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 xxxx

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.

© 2016 Elsevier B.V. All rights reserved.

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/© 2016 Elsevier B.V. All rights reserved.

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

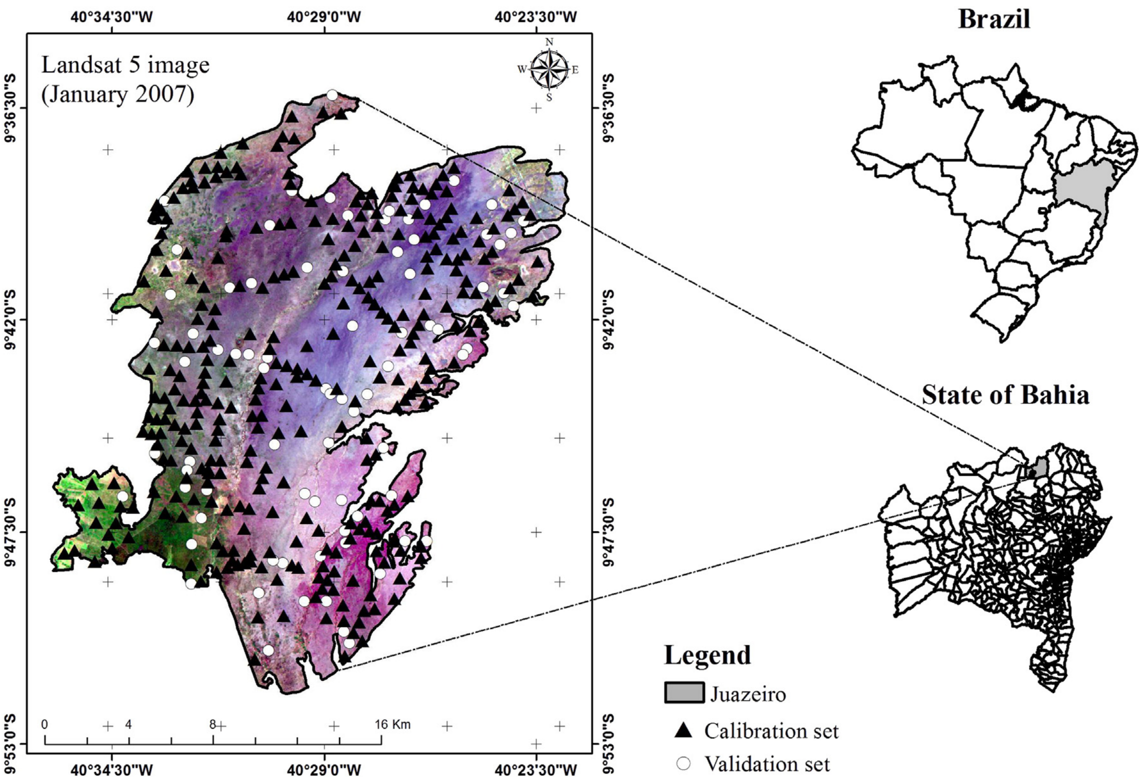

Fig. 1. Location of the study area in the state of Bahia and the spatial distribution of the studied soil profiles.

follows: sand (2–0.05 mm), silt (0.05–0.002 mm), and clay (<0.002 mm), as described by Embrapa (1979).

As environmental covariates, we used data from the Landsat 5 TM sensor (digital numbers) with a spatial resolution of 30 m, obtained from an image captured during the rainy season (January 2007), as follows: band 1 (0.450–0.515 μm), band 2 (0.525–0.605 μm), band 3 (0.630–0.690 μm), band 4 (0.755–0.900 μm), Band 5 (1.550–1.750 μm), band 7 (2.090–2.350 μm), NDVI (band 4 − band 3 / band 4 + band 3), and GSI ((band 3 − band 1) / (band 3 + band 2 + band 1)), as described by Xiao et al. (2006), and the relationships between band 3 and band 2 (b3/b2), between band 3 and 7 (b3/b7) and between band 5 and 7 (b5/b7), as described by Malone et al. (2009) and Carvalho Junior et al. (2014).

According to Xiao et al. (2006), the GSI was specifically designed for using Landsat TM/ETM + data. Considering the reflectance spectra of different soil surfaces and vegetation, the difference between bands R (band 3) and B (band 1) in the GSI equation is designed to distinguish between the vegetated or water surface and bare soil; meanwhile, the accumulation of the reflectance in the R, G (band 2) and B bands can discriminate among topsoils composed of different grain sizes. Therefore, the designed GSI can potentially detect surface soil texture or grain size composition. The GSI value is close to 0 in vegetated and water areas, and it can sometimes even be negative.

The importance of the environmental covariates was estimated using Pearson's correlation coefficients, which are normally used to measure the linear association between variables, implemented in R (R Development Core Team, 2007) using the cor.test function, as proposed by Ciampalini et al. (2012) and Carvalho Junior et al. (2014). In Pearson's correlation, the p-value defines whether two variables are statistically correlated; in our study, p values below 0.05 were assumed to indicate significant correlations.

2.3. Prediction models

In this study, we used multiple linear and random forest regressions. MLR is a classical method that has been widely used to predict values of a (dependent) response variable from (independent) predictor variables, to recognize the interaction between these variables and to explore the forms in which they are correlated. MLR was implemented in R (R Development Core Team, 2007), using the *lm* function, which is associated with the *step* function that can be used to perform backward stepwise regression and select the best regression variables.

The random forest regression model is a non-parametric technique that was developed by Breiman (2001) as an extension of the CART (Classification and Regression Trees) program to improve the prediction performance of the model, and it consists of a combination of many predictor trees (i.e., a forest), in which each tree is generated from a random vector that is sampled independently and that has the same distribution for all trees in the forest. The subdivisions within each tree are determined based on a subset of predictor variables chosen randomly from all existing predictors. In the case of RFM application for regression, the final result represents the mean of the results of all trees (Breiman, 2001; Cutler et al., 2009).

