Digital Soil Mapping using Artificial Neural Networks – Sampling Issues

RICARDO BRASIL [1]; INÊS FONSECA [1]; SÉRGIO FREIRE [2]; JORGE ROCHA [1]; JOSÉ TENEDÓRIO [2]

[1] RISKam, Centre for Geographical Studies, Institute of Geography and Spatial Planning - University of Lisbon, *Edf. Fac. Letras, Universidade de Lisboa, Alameda da Universidade, 1600-214 Lisbon, Portugal.* rmsb@campus.ul.pt

[2] *Centro de Estudos de Geografia e Planeamento Regional, Faculdade de Ciências Sociais e Humanas da Universidade Nova de Lisboa, Av. De Berna 26-C, 1069-061 Lisbon, Portugal*

Introduction

Digital Soil Mapping (DSM) is an advanced technique for mapping soil classes (Dobos *et al*., 2006) which has been developed to bridge the gap between existing soil maps based on traditional soil survey and the increasing demand for soil information. Indeed, at the European level, DSM has been driven by the urgent need to address the importance of soils and the growing concern about environmental disasters, the impact of human activities on soils and the role that soil has on global change.

Artificial Neural Networks (ANNs) are sophisticated computer programs which are able to model complex functional relationships. As such, ANNs provide the means to predict soil types at locations without soil spatial data by combing existing soil maps with factors known to be responsible for the spatial variation of soils (McBratney *et al*., 2003). Thus, a set of variables related to soil forming factors and the respective soil type are used as training data for the ANNs, which construct rules (Tso & Mather, 2001) that can be extended to the unmapped areas.

Whilst the literature provides a number of examples where DSM is presented as an efficient surveying technique and soil spatial variation is 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 training sites have on the predictive accuracy of the models.

Indeed, sampling method and location of training sites is particularly important for ANNs because their rate of learning, convergion to a solution, network performance and ability to generalise depend on the efficiency of the layout of the sampling pattern which, in turn, depends on the presence of spatial periodicity of the phenomena. Although all environmental variables exhibit spatial autocorrelation at some scale (Englund, 1988), high values found in the spatial distribution of the variables used to train an ANN is likely to affect its performance. Thus, the main objective of this work is to assess the impact that sampling methods used to select training areas for an ANN have on their predictive accuracy.

Study area

The study area is a catchment in Mondim de Basto, north-western Portugal, approximately 900km2 in area. The catchment was chosen because it presents a varied geomorphological and ecological setting and a number of soil types that are well representative of the soil types found in the region between the Douro and Minho rivers.

Mondim de Basto, Entre-Douro-e-Minho (Portugal)

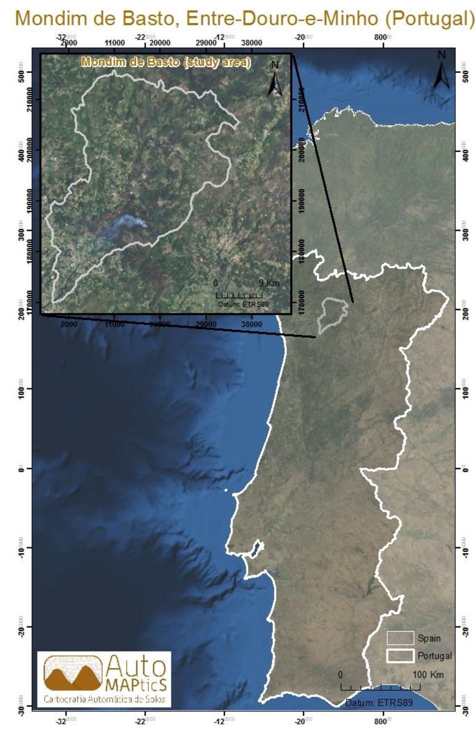
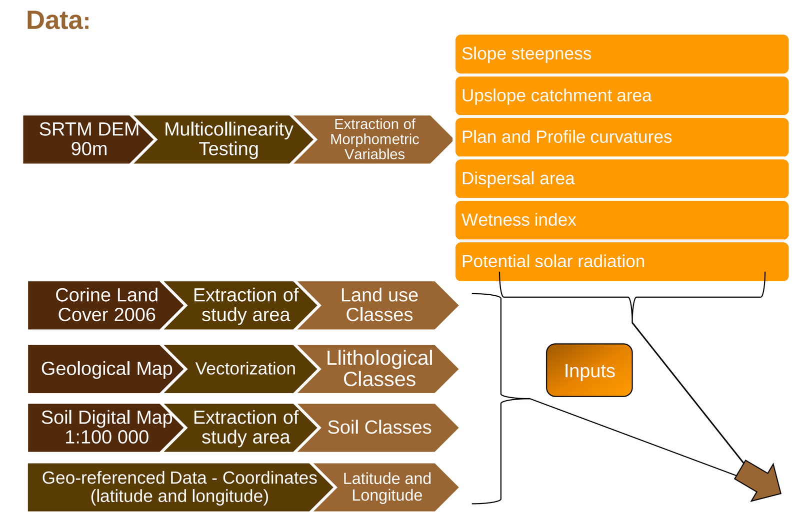


