Revista Brasileira de Ciência do Solo

Division - Soil In Space and Time | Commission – Pedometrics

Mapping soil properties in a poorly-accessible area

Elias Mendes Costa[(1)*], Helena Saraiva Koenow Pinheiro[(2)], Lúcia Helena Cunha dos Anjos[(2)], Robson Altiellys Tosta Marcondes[(1)] and Yuri Andrei Gelsleichter[(3)]

[(1)] Universidade Federal Rural do Rio de Janeiro, Departamento de Solos, Pós-Graduação em Agronomia Ciência do Solo (PPGAS-UFRRJ), Seropédica, 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, Programa de Pós-graduação em Ciência, Tecnologia e Inovação em Agropecuária, Seropédica, Rio de Janeiro, Brasil.

*** Corresponding author:**
E-mail: eliasmccosta@gmail.com

Received: August 21, 2019

Approved: November 5, 2019

How to cite: Costa EM, Pinheiro HSK, Anjos LHC, Marcondes RAT, Gelsleichter YA. Mapping soil properties in a poorly-accessible area. Rev Bras Cienc Solo. 2020;44:e0190107.
https://doi.org/10.36783/18069657rbcs20190107

ABSTRACT: Soil maps are important to evaluate soil functions and support decision-making process, particularly for soil properties such as pH, carbon content (C), and cation exchange capacity (CEC), but the spatial resolution and soil depth should meet the needs of users. On another hand, the efficiency of statistical models to create soil maps, with an acceptable level of accuracy, often require a large number of samples with an appropriate distribution across the area of interest. However, accessibility for sampling can be a trouble in remote areas, such as the Itatiaia National Park (INP). The hypothesis of this work is that it is possible to obtain a viable result in soil mapping of areas with limited access by using DSM tools. The general objective of this paper was to create 2- and 3-D maps of the soil properties pH, carbon content, and CEC, with the correspondent spatial uncertainty, in the INP plateau. The sampling strategy was designed using conditioned Latin Hypercube Sample (cLHS), and different methods were tested to produce the soil properties maps. For calibration of the models: linear (MLR, multiple linear regression) and nonlinear (GAM, Generalised Additive Models). The results showed differences in predictive performance for all statistical methods and covariate selection approaches. The GAM, with covariates selection based on soil formation factors, was the best method for the limited number of soil samples. The greatest uncertainty was associated with areas with the lowest accessibility and, consequently, with low sampling density and/or noises in covariates. Even though the 2- and 3-D maps of soil properties, each associated with explicit uncertainty, can contribute to the INP decision makers/managers by providing information not available before.

Keywords: depth function, generalized additive models, uncertainty propagation, predictor selection.

INTRODUCTION

Soil is a vital part of the natural environment and it has a crucial role in ecosystem functioning (Adhikari and Hartemink, 2016). The soil functions can be derived from interactions of soil properties. To predict soil properties to assess their functions can provide detailed spatial information particularly useful in complex mountain terrain (Jeong et al., 2017). Soil information is an extremely important factor for conservation and sustainable management, and is essential in the formulation of sustainable agricultural policies and monitoring impacts of inappropriate use of resources (Carvalho Júnior et al., 2016), especially in mountain areas.

A relatively new approach that gives useful soil information is the 3-D modeling of soil properties. The modeling of soil properties in three dimensions has been evaluated in several studies (Kidd et al., 2015; Mulder et al., 2016; Amirian Chakan et al., 2017), including the assessment of associated uncertainty (Kempen et al., 2011; Poggio and Gimona, 2017a, b). With the progress of digital soil mapping (DSM), there is rising use of 3-D modeling to provide information on soil patterns for applications, from agricultural management to ecosystem services (Zhang et al., 2017) and so on. In terms of the evaluation of predicted results besides the estimation of the errors, such as Accuracy and kappa in a categorical map and R2, RMSE, ME in the continuous map, it is important to evaluate the uncertainty associated with the prediction. This is another important information to guide land management choices and decision-makers (Poggio and Gimona, 2017a).

In recent years, there was a considerable advance in DSM due to new approaches, among them, powerful predictive algorithms (Beguin et al., 2017; Sindayihebura et al., 2017); models combining machine-learning and geostatistical tools (Poggio and Gimona, 2017a,b); expert knowledge-based methods (Menezes et al., 2014, 2018); and high-resolution soil maps (Nussbaum et al., 2017). However, the limiting factor is often the number of soil data used for model calibration (Samuel-Rosa et al., 2015; Somarathna et al., 2017). It was suggested that more data was more important than a better model (Somarathna et al., 2017). However, obtaining more data can be a problem because of the size and/or the accessibility of some test areas.

To facilitate DSM in poorly-accessible areas, Cambule et al. (2013) proposed a methodology of sampling in the area of greater accessibility, which is representative of the total area, and to evaluate the representativeness using, e.g., the similarity between the histogram of the covariates for the total and accessible areas. Other studies considered the costs of accessibility in soil sampling (Roudier et al., 2012; Carvalho Júnior et al., 2014; Stumpf et al., 2016) using a variation/optimization of the method known as conditioned Latin Hypercube Sampling (cLHS), proposed by Minasny and McBratney (2006). The cLHS is a robust tool for the allocation of sampling points by means of a set of auxiliary covariates. The idea is to be able to capture the maximum of soil variation, and its properties, by using environmental covariates as auxiliary information.

The Itatiaia National Park (INP) (Brazil) has limited access due to the steep slopes, dense forest cover, rocky outcrops, and altitude vegetation fields in the upper plateau (Barreto et al., 2013), all that making the INP an excellent case study. To obtain a viable result with a low cost, it is important to use DSM tools, ranging from optimization of the sampling site (Minasny and McBratney, 2006; Roudier et al., 2012; Stumpf et al., 2016) to the covariate selection using powerful predictive algorithms (Beguin et al., 2017; Jeong et al., 2017). Based on the above and considering the advancements and challenges of DSM, it is necessary to optimize financial and human resources to produce quality information that can be useful for decision-makers.

The hypothesis of this work is that it is possible to obtain a viable result to map soil properties in areas with limited access by using DSM tools.The general objective of this paper was to create 2- and 3-D models of soil properties (pH, carbon content, and cation exchange capacity), with the correspondent spatial uncertainty, with a resolution of 25 m and in a poorly accessible area of the INP plateau. For that, it was necessary to design a sampling strategy that balanced accessibility, costs, area, and environmental covariates; and to model soil properties, with the number of samples available. A goal of this study was to contribute with information for the management plan of the INP, regarding the soil properties studied and their potential relationship to the ecosystems in the park.

MATERIALS AND METHODS

Study area

The INP has an area of 225.54 km^2, and it is located in the Serra da Mantiqueira, the border region between the Minas Gerais (MG) and Rio de Janeiro (RJ) States (Barreto et al., 2013). According to Tomzhinski et al. (2012), the INP can be divided into three broad areas: the "Lower part", which comprises the southern part of the park, the "upper part" of the plateau (Figure 1) and Visconde de Mauá in the east side.

Data sources and environmental covariates

The environmental covariates used to model the soil properties were derived from three data sources: a digital elevation model, remote sensing data (orbital image), and a geology map (Table 1). They were chosen to describe the main soil-forming factors, according to the *scorpan* approach (McBratney et al., 2003).