The RFMs were implemented using the package *randomForest* in R (R Development Core Team, 2007). To use an RFM, three parameters must be defined: the number of trees in the forest (n_{tree}), the minimum amount of data per terminal node (*nodesize*) and the number of variables used per tree (m_{try}) (Liaw and Wiener, 2002). The standard for n_{tree} defined in the package is 500. Grimm et al. (2008) reported stable results with a larger number for n_{tree}, but preliminary tests did not confirm this finding, showing that a higher n_{tree} did not improve the model's performance. Therefore, the standard number (500) was used. The standard value for regression studies was used for the *nodesize* value, which is five for each terminal node. In regression problems, the standard value for m_{try} is one third of the total number of predictor variables (Liaw and Wiener, 2002); thus, an m_{try} value of three was used for the eight predictor variables.

The RFM provides reliable estimates of the errors using so-called out-of-bag (OOB) data, which is a random subset of the data that is not used by the algorithm to build the trees. From the OOB predictions of every tree in the forest, the mean square error ($\mathrm{MSE_{OOB}}$) is calculated, as described by Eq. (1) (Liaw and Wiener, 2002):

$$\mathrm{MSE_{OOB}} = n^{-1}\sum_{i=1}^{n}\left(\mathrm{z_i}-\mathrm{z}_i^{oob}\right)^2 \tag{1}$$

where $\mathrm{z_i}$ is the measured value of the variable and $\hat{\mathrm{z}}_i^{oob}$ is the average of all OOB predictions. However, as the MSE depends on the measurement scale, it cannot be used to compare the performance of different models. Therefore, the percentage of variance explained by the model ($\mathrm{Var_{ex}}$) was calculated using Eq. (2), as proposed by Liaw and Wiener (2002):

$$\mathrm{Var_{ex}} = 1-(\mathrm{MSE_{OOB}}/\mathrm{Var_z}) \tag{2}$$

where $\mathrm{Var_z}$ is the total variance of the variable.

2.4. Model validation

The performance of prediction models is ideally assessed using an independent set of validation data that was not used in the calibration process. Thus, the 399 profiles were divided into two independent sets, one for calibration (319 samples) and the other for validation (80 samples) of the tested models, randomly selected using the R statistical package (R Development Core Team, 2007). The performance of each model was computed from the validation samples by calculating the correlation between the observed and estimated values based on the coefficient of determination (R^2) and the RMSE (root mean squared error), as described by Eq. (3).

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{n}}\sum_{i=1}^{\mathrm{n}}\mathrm{d}_i^2} \tag{3}$$

where d is the difference between the observed and predicted values, and n is the total number of observations. As proposed by Holmes et al. (2000), the RMSE is normally used to estimate the error or uncertainty associated with estimates where data were not measured. The RMSE indicates the discrepancy between the observed and calculated values. The lower the RMSE, the more accurate the prediction.

3. Results and discussion

3.1. Descriptive statistics

The descriptive statistics of the physical properties of the calibration and validation samples of the soil surface layer (0–20 cm) are shown in Table 1, and the descriptive statistics of the environmental covariates are presented in Table 2.

The results indicate high similarity between the calibration and validation samples, and no significant differences in the analyzed

Table 1
Descriptive statistics of the samples used in the prediction of the soil properties.

Properties	Calibration					Validation				
	Max	Min	Mean	SD	CV	Max	Min	Mean	SD	CV
	$\mathrm{g{\cdot}kg^{-1}}$				%	$\mathrm{g{\cdot}kg^{-1}}$				%
Sand	790	116	317	138	44	801	140	333	147	44
Silt	344	54	221	51	23	335	41	217	58	27
Clay	636	133	462	107	23	647	158	450	111	25

SD — standard deviation; CV — coefficient of variation.

properties were detected at the 95% probability level based on the analysis of variance (ANOVA) results. This similarity indicates that the validation samples adequately represent the calibration samples.

The coefficient of variation (CV) was equal to or greater than 23% in all cases, which indicates the heterogeneity of the sample sets. The sand fraction had the highest CV, for both the calibration and validation samples, indicating that sand was the most heterogeneous fraction in this area.

Similar to the physical properties, the covariate calibration and validation samples did not differ significantly at the 95% probability level based on the analysis of variance (ANOVA) results. Except for the NDVI, all covariates had a CV $<$18%, indicating high data homogeneity.

Demattê et al. (2007) attributed the good prediction results for sand, silt and clay in their study to the high variance and standard deviation of the soil properties and the spectral data they used, which allowed a better calibration of MLR models compared with another area with lower data variance. In our study, the variance of the soil properties was high (CV $>$ 23%), while the spectral data had lower variance, except for the NDVI.

3.2. Importance of covariates

The Pearson's correlation coefficients (Table 3) showed that in general, the environmental covariates used were significantly correlated with the soil properties ($p < 0.05$).

The sand content was significantly correlated with most of the covariates, except with b1 and b5/b7 (Table 3), which is does not agree with the results presented by Carvalho Junior et al. (2014), who found no correlation between the sand content and the Landsat 5 TM image data (NDVI and the relationships between the bands b3/b2, b3/b7 and b5/b7). Only the covariate b3/b7 was strongly correlated ($r = -0.58$) with the sand content (Cohen, 1988), whereas the NDVI and GSI were moderately correlated ($r = 0.38$ and 0.43, respectively) and the other covariates weakly correlated, with r values below 0.26 (positive or negative).

The highest number of non-significant correlations (i.e., b1, b4 and b5/b7) was observed for silt; the other covariates were weakly (b2, b3, b5, b7, NDVI, and b3/b2) or moderately correlated (b3/b7 and GSI) with this soil property (Table 3). Carvalho Junior et al. (2014) found no significant correlation between silt and the environmental covariates derived from a Landsat 5 TM image. In contrast, Souza Junior et al. (2011) reported high correlation coefficients between silt and bands derived from ASTER satellite-derived data (bands 1–8). The lower correlations of silt with the covariates are consistent with previous studies conducted by Islam et al. (2003) and by Wettlerlind and Stenberg (2010). The latter suggested that silt is more difficult to distinguish in the visible and near-infrared regions.

Inverse relationships were observed between clay and the environmental covariates, but the magnitudes were the same as those observed for sand (Table 3); similar results were reported by Souza Junior et al. (2011). Only the covariates b1 and b5/b7 were not significantly correlated with clay, and the most relevant covariates were b3/b7 ($r = 0.56$), GSI ($r = 0.40$), NDVI ($r = -0.39$) and b3/b2 ($r = 0.25$). Carvalho Junior et al. (2014) found significant correlations between clay and the NDVI index and between clay and the bands b3/b2 and b5/b7, whereas no correlation was observed between clay and the b3/b7 band. On the other hand, Ahmed and Iqbal (2014) only detected significant correlations between clay and bands 4 and 6 using Landsat 5 TM.

