Geoderma 267 (2016) 65–77

Contents lists available at ScienceDirect

Geoderma

journal homepage: www.elsevier.com/locate/geoderma

# Retrieving pedologist's mental model from existing soil map and comparing data mining tools for refining a larger area map under similar environmental conditions in Southeastern Brazil

Sérgio Henrique Godinho Silva [a],*, Michele Duarte de Menezes [b], Phillip Ray Owens [c], Nilton Curi [b]

[a] Soil Science Department, Federal University of Lavras, P.O. BOX: 3037, 37200-000, Brazil
[b] Soil Science Department, Federal University of Lavras, P.O. BOX: 3037, 37200-000, Brazil
[c] Department of Agronomy, Purdue University, Lily Hall of Life Sciences, 915 West State Street, West Lafayette, IN, 47907-2054, United States

ARTICLE INFO

*Article history:*
Received 15 June 2015
Received in revised form 21 December 2015
Accepted 22 December 2015
Available online 7 January 2016

*Keywords:*
Digital soil mapping
Knowledge acquisition
Fuzzy logic
Data mining
Tropical soils
Cost-constrained conditioned Latin hypercube sampling scheme

ABSTRACT

Diverse projects are being carried out worldwide focusing on development of more accurate soil maps and one of the most valuable sources of data are the existing soil maps. This work aimed to (i) compare two data mining tools, KnowledgeMiner and decision trees, to retrieve legacy soil data from a detailed soil map, (ii) to create and validate the predicted soil maps in the field with the objective to identify the best method for modeling and refining soil maps, (iii) extrapolating soils information to the surrounding similar areas and (iv) to assess the accuracy of this soil map. The study was carried out in Minas Gerais state, Southeastern Brazil. From a detailed soil map, information of 12 terrain attributes was retrieved from the entire polygon of each mapping unit of the map (MUP) and from a circular buffer around the sampled points (CBP). KnowledgeMiner and decision trees were employed to retrieve information per soil class and soil maps were created per method. A field validation of 20 samples was chosen by a cost-constrained conditioned Latin hypercube sampling scheme and the accuracy of all maps was assessed using a global index, Kappa index, and errors of omission and commission. The KnowledgeMiner MUP map had a greater accuracy than the other methods, being even more accurate than the original map, accounting for 80% of global index and a Kappa index of 0.6524. The information extracted by KnowledgeMiner provided rules for mapping the watershed surroundings with 70.97% of global index and a kappa index of 0.5586. Legacy soil data extracted by KnowledgeMiner from a detailed soil map and used to model soil class distribution outperformed decision trees, promoted improvements on the existing soil map, and allows for the creation of a low cost soil map for the surroundings of the study area.

## 1. Introduction

The global search for more detailed soil maps has gained increasing importance in the last two decades (Mendonça-Santos and Santos, 2007; McBratney et al., 2006; Hartemink and McBratney, 2008). Diverse projects are being conducted and focusing on the development of more accurate soil maps than existing ones, such as the AfSoilGrids250m (Hengl et al., 2015) that created soil property maps for Africa at 250 m resolution, GlobalSoilMap (Arrouays et al., 2014), which aims to make a new digital soil map of the world at a fine resolution, and SoilGrids1Km (Hengl et al., 2014), the first output for a series of finer resolution maps of soil properties and classes to be produced in the future. This fact is associated with diverse technological advances in recent years, such as powerful electronic devices, the ease of accessing digital information, and satellite data availability, from which pedologists can utilize to their advantage.

Some of the most useful tools available are digital elevation models (DEM) 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 these parameters to predict soil properties and classes (Moore et al., 1993; McBratney et al., 2000, 2003; Behrens et al., 2010; Jafari et al., 2014; Vaysse and Lagacherie, 2015). These works consisted of studying the relief as major driver for soil differentiation, considering the other soil forming factors (climate, organisms, parent material and time) (Jenny, 1941) as relatively constant in the study area.