Figure 1 – Mondim de Basto catchment, in NW Portugal

Material and methods

Data:



Digital soil data at 1:100000 were provided by DRAEM, the regional agriculture department of North West Portugal. In order to account for the possible effect of autocorrelation, the coordinates (latitude and longitude) were also included in the input vector to indicate location.

Sampling:

Two different sampling strategies were implemented for training a multi-layer perceptron (MLP) neural network model in IDRISI Taiga (Clark Labs), using a highly popular supervised method known as error back-propagating algorithm (Haykin, 1999). Thus, the ANN was trained by presenting it a number of different examples of the same soil type drawn either (i) randomly (RS), or (ii) in a stratified fashion (SS). For the latter, training pixel vectors were located by choosing (a) random coordinates within soil types strata (SRS), (b) random coordinates within soil types strata and chosen evenly in the frequency space (Figure 4) (SRPS) and, (c) nearest coordinates within soil types strata and chosen evenly in the frequency space (SNPS).

The experimental setup for each training set used fixed input specifications, as presented in table 1 (200 pixels per class for training and testing), two network topology settings (1 hidden layer, no layer two nodes and either 7 or 8 layer one nodes), with a variety of training parameters changes which included using automatic training and a dynamic learning rate which, if used, could vary between 0.01 and 0.2. The stopping criteria were achieving either a RMSE ≤ 0.01, an accuracy of 100% or a maximum number of iterations ranging from 1,000 to 100,000.

Classifier parameters		
Group	**Parameter**	**Default value**
Input specifications	Avg. training pixels per class	500
	Avg. testing pixels per class	500
Network topology	Hidden layers	1
	Layer 1 nodes	1
Training parameters	Use automatic training	no
	Use dynamic learning rate	no
	Learning rate	0.01
	End Learning rate	0.001
	Momentum factor	0.5
	Sigmoid constant "a"	1
Stopping criteria	RMS	0.01
	Iterations	10000
	Accuracy rate	1

Table 1 – Characteristics of the ANN.

Results and discussion

Spatial autocorrelation assessment, measured through Moran´s I, indicates that autocorrelation is significantly high for wetness index (0.65) and slope steepness (0.76) and very high for potential solar radiation (0.88) and altitude (0.99). Analysis of all the results obtained with the different model parameterisations shows that the predictive accuracy of the ANN models is highly dependent on the sampling method and highly correlated with RMSE but not so dependent on the number of iterations (Table 2).

Whilst random sampling did not achieve as good predictive accuracy results as the one possible to obtain with stratified sampling (65% *vs.* 75%), it is clear that spatial autocorrelation causes an oustanding drop-off in the number of iterations required to achieve similar levels of accuracy (71% and 75%) and RMSE (0.31). Thus, accounting for spatial autocorrelation by choosing pixels that are as close as possible to each other (SNPS) resulted in only 5,000 iterations being required (as opposed to 30,000) to achieve similar accuracy levels.

Sampling Method	Iterations	Testing		Accuracy (%)	RMSE
		Min RMSE	Max RMSE		
SRPS	50000	0.37	0.39	57.8	0.36
RS	40000	0.36	0.46	65.4	0.35
SNPS	5000	0.31	0.43	71.1	0.31
SRS	30000	0.3	0.37	74.7	0.31

Table 2 – Impact of sampling method on the performance of ANN models and no. of iterations required to achieve the best results obtained with each method, assessed through predictive accuracy level and minimum and maximum RMSE values in the testing set.