Table 2
Descriptive statistics of the covariates used in the prediction of the soil properties.

Environmental covariates	Calibration			Validation		
	Mean	Standard deviation	CV (%)	Mean	Standard deviation	CV (%)
Band 1	107	13.31	12	110	13.8	13
Band 2	55	8.00	15	53	7.73	15
Band 3	65	11.25	17	64	10.91	17
Band 4	66	6.07	9	66	6.24	9
Band 5	137	18.77	14	135	17.05	13
Band 7	71	12.92	18	70	11.82	17
NDVI	0.01	0.07	700	0.02	0.08	400
Band 3/band 2	1.20	0.05	4	1.20	0.05	4
Band 3/band 7	0.93	0.09	10	0.92	0.09	10
Band 5/band 7	1.94	0.11	6	1.95	0.11	6
GSI	−0.19	0.03	16	−0.20	0.03	15

CV — coefficient of variation; GSI — grain size index.

Table 3
Pearson's correlation between soil properties and environmental covariates.

Environmental covariates	Properties					
	Sand		Silt		Clay	
	p-Value	r	p-Value	r	p-Value	r
Band 1	0.66^{ns}	0.02	0.43^{ns}	−0.04	0.64^{ns}	−0.02
Band 2	0.00^{*}	−0.16	0.04^{*}	0.10	0.00^{*}	0.15
Band 3	0.00^{*}	−0.19	0.02^{*}	0.12	0.00^{*}	0.17
Band 4	0.00^{*}	0.19	0.37^{ns}	−0.05	0.00^{*}	−0.23
Band 5	0.00^{*}	0.20	0.00^{*}	−0.15	0.00^{*}	−0.19
Band 7	0.00^{*}	0.15	0.04^{*}	−0.10	0.00^{*}	−0.15
NDVI	0.00^{*}	0.38	0.00^{*}	−0.19	0.00^{*}	−0.39
Band 3/band 2	0.00^{*}	−0.26	0.00^{*}	0.15	0.00^{*}	0.25
Band 3/band 7	0.00^{*}	−0.58	0.00^{*}	0.37	0.00^{*}	0.56
Band 5/band 7	0.05^{ns}	0.10	0.08^{ns}	−0.09	0.13^{ns}	−0.08
GSI	0.00^{*}	−0.43	0.00^{*}	0.30	0.00^{*}	0.40

r — Pearson's correlation coefficient; ^{ns} — non-significant; NDVI — Normalized Difference Vegetation Index; GSI — grain size index.

* Significant at 5% probability.

According to Sabins (1997), the index *Clay minerals* (b5/b7) can be useful to identify areas with different types of clay minerals. The soils of the study area consisted predominantly of 2:1 clay minerals; therefore, the small variability in terms of clay minerals may explain the absence of a significant correlation between the properties sand, silt and clay and covariate b5/b7. Similarly, covariate b1 was not significantly correlated with any of the properties evaluated in the study by Ahmed and Iqbal (2014).

Most of the NDVI index values were below 0.1, indicating little vegetation cover in the study area (Liao et al., 2013). Relatedly, Bartholomeus et al. (2007) noted that an accurate estimation of soil properties is hampered by vegetation cover of more than 20%. The high variance of this index is noteworthy and may help explain its importance in capturing the variations of the studied properties.

Demattê et al. (2009) and Liao et al. (2013) highlighted the significant correlation of band 7 with soil texture in their studies and related this to the greater sensitivity of this band to available soil water levels. In contrast, in this study, band 7 was weakly correlated with sand, clay and silt (Table 3). This result may be related to the drier conditions of the area (400 mm of rain per year) compared with the cited studies.

Demattê et al. (2009) explained that the presence or absence of a particular band is directly related to the specific characteristics of the soils of a region, which probably explains the differences between the results of the above-cited studies. On the other hand, although the observed correlations are not strong for most covariates, these can be used to improve the prediction performance of the different models.

3.3. Multiple linear regression

The MLR analysis showed a moderate correlation between the sand and clay and the covariates, with R^2 values close to 0.50 (Table 4); the RMSE results indicated a better performance of the model for clay. The silt was weakly correlated with the covariates, with an R^2 value of 0.20 and an RMSE value of 44.5 $g \cdot kg^{-1}$, which is considered an unsatisfactory result according to Nanni and Demattê (2006).

Among the covariates, the most relevant for predicting sand and clay using the stepwise model were b2, b3, b4, b5, NDVI, and GSI. The

Table 4
Models of multiple linear regression and the respective coefficients of determination.

Property	Regression equation	R^2	RMSE ($g \cdot kg^{-1}$)
Sand	$y = -574.81 + (-27.40 * \text{band } 2) + (43.78 * \text{band } 3) + (-24.37 * \text{band } 4) + (5.18 * \text{band } 5) + (3634.96 * \text{NDVI}) + (-1975.74 * \text{GSI})$	0.51	93.28
Silt	$y = 430.40 + (7.15 * \text{band } 2) + (-4.44 * \text{band } 3) + (-1.29 * \text{band } 5) + (673.66 * \text{GSI})$	0.20	44.50
Clay	$y = 1146.23 + (20.38 * \text{band } 2) + (-35.01 * \text{band } 3) + (20.14 * \text{band } 4) + (-3.88 * \text{band } 5) + (-3060.04 * \text{NDVI}) + (1408.65 * \text{GSI})$	0.49	75.17

stepwise model eliminated b3/b7 for both sand and clay, although Pearson's correlation analysis indicated that this covariate had the highest values ($r = -0.58$ and 0.56, respectively, for sand and clay) (Table 3). This result may be related to the effect of intercorrelation with the other covariates. For silt, the covariates b2, b3, b5 and GSI were selected, whereas the NDVI was not selected despite a significant Pearson correlation ($r = -0.19$).