Digital Elevation Model (DEM)

The DEM used, with a spatial resolution of 25 m, was generated from contour lines, with 20 m equidistance, and hydrography extracted from plani-altimetric charts, both in the

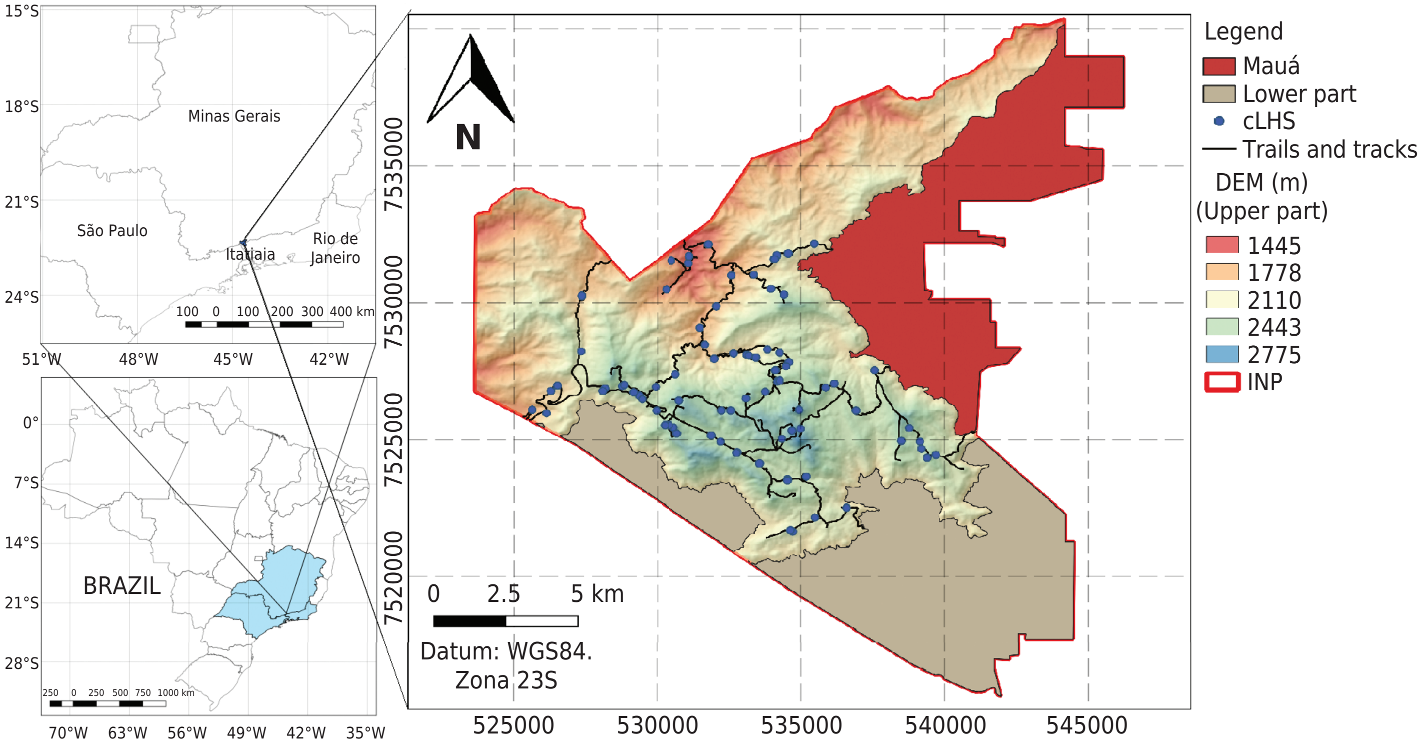


Figure 1. The upper part of the Itatiaia National Park in the southeastern region of Brazil. Major roads and trails are in black. Blue points are soil sampling points selected according to cLHS method (Minasny and McBratney, 2006).

Table 1. Environmental covariates, soil formation factor that represents their sources, resolution, and definition

Formation factor	Covariate	Source	Spatial resolution	Definition
Organism (O)	Bands (1,2,3,4, and 5)	RapidEye (2011)	5 m	Bands in the spectrum of 440-510 nm (Blue), 520-590 nm (Green), 630-685 nm (Red), 690–730 nm (Red Edge), 760-850 nm (Near IR)
	NDVI	RapidEye (2011)	5 m	NDVI=(NIR–Red)/(NIR+Red)
	SAVI	RapidEye (2011)	5 m	SAVI= (1+0.5) (NIR–Red)/(NIR+Red+0.5)
Relief (R)	DEM	INP managers	25 m	Digital elevation model of the area-representation of the terrain's surface made by contour lines and hydrology (scale 1:50,000, IBGE data)
	Slope	DEM	25 m	Gradient or rate of change of elevation between neighboring cells
	Aspect	DEM	25 m	Represents exposure faces, values in degrees (0 to 360°)
	Northernness	DEM	25 m	Indicates the direction of the slope relative to the northern. Northernness = abs (180°−Aspect)
	Plan_curv	DEM	25 m	The shape of the hillside on the horizontal plane (concave, rectilinear or convex).
	Prof_curv	DEM	25 m	The shape of the hillside on the vertical plane (concave, rectilinear or convex).
	Convergence	DEM	25 m	The general shape of the hillside in all directions (concave, rectilinear or convex)
	Cat_area	DEM	25 m	Related to volume of flooding that reaches a certain cell
	TWI	DEM	25 m	Describes a tendency for a cell to accumulate water
	LS_factor	DEM	25 m	Attribute equivalent to the topographic factor of the Revised Universal Soil Loss Equation (RUSLE)
	RSP	DEM	25 m	Represents relative slope position based on the base channel network
	CHND	DEM	25 m	Altitude above the channel network (CHNB- original elevation)
	CHNB	DEM	25 m	Interpolation of a channel network base level elevation
Parent material (P)	Geology	Santos et al. (2000)	25 m	Categorical map with geological information (scale 1:50,000)
Spatial position (N)	X, Y	Grid data	-	X = longitude, Y = latitude in UTM system, zone 23S, projection Sirgas 2000
	XY	Grid data	-	XY = polygon of second order of X and Y, XY = $(X^2+Y^2+X*Y)/10^6$

Geology classes: alluvial sediments, colluvium sediments, nepheline syenite, quartz syenite, alkaline granite, magmatic breccia, homogeneous gneisses. NDVI: normalized difference vegetation index; SAVI: soil-adjusted vegetation index; DEM: digital elevation model; Plan_curv: plan curvature; Prof_curv: profile curvature; Convergence: convergence index; Cat_area: catchment area; TWI: topographic wetness index; LS_factor: LS factor; RSP: relative slope position; CHND: channel network distance; CHNB: channel network base level.

1:50,000 scale. The sheets used were SF-23-ZA-I-2 Alagoa, SF-23-ZA-I-3 Passa Quatro, and SF-23-ZA-I-4 Agulhas Negras. They were obtained from the in vector format from the cartographic base of the Brazilian Institute of Geography and Statistics (IBGE). The dataset was provided by the INP administration.

Satellite image

Two scenes from the RapidEye sensor were used, scene 1 (from 02/07/2011) and scene 2 (16/08/2011). The two scenes were used to cover the entire study area. They have a 12-bit radiometric resolution, 6.5 m spatial resolution, and were orthorectified to 5 m spatial resolution (RapidEye, 2012). Both images were atmospherically corrected using the 6S model (Vermote et al., 1997).

Geology map

It was obtained from Santos et al. (2000), and it was scanned, vectorized, and georeferenced. The file was rasterized at the same spatial resolution as the DEM (25 m).

Soil dataset and sampling strategy

The soil dataset used for the DSM modeling had a total of 90 soil profiles, being 359 horizons with a morphological description, and 346 horizons with analytical data. The approach used in this study to select the sampling points followed the principles proposed by Minasny and McBratney (2006). The points were selected by using cLHS with auxiliary covariates, and considering the access costs (Roudier et al., 2012; Carvalho Júnior et al., 2014; Stumpf et al., 2016). As a constraint, three buffer sizes were created in relation to roads and trails, as proposed by Carvalho Júnior et al. (2014). After testing the distances of 100, 200, and 400 m, with no significant differences between buffers (data not shown here), the distance of 100 m was selected to represent the accessible area.