*Abbreviations:* DEM, Digital Elevation Model; Hx, Hapludox; Ax, Acrudox; Dt, Dystrudept; Et, Endoaquent; TWI, topographic wetness index; SWI, SAGA wetness index; mrvbf, multiresolution index of valley bottom flatness; mrrtf, multiresolution index of top ridge flatness; VDCN, vertical distance to channel network; MUP, entire mapping unit polygon; CBP, circular buffer of 100 m from the sampled points; SoLIM, Soil Land Inference Model; ArcSIE, Soil Inference Engine; TA, Terrain attributes; CCLH, cost-constrained conditioned Latin hypercube sampling scheme.

* Corresponding author.
*E-mail addresses:* sergiohgsilva@gmail.com (S.H.G. Silva), michele.menezes@dcs.ufla.br (M.D.d. Menezes), prowens@purdue.edu (P.R. Owens), niltcuri@dcs.ufla.br (N. Curi).

http://dx.doi.org/10.1016/j.geoderma.2015.12.025
0016-7061/© 2015 Elsevier B.V. All rights reserved.

A more recent advance from the Jenny's model for soil formation (clorpt) is the SCORPAN (*soil* = *f* (*soil*, *climate*, *organisms*, *relief*, *parent material*, *age*, *n*), in which soils can be predicted from the classic five factors proposed by Jenny (1941) plus available information about the soils (s), such as existing maps, and soils spatial position (n). This model proposed by McBratney et al. (2003) allows for a more quantitative description of the relationships between soil and other referenced factors and it stresses that existing soil information (legacy data) could also be used to refine soil maps. In accordance with this fact, Silva et al. (2014) suggested that maps are made with the best tools and data available at the time they are created, but it does not impede that they can be updated as soon as more information is acquired in the future.

From this point of view, existing soil maps in Brazil, of which most were created prior to the advent of digital soil mapping, could be refined with current available tools. Soil maps represent the pedologist's mental model about soils variability across the landscape (Bui, 2004). Many digital mapping tools can retrieve this knowledge from existing maps and correlate it with environmental factors, such as geology, topography, and vegetation (Taghizadeh-Mehrjardi et al., 2015).

Among data mining tools, decision trees are one of the most commonly used for digital soil mapping (Lagacherie and Holmes, 1997; Giasson et al., 2011; Häring et al., 2012; Kempen et al., 2015). They can identify the conditions that characterize each soil class according to different environmental variables with reasonable accuracy (e.g. 65–88% of overall accuracy, and a Kappa index of 0.44–0.51, according to Scull et al. (2005)), in which the data set is divided into more homogeneous subsets (Moran and Bui, 2002).

The procedure of decision trees starts at a root node, where the algorithm identifies the optimal split based on an exhaustive search of all possibilities, in order to maximize the average purity of the two nodes, employing the splitting or impurity function called Gini index (Loh, 2011). Nodes are locales where trees split the data set; terminal nodes are called leaves. A leaf node (predicted soil class, for example) is created when the decision tree reaches a stopping criterion (condition defined by the algorithm implemented in the trees, e.g. when the maximum tree depth is reached, when the splitting criteria are not greater than a threshold, among others (Rokach and Maimon (2008)), which makes the tree stops splitting nodes. Otherwise, the aforementioned step is, in turn, applied to each child node.

Decision trees are simple to understand and can identify the most representative variables for prediction (Bou Kheir et al., 2010). This results in a consistent supervised way to aid in the comprehension of the pedologist's mental model encrypted in soil maps. Also, no assumptions are made regarding the underlying distribution of values of the predictor variables (non-parametric) (Friedman et al., 2000) and decision trees are able to search all possible covariates as splitters in the decision nodes. However, this exhaustive search approach has disadvantages, such as the one reported by Loh (2011), which is the greater chance of selecting the covariates that have more distinct values, if everything else is equal, affecting the integrity of inferences drawn from the tree structure. Henderson et al. (2005) used decision trees in Australia to predict soil pH and other properties based on terrain and climatic variables, at a 250 m resolution, and Lacarce et al. (2012) combined regression trees with geostatistics to predict Pb stocks in soils in France.