The results obtained in this study for sand and clay, 0.52 and 0.67, respectively, were lower than those of Nanni and Demattê (2006), who used six bands of Landsat 5 TM and MLR analysis. In a predominantly sandy area in Paraguaçu Paulista (SP), Demattê et al. (2007, 2009) also found higher values. On the other hand, in these same studies, results for sand, silt and clay were considerably lower (0.08, 0.12 and 0.18, respectively) for an area with predominantly iron-rich, clayey soils in Rio Brilhante (MS). In this respect, the greater homogeneity of the spectral data may have contributed to the lower results in the present study than those reported by Demattê et al. (2007) for Paraguaçu Paulista.

Liao et al. (2013) studied soils rich in high-activity clay and found lower values than in the current study for sand (0.32), silt (0.21) and clay (0.36), which they attributed to the negative influence of atmospheric, topographic and solar effects. In addition, they noted that the spatial variability of the surface texture with a 30-m pixel resolution also affected the accuracy of the regression models. Ahmed and Iqbal (2014) used the same approach as earlier studies and found R^2 values of 0.51 and 50.21 $g \cdot kg^{-1}$ and RMSE values of 0.72 and 42.15 $g \cdot kg^{-1}$ for clay and silt, respectively, using only bands 4 and 6.

The differences between the cited studies may be attributed to the satellite data being influenced by factors such as geometric and atmospheric variations, surface roughness, water content, light angle and intensity, cover and type of vegetation, and the characteristics of each sensor (Moran et al., 1997; Demattê et al., 2007).

3.4. Random forest model

The importance of the environmental covariates in each tested RFM is shown in Fig. 2, in which the percentage of variance explained (Var_{ex}), derived from the out-of-bag data (MSE_{OOB}), can be observed. This variance was considered moderately satisfactory for sand (47.65%) and clay (48.94%) and unsatisfactory for silt (8.61%). The same trend was observed when using stepwise MLR. The RMSE results for sand (99.65 $g \cdot kg^{-1}$) and silt (48.93 $g \cdot kg^{-1}$) were slightly lower than those obtained using stepwise MLR, while the two models produced similar results for clay (76.44 $g \cdot kg^{-1}$).

Thus far, few studies have used RFMs for the prediction of soil texture, and none have used only remote sensing data as the main covariates (Table 5). Ließ et al. (2012) used terrain properties derived from a digital elevation model, combined with two determination methods of particle-size distribution (pipette and laser), and found lower values for Var_{ex} than in this study for sand (30%) and clay (43%) and a higher value for silt (26%) in the soil surface layer. The poor performance of the RFM in this study was attributed to the small size of the dataset.

In a study in Nigeria, Akpa et al. (2014) reported percentages of Var_{ex} for RFMs, i.e., 48–49% for sand, 26–27% for silt and 53–56% for clay in the top soil layer (0–15 cm). Their results are quite similar to those obtained in this study for sand and superior to those for silt and clay. Therefore, the RMSE results for sand (19.26–19.67%), silt (11.72–12.22%) and clay (13.11–13.59%) (Table 5) were lower than those obtained in the present study.

Vaysse and Lagacherie (2015) reported lower Var_{ex} values for sand (33 to 35%) and clay (31 to 35%) and higher values for silt (23 to 29%). The low performance of the RFMs was attributed to the small-scale variation of the source material and to the relative erosion/deposition along

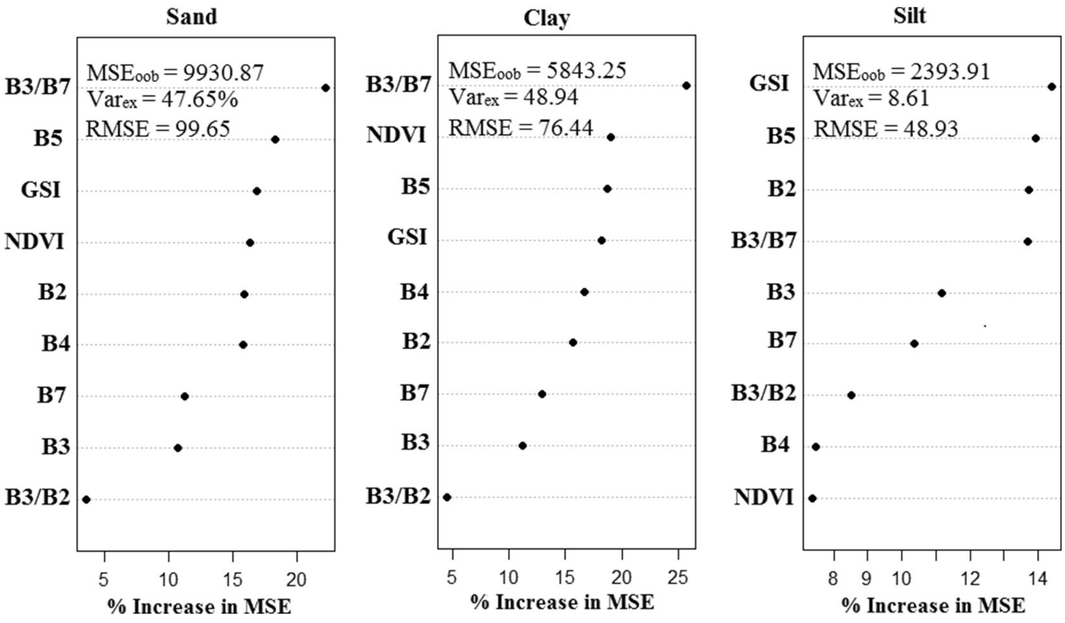

Fig. 2. Importance of the environmental covariates derived from the RFMs for sand, silt and clay. B2 — band 2; B3 — band 3; B4 — band 4; B5 — band 5; B7 — band 7; NDVI — Normalized Difference Vegetation Index; GSI — grain size index; MSE — mean squared error.

Table 5
Results from previous studies for the prediction of sand, silt and clay using random forest models.

Author	Explained variance (%)			RMSE		
	Sand	Silt	Clay	Sand	Silt	Clay
Ließ et al. (2012)	30	26	43	–	–	–
Akpa et al. (2014)	48–49	26–27	53–56	19.26–19.67	11.72–12.22	13.11–13.59
Vaysse and Lagacherie (2015)	33–35	23–29	31–35	139.80–140.23	97.56–98.19	99.70–99.97
This study	44	8	46	99.65	48.93	76.44

the slope, which could not be captured by the spatial resolution of the covariates used (100 m). Moreover, these authors claimed that the quality of the soil dataset must be improved to improve prediction performances.

