Revista Brasileira de Ciência do Solo

Division - Soil in Space and Time | Commission – Pedometric

Mapping Soil Cation Exchange Capacity in a Semiarid Region through Predictive Models and Covariates from Remote Sensing Data

César da Silva Chagas[(1)], Waldir de Carvalho Júnior[(1)], Helena Saraiva Koenow Pinheiro[(2)], Pedro Armentano Mudado Xavier[(3)*], Silvio Barge Bhering [(1)], Nilson Rendeiro Pereira [(1)] and Braz Calderano Filho[(1)]

[(1)] Empresa Brasileira de Pesquisa Agropecuária, Embrapa Solos, Rio de Janeiro, Rio de Janeiro, Brasil.
[(2)] Universidade Federal Rural do Rio de Janeiro, Departamento de Solos, Seropédica, Rio de Janeiro, Brasil.
[(3)] Universidade Federal Rural do Rio de Janeiro, Departamento de Solos, Programa de Pós-Graduação em Agronomia – Ciência do Solo, Seropédica, Rio de Janeiro, Brasil.

ABSTRACT: Planning sustainable use of land resources requires reliable information about spatial distribution of soil physical and chemical properties related to environmental processes and ecosystemic functions. In this context, cation exchange capacity (CEC) is a fundamental soil quality indicator; however, it takes money and time to obtain this data. Although many studies have been conducted to spatially quantify soil properties on various scales and in different environments, not much is known about interactions between soil properties and environmental covariates in the Brazilian semiarid region. The goal of this study was to evaluate the efficiency of random forest and cokriging models applied to predict CEC in the Brazilian semiarid region. The covariates used to predict CEC consist of images from Landsat 5 TM and a legacy soil map (scale 1:10,000). The sample set comprises 499 samples from the topsoil layer (0.00-0.20 m), where 375 samples were used in training processes and 124 as validation samples. The cokriging model (R^2 = 0.57 and RMSE = 7.22 $cmol_c\ kg^{-1}$) performed better in predicting CEC than the random forest model (R^2 = 0.47 and RMSE = 7.89 $cmol_c\ kg^{-1}$). The approach used showed potential for estimating CEC content in the Brazilian semiarid region by using covariates obtained from orbital remote sensing and the legacy soil map.

Keywords: data mining, geostatistics, Landsat 5, legacy data, soil survey.

*** Corresponding author:**
E-mail: pedroarmentano@hotmail.com

Received: June 4, 2017
Approved: December 4, 2017

How to cite: Chagas CS, Carvalho Júnior W, Pinheiro HSK, Xavier PAM, Bhering SB, Pereira NR, Calderano Filho B. Mapping soil cation exchange capacity in a semiarid region through predictive models and covariates from remote sensing data. Rev Bras Cienc Solo. 2018;42:e0170183.
https://doi.org/10.1590/18069657rbcs20170183

INTRODUCTION

Sustainable agriculture and environmental management implies understanding the variability of soil properties on an appropriate scale to support decision-making. Cation exchange capacity (CEC) is an important soil property and a fundamental indicator of soil quality and environmental decontamination potential (Tang et al., 2009). The CEC affects ionic adsorption and desorption in elements such as copper, zinc, and lead, as well as in organic pollutants such as atrazine, phenanthrene, diquat, and paraquat (Liao et al., 2011). Understanding CEC spatial distribution is important for supporting decisions regarding crops and soil management; however, obtaining a large soil dataset is expensive and time consuming.

The use of covariates in soil surveys, such as information related to soil-landscape relationships, makes it possible to achieve a good cost-benefit ratio and quality of results. Reflectance and emission data can be analyzed to extract information about the Earth and its resources, as the physical and chemical properties of different surfaces vary across the electromagnetic spectrum. In this context, remote sensing data can be used as environmental covariates in digital mapping of soil properties (Boettinger et al., 2008), especially in arid and semi-arid regions, due to the absence of interference from vegetation. Images from orbital remote sensing are an important source of environmental data used in digital soil mapping, due to the relation of soil images with soil forming factors, such as organisms (vegetation) and parent material (McBratney et al., 2003).

Different methods have been used to predict soil CEC, including geostatistics, e.g., cokriging models (Tang et al., 2009; Bilgili et al., 2011; Ciampalini et al., 2012), and data mining methods, e.g., random forest models (Lagacherie et al., 2013; Hengl et al., 2015). A study conducted by Liao et al. (2011) used kriging models and principal component analysis to predict soil properties; they observed that cokriging had better performance than ordinary kriging in predicting CEC spatial distribution.

A combination of hyperspectral spectroscopy data from the visible and near-infrared spectrum and cokriging models to predict soil properties (including CEC) was exemplified by Bilgili et al. (2011) in a precision agriculture area in northern Turkey. The authors indicated that this approach better predicted soil properties compared to partial least squares regression and ordinary kriging. Moreover, it allowed reduction in the sample set and, consequentially, in laboratory analysis expenses.

Cokriging and hyperspectral data from the visible and near-infrared spectrum was used by Ciampalini et al. (2012) to predict CEC (among other properties) of the topsoil layer in Cap Bon, Tunisia. The results of CEC prediction are considered relatively good performance due to the low density of the dataset, which was not sufficient to capture short distance variations related to the lithology of that area. Furthermore, the results explain great part of variability considering the intrinsic dynamic of the measured property (CEC).