Another tool more recently created is the KnowledgeMiner that is part of the Soil Land Inference Model (SoLIM) software (Zhu et al., 2001). It employs Kernel density to extract environmental variables information from each polygon on the map and then provides statistical indexes, such as minimum, maximum, mean, mode, median and standard deviation, in order to characterize those polygons (map units) and help the user to define the optimal environmental conditions for each map unit to occur.

It considers each cell value of a terrain attribute raster and a numerical interval that contains that cell value. Then, it counts the number of cells within each polygon that has values contained in that interval (SoLIM, 2007). This number of cells will be used to generate the frequency distribution curves (Kernel density), which allow the user to identify the most appropriate value of terrain attributes (e.g. slope gradient values) to individualize each soil class. These data mining procedures may contribute to disaggregate polygons of the original map to create more detailed soil maps (Bui and Moran, 2001; Thompson et al., 2010; Nauman and Thompson, 2014; Subburayalu et al., 2014), however, a soil map whose polygons present more than one soil class (inclusions) may hinder these inferences.

Combining the need for more detailed soil maps in Brazil, where most of them are at a 1:750,000 scale due to increased funding limitations (Giasson et al., 2006), with the feasibility of using digital soil mapping tools, it has brought to light an economical alternative to obtain soil data: the usage of data mining tools to rescue information embedded on existing soil maps to improve those maps in a digital environment at a lower cost. Thus, this work had as objectives: (i) to compare two data mining tools, KnowledgeMiner and decision trees, to retrieve legacy soil data from a detailed soil map of a watershed in Minas Gerais, Southeastern Brazil; (ii) to create and validate these soil maps in the field, identifying the best method for refinement of soil maps; (iii) to extrapolate that extracted legacy data to the surrounding similar areas of this watershed, which present similar environmental conditions; and (iv) to validate this map.

## 2. Material and methods

### *2.1. Study area and source of data*

The study was developed at Marcela Creek Watershed, located in Nazareno county, state of Minas Gerais, Southeastern Brazil (Fig. 1), 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 (warm temperate), according to Köppen classification, having dry winters and warm and rainy summers, presenting a mean annual precipitation of 1300 mm, a mean annual temperature of 19.7 °C and area of 485 ha.

This watershed was chosen due to its great agricultural potential (Silva et al., 2013), high water yield capacity and potential for electric energy generation (Beskow et al., 2013), and for being representative of the Mantiqueira Fields physiographical region. Its water drains into the Itutinga/Camargos hydroelectric power plant reservoir, which is a very important source of electric energy for Southeastern Brazil. Whereas water management is a governmental concern, the knowledge of soils and their distribution are important since soils exert an influence on water movement in different ways (Mello and Curi, 2012). Additionally, there is an important environmental issue in this region: the native vegetation (cerrado and forest) has been rapidly replaced by extensive pasture or crops (more recently) promoting intense land degradation (Alvarenga et al., 2012). This fact could impair the maintenance of hydrological functions of both the study area (Beskow et al., 2013) and its surroundings.

This watershed was mapped by very experienced pedologists and published by Motta et al. (2001), at a scale of 1:12,500, through intensive field work, including description of soil profiles, collection and laboratory analyses of soil samples, making up the basic source of information for the development of this current work. The soil classes found were Hapludox (Hx), Acrudox (Ax), Dystrudept (Dt) and Endoaquent (Et), according to Soil Taxonomy (Soil Survey Staff, 1999).

A 30 m Aster (version 2) DEM, which is the best DEM resolution freely available for Brazil, was obtained from the website http://gdem.ersdac.jspacesystems.or.jp/, preprocessed in order to make it hydrologically consistent, and used to create 12 terrain attributes (TA) on SAGA GIS software (Bohner et al., 2006): slope gradient, topographic wetness index (WI) (Beven and Kirkby, 1979), longitudinal curvature, cross-sectional curvature, multiresolution index of valley bottom flatness