Conclusions

The main conclusions of this work are:

(1) sampling strategy has a very important impact on the accuracy of soil predictive maps developed using ANNs and different strategies should be tested, and

(2) sampling strategy benefits from reflecting high autocorrelation of factors of soil formation because the ANN learns faster that close neighbouring positions are more likely to have similar soil types, allowing the ANN to converge faster to a better solution.

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. Office for Official Publications of the European Communities, Luxemburg, 68pp.

ENGLUND E.J. (1988) *Spatial Autocorrelation: Implications for Sampling and Estimation*. In Liggett W. (Ed.) Proceedings of the ASA/EPA Conferences on Interpretation of Environmental Data, III Sampling and Site Selection in Environmental Studies, EPA 230/8-88/035, 31-39

HAYKIN S. (1999) *Neural Networks – a Comprehensive Foundation*. Prentice Hall, New Jersey, 842pp.

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, 332pp.

Training set and model

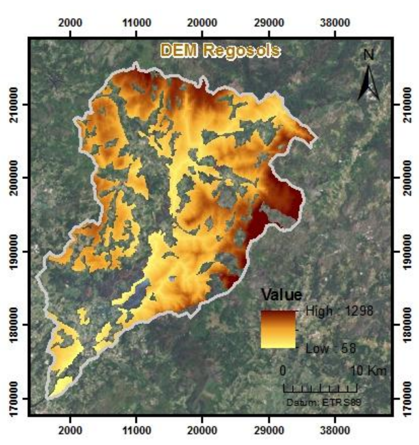


Figure 2 - DEM for the Regosols soil class

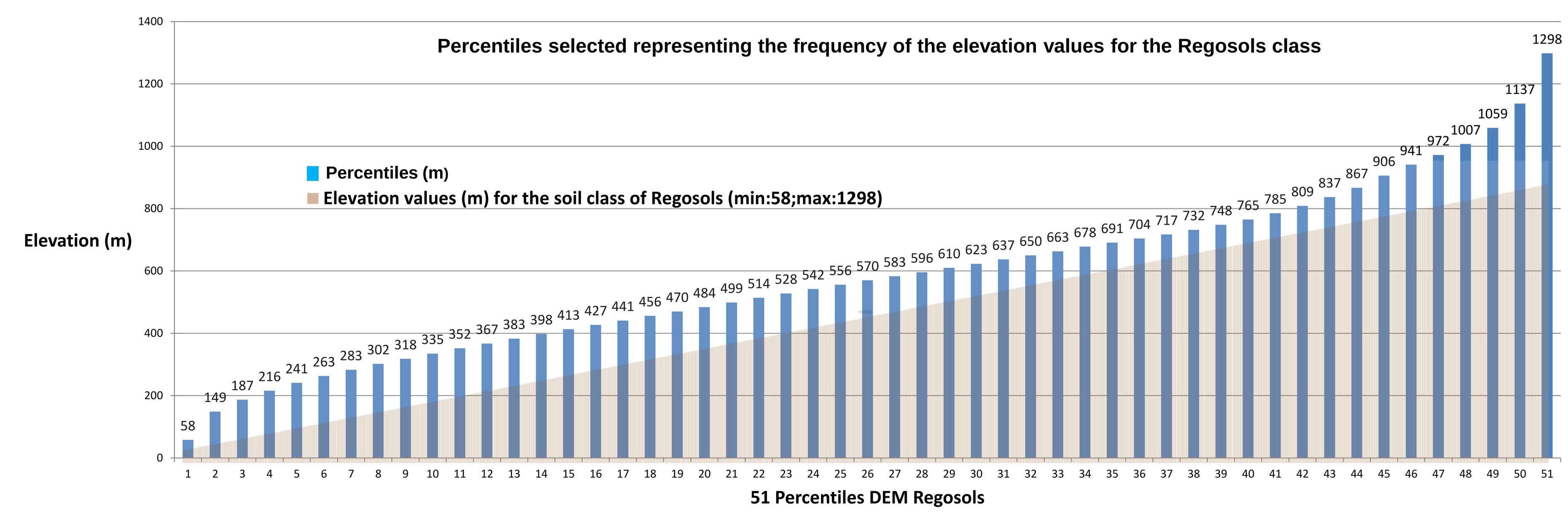


Figure 3 - Example of the 51 values (50 percentiles + min value) selected to represent the frequency distribution of the elevation values for Regosols in Mondim de Basto