Random forest (RF) is a data mining method that has shown some advantages over most modeling methods, as highlighted by Breiman (2001) and Liaw and Wiener (2002). The RF model was also used by Lagacherie et al. (2013) to predict soil CEC in the soil subsurface using legacy data as input, and as covariate hyperspectral data from the visible and near-infrared spectrum and terrain attributes from a digital elevation model (DEM) with a resolution of 30 m in the Cap Bon region (Tunisia). The results obtained in the present study were considered satisfactory for the 0.15-0.30 and 0.30-0.60 m soil layers; lower accuracy was found in the 0.60-1.00 and 0.30-1.00 m layers. The resulting maps improved the existing soil maps for that region and were consistent with the tacit pedological knowledge. In contrast, Vaysse and Lagacherie (2015) did not obtain satisfactory results applying RF models to predict CEC through the use of soil legacy data and covariates derived from SRTM, images from Landsat 7, and geology maps as inputs. The poor performance was attributed to the small-scale variation of the parent material,

and the erosion/deposition rates along the catena system were not well captured by the spatial resolution of the covariates used as input data (100 m).

A study performed by Hengl et al. (2015) compared RF and multiple linear regression (MLR) to predict several soil properties (including CEC) for the African continent using a large number of covariates [MODIS, SRTM, land cover map, and soil maps (SoilGrids 1 km)] and 28,000 soil samples as input. The RF models performed better than MLR for all soil properties and depths predicted.

Although mapping spatial variation of soil properties in flat areas remains a challenge since the terrain covariates related to current topography have low predictive power, remote sensing data can be useful to represent surface variability that may better correlate with soil properties predicted by using digital mapping techniques. Given the context of the issues presented above, the goal of this study was to compare the accuracy of cokriging and RF models in predicting CEC in the topsoil layer of a flat area in a Brazilian semiarid region under *Caatinga* (xeric shrubland) vegetation by using Landsat 5 TM data as a covariate.

MATERIALS AND METHODS

Study area

The study was conducted in an area of 34,437.82 ha located between 9° 53’ 0’’ and 9° 36’ 30’’ S and 40° 34’ 30’’ and 40° 23’ 30’’ W, in the municipality of Juazeiro in the state of Bahia, Brazil (Figure 1).

The Köppen climate classification of the area is BSwh’ (semi-arid with dry winter and rainy summer, and the coldest mean monthly temperature is above 18 °C). The average annual rainfall is 400 mm, with the rainy season extending from November to April (highest rainfall in March); and the average annual temperature is around 26 °C. Xerothermic indexes are between 200 and 150, comprising seven to eight dry months. The area has a unique vegetation formation, known as hyperxerophilic Caatinga, characterized by shrubs with a high degree of xerophytism. However, part of the original vegetation

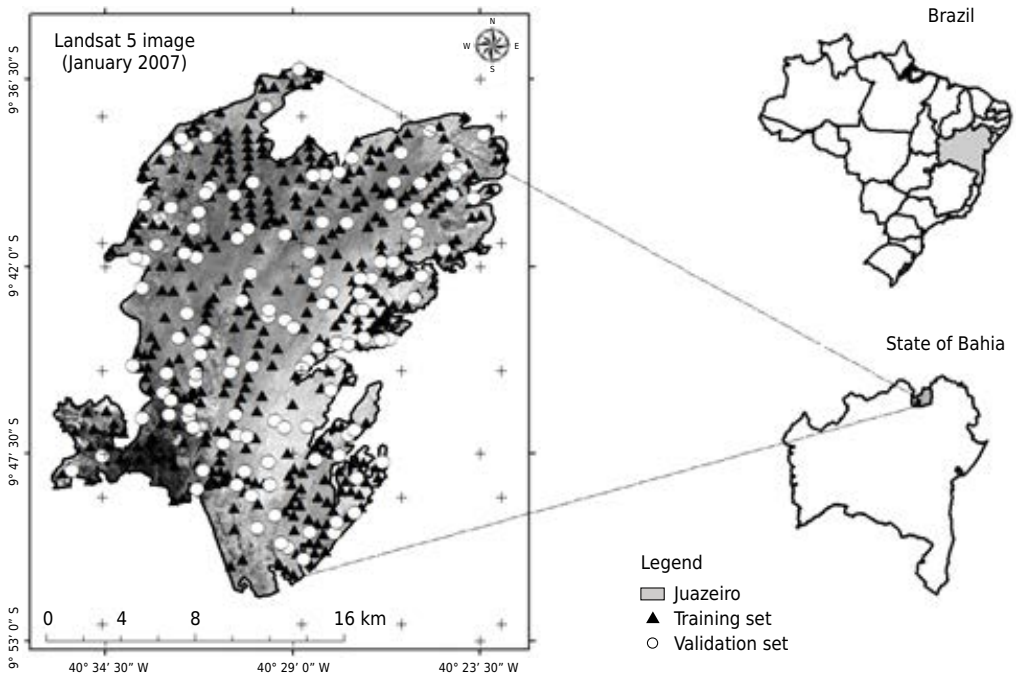

Figure 1. Study area and location of soil profiles, over Landsat image (band3).

has been removed, for various purposes, and currently shows signs of degradation, sometimes exposing the soil surface. The landscape is composed of flat surfaces, with the lithology mainly composed of limestone rocks from the Caatinga Formation (old Tertiary -Quaternary) and gneiss-granite rocks of the Caraiba-Paramirim Complex (Souza et al., 2003). The most representative soil types in this area are *Vertissolos*, *Cambissolos*, and *Planossolos* (Santos et al., 2013), which corresponding to Vertisols, Cambisols, and Planosols, respectively (IUSS, 2015).