The covariates used as predictors in the RFM (Fig. 2) were the same as those used for the MLR models. One of the main advantages of RFMs over MLR models is that the former provides an estimate of the relative importance of the covariates in the model, unlike MLR, in which only highly correlated predictive covariates are maintained in the model through stepwise selection (Cutler et al., 2009). The RFM avoids the elimination of predictive covariates that may be relevant for soil, even if there are correlations between them (Akpa et al., 2014).

In this study, a threshold of importance was defined, below which the covariates were considered unimportant. These values were set at 15% for sand, 12% for silt and 20% for clay. The results showed different combinations of covariates based on the analyzed properties. Thus, the most relevant covariates for sand were b3/b7 > b5 > GSI > NDVI > b2 > b4. For silt, the most important covariates were GSI > b5 > b2 > b3/b7, and for clay, b3/b7 > NDVI > b5 > GSI > b4 > b2. This confirmed the statement made by Akpa et al. (2014), i.e., that the relative importance of the covariates estimated by the RFMs is affected by the considered property as well as by other variables.

The major importance of covariate b3/b7 with respect to the prediction of all properties was consistent with the Pearson correlation results,

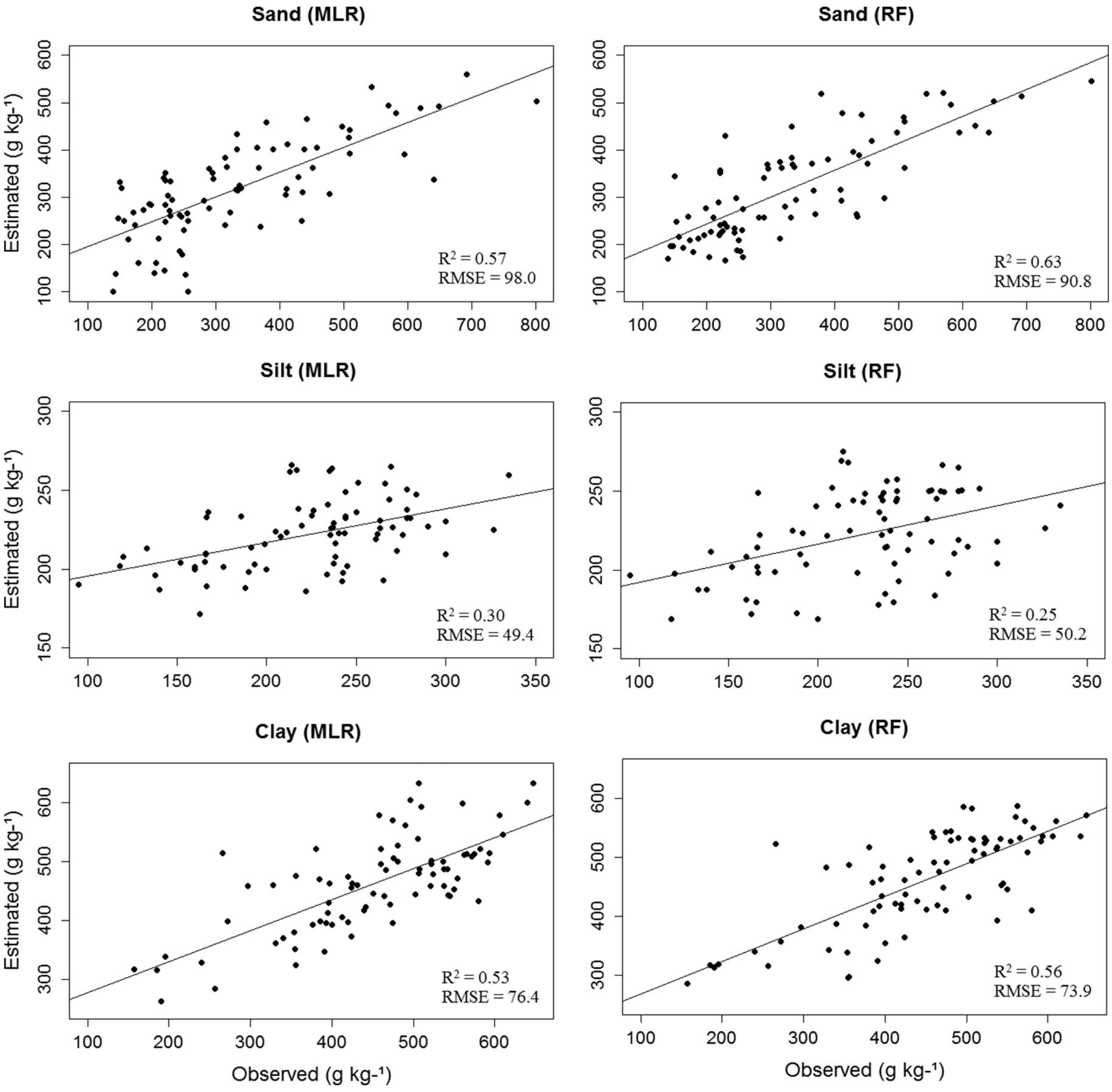

Fig. 3. Results of the coefficient of determination (R^2) and RMSE of the prediction models using the validation samples.

which also showed the highest correlation coefficients for this covariate, i.e., −0.58 for sand, 0.37 for silt and 0.56 for clay (Table 3). The NDVI index was the second most important covariate for clay and the fourth for sand, which is also in line with the Pearson correlation coefficients (0.38 and −0.39 for sand and clay, respectively). In contrast, this covariate was rather unimportant for silt. However, as the prediction performance for this property was very poor, the interpretation of the significance of the covariates is irrelevant (Grimm et al., 2008).

3.5. Validation of prediction models

The results of the predictive models (MLR and RFM) were validated using an independent dataset (Fig. 3). The RFM performed slightly better than the stepwise MLR in predicting sand (R^2 values of 0.63 and 0.57, and RMSE values of 90.77 and 98.00 $g \cdot kg^{-1}$) and clay (R^2 values of 0.57 and 0.52, and RMSE values of 73.94 and 76.39 $g \cdot kg^{-1}$ for the RFM and MLR, respectively). For silt, however, the stepwise MLR model performed better, with an R^2 value of 0.30 and an RMSE value of 49.42 $g \cdot kg^{-1}$ versus 0.25 (R^2) and 50.18 $g \cdot kg^{-1}$ (RMSE) for the RFM.

