Comparison of four machine learning algorithms for digital soil mapping in the Vale dos Vinhedos, RS, Brazil

Sarmento, E. C.¹, Giasson, E.¹, Weber, E.², Flores, C. A.³, Hasenack, H.²

¹ Universidade Federal do Rio Grande do Sul, Faculdade de Agronomia. Porto Alegre, RS, Brasil. eliana.sarmento@ufrgs.br, giasson@ufrgs.br
² Universidade Federal do Rio Grande do Sul, Centro de Ecologia, Porto Alegre, RS, Brasil. eweber@portoweb.com.br, hasenack@ecologia.ufrgs.br
³ Embrapa Clima Temperado, Pelotas, RS, Brasil. flores@cpact.embrapa.br

Abstract

This work compares three neural networks (Fuzzy ARTMap, SOM, and MLP) and a decision tree (Gini) to predict soil classes in Rio Grande do Sul State, Brazil. A digital elevation model (DEM) with spatial resolution of 5 meters, a digital cartographic base at scale 1:5,000, and a detailed conventional soil map were used on Idrisi GIS software. Seven variables were calculated from the DEM and the cartographic base. Variable values and soil types were read in randomly distributed points with sampling densities of 0.5, 1, 1.5, 2 and 4 points per hectare. Data from sampling points were used to train the algorithms and to estimate soil orders in the whole study area. Accuracy assessment used the conventional soil map as reference. The Gini decision tree presented highest overall accuracy (71%) and Kappa Index (0.58) with 2 points per hectare, and was less sensitive to sampling density, obtaining Kappa above 0.5 in all sampling densities starting from 1 point per hectare. MLP neural network showed similar performance, but was more sensitive to sampling density, obtaining Kappa above 0.5 only with sampling densities starting from 1.5 points per hectare. Both algorithms showed potential to predict soil orders at detailed scale in the study area.

Keywords: digital soil mapping, machine learning, geographic information systems, DEM, detailed scale

1. Introduction

The unavailability of maps against the increased demand for soil information led to the development of modeling techniques to spatially predict soil properties or soil classes. The diffusion of technologies such as Global Positioning System (GPS), remote sensing imagery and Geographic Information Systems (GIS) facilitated the integration of spatial information from different sources and made possible to run complex analysis, allowing the emergence of digital soil mapping (Morris et al., 2000; McBratney et al., 2003; Lagacherie & McBratney, 2007).

As in conventional surveys, many studies on digital soil mapping are based on relationships between soils and landscape features. However, as the first soil-landscape relations are established in a qualitative way, in the digital soil mapping they are evaluated quantitatively from variables derived from Digital Elevation Models (DEM). Among the most used approaches to quantify these relationships are parametric methods such as linear and logistic regression, geostatistical analysis and fuzzy logic.

More recently non-parametric methods based on Machine Learning Algorithms (MLA) have been applied to digital soil mapping (Grinand et al., 2008). MLA are capable to process large amounts of multidimensional data with low level of human intervention, reduced processing time and accuracy equivalent or higher than parametric methods (Lippitt et al., 2008). However, there is still need to know more about their capabilities, limitations and potential application for digital soil mapping.

This study compares four machine learning algorithms for estimating occurrence of soil orders from variables related to relief in Serra Gaúcha, Rio Grande do Sul, Brazil. The main objective is to evaluate the performance of algorithms and their sensitivity to sampling density.
2. Material and methods

The study area is located in the Vale dos Vinhedos, in the wine production region of the Serra Gaucha, northeast of Rio Grande do Sul State, in Southern Brazil (Figure 1). The Vale dos Vinhedos covers an area of 8,279 ha and is the first Geographical Indication of Brazil. Currently the wine sector wants to raise to the category of Apellation of Origin, which requires detailed surveys of the factors that affect wine production, including soil types. The study area is defined by a rectangle with a surface of 673.5 ha between longitudes 51°34’31.86”W and 51°33’1.86”W and latitudes 29°10’31.78”S and 29°09’1.78”S and corresponds to a map sheet on scale 1:5,000 of the detailed soil survey of the Vale dos Vinhedos.

According to Köppen’s classification, the region has a climate Cfb (Moreno, 1961), subtropical with mild summer, mean temperature of the coldest month between -3°C and 18°C, mean temperature of warmest month below 22°C and rainfall evenly distributed throughout the year. The natural conditions of the region are heterogeneous, with large variations in elevation, slope and aspect and large variability of soil types.

The material consists of a Digital Elevation Model (DEM) with 5 meters spatial resolution, a digital cartographic base at scale 1:5,000 and a conventional detailed soil map (Sarmento et al., 2008). The conventional soil map contains 155 polygons and 37 map units belonging to four soil orders according to the Brazilian Soil Classification System (Embrapa, 2006): 10 Argissolos/Ultisols (15.08%), 16 Cambissolos/Inceptisols (41.77%), 4 Chernossolos/Molisols (33.96%) and 7 Neossolos/Entisols (9.19%). The GIS software used to perform all calculations and analysis was Idrisi Taiga (Clarklabs©).

The variables used as predictors were elevation, slope, aspect, profile curvature, flow accumulation, flow direction and distance to water streams. The first six were derived from the DEM and the last from the stream network. These variables describe important aspects of soil genesis and distribution of its occurrence throughout the landscape, such as water regime, erosion and deposition of sediments, organic matter concentration, depth of A horizon, among others (Florinsky et al., 2002). As relief is a major factor influencing soil formation in the study area, it is
expected that variables that represent the form and energy of the relief act as good soil type predictors.