The auxiliary covariates used to select sampling locations were: geology, elevation, slope, northernness, and soil-adjusted vegetation index. From the 90 points, 74 were selected for soil profile description. Also, legacy data (Soares et al., 2016) and data unpublished from field trips in the area were added to the database. In addition, other samples were selected based on the expert pedological knowledge and the relationship between soil and landscape, as recommended in the methods for conventional soil surveys (IBGE, 2015). These additional data (n=16) were obtained from places inside and outside the area in which the sampling was considered of higher accessibility (buffer of 100 m).

Covariates selection approach

To evaluate the relationships between soil properties such as pH, total carbon content (C), and cation exchange capacity (CEC), and environmental covariates, Multiple Linear Regression (MLR) and Generalized Additive Models (GAM) were tested (Figure 2). The MLR is a parametric method, and it assumes that the relationships between dependent variable and covariates are linear (Hastie et al., 2009). The GAM is a flexible statistical method that may be used to identify and characterize nonlinear regression effects through smoothing functions (Hastie et al., 2009; Wood, 2017).

The selection of the covariates was carried out to produce simpler models with the minimum number of covariates, and still able to explain the maximum of the data variability. Different strategies were used and they are described below.

Step one - correlation cut off: it evaluated the correlation between covariates. If two covariates had a coefficient greater than 0.85 (the cutoff value considered for this study), only one was maintained. The covariate used in the model was presumed to have a greater relationship with the *SCORPAN* (McBratney et al., 2003) model, i.e., greater pedological information.

Step two - selection using different approaches: it involved fitting models using the covariates maintained in the first step. But whereas fitted MLR models were fitted with all

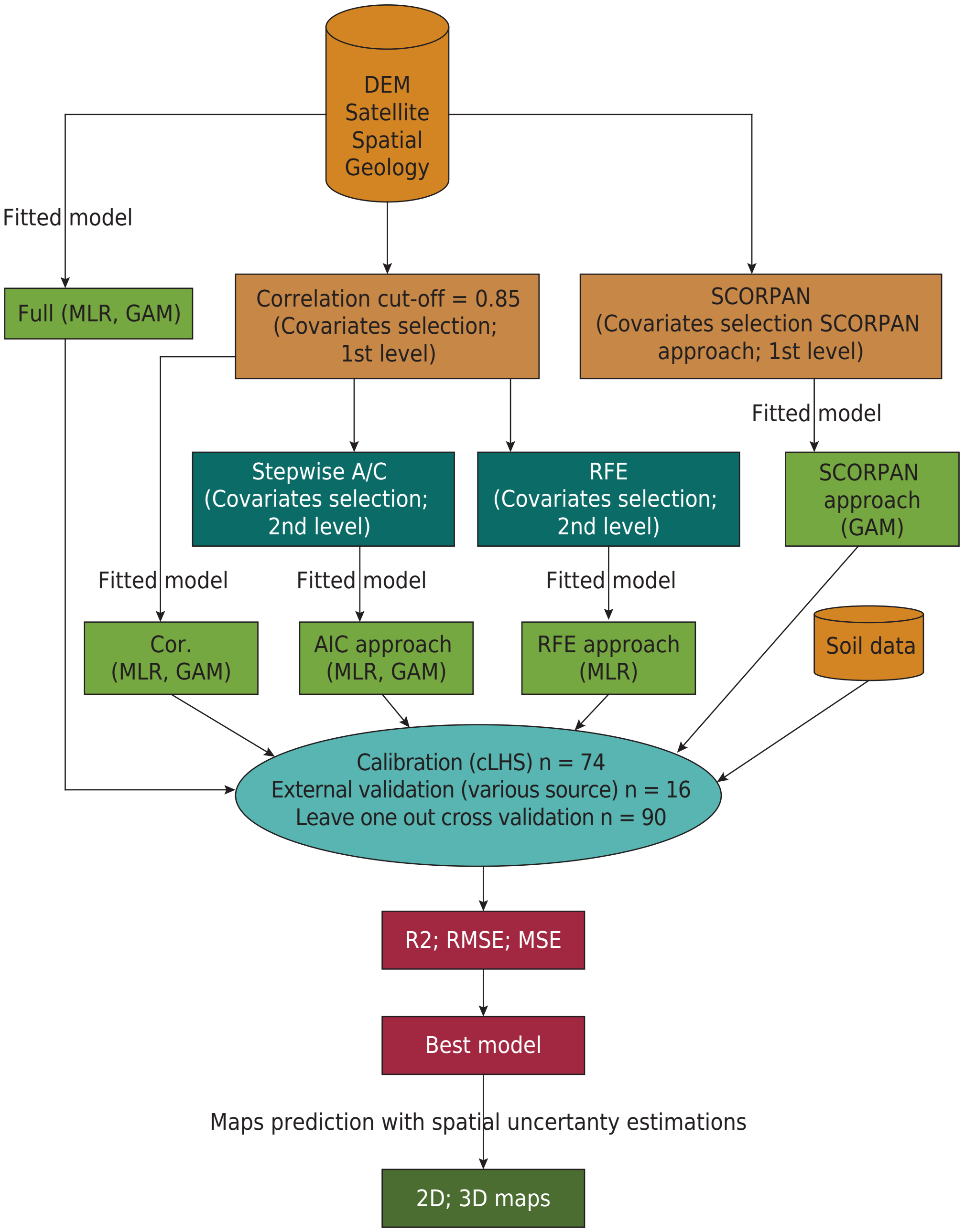

DEM
Satellite
Spatial
Geology
Fitted model
Full (MLR, GAM)
Correlation cut-off = 0.85 (Covariates selection; 1st level)
SCORPAN (Covariates selection SCORPAN approach; 1st level)
Fitted model
Stepwise A/C (Covariates selection; 2nd level)
RFE (Covariates selection; 2nd level)
SCORPAN approach (GAM)
Fitted model
Fitted model
Fitted model
Cor. (MLR, GAM)
AIC approach (MLR, GAM)
RFE approach (MLR)
Soil data
Calibration (cLHS) n = 74
External validation (various source) n = 16
Leave one out cross validation n = 90
R2; RMSE; MSE
Best model
Maps prediction with spatial uncertanty estimations
2D; 3D maps

Figure 2. Covariate selection approach, model fitting, validation, and prediction workflow. AIC: Akaike's information criterion; RFE: recursive feature elimination; R^2: coefficient of determination; RMSE: root mean square error; MSE: mean square error; LOO-CV: leave-one-out cross-validation.

covariates, the same was not possible for GAM, due to limitation of degrees of freedom for many covariates and the few soil samples (Poggio et al., 2013). The purpose of the model with all covariates is to have a basis of comparison with different methods of selection commonly used.

MLR: four models were fitted: one with all covariates (MLR_full); with covariates selection by correlation less than 0.85 between covariates (MLR_cor); other with the popular technique used in regression models, AIC (Akaike's Information Criterion) stepwise selection (Carvalho Junior et al., 2016; Chagas et al., 2016; Vermeulen and Niekerk, 2017) (MLR_step); and the technique of Recursive Feature Elimination (RFE) (MLR_RFE).

This last has recently been used in soil science for variable selection in machine-learning algorithms, and it is a backward selection using rank (Jeong et al., 2017; Vašát et al., 2017). The backward selection algorithm iteratively eliminates the least promising predictors from the model based on an initial predictor importance measure. When the full model has created a measure of variable importance is computed and shows the ranks of predictors from most to least important (Kuhn, 2017).