In general, the models had a satisfactory predictive capacity for sand and clay, with a slight superiority of the RFM, but both models were unsatisfactory for silt. The superior results obtained using the RFM compared with MLR for the prediction of the surface texture of soils are detailed in the study conducted by Hitziger and Ließ (2014) and that conducted by Guo et al. (2015) for the prediction of soil organic matter. Additionally, the RF-modeled maps of the spatial distribution of the properties were more realistic than the maps produced by the MLR models (Fig. 4), as was also reported by Guo et al. (2015).

The performance of MLR, with slightly better predictions for silt than those obtained using the RFM, was contrary to expectations because the RFM is a more robust approach than MLR. This result may be related to the specific conditions of the study area, in combination with the considerable difficulty in distinguishing silt in the visible and infrared spectral regions (Wettlerlind and Stenberg, 2010).

Demattê et al. (2007) emphasized that the quantification of soil properties from satellite sensor data is by no means a simple task because of the complexity of soils. Relatedly, the results observed in this study for sand and clay using both methods (RFM and MLR) can be considered satisfactory, similar to the conclusions drawn by Nanni and Demattê (2006) and Demattê et al. (2007). These results can be explained by the probable physical interference of these constituents with the energy incident on and reflected from the soil.

3.6. Spatial prediction

Stepwise MLR and RFM were used in the spatial modeling of the studied properties (Fig. 4). In the MLR model, the sand content varied the most among the properties (−39–855 $g \cdot kg^{-1}$), whereas the variation was smaller (−150–600 $g \cdot kg^{-1}$) in the RFM. The silt content had a wider variation (48–403 $g \cdot kg^{-1}$) when the MLR was used than when the RFM was used (−123–301 $g \cdot kg^{-1}$), while the clay content ranged from 32 to 987 $g \cdot kg^{-1}$ based on the MLR and from 221 to 603 $g \cdot kg^{-1}$ based on the RFM.

The maps generated by the two models showed a higher clay concentration in the North Central part of the area (Fig. 4), which is predominated by Vertisols. The predominance of limestone in the Caatinga formation, which covers 94.5% of the area, explains the

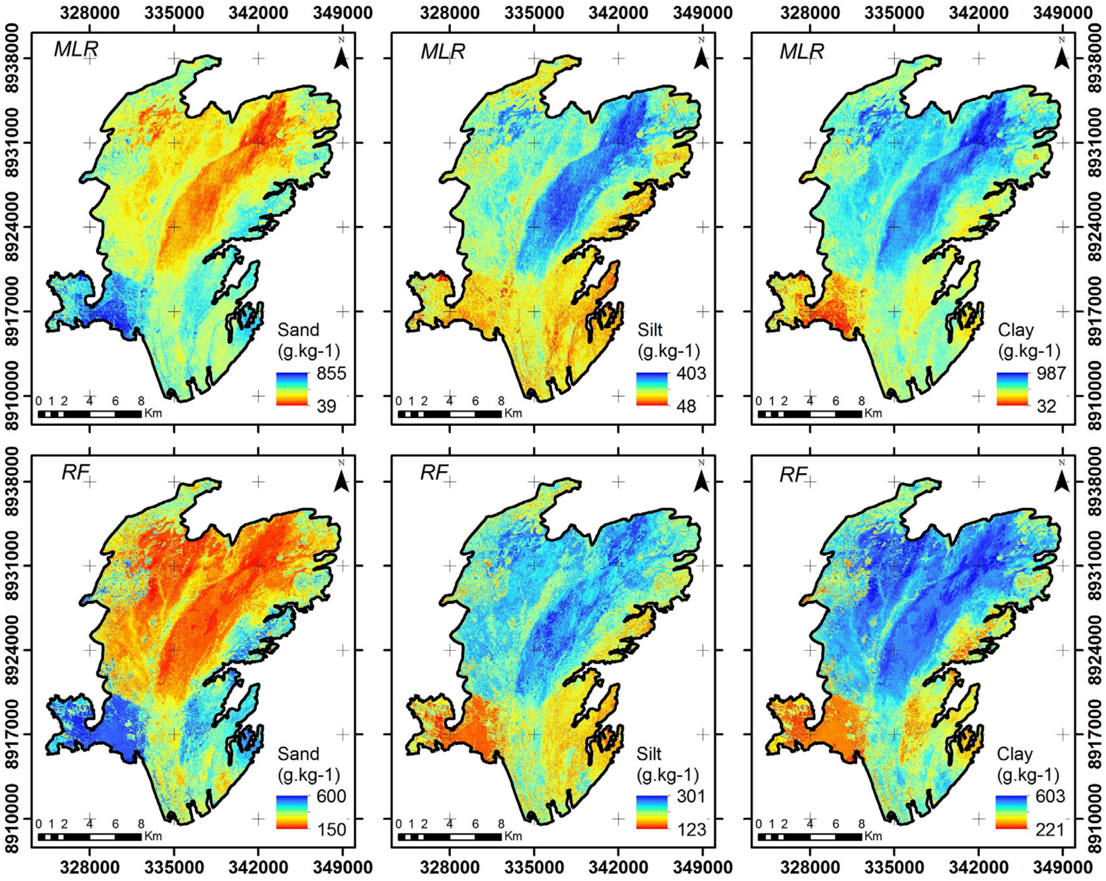

Fig. 4. Spatial distributions of the physical properties estimated using the studied methods.

prevalence of the clay fraction in these soils. The Vertisols in the study area have a weak or moderate A horizon, with an average thickness of 13.8 cm and a medium to very clayey surface texture, with high base saturation.

In the southwestern and southeastern parts of the area, the sand fraction is predominant (Fig. 4). These areas are associated with Cambisols and Planosols. The Cambisols have a weak or moderate A horizon with an average thickness of 14.9 cm. For the most part, they contain high-activity clay, are eutrophic and have a medium and more rarely clayey texture. The Planosols are characterized by only a weak A horizon with an average thickness of 16.6 cm and a medium surface texture and very high base saturation. Silt has a distribution similar to clay, being predominant in the area of the Vertisols.