Covariates and soil input data

The values of CEC were determined according to Oliveira (1979) and Embrapa (1979). The input dataset comprised the topsoil layer (0.00-0.20 m) for 499 soil profiles from a soil survey (legacy data), performed by the *Companhia de Desenvolvimento do Vale do São Francisco* (Codevasf) in 1989, which were subdivided for model training and validation. Sampling was performed according to the requirements of the National Soil Survey Service, regarding level of detail, as recommended by IBGE (2015) and Carvalho (1988). The following covariates were used: Landsat 5 TM data with 30 m spatial resolution (from January 2007, available at INPE website), composed of spectral bands 1 (0.450-0.515 µm), 2 (0.525 to 0.605 µm), 3 (0.630-0.690 µm), 4 (0.755-0.900 µm), 5 (1.550-1.750 µm), 7 (2.090-2.350 µm), and the NDVI (band 4 - band 3/band 4 + band 3), as well as the ratio between band 3 and band 2 (b3/b2), the ratio between band 3 and band 7 (b3/b7), and the ratio between band 5 and band 7 (b5/b7), according to Malone et al. (2009) and Carvalho Junior et al. (2014).

Additionally, a categorical covariate, represented by the detailed soil map (legacy data) on a 1:10,000 scale (Codevasf, 1989), was used on the first taxonomic level according to the Brazilian soil taxonomic system (Santos et al., 2013). Most of the soils are Vertisols. This categorical covariate (soil map) was used only in the RF model.

Predictive models

The modelling procedures to execute the random forest (RF) and cokriging (CK) prediction were performed on R software (R Development Core Team, 2007) through the randomForest (RF) and gstat (CK) packages.

Random forest is a non-parametric technique developed by Breiman (2001) as an extension of CART (Classification and Regression Tree) systems to improve the performance of predictors. The random forests are a combination of many predictive trees (forest), where each tree is derived from a random vector sampled independently and with the same distribution for all trees in the forest. The subdivisions within each tree are determined based on a subset of predictor variables randomly chosen from the total of predictors provided. The results of RF, when applied to predict continuous data by regression, consist of the average of the results from all the trees in the forest (Breiman 2001; Cutler et al., 2007).

To implement the RF models, three parameters are necessary: ntree - number of trees in the forest; nodesize - minimum amount of data in each terminal node; and mtry - number of covariates used in each tree (Liaw and Wiener, 2002). The ntree value was set to system default (500), although more stable results can be achieved with a larger number, according to Grimm et al. (2008). Preliminary tests showed that increasing the ntree did not improve model performance, supporting use of the default value. The nodesize value was set to five for each terminal node, as usually selected in regression studies. The mtry value chosen in this study was according to Liaw and Wiener (2002), who propose an amount corresponding to one third the total number of predictor variables for regression problems.

The RF model provides error estimation known as Out-Of-Bag (OOB), which is based on observed data not used by the algorithm to create the trees. Based on the OOB predictions from all the trees in the forest, the mean square error (MSE_{OOB}) is calculated, according to equation 1 (Liaw and Wiener, 2002).

$$MSE_{OOB} = n^{-1} \sum (zi - \hat{z}_i^{oob})^2 \quad \text{Eq. 1}$$

in which zi represents the average value of the variable, and $\hat{z}i^{oob}$ is the average of all OOB predictions. However, as the MSE_{OOB} is dependent on the measurement scale of the target variable (CEC), this index is not appropriate to compare the performance of different models. Addressing this issue, the percent of variance explained by the model (VAR_{ex}) is calculated according to equation 2 (Liaw and Wiener, 2002), where Var_z is the total variance of the variable.

$$Var_{ex} = 1 - (MSE_{OOB}/Var_z) \quad \text{Eq. 2}$$

Cokriging is a geostatistical procedure where a regionalized variable can be estimated based on correlation with one or more secondary variables (co-regionalized) with spatial co-dependence (McBratney and Webster, 1983). It can be understood as an extended kriging method, which has a vector of values rather than a single value for each spot sampled. Therefore, the CK model allows prediction of a variable (soil CEC, in this case), using a known dataset and based on correlation with other covariates.

Validation

Performance of the models was evaluated based on an independent validation set that was not used in the training procedure. To do so, the 499 soil samples were randomly divided into two independent datasets in the R software; one of these was used in the training process (375 soil samples) and another in the validation process (124 soil samples). The analysis of model performance was based on the correlation between the observed values (validation samples) and the estimated values, calculated by the coefficient of determination (R^2) and the root mean square error (RMSE), using equation 3.

$$RMSE = \sqrt{n-1} \sum d_i^2 \quad \text{Eq. 3}$$

in which "d" is the difference between the observed and estimated values, and "n" is the number of samples used in the validation process. The RMSE is commonly used to estimate the error or uncertainty in places where the error was not measured directly (Holmes et al., 2000). The higher the values of RMSE, the greater the differences between the datasets.

RESULTS AND DISCUSSION

Descriptive statistics and covariate selection

The descriptive statistics of the CEC from the topsoil layer (0.00-0.20 m) for the training and validation datasets, and from the environmental covariates are presented in table 1.

Analysis of variance (ANOVA) showed similarity between the training and validation samples, and no significant difference was detected for the CEC and the covariates at 95 % probability. The results of this analysis indicate that the validation samples adequately represent the training samples.