GAM: the approach was different due to the degree of freedom limitation. In this case, the models are penalized by the low number of points (soil data) and a large number of covariates, i.e., large parametric parameters of the GAM model. The models were fitted based on the stepwise forward approach, where covariates are added according to AIC. All models began with geographic coordinates (X, Y) and geology as fixed covariates. Three models were fitted. The first using the base model, where each covariate was added in the base model individually, and then evaluated by its AIC. The model ran with all covariates, and the four with lower AIC value composed the final model termed GAM_one. The second model consisted of making all possible combinations of four covariates and then run the model with all possible combinations. The combination with a smaller AIC was termed GAM_comb. This approach seeks to capture the interaction between covariates when a predictor model is fitted. In both cases (GAM_one and GAM_comb), it was included as many covariates as possible; since the base model already had nine covariates X, Y, and seven different levels for geology, it was possible to include another four covariates totaling a model with a maximum of 13. The third model involved a more parsimonious model based on the *scorpan* approach (McBratney et al., 2003). In this case, in addition to the base model that already included the parent material (geology) and spatial position (X, Y), for 2-D prediction and geology, and X, Y and depth (Z) for 3-D prediction, different combinations of data derived from the satellite image (in this way adding the factor organism associated to the land coverage) and data derived from the DEM (mainly represent factor relief, topography) were tested.

In all, possible combinations were tested for each soil property using external validation, but the same procedure was repeated using cross-validation. The best model in both evaluations was selected and termed GAM_*scorpan*.

The GAM model was selected for the 3-D approach due to its simplicity, and being a flexible approach that can deal with both linear and non-linear relationships between soil properties and the considered covariates (Poggio and Gimona, 2017a). Also, the 3-D smooth can provide a better performance, considering non-linear relationships between covariates and soil properties (Poggio and Gimona, 2017a), which are frequent when modeling natural environments.

For all combinations of models and covariate selection, an extension of the scorpan-kriging approach, hybrid geostatistical model, i.e., GAM-kriging or MLR-kriging was tested, combining the models with spatially correlated errors. However, the short-range spatial structure combined with the sparse sampling along the roads and tracks led to the fact that the errors do not show spatial dependence. Thus, this analysis was not fully accomplished.

Validation and uncertainty

The model's performance was evaluated in two ways: the first by external validation, where points selected by the cLHS n = 74 soil profiles (Minasny and McBratney, 2006) were used to fit the models. To validate the performance of the models, there were used n = 16 profiles, from the legacy data (retrieved from the literature), as well as extra points collected in the field based on the pedological knowledge and the soil-landscape relationship (without pre-selection). In training, samples were taken within a 100 m buffer in relation to roads and tracks. The validation samples include points inside and outside the buffer, defined as accessibility criterion; the second form of evaluation was leave-one-out cross-validation (LOO-CV) (Brus et al., 2011). In both cases, the Mean Square Error (MSE) and Root Mean Square Error (RMSE) were computed.

And a coefficient of determination was derived from a linear model between observed and predicted data (R^2). For 3-D soil modeling, the results of the modeling were summarized for the whole profile and at five depth layers, according to Global- Soil Map project specification (Arrouays et al., 2014), and compared with observed values from corresponding depths. Uncertainty propagation was analyzed through simulation (N = 1000) from the posterior distribution of fitted GAMs to derive simultaneous confidence intervals for the derivatives of penalized splines (Ruppert et al., 2003). All algorithms implementation, spatial prediction, and uncertainty analysis were done using the R program (R Core Team, 2018).

RESULTS

Correlation analysis between covariates

Strong relationships were observed between covariates derived from the satellite image, most of them with a correlation greater than 0.85 (Figure 3). They contributed in a similar way as information of vegetal coverage or land use, and their use may impair the model's fitness due to multicollinearity problems. The covariates CHND, band1, band2, band3, and SAVI were excluded, due to a correlation greater than 0.85 with one or more covariates. The CHND showed a strong correlation with elevation values.

It is usually necessary to decrease the number of covariates in the GAM models, especially when there is limited number of soil samples, as in this study. Since there was no high relation (no greater than the cut-off value 0.85) between covariates derived from satellite image and DEM (Figure 3), all possible combinations were tested to build the *scorpan* GAM model.

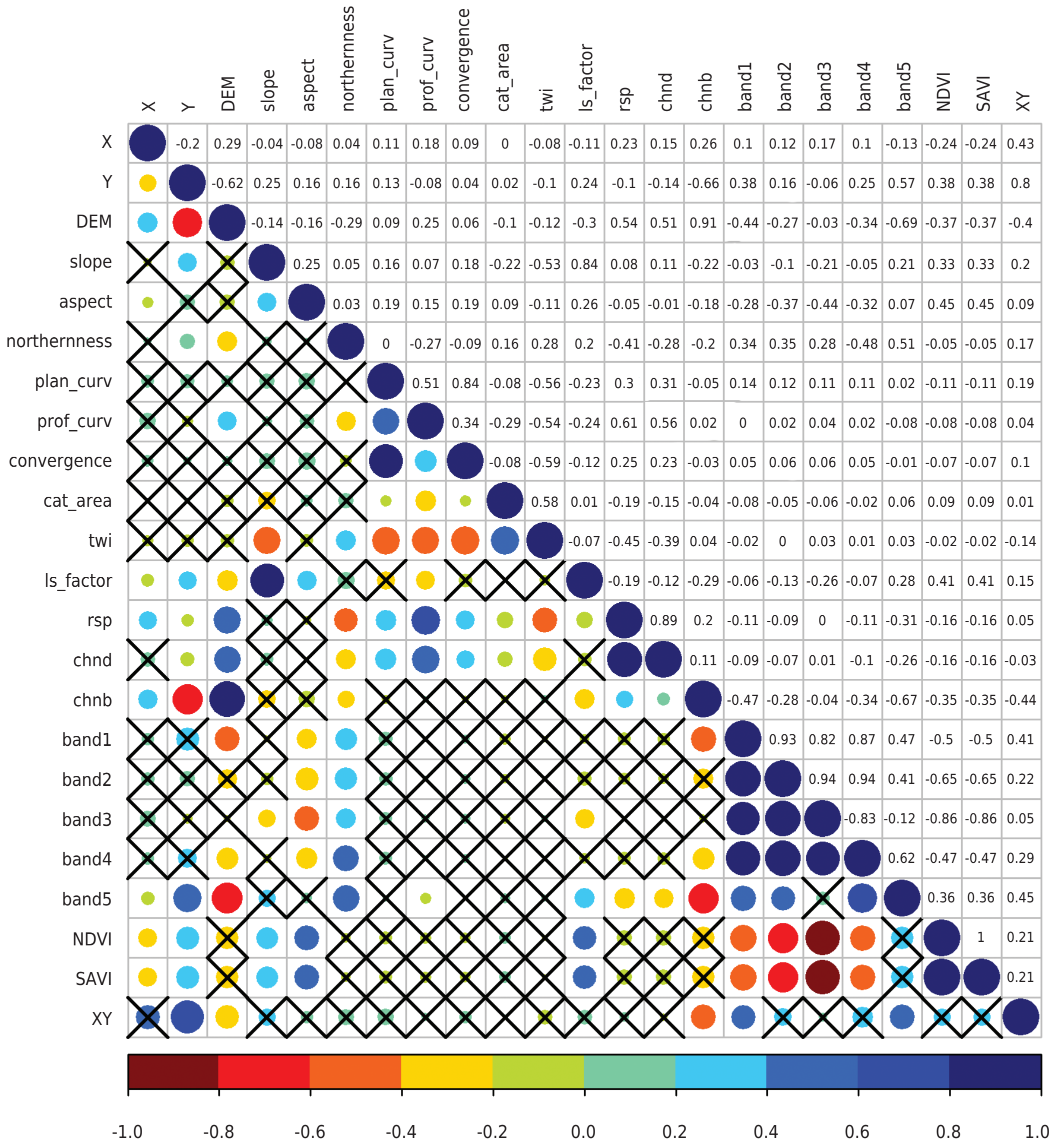
The 2-D approach

For the model's comparison by soil properties, for pH prediction using linear models, the best method of covariate selection was the RFE, in both cases, using external and cross-validation (Table 2). However, this pattern was not repeated for the other soil properties, C and CEC (Tables 3 and 4). For the total soil C and CEC, the RFE method presented the worst performance for the linear model (Table 2). In the case of the linear model, the most commonly used method for covariate selection, the stepwise method increased the R^2 coefficients when compared to other linear methods, for both validation methods - external validation and cross-validation, for C and CEC prediction and also RFE for pH.

