XXXV Congresso Brasileiro de Ciência do Solo
CENTRO DE CONVENÇÕES - NATAL / RN
O SOLO E SUAS MÚLTIPLAS FUNÇÕES
02 a 07 DE AGOSTO DE 2015

Modeling and refining soil maps from legacy data comparing KnowledgeMiner with decision trees[1].

Sérgio Henrique Godinho Silva[2]; Michele Duarte de Menezes[3]; Phillip Ray Owens[4]; Nilton Curi[5]

[1] Trabalho executado com recursos do CNPq, Capes e FAPEMIG.
[2] Doutorando em Ciência do Solo; Universidade Federal de Lavras (UFLA); Lavras, MG; sergiohgsilva@gmail.com; [3] Professora Adjunta; Depto. de Ciência do Solo UFLA; [4] Professor Associado; Depto. de Agronomia da Purdue University; [5] Professor Titular; Depto. de Ciência do Solo UFLA.

RESUMO: Diverse projects are being carried out worldwide focusing on the development of more accurate soil maps. This work aimed to compare two data mining tools, KnowledgeMiner and decision trees, to retrieve the soil legacy data from a detailed soil map of a watershed in Minas Gerais, Brazil, and then to create and validate the soil maps in the field to identify the best method for future refining of soil maps. The study area is a watershed located in Nazareno county, state of Minas Gerais, Brazil. From the existing detailed soil map, terrain attributes information was retrieved by each polygon of the map. KnowledgeMiner and decision trees were employed to identify the pattern of each soil class according to 12 terrain attributes and to create a new soil map by method. Validation was performed in the field at 20 places chosen by cost-constrained conditioned Latin hypercube scheme. KnowledgeMiner maps had an accuracy of 80% and 0.6524 kappa index against 55% and 0.0674 for decision trees. Digital mapping tools are contributing to improve existing soil maps using legacy data. KnowledgeMiner had a better performance than decision trees to retrieve knowledge and map the soils of the study area.

Index terms: Digital soil mapping, Pedology, soil classes prediction.

INTRODUCTION

The global search for more detailed soil maps has gained increasing importance. Diverse projects are being carried out worldwide focusing on the development of more accurate soil maps. This fact is associated with diverse technological advances in the recent years, such as the powerful electronic devices, the ease of accessing information, satellite data availability, and so forth, from which pedologists can usufruct.

Some of the most useful tools available are digital elevation models (DEMs) found at different resolutions that provide great information and from which terrain attributes, such as slope, curvature and topographic wetness index, can be derived. Many works have applied them to predict soil properties and classes (Vaysse & Lagacherie, 2015; Adhikari et al., 2014; Menezes et al., 2014). They consist of studying the relief as a major driver for soil differentiation, considering that the other soil forming factors (climate, organisms, parent material and time) (Jenny, 1941) are relatively constant in the study area.

The SCORPAN model (McBratney et al., 2003) includes soils existing information (legacy data) (s) and their spatial location (n) to the Jenny's model, which allows for more quantitative descriptions of relationships between soils and other factors.

Among the data mining tools to use legacy data, decision trees are one of the most common. They are simple to understand and can identify the most representative variables to prediction (Kheir et al., 2010), consisting of a supervised way of discovering the mental model encrypted in the soil map. Another tool more recently created is the KnowledgeMiner that is part of the SoLIM software (Zhu et al., 2001). It employs Kernell density to extract environmental variables information and then provide several statistical indexes to characterize each polygon on the map. It also generates frequency distribution curves that allow the user to identify the most appropriate variables to individualize each soil class.

Combining the need for more detailed soil maps in Brazil, where most of them are at a 1:750,000 scale due to funding limitations (Giasson et al., 2006), with the feasibility of using digital soil mapping tools to rescue information embedded on maps it has brought to light an economic alternative to improve those maps in a digital environment. Thus, this work aimed to compare two data mining tools, KnowledgeMiner and decision trees, to retrieve the soil legacy data from a detailed soil map of a watershed in Minas Gerais, Brazil, and then to create and validate the soil maps in the field to identify the best method for future refining of soil maps.

MATERIAL AND METHODS

The study was developed at Marcela Creek

Watershed, located in Nazareno county, state of Minas Gerais, between the latitudes 21°14'27'' and 21°15'51'' S and longitudes 44°30'58'' and 44°29'29'' W. The climate of the study area is Cwa, according to Köppen classification, being characteristic of dry winters and warm and rainy summers, presenting a mean annual precipitation of 1,300 mm and mean annual temperature of 19.7°C.

This area was mapped by Motta et al. (2001), at a scale of 1:12,500, through intensive field work, including description of 5 soil profiles and collection of 4 samples, and aerial photographs and contour lines were analyzed on a stereoscope, making up the basic source of information for the development of the current work. The soil classes found were Red-Yellow Latosol (LVA), Red Latosol (LV), Haplic Cambisols (CX) and indiscriminate hydromorphic soils (SIV).