High values for the coefficient of variation (CV) for the soil CEC indicate heterogeneity in training and validation datasets, with values corresponding to 34 and 32 %, respectively. In contrast, all other covariates had CV values lower than 18 %, representing more homogeneity, except for NDVI.

Random forest model

The RF model also has the ability to estimate the relative importance of the covariates (Figure 2). In the training process, the RF model maintains all input covariates, preventing even those weakly correlated but with important pedological meaning from being

discarded from the model (Akpa et al., 2014). In contrast, predictive methods, such as stepwise linear regression, maintain only highly correlated covariates in the model (Cutler et al., 2007).

The importance ranking of the covariates for CEC prediction was as follows: Soil Map > b3/b7 > NDVI > b5/b7 > b1 > b7 > b3 > b4 > b2 > b5 > b3/b2. Removing these covariates from the model tended to increase the MSE_{OOB}, which ranged from 12.45 (b3/b2) to 25.29 % (Soil Map). The importance of soil legacy data for mapping soil properties (including CEC) was highlighted by Hengl et al. (2015), based on the application of random forest models in Africa.

The importance of those variables may be related to types of land use, which affected soil CEC values through the contribution of organic matter.

Table 1. Descriptive statistics of soil samples, training and validation datasets, and covariates used to predict soil CEC

Covariates	Training					Validation				
	Max.	Min.	Avg.	SD	CV	Max.	Min.	Avg.	SD	CV
	g kg^{-1}				%	g kg^{-1}				%
CEC	61.41	5.71	34.29	11.55	34	60.08	10.63	33.99	10.91	32
b1	159	80	107.56	13.82	13	141	81	109.10	14.19	13
b2	81	34	53.68	7.88	15	76	36	54.16	8.18	15
b3	100	34	64.43	11.04	17	96	37	64.91	11.79	18
b4	91	51	65.83	6.51	10	86	55	66.12	6.02	9
b5	191	86	135.60	17.91	13	181	94	136.54	18.31	13
b7	112	37	70.35	12.42	18	106	41	70.90	12.56	18
NDVI	0.27	-0.09	0.02	0.07	350	0.26	-0.08	0.02	0.08	400
b3/b2	1.42	1	1.20	0.05	4	1.43	1.03	1.19	0.06	5
b3/b7	1.18	0.71	0.92	0.09	10	1.16	0.75	0.92	0.09	10
b5/b7	2.32	1.71	1.94	0.11	6	2.29	1.71	1.94	0.10	5

Max. = Maximum; Avg. = Average; Min. = Minimum; SD = Standard Deviation; CV = Coefficient of Variation.

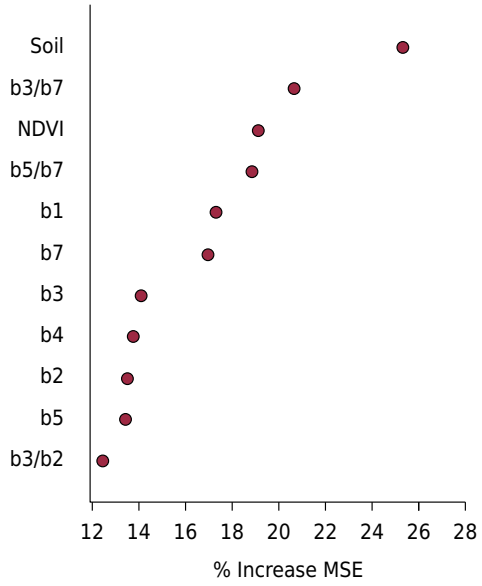

Figure 2. Importance of the predictors in models derived from the Random Forest. MSE = Mean Square Error.

The VAR_{ex} obtained from the Out-of-Bag (MSE_{OOB}) data using the training samples was 51.07 %, and the goodness of fit estimated by RMSE was 8.07 $cmol_c\ kg^{-1}$ (Table 2). The results can be considered satisfactory for performance of the models. Few studies in the literature have used RF to predict soil CEC, and no one has used only remote sensing data as main covariates alongside legacy soil maps (Table 2).

The results from this study were lower than those obtained by Lagacherie et al. (2013). These authors obtained 79 % for VAR_{ex} for the layer between 0.15 and 0.30 m and 3.4 $cmol_c\ kg^{-1}$ for RMSE (Table 2) using terrain attributes and hyperspectral data in the visible and near-infrared spectrum (AISA-Dual) with 5 m of spatial resolution as input data. The lower results obtained in this study may be correlated with the coarser spatial resolution from Landsat 5 (30 m) in comparison with Lagacherie et al. (2013), who used hyperspectral data (5 m). The effect of spatial resolution on predicting soil properties was also reported by Smith et al. (2006) and Ruiz Navarro et al. (2012).

The results obtained by Hengl et al. (2015) were higher than those obtained in the present study (Table 2). Those authors used RF in combination with a large number of covariates and 28,000 soil profiles as input data.

However, the performance metrics for CEC prediction in this study were better than those achieved by Vaysse and Lagacherie (2015), who did not have satisfactory results predicting this soil property (Table 2). The poor performance in that case may be associated with the small variation from the parent material and the erosion/deposition rates, which were not captured by the spatial resolution of the covariates (100 m). Moreover, the authors highlighted the importance of the input dataset to improve performance of the models.

