Catena 139 (2016) 232–240

Contents lists available at ScienceDirect

Catena

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

Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions

César da Silva Chagas, Waldir de Carvalho Junior *, Silvio Barge Bhering, Braz Calderano Filho

Embrapa Soils, Rua Jardim Botânico, 1024 Rio de Janeiro, RJ, CEP 22460-000, Brazil

ARTICLE INFO

Article history:
Received 6 July 2015
Received in revised form 29 December 2015
Accepted 1 January 2016
Available online 16 January 2016

Keywords:
Pedometrics
Landsat dataset
Digital soil mapping
Granulometric composition
Grain size

ABSTRACT

Soil texture is an essential and extremely variable physical property that strongly influences many other soil properties that are highly relevant for agricultural production, e.g., fertility and water retention capacity. In plain areas, terrain properties derived from a digital elevation model are not effective for digital soil mapping, and the variation in the properties of such areas remains a challenge. In this regard, remote sensing can facilitate the mapping of soil properties. The purpose of this study was to evaluate the efficiency of using of data obtained from the Thematic Mapper (TM) sensor of Landsat 5 for digital soil mapping in a semi-arid region, based on multiple linear regression (MLR) and a random forest model (RFM). To this end, 399 samples of the soil surface layer (0–20 cm) were used to predict the sand, silt and clay contents, using the bands 1, 2, 3, 4, 5 and 7, the Normalized Difference Vegetation Index (NDVI), the grain size index (GSI), and the relationships between bands 3 and 2, bands 3 and 7, and bands 5 and 7 (clay index) of the Landsat 5 TM sensor as covariates. Among these covariates, only band 1 (b1), the relationship between bands 5 and 7 (b5/b7) for sand, silt and clay, and band 4 (b4) for silt were not significantly correlated according to Pearson's correlation analysis. The validation of the models showed that the properties were best estimated using the RFM, which explained 63% and 56% of the spatial variability of sand and clay, respectively, whereas the MLR explained 30% of the spatial variation of silt. An analysis of the relevance of the variables predicted by the RFM showed that the covariates b3/b7, b5, NDVI and b2 explained most of the variability of the considered properties. The RFM proved to be more advantageous than the MLR with respect to insensitivity to overfitting and the presence of noise in the data. In addition, the RFM produced more realistic distribution maps of the soil properties than did the MLR, taking into account that the estimated values of the soil attributes were in the same range as the calibration data, while the MLR model estimates were out of the range of the calibration data.

1. Introduction

Information on soils, including knowledge of the variability in soil properties, is critical for the formulation of agricultural policies, soil management and monitoring of environmental impacts arising from land use. Indeed, the lack of such information can result in the adoption of inadequate public policies, which may increase the risk of ecosystem degradation and the emission of carbon into the atmosphere (Mulder et al., 2011).

According to Boettinger et al. (2008) and Ben-Dor et al. (2008), orbital remote sensing data can be used as environmental covariates in digital soil mapping, especially in arid and semi-arid regions, thus facilitating mapping and reducing the need for costly and time-consuming field surveys (Mulder et al., 2011). Several studies have demonstrated relationships between different soil properties and remote sensing data. Among these, studies on organic carbon (Gomez et al., 2008; Stevens et al., 2010) and particle size composition (Demattê et al., 2007; Breunig et al., 2008; Liao et al., 2013) are particularly noteworthy.

The most commonly used methods for the prediction of soil properties, using remote sensing data as environmental covariates, are MLR (Ben-Dor et al., 2002; Nanni and Demattê, 2006), partial least squares regression (Stevens et al., 2008; Gomez et al., 2008), geostatistical and hybrid methods (Lark and Bishop, 2007; Lark et al., 2012; Rivero et al., 2007; Eldeiry and Garcia, 2010) and boosting regression tree models (Ciampalini et al., 2014a, 2014b). By contrast, data mining methods such as the random forest model (RFM) are less commonly used.

Random forest regression is a data mining method that has some advantages over most statistical modeling methods, as noted by Breiman (2001) and Liaw and Wiener (2002). These advantages include the ability to model highly nonlinear dimensional relationships; the use of categorical and continuous variables; resistance to “overfitting”; relative robustness with respect to the presence of noise in the data; the establishment of an impartial measure of the error rate; the capacity to determine the relevance of the variables used; and the requirement of few parameters for implementation. The main disadvantage of this method

* Corresponding author.
E-mail addresses: cesar.chagas@embrapa.br (C.S. Chagas), waldir.carvalho@embrapa.br (W. de Carvalho Junior), silvio.bhering@embrapa.br (S.B. Bhering), braz.calderano@embrapa.br (B. Calderano Filho).

http://dx.doi.org/10.1016/j.catena.2016.01.001
0341-8162/

is the limited interpretability of the results because the relationship between the predictors and the responses cannot be examined individually for each tree in the forest, which is why this technique is often called a "black box" approach (Grimm et al., 2008).

Random forest regression was used by Grimm et al. (2008) for the spatial prediction of soil organic carbon in a region in Panama. These authors used the following environmental covariates: topographic properties, soil units, soil parent material and forest history of the area. Based on this approach, digital mapping was used to predict soil organic carbon with high spatial resolution, to provide an estimate of the prediction error, and to identify the importance of the predictor variables.

Viscarra Rossel and Behrens (2010) compared different data mining algorithms, including an RFM, for the prediction of soil organic carbon, clay content and soil water pH, using diffuse reflectance data from the visible to the near-infrared region (350–2500 nm), based on a dataset of 1104 samples of Australian soils. Wiesmeier et al. (2011) used an RFM to predict soil organic carbon in a semi-arid region of northern China, using the following predictive variables: use of land units, reference soil units, geological units and 12 terrain properties derived from a digital elevation model. According to the authors, the prediction accuracy and maps were acceptable and explained 42 to 62% and 66 to 75%, respectively, of the data variation.