A 30 m Aster Digital Elevation Model was used to create 12 terrain attributes: slope gradient, topographic wetness index (TWI), SAGA wetness index (SWI), longitudinal curvature, cross-sectional curvature, multiresolution index of valley bottom flatness (mrrtf), multiresolution index of top ridge flatness (mrrtf), vertical distance to channel network (VDCN), hillshade, aspect, and valley depth on SAGA GIS software (Böhner et al., 2006). From them, knowledge extraction was done for each soil map polygon using both KnowledgeMiner, after a selection of the most representative ones by analysis of boxplots, and decision trees. From the rules obtained for each MUP, soil maps were created by algebra of maps for decision trees and ArcSIE (Shi, 2013) for KnowledgeMiner.

Validation was perfomed by observing 20 places in the study area, chosen by cost-constrained conditioned Latin Hypercube sampling scheme (Roudier et al., 2012). Global index and Kappa index were calculated to assess their accuracy.

RESULTS AND DISCUSSION

KnowledgeMiner for mapping soils

Figure 1 represents the curves generated by KnowledgeMiner for the terrain attributes found to be more relevant to separate the soil classes accordingo to a boxplot analysis. These curves aid to identify the degree of overlapping of values for each TA and soil classes. The more individualized is one of the curves, the better that TA is to distinguish a soil class. It is observed that VDCN is one of the TAs that has more different values for separating the soil classes, while slope and WI are more adequate to individualize the SIV, since the curves of these two TAs have great overlapping for the other soil classes.

Possessing the characteristic values for each soil class **(Table 1)** by each method, they were inserted in ArcSIE (Shi, 2013), an ArcGIS extension, to create the predicted map (predicted maps presented in future section).

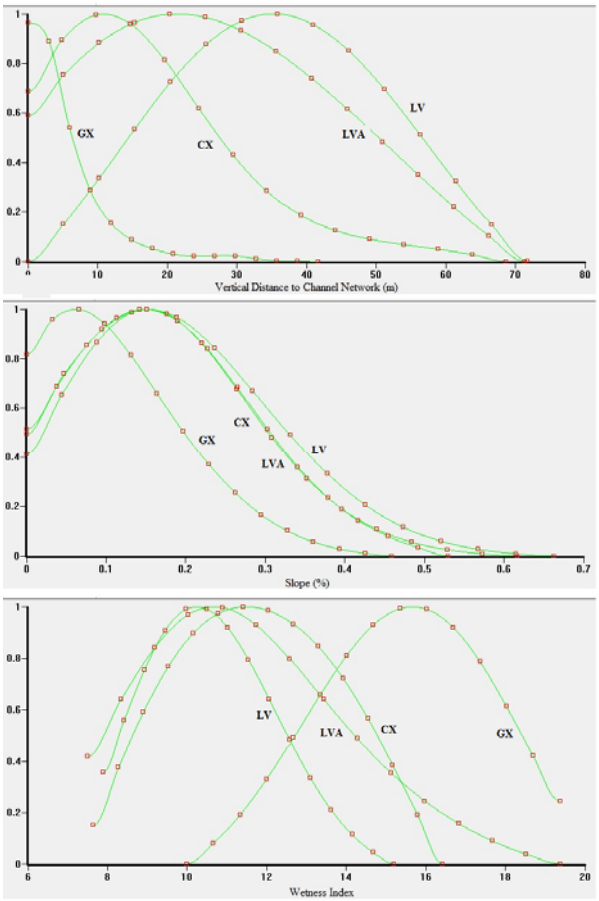

Figure 1 - Curves generated in KnowledgeMiner to identify the best attributes to separate each soil class.

Table 1 - Mean values for each soil class generated by KnowledgeMiner.

Soil Class	VDCN	Slope (%)	WI
CX	16.6	13.5	11.8
SIV	3.7	7.0	15.4
LVA	25.5	14.0	11.4
LV	35.7	13.8	10.5

Decision Trees for mapping soils

Although the use 12 TAs as input data, only 5

TAs were employed by the decision tree to separate the soil classes, being them TWI, VDCN, longitudinal curvature, aspect and valley depth. Kheir et al. (2010) and Jafari et al. (2014), employing many environmental covariates for modeling and predicting soil properties, also found that the covariates have different importance on modeling and for this reason it is normal that some are more common to be used in the predictions than others. It is noticed that the Inceptisols were not predicted by the decision tree **(Figure 2)**. It was probably found because this soil class occurs in similar landscape of that where Latosols are common.

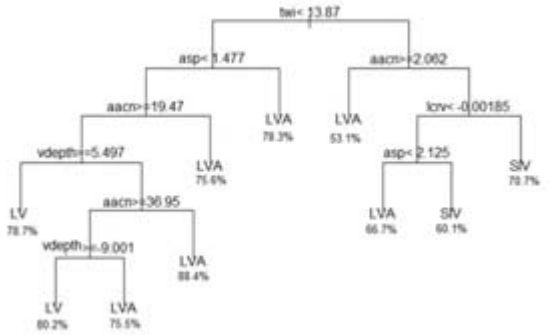

Figure 2 - Decision tree for predicting the soil classes.