Cokriging

The semivariogram obtained for soil CEC (Figure 3) provides a description of spatial dependence and indicates processes related to spatial distribution (Liao et al., 2013).

As shown in figure 3, the exponential model satisfactorily fits the semivariogram, as in the study performed by Liao et al. (2011). However, Bilgili et al. (2011) and Ciampalini et al. (2012) obtained a better fit of the semivariogram with a spherical model. The nugget effect is an important parameter that indicates unexplained variation based on sampling distance (McBratney and Webster, 1983). In this sense, the value of the nugget effect greater than zero (12.53) may be related to measurement error or to unexplained spatial variation of the soil property (Liao et al., 2013).

Table 2. Pearson's correlation between the cation exchange capacity (CEC) and the environmental covariates

Covariates	CEC		
	p-value	r	Correlation
b1	0.00*	-0.14	Weak
b2	0.17^{ns}	-0.06	not correlated
b3	0.07 ^{ns}	-0.08	not correlated
b4	0.00*	-0.34	Moderate
b5	0.00*	-0.32	Moderate
b7	0.00*	-0.31	Moderate
NDVI	0.00*	-0.18	Weak
b3/b2	0.09^{ns}	-0.08	not correlated
b3/b7	0.00*	0.44	Moderate
b5/b7	0.00*	0.17	Weak
Soil Map	0.00*	0.50	Strong

r = values for Pearson's correlation; * = significance at 5 %; ns = no significance.

The strength of spatial autocorrelation for a soil property can be determined by the ratio between the nugget effect and sill (Cambardella et al., 1994; Rossi et al., 2009; Bilgili et al., 2011; Liao et al., 2011). For this ratio, values below 25 % are considered a strong spatial dependence; whereas, this dependence is considered moderate for values between 25 and 75 %. Values greater than 75 % are considered poorly dependent. Therefore, the nugget/sill ratios of 15 % observed in this study (Figure 3) indicate a strong spatial dependence of the CEC in this area. In this sense, it is possible that the spatial variability of this property may be driven by intrinsic soil forming factors, such as the parent material (Cambardella et al., 1994). In the case of this study, the area is composed of limestone and gneiss-granite rocks and has a flat surface, showing low relevance of relief as a formation factor. Strong spatial dependence of the soil CEC was also observed by Liao et al. (2011), Bilgili et al. (2011), and Ciampalini et al. (2012) in their respective studies.

The sill parameter expresses the inflection point where the distance between samples (1,129 m in that case) shows no spatial autocorrelation. In a study conducted in Tunisia, Ciampalini et al. (2012) observed values for this parameter ranging from 250 to 2,000 m, which can be considered similar to the values obtained in this study. Liao et al. (2011) observed higher values (37,500 to 40,100 m), whereas a lower range of values (379-427 m) was reported by Bilgili et al. (2011). Furthermore, these results reflect the characteristics of different case studies due to sampling density (Trangmar et al., 1986).

The semivariogram used to estimate the CEC values has an R^2 value of 0.57 and an RMSE of 7.22 $cmol_c\ kg^{-1}$ for the validation samples. These performance metrics were better than those found by Liao et al. (2013), who used cokriging and principal components derived from soil properties to predict soil CEC in China. As observed by Ciampalini et al. (2012), the values for the coefficient of determination (R^2 = 0.50) and the RMSE (5.16 $cmol_c\ kg^{-1}$) were slightly lower compared to those obtained in this study. In contrast, Bilgili et al. (2011) observed lower RMSE values for all sample sets (1.41 to 1.83 $cmol_c\ kg^{-1}$) when predicting CEC values by cokriging using hyperspectral data in the visible and near-infrared spectrum as covariates.

Comparison of predictive models

The descriptive statistics of soil CEC prediction for the RF and CK models are presented in table 3.

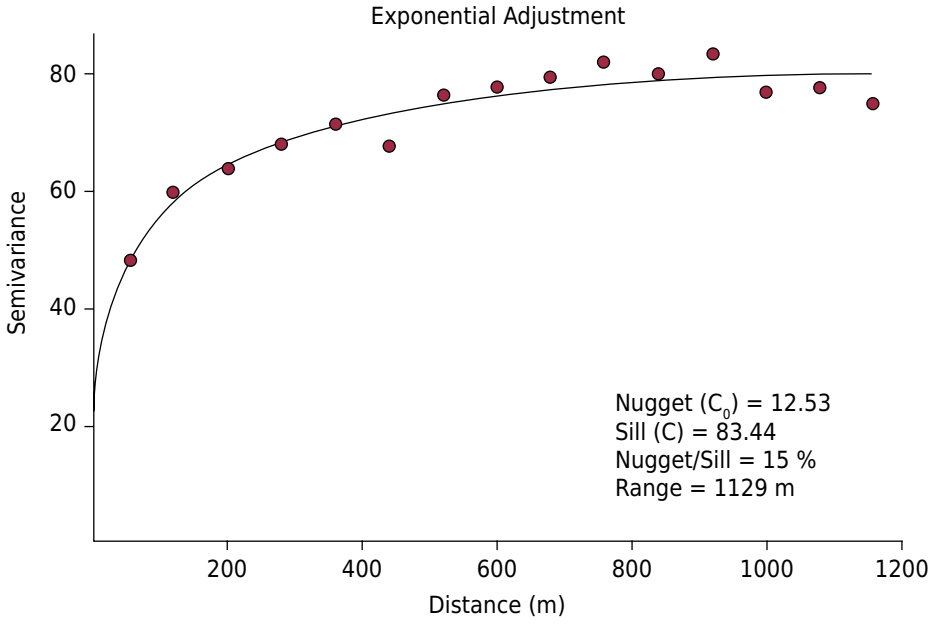

Figure 3. Semivariogram of cation exchange capacity obtained by cokriging.