For the GAM models, the best method of covariate selection was the *scorpan* approach, in both cases, using external, and cross-validation (Tables 2, 3, and 4). It was also the best model when compared with different approaches to select covariates in MLR models (Tables 2, 3, and 4). The *scorpan* model remained the best approach for covariate selection in GAM models.

For CEC prediction, all the best models in each approach had better performance when used external validation for evaluating. However, the validation samples are not completely random, and this may overestimate the model's performance for the external validation approach. Regardless of the differences, the best model selected in the external validation was also the best model in the cross-validation.

About the spatial prediction and uncertainty propagation, the predicted values (in the grid) and the observed values using the best covariate selection approach for MLR and GAM models are presented in the figure 4. For carbon content, the extreme values (minimum and maximum) were similar to the prediction made by MLR, with maximum values close to measured values, but the minimum showed negative values. The uncertainty in the predictions of soil properties for the superficial layer was mainly associated with



	X	Y	DEM	slope	aspect	northerness	plan_curv	prof_curv	convergence	cat_area	twi	ls_factor	rsp	chnd	chnb	band1	band2	band3	band4	band5	NDVI	SAVI	XY
X		-0.2	0.29	-0.04	-0.08	0.04	0.11	0.18	0.09	0	-0.08	-0.11	0.23	0.15	0.26	0.1	0.12	0.17	0.1	-0.13	-0.24	-0.24	0.43
Y			-0.62	0.25	0.16	0.16	0.13	-0.08	0.04	0.02	-0.1	0.24	-0.1	-0.14	-0.66	0.38	0.16	-0.06	0.25	0.57	0.38	0.38	0.8
DEM				-0.14	-0.16	-0.29	0.09	0.25	0.06	-0.1	-0.12	-0.3	0.54	0.51	0.91	-0.44	-0.27	-0.03	-0.34	-0.69	-0.37	-0.37	-0.4
slope					0.25	0.05	0.16	0.07	0.18	-0.22	-0.53	0.84	0.08	0.11	-0.22	-0.03	-0.1	-0.21	-0.05	0.21	0.33	0.33	0.2
aspect						0.03	0.19	0.15	0.19	0.09	-0.11	0.26	-0.05	-0.01	-0.18	-0.28	-0.37	-0.44	-0.32	0.07	0.45	0.45	0.09
northerness							0	-0.27	-0.09	0.16	0.28	0.2	-0.41	-0.28	-0.2	0.34	0.35	0.28	-0.48	0.51	-0.05	-0.05	0.17
plan_curv								0.51	0.84	-0.08	-0.56	-0.23	0.3	0.31	-0.05	0.14	0.12	0.11	0.11	0.02	-0.11	-0.11	0.19
prof_curv									0.34	-0.29	-0.54	-0.24	0.61	0.56	0.02	0	0.02	0.04	0.02	-0.08	-0.08	-0.08	0.04
convergence										-0.08	-0.59	-0.12	0.25	0.23	-0.03	0.05	0.06	0.06	0.05	-0.01	-0.07	-0.07	0.1
cat_area											0.58	0.01	-0.19	-0.15	-0.04	-0.08	-0.05	-0.06	-0.02	0.06	0.09	0.09	0.01
twi												-0.07	-0.45	-0.39	0.04	-0.02	0	0.03	0.01	0.03	-0.02	-0.02	-0.14
ls_factor													-0.19	-0.12	-0.29	-0.06	-0.13	-0.26	-0.07	0.28	0.41	0.41	0.15
rsp														0.89	0.2	-0.11	-0.09	0	-0.11	-0.31	-0.16	-0.16	0.05
chnd															0.11	-0.09	-0.07	0.01	-0.1	-0.26	-0.16	-0.16	-0.03
chnb																-0.47	-0.28	-0.04	-0.34	-0.67	-0.35	-0.35	-0.44
band1																	0.93	0.82	0.87	0.47	-0.5	-0.5	0.41
band2																		0.94	0.94	0.41	-0.65	-0.65	0.22
band3																			-0.83	-0.12	-0.86	-0.86	0.05
band4																				0.62	-0.47	-0.47	0.29
band5																					0.36	0.36	0.45
NDVI																						1	0.21
SAVI																							0.21
XY																							

Figure 3. Matrix of correlation between environmental covariates. Correlations with "X" are not significant at 5 % of confidence.

extrapolation of values for regions not sampled, and the INP boundaries with greater limitation of access (Figure 4).

The 3-D approach

On the continuous depth function using GAM *scorpan*, based on results for topsoil layer prediction, the GAM *scorpan* approach was chosen to predict the soil properties for the whole profile. In this case, besides the base model of 2-D GAM with covariates X, Y, and geology, the soil depth (Z) was added as a covariate in the base model to create a smoother 3-D. As with 2-D modeling, the base model was used to test different combinations of properties derived from DEM and satellite images. Since most of the soils in the INP have shallow profiles, it was considered for prediction the maximum depth limit of 1.00 m. A greater soil profile depth than that represents less than 23 % of the total data (Figure 5).

For model evaluation, it was used the descriptive statistic for the whole soil profile, and predictions for soil pH are very close (Tables 3) to observed values (Table 3), especially when using the cross-validation approach (Table 3). The values of the determination

Table 2. Performance of MLR and GAM models to predict soil pH, carbon content, and CEC

pH	**External validation**			**LOO-CV**		
Model	**R²**	**MSE**	**RMSE**	**R²**	**MSE**	**RMSE**
MLR_full	0.15	0.66	0.43	0.26	0.43	0.19
MLR_cor	0.21	0.62	0.39	0.30	0.41	0.17
MLR_step	0.22	0.48	0.23	0.24	0.42	0.18
MLR_rfe	**0.32**	**0.45**	**0.20**	**0.31**	**0.40**	**0.15**
GAM_one	0.20	0.48	0.23	0.35	0.38	0.14
GAM_comb	0.50	0.39	0.15	0.31	0.39	0.15
GAM_*scorpan*	**0.52**	**0.39**	**0.15**	**0.35**	**0.38**	**0.14**
Carbon	**External validation**			**LOO-CV**		
Model	**R²**	**MSE**	**RMSE**	**R²**	**MSE**	**RMSE**
MLR_full	0.06	7.56	57.10	0.14	6.55	42.87
MLR_cor	0.10	6.71	44.95	0.13	6.44	41.47
MLR_step	**0.17**	**5.29**	**27.97**	**0.24**	**5.25**	**27.59**
MLR_rfe	0.04	5.44	29.56	0.09	5.44	29.59
GAM_one	0.33	3.95	15.59	0.42	4.35	18.96
GAM_comb	0.31	4.09	16.71	0.43	4.31	18.56
GAM_*scorpan*	**0.49**	**3.85**	**14.83**	**0.45**	**4.21**	**17.74**
CEC	**External validation**			**LOO-CV**		
Model	**R²**	**MSE**	**RMSE**	**R²**	**MSE**	**RMSE**
MLR_full	0.20	14.96	223.69	0.05	12.92	166.81
MLR_cor	0.18	15.28	233.50	0.03	13.25	175.67
MLR_step	**0.19**	**14.78**	**218.37**	**0.04**	**11.18**	**124.97**
MLR_rfe	0.00	16.78	281.61	0.02	9.283	86.17
GAM_one	0.38	13.71	187.92	0.22	8.289	68.72
GAM_comb	0.32	13.17	173.41	0.17	8.581	73.63
GAM_*scorpan*	**0.41**	**13.61**	**185.06**	**0.27**	**7.764**	**60.29**