The 37 classes of conventional map were grouped for the first level (soil orders), resulting in a map of 4 soil classes (orders). This map was used as reference to characterize soil-landscape relationships from randomly distributed points using five sampling densities: 0.5, 1, 1.5, 2 and 4 points.ha$^{-1}$. At each point were extracted information of the corresponding predictor variables previously mentioned and the identification of soil order.

Data collected was used to train the algorithms, establishing relationships between variables and the spatial distribution of soils, and then to predict occurrence of soil orders in the whole study area. Four classification algorithms based on the concept of machine learning were used, three implementations of artificial neural networks (MLP - Multi-layer Perceptron, Fuzzy ARTMap - Adaptive Resonance Theory, and SOM - Self-Organizing Map) and one decision tree (Gini). These two types of MLA are abstractions of the process of human learning, but differ fundamentally in their approach: neural networks simulate the operation of the structure of neurons and connections of the human brain, while decision trees simulate the human process of abstraction through a hierarchical categorization (Lippitt et al., 2008).

All four MLA algorithms were trained with data from the same sets of sample points. The agreement between the estimated maps and the conventional soil map used as reference was assessed using confusion matrix, overall accuracy and Kappa index.

3. Results and discussion

All estimated maps were spatially more detailed than the conventional soil map, showing that the use of these algorithms individualized smaller spatial units than those present in conventional soil map, which agrees with previous studies (Zhu, 2000; MacMillan, 2008).

Commission errors ranged from 0.1646 to 0.8182, and shows that Gini decision tree had the lowest commission errors for classes Ultisol, Inceptisol and Molisol, respectively with densities of 2, 4 and 4 points.ha$^{-1}$, while the MLP neural network presented the lowest commission error for Entisols using density of 4 points.ha$^{-1}$. Omission errors ranged from 0.0924 to 0.9489 and Gini decision tree had the lowest omission errors for classes Ultisol and Entisol with sampling density of 4 points.ha$^{-1}$ and MLP neural network for classes Inceptisol and Molisol, respectively for the densities of 1.5 and 4 points.ha$^{-1}$. Omission and commission errors tend to decrease with increasing density of sampling, i.e., with more training samples the relationships between the predictor variables and the soil classes are better characterized, making the rules more consistent and reducing confusion between classes.

Overall accuracy ranged from 42% to 71%, and Gini decision tree and MLP neural network had similar values for all densities of sampling, being superior to fuzzy ARTMap and SOM neural networks (Figure 2a). However, MLP neural network estimated only the two classes of larger
occurrence in the study area in the sampling density of 0.5 points.ha$^{-1}$, suggesting that this algorithm has a critical lower limit for number of training samples per class while the others only reduce the accuracy when sampling density decreases.

Kappa index ranged from 0.214 to 0.5813, and Gini decision tree presented Kappa higher than the three algorithms of neural networks in all sampling densities (Figure 2b). Moreover, the decision tree presented Kappa index above the threshold of 0.5, considered satisfactory, in all sampling densities from 1 point.ha$^{-1}$ while the MLP neural network presented Kappa index above 0.5 with the sampling density from 2 points.ha$^{-1}$, although with 1.5 points.ha$^{-1}$ had almost reached that threshold. Fuzzy ARTMap and SOM neural networks showed Kappa above 0.5 only for sampling density of 4 points.ha$^{-1}$.

Gini decision tree and MLP neural network were less sensitive to sampling density, while the Fuzzy ARTMap neural network showed the highest sensitivity and the SOM neural network showed intermediate values (Figure 2). The results agree with studies that had superior performance for decision trees over other machine learning algorithms, even with reduced sets of training samples, with the advantage of producing easily comprehensible classification rules (Foody, 1995; Lippitt et al., 2008). However, the performance of neural networks disagrees with studies in which SOM neural networks showed higher performance than the MLP neural network (Lippitt et al., 2008). The difference may be related to intrinsic characteristics of data or to differences in configuration of the neural networks structure, since MLP, though simpler in design, allows more modifications in the structure than SOM in both, the number of hidden layers and the number of neurons, what affects their performance.

With the higher sampling density (4 points.ha$^{-1}$) the four algorithms tend to converge, with overall accuracy between 66% and 71% (Figure 2a) and Kappa index between 0.5069 and 0.5566 (Figure 2b), suggesting that the number of samples is approaching a level from which the increase in the number of samples fails to add useful information and can generate overfitting. This sampling density matches the upper limit of recommended range for detailed soils surveys in Brazil, varying from 0.2 to 4 points.ha$^{-1}$ (IBGE, 2007). Therefore, it is reasonable to assume that the sampling density recommended for conventional soil surveys can also be used as reference for digital soil mapping, what can support the definition of standardized procedures.

It is important to stand out that results express the ability of MLA to reproduce the conventional soil map used as reference. Figure 3, for example, shows the two estimated maps that best fits the convencional soil map. Due to inherent uncertainties in the conventional map, it is likely to find higher accuracy with reference observations made in the field (Zhu, 2000; Zhou et al., 2004), as well as some difference between algorithms performance. Despite this, results provide valuable information to optimize field data collection in future studies.

Figure 3. (a) Conventional soil map; (b) Gini decision tree (2 points.ha$^{-1}$); (c) MLP Neural network (4 points.ha$^{-1}$).
4. Conclusions

The decision tree Gini has higher accuracy and is less sensitive to sampling density than neural networks Fuzzy ARTMap, MLP and SOM;

The MLP neural network has potential to predict soil orders in detailed scale, but the effect of number of neurons and hidden layers on their performance should be studied;

The sampling density range recommended for conventional detailed soil surveys can be used as reference to sample size in digital soil mapping at the same scale.

5. Acknowledgements

To FINEP and CNPq, for financial support.

6. References