The RF model produced predicted values for CEC within the range of the original values, with a smaller standard deviation and coefficient of variation, as expected for this model (Table 3). In contrast, the CK model produced a map with a greater range of values than RF, with more similarity to the descriptive statistics for CEC from the soil samples, which, according to Liao et al. (2011), means that this model had greater accuracy in describing CEC spatial variation than the RF model. The mean values (34.11 $cmol_c\ kg^{-1}$) and standard deviation (9.58 $cmol_c\ kg^{-1}$) are also closer to the values for soil samples (34.22 and 11.37 $cmol_c\ kg^{-1}$, respectively). The CV for both models were similar and smaller than the CV from the soil samples, which was also observed by Liao et al. (2011).

Performance of the models based on the independent validation dataset (124 samples) is presented in figura 4. The CK model had better performance (R^2 of 0.57 and 7.22 $cmol_c\ kg^{-1}$ for RMSE) than the RF model (R^2 of 0.47 and 7.89 $cmol_c\ kg^{-1}$ for RMSE) in predicting soil CEC.

Figure 4 corroborates this since the dispersion of points in the RF model is greater than in the CK model. A notable point of this study regards the comparison between the random forest and cokriging models in predicting soil CEC; this comparison was not reported in the literature.

Quantification of soil properties using an orbital sensor is not an easy task, due to the complexity of soil dynamics and properties, as pointed out by Demattê et al. (2007). In this sense, the results obtained in this study can be considered satisfactory and are probably related to physical interference of the soil in incident and reflected energy.

The spatial distribution of CEC predicted by the RF and CK models, as well the statistics regarding modeling by RF and CK, are presented in figure 5 and table 4, respectively. Higher CEC can be observed in the central portion of the area for both models, corresponding to *Vertisols* developed from limestone, which represent 94.5 % of the study area. These soils have a surface horizon with average thickness of 0.138 m and topsoil texture ranging from clay loam to clay, with high cation saturation in the exchange complex. The lowest levels for CEC were observed in the eastern, southwestern, and northwestern portions of the area and are related to *Cambisols* and *Planosols*. *Cambisols* have a topsoil horizon (A) with an average thickness of 0.149 m. Most of these soils have high activity clay and are eutrophics, with soil texture of clay loam and, infrequently, clay. *Planosols* had a surface horizon with an average thickness of 0.166 m, clay loam as topsoil texture, and very high cation saturation.

Table 3. Results from previous studies that used the Random Forest model to predict soil cation exchange capacity

Author	Explained variance	RMSE
	%	
Lagacherie et al. (2013)	35-79	3.4-6.5 $cmol_c\ kg^{-1}$
Vaysse and Lagacherie (2015)	-	9.49 $cmol_c\ kg^{-1}$
Hengl et al. (2015)	66.3	7.92 $cmol_c\ kg^{-1}$
This study	48.57	8.27 $cmol_c\ kg^{-1}$

RMSE = root mean square error.

Table 4. Descriptive statistics from soil samples, Random Forest, and Cokriging models

Models	Minimum	Maximum	Average	SD	CV
	$cmol_c\ kg^{-1}$				%
Training dataset	5.71	61.41	34.22	11.37	39
Random Forest	11.98	53.32	34.69	8.32	24
Cokriging	9.81	70.77	34.11	9.58	28

SD = Standard Deviation; CV = Coefficient of Variation.

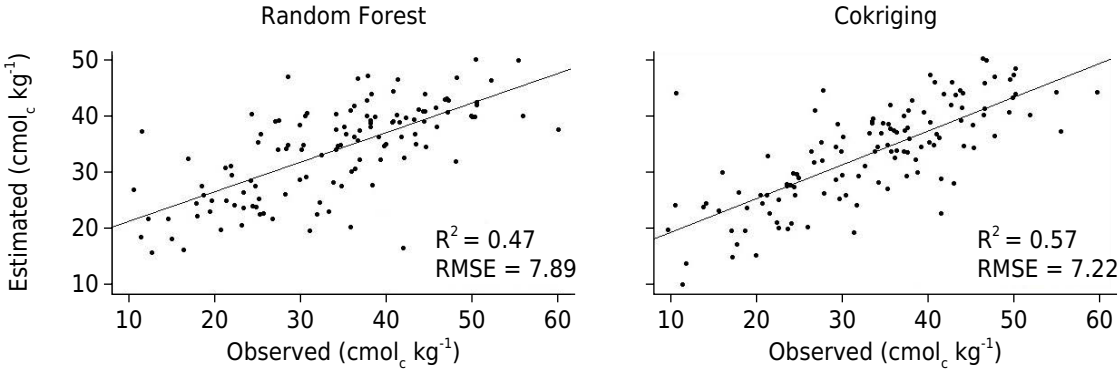

Figure 4. Plot of observed and estimated values, coefficient of determination (R^2), and the root mean square error (RMSE) obtained from models and validation samples.

