	**73.5**	47.0	45.6	72.3	**74.0**	37.8	20.9
	SOM	Sample	68.2	67.7	63.1	70.2	63.6	67.7	68.7	73.0	73.5	73.4	65.8	65.7	75.0	74.3	81.0	86.0
		Catchment	52.0	59.9	31.0	23.7	60.6	62.1	38.4	40.4	70.5	70.5	51.3	51.3	72.6	72.2	42.2	22.4
Standard	MLP	Sample	65.6	62.9	46.0	error	61.1	59.4	50.1	57.1	66.2	66.6	42.1	45	72.8	68.4	37.2	68.4
		Catchment	58.0	61.2	45.0	error	65.0	65.1	39.3	54.4	68.4	67.5	30.0	30.4	73.7	66.2	67.2	30.9
	SOM	Sample	73.3	70.8	75.3	77.7	68.0	69.0	75.0	80.0	76.6	78.1	64.8	70.1	67.0	67.0	64.0	64.0
		Catchment	58.0	62.8	45.0	30.0	64.8	63.7	45.9	36.7	72.5	***74.6***	57.0	54.3	60.9	60.9	39.4	39.4

Accuracy levels tend to be on average over 10% higher if models are validated at the training sample rather than the catchment level and if the sampling strategy also uses training sites that are in the vicinity (SN data) because soils are spatially autocorrelated. Indeed, the results of the "sample validation" are shown mainly to illustrate the fact that validation procedures should not be limited by software because validation of soil-landscape models should be performed at the spatial extent in which models are meant to make predictions, and this is clearly the catchment. Therefore, it is interesting to note that predictions at the catchment scale are substantially higher if using random (RD) data rather than shortest-distance (SN) data (66% *vs.* 40%, respectively). This is due to the fact that the latter influences the models to create too large homogeneous patches which are generally correct locally, but fail to predict accurately globally, i.e. over the whole catchment.

4. CONCLUSIONS

Considering accuracy levels measured at the catchment scale, most accurate predictions were consistently made using the MLP network with 12-variable random data, and marginally better with 90m pixels (66% –74% accuracy levels, highlighted in bold in Table 1). Thus, the addition of location data as input data clearly improved models because ANNs learnt that distance influences soil class, i.e. close neighbouring positions are more likely to be similar and vice-versa, indicating that the flow of water and sediment across the hillslopes gradually changes soils. Therefore, the spatial distribution of soils in the studied catchments conforms to the principle of soil catena, which means that hillslopes have some degree of connectivity. Thus, the higher the connectivity of the landscape, where there are significant lateral water and sediment flows, the more important it is to use "location" as input data in landscape models for the prediction of soils over large areas.

ACKNOWLEGMENTS

The authors acknowledge with gratitude the DRAEDM for providing digital soil data. This work was produced for AutoMAPticS, a project supported by a grant from FCT - Portugal (PTDC/CS-GEO/111929/2009), whose PI is employed under the FCT Science 2008 Programme.

REFERENCES

Dobos, E., Carré, F., Hengl, T., Reuter, H.I. & Tóth, G. (2006) *Digital Soil Mapping as a Support to Production of Functional Maps*. EUR 22123 EN, OFPEC, Luxemburg, 68 p.

Haykin, S. (1999): *Neural Networks – a Comprehensive Foundation*. Prentice Hall, New Jersey, 842p.

McBratney, A.B., Mendonça Santos, M.L., Minasny, B. (2003) On digital soil mapping. *Geoderma*, 117, 3-52.

Mora-Vallejo, A., Claessens, L., Stoonvogel, J. & Heuvelink, G.B.M. (2008) Small scale digital soil mapping in southeastern Kenya. *Catena*, 76, 44-53.

Tso, B. & Mather, P.M. (2001) *Classification Methods for Remotely Sensed Data*. Taylor and Francis, London, 332 p.