The map generated by the RFM shows the predicted fitted values of the distribution of this fraction in the range of the observed data, while in the MLR-modeled map, the predicted values are extrapolated.

4. Conclusions

- Remote sensing data combined with the random forest models to estimate grain size, particularly the sand and clay contents of soils under semi-arid climate conditions, produced satisfactory results.
- According to Pearson's correlation analysis, the highest correlations with the soil variables were obtained with the covariates b3/b7, GSI, NDVI and b3/b2, in that order. The following covariates were selected in the MLR model for the sand and clay contents: b2, b3, b4, b5, NDVI and GSI. For the silt prediction, the covariates b2, b3, b5 and GSI were selected in the stepwise model.
- The assessment of the importance of the covariates for the random forest model (RFM) showed that the most significant covariates were b3/b7 > b5 > GSI > NDVI > b2 > b4 for the prediction of the sand, GSI > b5 > b2 > b3/b7 for the prediction of the silt and b3/b7 > NDVI > b5 > GSI > b4 > b2 for the prediction of the clay contents.
- Based on the validation samples, the coefficients of determination (R^2) and the RMSE values were more favorable for the RFM than for MLR for the prediction of sand (R^2 values of 0.63 and 0.57 and RMSE values of 90.77 and 98.00, respectively) and clay (R^2 values of 0.56 and 0.53 and RMSE values of 73.94 and 76.39, respectively). For silt, the prediction performance was better for the MLR model, with an R^2 value of 0.30 and an RMSE value of 49.42 versus 0.25 (R^2) and 50.18 (RMSE) for the RFM. These results are similar to those of other studies.

References

Ahmed, Z., Iqbal, J., 2014. Evaluation of Landsat TM5 multispectral data for automated mapping of surface soil texture and organic matter in GIS. Eur. J. Remote Sens. 47, 557–573.

Akpa, S.I.C., Odeh, I.O.A., Bishop, T.F.A., Hartemink, A.E., 2014. Digital mapping of soil particle-size fractions for Nigeria. Soil Sci. Soc. Am. J. 78, 1953–1966.

Bartholomeus, H., Epema, G., Schaepman, M.E., 2007. Determining iron content in Mediterranean soils in partly vegetated areas, using spectral reflectance and imaging spectroscopy. Int. J. Appl. Earth Obs. Geoinf. 9, 194–203.

Ben-Dor, E., Patkin, K., Banin, A., Karnieli, A., 2002. Mapping of several soil properties using DAIS-7915 hyperspectral scanner data — a case study over clayey soil in Israel. Int. J. Remote Sens. 23, 1043–1062.

Ben-Dor, E., Taylor, R.G., Hill, J., Demattê, J.A.M., Whiting, M.L., Chabrillat, S., Sommer, S., Donald, L.S., 2008. Imaging spectrometry for soil applications. Adv. Agron. 97, 321–392.

Boettinger, J.L., Ramsey, R.D., Bodily, J.M., Cole, N.J., Kienast-Brown, S., Nield, S.J., Saunders, A.M., Stum, A.K., 2008. Landsat spectral data for digital soil mapping. In: Hartemink, A.E., Mcbratney, A.B., Mendonça-Santos, M.L. (Eds.), Digital Soil Mapping with Limited Data. Springer-Verlag, New York, pp. 192–202.

Breiman, L. Technical report for Version 3. 2014 2001. Available in: http//oz.berkeley.edu/users/breiman/randomforest2001.pdf (Accessed at: 12/28/14).

Breunig, F.M., Galvão, L.S., Formaggio, A.R., 2008. Detection of sandy soil surfaces using ASTER-derived reflectance, emissivity and elevation data: potential for the identification of land degradation. Int. J. Remote Sens. 29, 1833–1840.

Carvalho Junior, W., Lagacherie, P., Chagas, C.S., Calderano Filho, B., Bhering, S.B., 2014. A regional-scale assessment of digital mapping of soil attributes in a tropical hillslope environment. Geoderma 232, 479–486.

Ciampalini, R., Lagacherie, P., Hamrouni, H., 2012. Documenting GlobalSoilMap.net grid cells from legacy measured soil profile and global available covariates in Northern Tunisia. In: Minasny, B., Malone, B.P., McBratney, A.B. (Eds.), Digital Soil Assessments and Beyond. CRC Press/Balkema, London, pp. 439–444.

Ciampalini, R., Martin, M.P., Saby, N.P., de Forges, A.C.R., Arrouays, D., Nehlig, P., Martelet, G., 2014a. Soil texture GlobalSoilMap products for the French region "Centre". In: Arrouays, D., McKenzie, N., Hempel, J., de Forges, A.C.R., McBratney, A. (Eds.), GlobalSoilMap: Basis of the Global Spatial Soil Information System. CRC Press, pp. 121–126.

Ciampalini, R., Martin, M., Saby, N., de Forges, A.C.R., Nehlig, P., Martelet, G., Arrouays, D., 2014b. Modelling soil particle-size distribution in the region "Centre" (France). In: Arrouays, D., McKenzie, N., Hempel, J., de Forges, A.C.R., McBratney, A. (Eds.), GlobalSoilMap: Basis of the Global Spatial Soil Information System. CRC Press, pp. 325–331.

Cohen, J., 1988. Statistical Power Analysis for the Behavioral Sciences. Academic press, New York.

Cutler, D.R., Edwards Jr., T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., Lawler, J.J., 2009. Random forests for classification in ecology. Ecology 88, 2783–2792.

Demattê, J.A.M., Galdos, M.V., Guimarães, R.V., Genú, A.M., Nanni, M.R., Zullo Jr., J., 2007. Quantification of tropical soil attributes from ETM+/LANDSAT-7 data. Int. J. Remote Sens. 28, 3813–3829.

Demattê, J.A.M., Fiorio, P.R., Ben-Dor, E., 2009. Estimation of soil properties by orbital and laboratory reflectance means and its relation with soil classification. Open Remote Sens. J. 2, 12–23.