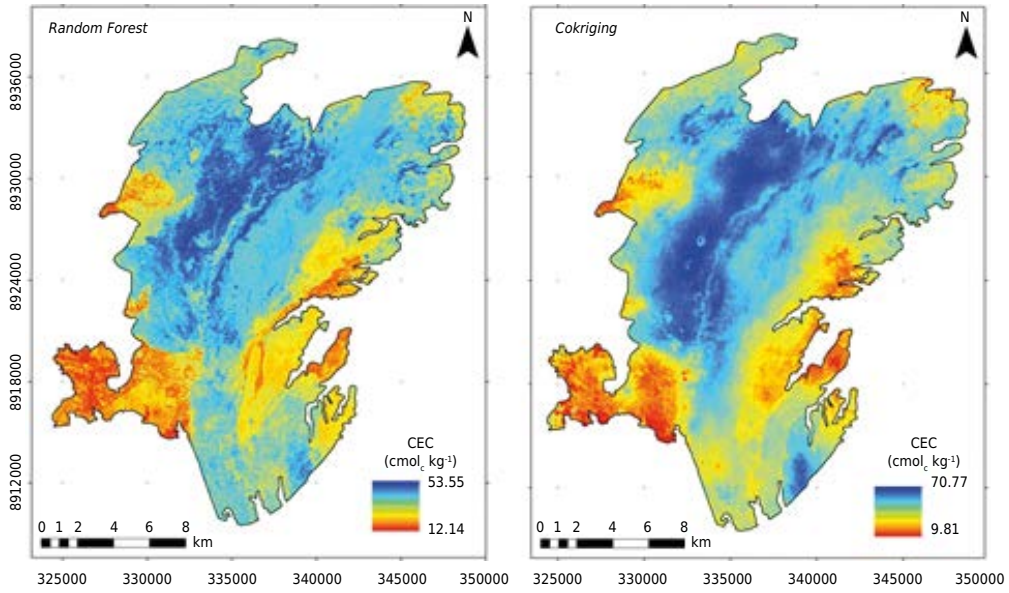

Figure 5. Spatial distribution of soil cation exchange capacity (CEC) estimated by the Random Forest and Cokriging models.

CONCLUSIONS

The approach based on the combination of orbital remote sensing data, with Random Forest (RF) and Cokriging (CK) methods to predict soil Cation Exchange Capacity (CEC), in Brazilian semi-arid soils provides satisfactory results.

The semivariogram showed that soil CEC has strong spatial dependence in the study area. Although the cokriging results were satisfactory, improvement in the spatial resolution of covariates may achieve better performance of models in predicting soil CEC in the area.

More research is needed to improve the quality of the covariates used as input data in digital soil mapping.

ACKNOWLEDGMENTS

This study was supported by Embrapa Solos (National Center of Soil Research) and the Federal Rural University of Rio de Janeiro (Soil Department - Agronomy Institute).

REFERENCES

Akpa SIC, Odeh IOA, Bishop TFA, Hartemink AE. Digital mapping of soil particle-size fractions for Nigeria. Soil Sci Soc Am J. 2014;78:1953-66. https://doi.org/10.2136/sssaj2014.05.0202

Bilgili AV, Akbas F, van Es HM. Combined use of hyperspectral VNIR reflectance spectroscopy and kriging to predict soil variables spatially. Precision Agric. 2011;12:395-420. https://doi.org/10.1007/s11119-010-9173-6

Boettinger JL, Ramsey RD, Bodily JM, Cole NJ, Kienast-Brown S, Nield SJ, Saunders AM, Stum AK. Landsat spectral data for digital soil mapping. In: Hartemink AE, McBratney AB, Mendonça-Santos ML, editors. Digital soil mapping with limited data. New York: Springer; 2008. p. 193-202.

Breiman L. Random forests. Technical report for version 3. Berkeley: Statistics Department University of California Berkeley; 2001 [access 2014 Dec 28]. Available from: http//oz.berkeley.edu/users/breiman/randomforest2001.pdf.

Cambardella CA, Moorman TB, Parkin TB, Karlen DL, Novak JM, Turco RF, Konopka AE. Field-scale variability of soil properties in center Iowa soils. Soil Sci Soc Am J. 1994;58:1501-11. https://doi.org/10.2136/sssaj1994.03615995005800050033x

Carvalho AP, Larach JOI, Jacomine PKT, Camargo MN. Critérios para distinção de classes de solos e de fases de unidades de mapeamento: normas em uso pelo SNLCS. Rio de Janeiro: Embrapa-SNLCS; 1988. (Documentos, 11).

Carvalho Junior W, Lagacherie P, Chagas CS, Calderano Filho B, Bhering SB. A regional-scale assessment of digital mapping of soil attributes in a tropical hillslope environment. Geoderma. 2014;232-234:479-86. https://doi.org/10.1016/j.geoderma.2014.06.007

Ciampalini R, Lagacherie P, Hamrouni H. Documenting GlobalSoilMap.net grid cells from legacy measured soil profile and global available covariates in Northern Tunisia. In: Minasny B, Malone BP, McBratney AB, editors. Digital soil assessments and beyond. London: CRC Press; 2012. p. 439-44.

Companhia de Desenvolvimento dos Vales do São Francisco - Codevasf. Projeto Salitre: levantamento detalhado de solo e classificação de terras para irrigação. Recife: Codevasf; 1989. (Projetos técnicos, 15).