Validation of the maps

In the map created from KnowledgeMiner procedure, 16 out of the 20 samples (80%) were correctly predicted by the soil map and resulted in a Kappa index of 0.6537, equivalent to a substantial classification according to Landis & Koch (1977), while the decision tree correctly predicted 11 out of 20 samples (55%) and had a Kappa index of 0.0674, corresponding to a slight agreement between this map and the original **(Figure 3)**.

CONCLUSIONS

The use of digital soil mapping tools are contributing to refine existing maps in an economic and efficient way.

KnowledgeMiner had a better performance to extract knowledge from an existing map and to support the generation of rules to refine the soil map than decision trees, although being more time-consuming.

ACNKOWLEDGEMENTS

The authors thank CNPq, Capes and FAPEMIG for providing financial support for this research.

REFERENCES

ADHIKARI, et al. Constructing a soil class map of Denmark based on the FAO legend using digital techniques. Geoderma, 214-215:101-113, 2014.

BÖHNER, J.; MCCLOY, K.R.; STROBL, J. SAGA - analysis and modeling applications. Gottinger Geographische Abhandlungen. 2006. v.115, 130p.

GIASSON et al. Digtal Soil mapping using multiple logistic regressions on terrain parameters in southern Brazil. Science Agriculture, 63: 262-268, 2006.

JAFARI, et al. Spatial prediction of soil great groups by boosted regression trees using a limited dataset in an arid region, southeastern Iran. Geoderma, 234:1-27, 2014.

JENNY, H. Factors of soil formation: A system of quantitative pedology. New York: McGraw Hill Book Company, 1941. 281p

KHEIR, et al. Spatial soil zinc content distribution from terrain parameters: A GIS-based decision-tree model in Lebanon. Environmental Pollution, 158:520-528, 2010.

LANDIS, J.R. & KOCH, G.G. The measurement of observer agreement for categorical data. Biometrics, 33:159-174, 1977.

MCBRATNEY, A.B.; MENDONÇA-SANTOS, M.L.; MINASNY, B. On digital soil mapping. Geoderma Regional, 117:3-52, 2003.

MOTTA, et al. Levantamento Pedológico Detalhado, Erosão dos Solos, Uso Atual e Aptidão Agrícola das Terras de Microbacia Piloto na Região sob Influencia do Reservatório da Hidrelétrica de Itutinga - Camargos - MG. Lavras: UFLA;CEMIG, 2001. v.1, 51p.

VAYSSE, K. & LAGACHERIE, P. Evaluating Digital Soil Mapping approaches for mapping GlobalSoilMap soil properties from legacy data in Languedoc-Roussillon (France). Geoderma Regional, 4:20-30, 2015.

ZHU, A. X. et al. Soil Mapping Using GIS, Expert Knowledge, and Fuzzy logic. Soil Science Society of America Journal, 65:1463-1472, 2001.

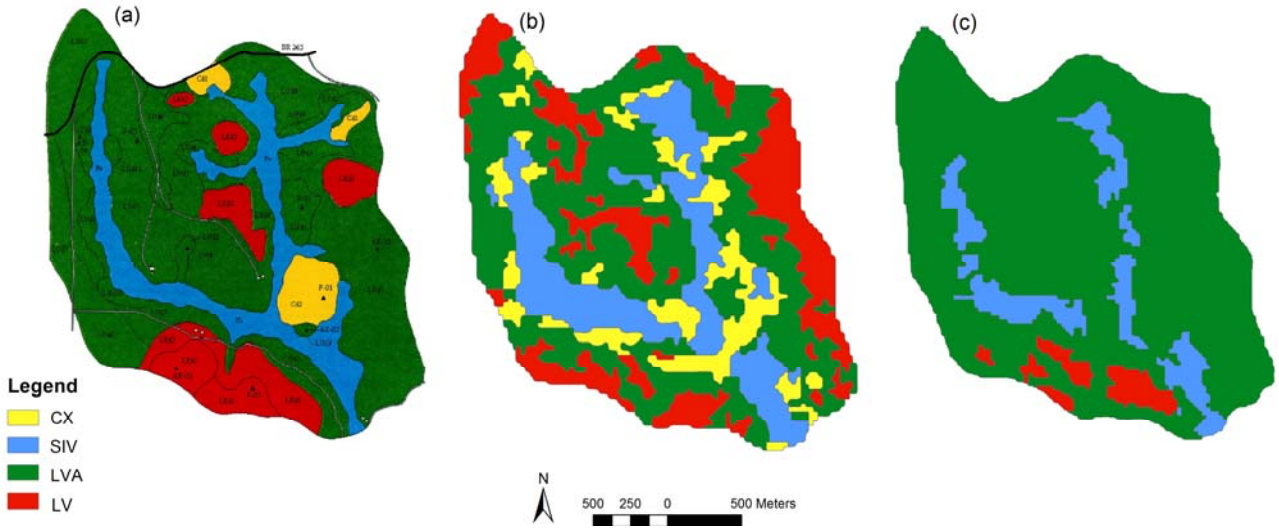

Figure 3 - Original soil map (a) and predicted maps generated from KnowledgeMiner (b) and decision trees (c) of Marcela Creek Watershed.