Ließ et al. (2012) compared the efficiency of regression trees and an RFM for the spatial prediction of soil texture from soil properties, using data of 56 soil profiles in the southern Ecuadorian Andes. The results obtained showed that the RFM performed better than the regression trees and explained 30 to 40% of the variation in the texture of the soil surface. Among the terrain properties, elevation had the strongest influence on the results during the construction of the model.

In this paper, we evaluated the potential of Landsat 5 TM data and modern statistical models and techniques for the purpose of predicting the texture of the A horizon of soils. The purpose of this study was to compare the efficiency of MLR and an RFM in predicting the texture of the A horizon of soils in an area of the Brazilian semi-arid region that is characterized by sparse savanna vegetation and high-activity clay soils.

2. Materials and methods

2.1. The study area

The study was carried out in part of an area belonging to the irrigation project Salitre, in Juazeiro, State of Bahia. The selected area covers approximately 35,000 ha (Fig. 1).

According to the Köppen climate classification, the climate in this region is BSwh' (semi-arid climate with dry winters and rainy summers; mean temperature of the coldest month > 18 °C). Annual rainfall reaches approximately 400 mm, and the rainy season lasts from November to April; March is the wettest month, and the average annual temperature is approximately 26 °C. The xerothermic indices vary from 200 to 150, and the dry period lasts 7 to 8 months. Originally, the area had hyperxerophilic shrub-tree Caatinga vegetation with a marked degree of xerophytism, much of which was highly degraded due to timber extraction for various purposes. The relief of the area is essentially flat. The geologic components of the area consist mainly of limestone of the Caatinga formation of the Tertiary–Quaternary and of gneiss–granitic rocks of the Caraíba–Paramirim complex (Souza et al., 2003). In this area, the most representative soil types are Vertisols, Cambisols and Planosols, according to the Brazilian Soil Classification System (Embrapa, 2013).

2.2. Soil properties and environmental covariates

For soil analysis and the prediction of the sand, silt and clay contents, we used data of the surface layer (0–20 cm) of 399 soil profiles, collected in a detailed soil survey of the Salitre project and provided by the Companhia de Desenvolvimento dos Vales do São Francisco e do Parnaíba (Codevasf). These soil properties were chosen in view of their importance for local irrigation management. Particle size distribution was determined by a hydrometer, using sodium hexametaphosphate or hydroxide as a dispersing agent and separating the fractions as

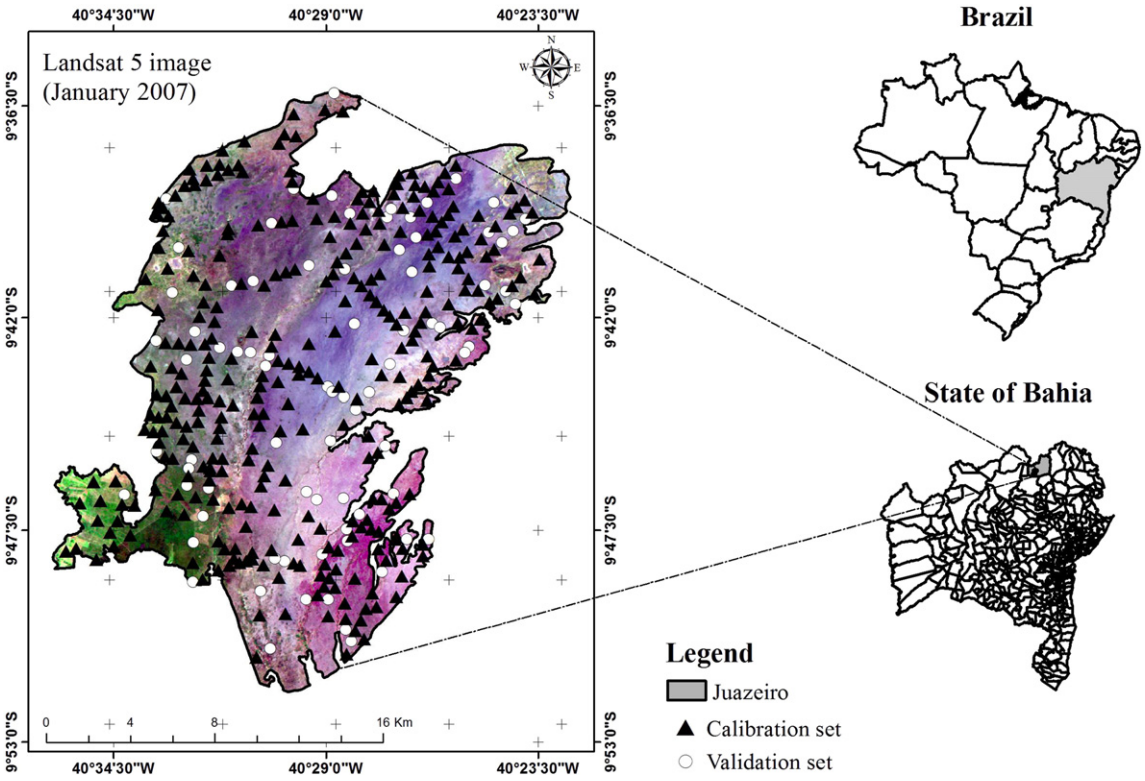

Fig. 1. Location of the study area in the state of Bahia and the spatial distribution of the studied soil profiles.