MLR_full: all covariates (Full model); MLR_cor: covariates selected with correlation smaller than 0.85 with each other's. Selected covariates for pH: MLR_step: covariates selected: DEM, Northernness, geology, X, NDVI; MLR_rfe: covariates selected: plan_curv, prof_curv, NDVI, Y, X, twi; GAM_one: covariates selected: X, Y, geology, DEM, northernness, chnd, band5); GAM_comb: covariates selected: X, Y, geology, DEM, aspect, plan_curv, cat_area; GAM_*scorpan*: covariates selected: X, Y, geology, prof_cuv, band3. Selected covariates for carbon content: MLR_step: covariates selected: DEM, Northernness, geology, X, NDVI; MLR_rfe: covariates selected: plan_curv, prof_curv, NDVI, Y, X, twi; GAM_one: covariates selected: X, Y, geology, DEM, northernness, chnd, band5); GAM_comb: covariates selected: X, Y, geology, DEM, aspect, plan_curv, cat_area; GAM_*scorpan*: covariates selected: X, Y, geology, prof_cuv, band3. Selected covariates for CEC: MLR_step: covariates selected: band5, northernness, DEM, X, chnd, geology; MLR_rfe: covariates selected: plan_curv, prof_curv, NDVI, ls_factor, twi, slope, convergence; GAM_one: covariates selected: X, Y, geology, band5, northernness, DEM, NDVI; GAM_comb: covariates selected: X, Y, geology, plan_curv, twi, band5, NDVI; GAM_*scorpan*: covariates selected: X, Y, geology, chnb, band3.

coefficient for carbon content and CEC are higher among observed and predicted values than for the pH, especially in cross-validation (Table 3). The magnitude of errors, RMSE, and MSE shows a tendency to extrapolate low carbon contents.

Although there is a positive relationship between predicted and observed carbon values, they decrease in depth (Figure 5) and there is a tendency for very low values for depths greater than 0.30 m (Table 4). This is especially true for soils that begin with low levels of soil carbon, such as mineral soils. Particularly in the deeper layers, the low number of points to represent these layers affected the prediction (Tables 4).

The form of the model evaluation, external data or cross-validation, leads to different results by layer. For cross-validation, in which better results were obtained, the best performance was 0.05-0.15 m, for pH and C, and 0.60-1.00 m for CEC (Table 4).

On the spatial prediction and uncertainty propagation, when comparing the spatial prediction of the models, the cross-validation showed better results, so these models

Table 3. Descriptive statistics of predicted values for the whole profile using external validation and LOO-CV dataset, and for observed values (all data and validation data set)

Property	**Predicted values on external validation**					
	R^2	**RMSE**	**MSE**	**Min**	**Mean**	**Max**
pH	0.27	0.38	0.15	3.80	4.49	5.00
C	0.26	5.73	32.82	-2.98	10.53	22.58
CEC	0.42	10.75	115.54	6.61	17.67	31.52
	Predicted values on LOO-CV					
pH	0.45	0.29	0.09	3.42	4.51	5.14
C	0.60	3.63	13.20	-3.81	6.43	20.96
CEC	0.59	5.95	35.36	-0.03	13.65	46.95
	Observed values					
Property	All data			Data validation		
	Min	Mean	Max	Min	Mean	Max
pH	3.24	4.51	5.72	3.72	4.69	5.46
C	0.24	6.42	29.48	0.43	7.95	17.46
CEC	3.00	13.68	69.01	4.35	19.04	69.01

Table 4. Descriptive statistics of predicted values for each depth, with LOO-CV dataset

Property	**R^2**	**RMSE**	**MSE**	**Min**	**Mean**	**Max**	**Layer**	**n[(1)]**
							m	
pH	0.43	0.35	0.13	3.78	4.39	5.08	0.00-0.05	90
	0.46	0.33	0.11	3.81	4.41	5.00	0.05-0.15	90
	0.41	0.30	0.09	3.87	4.45	4.88	0.15-0.30	85
	0.32	0.27	0.07	4.00	4.50	4.88	0.30-0.60	73
	0.38	0.28	0.08	4.14	4.57	5.09	0.60-1.00	51
C	0.32	9.42	88.66	2.64	14.31	22.90	0.00-0.05	90
	0.35	9.50	90.33	2.02	13.38	21.93	0.05-0.15	90
	0.30	9.81	96.15	1.02	11.85	20.34	0.15-0.30	85
	0.27	9.91	98.14	-0.41	8.38	12.95	0.30-0.60	73
	0.28	9.03	81.49	-1.80	6.97	13.86	0.60-1.00	51
CEC	0.40	7.15	51.15	14.47	22.61	33.27	0.00-0.05	90
	0.41	7.15	51.17	13.62	21.59	32.08	0.05-0.15	90
	0.52	6.81	46.31	12.11	19.84	30.08	0.15-0.30	85
	0.58	6.29	39.59	9.01	16.60	26.62	0.30-0.60	73
	0.65	4.36	19.01	7.52	14.78	23.44	0.60-1.00	51

[(1)] Number of observations in each layer.

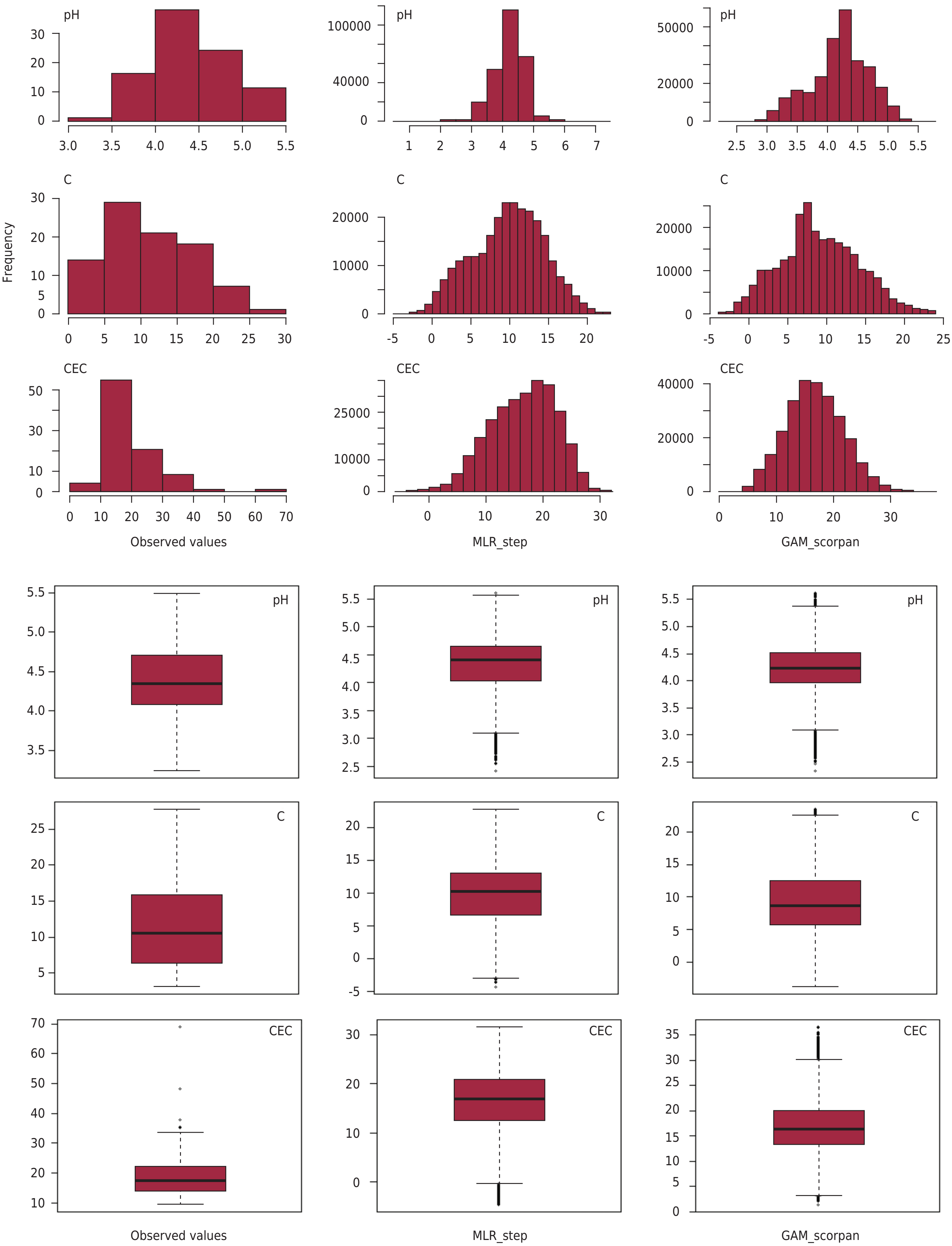


Figure 4. Descriptive statistics (histograms and boxplot) of the observed values and predicted values (grid) for soil properties using the best covariate selection approach for MLR and GAM models.