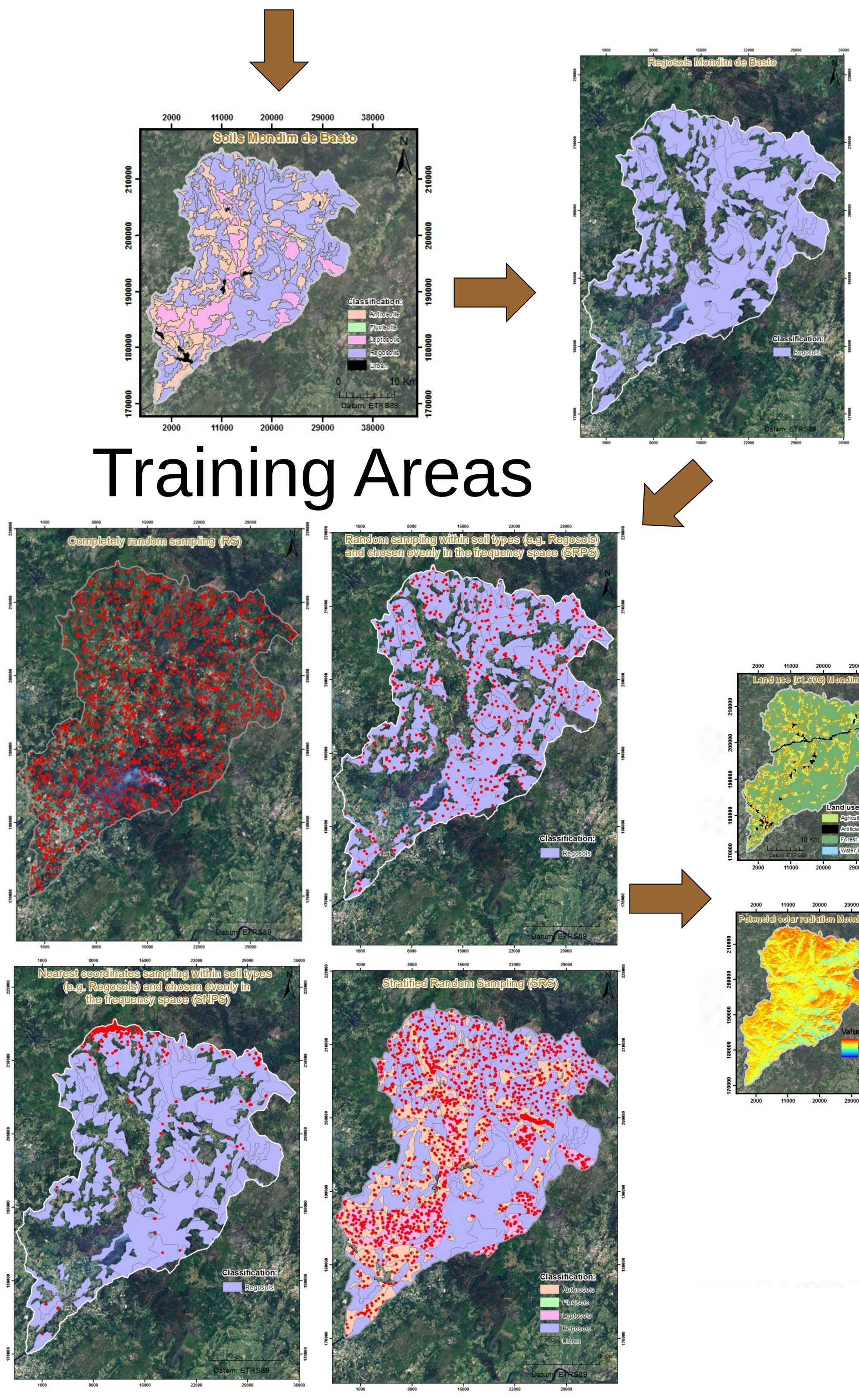


Figure 4 - Spatial distribution of training sites for the different sampling strategies