Cutler DR, Edwards Jr TC, Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ. Random forests for classification in ecology. Ecology. 2007;88:2783-92. https://doi.org/10.1890/07-0539.1

Demattê JAM, Galdos MV, Guimarães RV, Genú AM, Nanni MR, Zullo Junior J. Quantification of tropical soil attributes from ETM+/LANDSAT-7 data. Int J Remote Sens. 2007;28:3813-29. https://doi.org/10.1080/01431160601121469

Grimm R, Behrens T, Märker M, Elsenbeer H. Soil organic carbon concentrations and stocks on Barro Colorado Island - digital soil mapping using Random Forests analysis. Geoderma. 2008;146:102-13. https://doi.org/10.1016/j.geoderma.2008.05.008

Hengl T, Heuvelink GBM, Kempen B, Leenaars JGB, Walsh MG, Shepherd KD. Mapping soil properties of Africa at 250 m resolution: random forests significantly improve current predictions. PLoS One. 2015;10:e0125814. https://doi.org/10.1371/journal.pone.0125814

Holmes KW, Chadwick OA, Kyriakidis PC. Error in a USGS 30-meter digital elevation model and its impact on terrain modeling. J Hydrol. 2000;233:154-73. https://doi.org/10.1016/S0022-1694(00)00229-8

Instituto Brasileiro de Geografia e Estatística - IBGE. Manual técnico de pedologia. 3. ed. Rio de Janeiro, RJ: IBGE; 2015. (Manuais técnicos em geociências).

Lagacherie P, Sneep A-R, Gomez C, Bacha S, Coulouma G, Hamrouni MH, Mekki I. Combining Vis–NIR hyperspectral imagery and legacy measured soil profiles to map subsurface soil properties in a Mediterranean area (Cap-Bon, Tunisia). Geoderma. 2013;209-210:168-76. https://doi.org/10.1016/j.geoderma.2013.06.005

Liao K, Xu S, Wu J, Zhu Q. Spatial estimation of surface soil texture using remote sensing data. Soil Sci Plant Nutr. 2013;59:488-500. https://doi.org/10.1080/00380768.2013.802643

Liao K-h, Xu S-h, Wu J-c, Ji S-h, Lin Q. Cokriging of soil cation exchange capacity using the first principal component derived from soil physico-chemical properties. Agr Sci China. 2011;10:1246-53. https://doi.org/10.1016/S1671-2927(11)60116-8

Liaw A, Wiener M. Classification and regression by randomForest. R News. 2002;2/3:18-22.

Malone BP, McBratney AB, Minasny B, Laslett GM. Mapping continuous depth functions of soil carbon storage and available water capacity. Geoderma. 2009;154:138-52. https://doi.org/10.1016/j.geoderma.2009.10.007

McBratney AB, Mendonça Santos ML, Minasny B. On digital soil mapping. Geoderma. 2003;117:3-52. https://doi.org/10.1016/S0016-7061(03)00223-4

McBratney AB, Webster R. Optimal interpolation and isarithmic mapping of soil properties: V. Co-regionalization and multiple sampling strategy. Eur J Soil Sci. 1983;34:137-62. https://doi.org/10.1111/j.1365-2389.1983.tb00820.x

Oliveira LB. Manual de métodos de análise do solo. Rio de Janeiro; Embrapa-SNLCS; 1979.

R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria; 2007 [cited 2015 Aug 5]. Available from: http://www.R-project.org/.

Rossi J, Govaerts A, Vos BD, Verbist B, Vervoort A, Poesen J, Muys B, Deckers J. Spatial structures of soil organic carbon in tropical forests - a case study of Southeastern Tanzania. Catena. 2009;77:19-27. https://doi.org/10.1016/j.catena.2008.12.003

Ruiz Navarro A, Barberá GG, García-Haro J, Albaladejo J. Effect of the spatial resolution on landscape control of soil fertility in a semiarid area. J Soils Sediments. 2012;12:471-85. https://doi.org/10.1007/s11368-012-0470-8

Santos HG, Jacomine PKT, Anjos LHC, Oliveira VA, Oliveira JB, Coelho MR, Lumbreras JF, Cunha TJF. Sistema brasileiro de classificação de solos. 3. ed. rev. ampl. Rio de Janeiro: Embrapa Solos; 2013.

Smith MP, Zhu A-X, Burt JE, Stiles C. The effects of DEM resolution and neighborhood size on digital soil survey. Geoderma. 2006;137:58-69. https://doi.org/10.1016/j.geoderma.2006.07.002

Souza JD, Kosin M, Melo RC, Santos RA, Teixeira LR, Sampaio AR, Guimarães JT, Bento RV, Borges VP, Martins AAM, Arcanjo JB, Loureiro HSC, Angelim LAA. Mapa geológico do estado da Bahia, escala 1:1000000. Versão 1.1. Salvador: CPRM/Programas Carta Geológica do Brasil a milionésimo e levantamentos geológicos básicos do Brasil; 2003.

Tang L, Zeng GM, Nourbakhsh F, Shen GL. Artificial neural network approach for predicting cation exchange capacity in soil based on physico-chemical properties. Environ Eng Sci. 2009;26:137-46. https://doi.org/10.1089/ees.2007.0238

Trangmar BB, Yost RS, Uehara G. Application of geostatistics to spatial studies of soil properties. Adv Agron. 1986;38:45-94. https://doi.org/10.1016/S0065-2113(08)60673-2

Vaysse K, Lagacherie P. Evaluating digital soil mapping approaches for mapping GlobalSoilMap soil properties from legacy data in Languedoc-Roussillon (France). Geoderma Regional. 2015;4:20-30. https://doi.org/10.1016/j.geodrs.2014.11.003