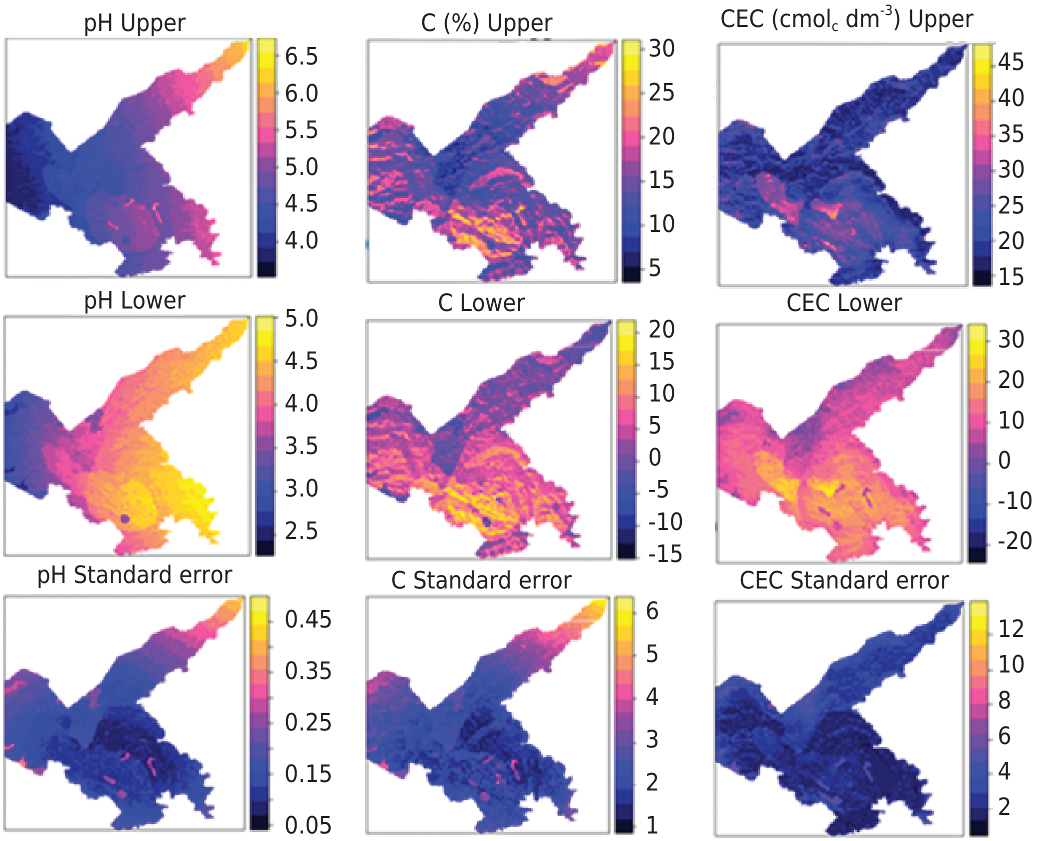


Figure 5. Standard error, lower and upper values derived from a Bayesian posterior distribution of each GAM_*scorpan* model fitted. C (%); CEC (cmol$_c$ dm^{-3})

were used to predict the soil properties on the grid and to produce the maps (Figure 7). Some extrapolated values for C and CEC are affected, among other factors, by the access limitations; consequently, we had a smaller number of points to calibrate the models and limited spatial distribution of the soil samples.

DISCUSSION

2-D approach

The performance gain in MLR models, already in the selection of covariates by correlation, excluding those with highly correlated (Figure 3), is probably due to the multicollinearity problem (Kempen et al., 2009; ten Caten et al., 2011), which significantly affects the model performance. The MLR with several covariates showed a tendency to have the worst performance because of its effect of harmful multicollinearity in the parametric models; thereby impairing the model. This leads to the problem of inflating the variance of parameters, model over-fitting, and even noise problems (Kempen et al., 2009; ten Caten et al., 2011; Nussbaum et al., 2018). This is especially important if we have a limited number of samples and a large number of covariates.

The MLR model presented regular performance, and it was worse than GAM (Table 2). In the RFE used in Random Forest, as it was done Jeong et al. (2017), the best results can be due to the parameters optimization using cross-validation inside the algorithm (using *caret* package), thus selecting an optimal value of the number of trees and *mtry* parameter. This pattern was not observed in the MLR using RFE. For linear models, the RFE selection method (using the RFE *lmFuncs* function on *caret* package) did not present good results; it was the model with the worst result, except for soil pH (Figure 5). Furthermore, it almost always selects the same covariates, regardless of the soil properties tested. This suggests that selection algorithm using the function (*rfe*) for linear models in the *caret* package should be used carefully. In general, the best way to select covariates for MLR is the stepwise selection with AIC criteria, a common method to select models in linear regression (Carvalho Junior et al., 2016; Chagas et al., 2016; Vermeulen and Niekerk, 2017).

The GAM_*scorpan* was the most appropriate model for the prediction of all soil properties. It presented the best performance, in both ways, when evaluated using external validation and with cross-validation. However, this was the only model where, in all soil properties, performance in the cross-validation was lower than in the external validation (Tables 3 and 4). This may be due to the fact that external validation samples are not probabilistic, as suggested by Brus et al. (2011), and do not include all geographic and attribute space. For these conditions, where the validation samples come from various sources and are not completely random, and there is a limited set of soil data, cross-validation is the best approach to evaluate the models.

For the three soil properties and selection methods evaluated in this work, the linear models showed inferior performance to GAM. This is probably because relationships between soil attributes and covariates are not linear, and models such as the MLR fail to capture the nonlinear relationships efficiently (Poggio et al., 2013; Jeong et al., 2017). Since, in general, soil properties do not have linear relationships with environmental covariates, the models that captured these relationships tend to be better. In contrast, GAM models, where it is possible to model nonlinear relationships (Jeong et al., 2017; Poggio and Gimona, 2017 a,b), had the best performance when combined a nonlinear modeling approach and the concept of soil formation factors for covariate selection (expert knowledge).

When analyzing the models separately according to their respective methods of covariates selection, it is observed that, in general, the best results use stepwise selection for MLR (except for predicting soil pH), and GAM using the *scorpan* approach. It is possible to separate the models and covariates selection approach. This appears to contradict Somarathna et al. (2017), who suggested investing in sampling (Figure 4) rather than more robust models. However, it agrees with Beguin et al. (2017), who tested different statistical approaches and found significant differences, thus suggesting that robust methods can enhance DSM capabilities and support existing efforts for improving digital soil products, even with limited data.

Pedological knowledge is crucial in DSM and was used by Nussbaum et al. (2018), to exclude covariates with low spatial variation and aggregate levels of categorical variables with low sample density per level. The knowledge of soil-forming factors as well as of the study area is a powerful tool, and when associated with computational tools, it may improve predictions of soil properties and classes.