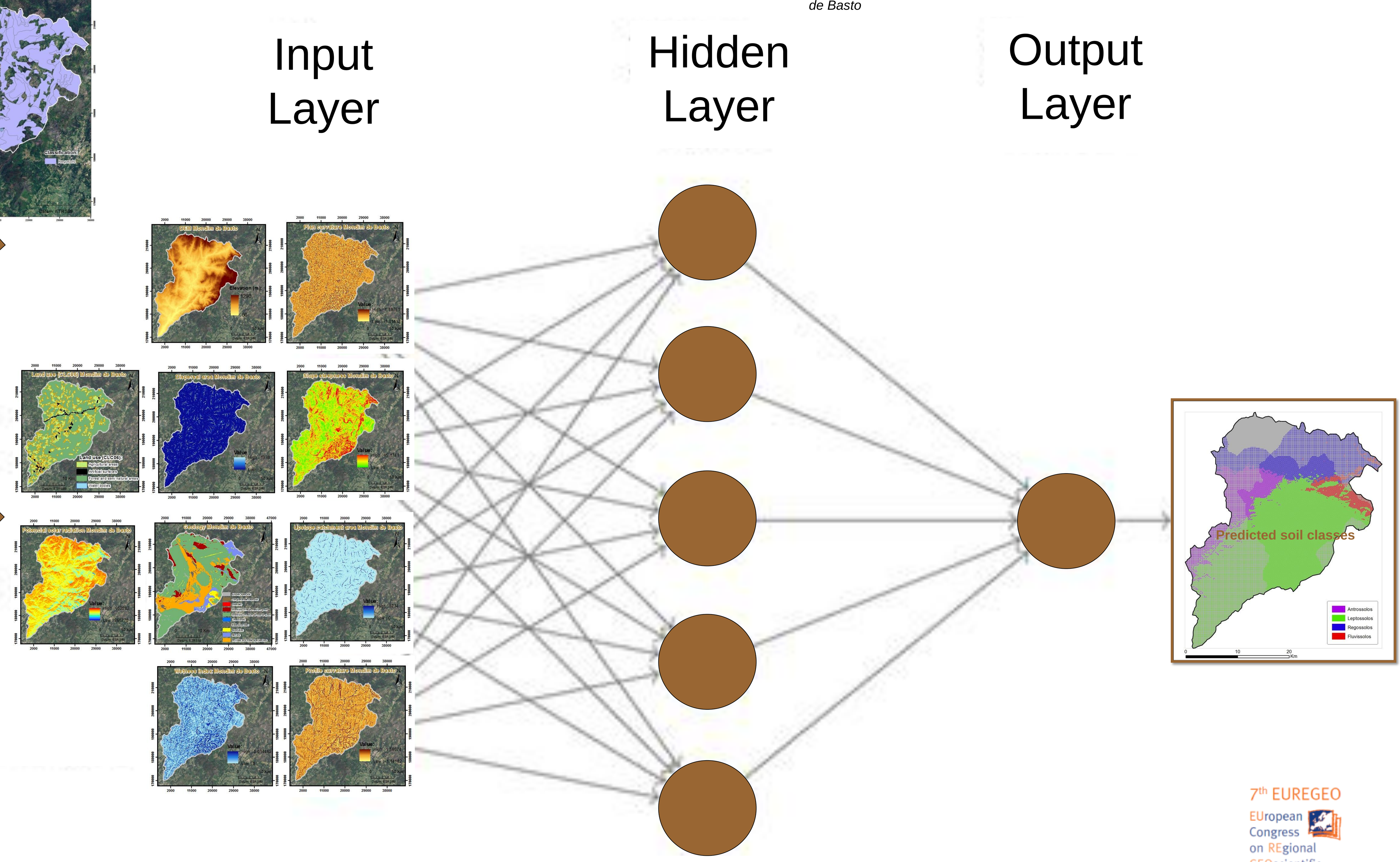


Figure 5 - Illustration of the configuration of an ANN showing the input maps, layer nodes and output map

Research Group on Environmental Hazard and Risk Assessment and Management
Centre of Geographical Studies
Institute of Geography and Spatial Planning, University of Lisbon
www.riskam.ul.pt

RISKam - Environmental Hazard and Risk Assessment and Management

7th EUREGEO European Congress on REgional GEOscientific Cartography and Information Systems