Eldeiry, A.A., Garcia, L.A., 2010. Comparison of ordinary kriging, regression kriging, and cokriging techniques to estimate soil salinity using LANDSAT images. J. Irrig. Drain. Eng. 136, 355–364.

Embrapa, 1979. Serviço Nacional de Levantamento e Conservação do Solo. Manual e métodos de análise do solo (Rio de Janeiro, 1v., in Portuguese).

Embrapa, 2013. Centro Nacional de Pesquisa de Solos. Sistema Brasileiro de Classificação de Solos, 3. ed. Embrapa, ver. ampl. - Brasília, DF (353 pp., in Portuguese).

Gomez, C., Viscarra Rossel, R.A., McBratney, A.B., 2008. Soil organic prediction by hyperspectral remote sensing and field VIS–NIR spectroscopy: an Australian case study. Geoderma 146, 403–411.

Grimm, R., Behrens, T., Märker, M., Elsenbeer, H., 2008. Soil organic carbon concentrations and stocks on Barro Colorado Island — digital soil mapping using Random Forests analysis. Geoderma 146, 102–113.

Guo, P.T., Li, M.F., Luo, W., Tang, Q.F., Liu, Z.W., Lin, Z.M., 2015. Digital mapping of soil organic matter for rubber plantation at regional scale: an application of random forest plus residuals kriging approach. Geoderma 237, 49–59.

Hitziger, M., Ließ, M., 2014. Comparison of three supervised learning methods for digital soil mapping: application to a complex terrain in the Ecuadorian Andes. Appl. Environ. Soil Sci. 1–12.

Holmes, K.W., Chadwick, O.A., Kyriakidis, P.C., 2000. Error in a USGS 30-meter digital elevation model and its impact on terrain modeling. J. Hydrol. 233, 154–173.

Islam, K., Singh, B., McBratney, A., 2003. Simultaneous estimation of several soil properties by ultra-violet, visible, and near infrared reflectance spectroscopy. Aust. J. Soil Res. 41, 1101–1114.

Lark, R.M., Bishop, T.F.A., 2007. Cokriging particle size fractions of the soil. Eur. J. Soil Sci. 58, 763–774.

Lark, R.M., Dove, D., Green, S.L., Richardson, A.E., Stewart, H., Stevenson, A., 2012. Spatial prediction of seabed sediment texture classes by cokriging from a legacy database of point observations. Sediment. Geol. 281, 35–49.

Liao, K., Xu, S., Wu, J., Zhu, Q., 2013. Spatial estimation of surface soil texture using remote sensing data. Soil Sci. Plant Nutr. 59, 488–500.

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R News 2 (3), 18–22.

Ließ, M., Glaser, B., Huwe, B., 2012. Uncertainty in the spatial prediction of soil texture: comparison of regression tree and Random Forest models. Geoderma 170, 70–79.

Malone, B.P., McBratney, A.B., Minasny, B., Laslett, G.M., 2009. Mapping continuous depth functions of soil carbon storage and available water capacity. Geoderma 154, 138–152.

Moran, M.S., Ionoue, Y., Barnes, E.J., 1997. Opportunities and limitations for image based remote sensing in precision crop management. Remote Sens. Environ. 61, 319–346.

Mulder, V.L., De Bruin, S., Schaepman, M.E., Mayr, T.R., 2011. The use of remote sensing in soil and terrain mapping — a review. Geoderma 162, 1–19.

Nanni, M.R., Demattê, J.A.M., 2006. Spectral reflectance methodology in comparison to traditional soil analysis. Soil Sci. Soc. Am. J. 70, 393–407.

R Development Core Team, 2007. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria 3-900051-07-0 (Available in: http://www.R-project.org/isbn 3-900051-07-0 (Accessed at: 05/08/2015)).

Rivero, R.G., Grunwald, S., Bruland, G.L., 2007. Incorporation of spectral data into multivariate geostatistical models to map soil phosphorus variability in a Florida wetland. Geoderma 140, 428–443.

Sabins, F.F., 1997. Remote Sensing: Principles and Interpretation. third ed. W. H. Freeman and Company, New York (432 pp.).

Souza Junior, J.G., Demattê, J.A., Araújo, S.R., 2011. Modelos espectrais terrestres e orbitais na determinação de teores de atributos dos solos: potencial e custos. Bragantia 70, 610–621 (in Portuguese).

Souza, D. J., Kosin, M., Melo, R. C., Santos, R. A., Teixeira, L. R., Sampaio, A. R., Guimarães, J. T., Bento, R. V., Borges, V. P., Martins, A. A. M., Arcanjo, J. B., Loureiro, H. S. C., Angelim, L. A. A., 2003. Mapa geológico do Estado da Bahia — escala 1:1000000. Salvador: CPRM, 2003. Versão 1.1. Programas Carta Geológica do Brasil a milionésimo e Levantamentos geológicos básicos do Brasil (PLGB). (in Portuguese).

Stevens, A., Van Wesemael, B., Bartholomeus, H., Rossillon, D., Tychon, B., Ben-Dor, E., 2008. Laboratory, field and airborne spectroscopy for monitoring organic carbon content in agricultural soils. Geoderma 144, 395–404.

Stevens, A., Udelhoven, T., Denis, A., Tychon, B., Lioy, R., Hoffmann, L., Van Wesemael, B., 2010. Measuring soil organic carbon in croplands at regional scale using airborne imaging spectroscopy. Geoderma 158, 32–45.

Vaysse, K., Lagacherie, P., 2015. Evaluating Digital Soil Mapping approaches for mapping GlobalSoilMap soil properties from legacy data in Languedoc-Roussillon (France). Geoderma Reg. 4, 20–30.

Viscarra Rossel, R.A., Behrens, T., 2010. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma 158, 46–54.

Wettlerlind, J., Stenberg, B., 2010. Near-infrared spectroscopy for within-field soil characterization: small local calibrations compared with national libraries spiked with local samples. Eur. J. Soil Sci. 61, 823–843.

Wiesmeier, M., Barthold, F., Blank, B., Kögel-Knabner, I., 2011. Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem. Plant Soil 340, 7–24.

Xiao, J., Shen, Y., Tateishi, R., Bayaer, W., 2006. Development of topsoil grain size index for monitoring desertification in arid land using remote sensing. Int. J. Remote Sens. 12, 2411–2422.