When evaluating soil-forming factors and pedological elicitation, the elevation, parental material, and covariates from the RapidEye sensor (Table 1) were the factors that most influenced the soil properties of the INP plateau, because they were frequently selected by the different approaches of covariate selection for both models, MLR and GAM. All soil factors are related, for example, the relief has spatial variation, and the highest part is in the center of the study area; in turn, the elevation influences the weather, that is cold and wet in the INP, and that leads to a distinct distribution of plant species. This environment favors accumulation and preservation of soil organic matter due to low temperatures, leading to the formation of the organic soils in the high altitudes (Soares et al., 2016), which explains why the organic soils predominate in the upper part of the INP.

This agreed with the results found in the GAM_*scorpan* models, the models selected as the best since they combine the most important covariates, related to the parent material, relief, organism, and position in the geographic space. The combined influence of these factors is the main reason for soil formation in the upper part of the INP. For example, the higher carbon and CEC contents, properties strongly related to each other, were predicted with highest values in the areas of INP with altitudinal fields coverage (Figure 1) (predominant species are Poaceae and Cyperaceae), which are concentrated in the plateau central region of INP, where the dominant geology is composed by quartz-syenites and related sediments (Santos et al., 2000; Soares et al., 2016).

When the predicted values (in the grid) and the observed values were compared, it was observed a tendency of the MLR to extrapolate results, especially lower values being more negative; as in the carbon and CEC (Figure 5). This can be related to the limited access to regions located to the North and West of INP, consequently a small number of soil samples, with some covariates having high variations (DEM). Similar results were observed by Chagas et al. (2016) and Carvalho Junior et al. (2016), when comparing Random Forest and MLR, regression seems to extrapolate the values. Like MLR, GAM models tend to extrapolate the extreme values.

Despite the best performance of the GAM model (Table 2), evaluated by the R2, RMSE, and MSE metrics, there was an extrapolation of values predicted in the grid. This reinforces the importance of evaluating the spatial propagation of uncertainty in DSMs (Stumpf et al., 2016; Vaysse and Lagacherie, 2017). Besides some geology classes occurred in small areas; consequently they had fewer soil samples. A similar pattern was observed by Cambule et al. (2014), predicting carbon stocks in the Limpopo National Park, where they observed high uncertainty values, and it was suggested that it was due to short-range spatial structure combined with the sparse sampling.

3-D approach

Due to limited accessibility conditions and hence the reduced number of points, the best approach to fit and validate GAM models are smoothing functions using cross-validation. Also, because when there are many covariates and few points, it is not possible to fit GAM models due to the limitation on the degrees of freedom of the model (Poggio et al., 2013). In this case, with limited ground points, the cross-validation seems to be more appropriate to fit the 3-D function (Taghizadeh-Mehrjardi et al., 2016; Amirian Chakan et al., 2017). Even if more data are available, this is a common approach in recent 3-D soil properties modeling (Veronesi et al., 2014; Mulder et al., 2016). As there are more samples for the topsoil, in general, this layer presented better prediction results as observed by Kempen et al. (2011), Mulder et al. (2016), and Poggio and Gimona (2017a) consequently, larger errors prediction and greater uncertainty are observed in the deepest depths (Figures 6 and 7). In this sense, LOO-CV is the best approach to obtain the results given that in depth the number of samples is even more limited.

The evaluation of spatial uncertainty propagation showed a greater uncertainty for the higher values given by the standard error, due to noise in the very steep areas (above 100 %) and/or shadowed regions detected in the satellite images. Also, areas that had greater uncertainty occurred when the predicted values are associated with higher access limitation, with little or no soil samples. As mentioned, in some areas of INP, the access is very limited, with unused/semi-closed tracks, or no tracks at all, for the North and West directions. In addition, no vehicles, except during an emergency, are allowed to travel inside the park. These reasons led to a low number or no soil sampling at all in some areas of INP (Figure 1). When the profile depth is considered, due to the dominance of shallow soils, the number of soil samples decreases, in the distance and with depth. Some covariates such as DEM also varied significantly (North and West). This contributes to the model's higher extrapolations and uncertainty.

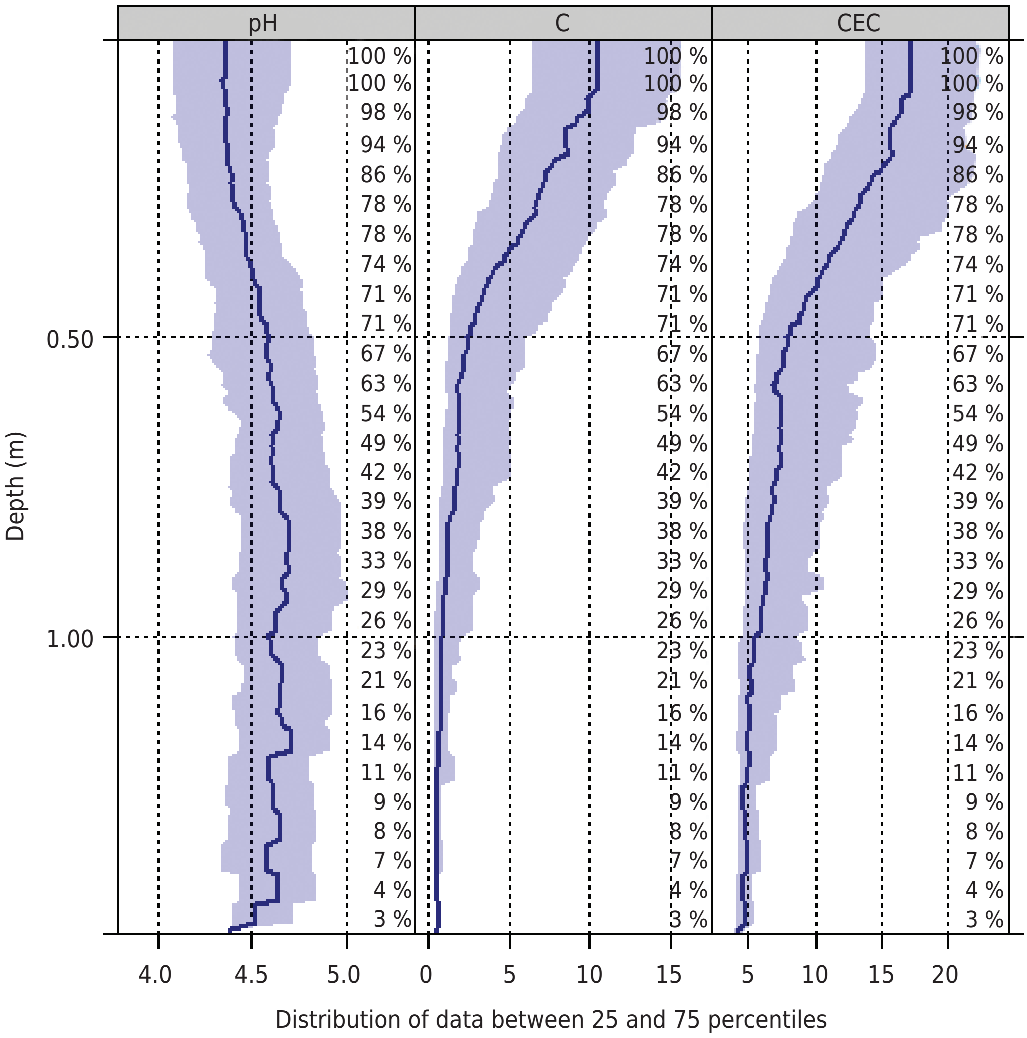


Figure 6. Distribution of pH, carbon content (%), and CEC (cmol$_c$ dm^{-3}) for the data collection. The percentage values represent the relative number of profiles that contributed to the estimates in each layer.

In terms of soil genesis, due to the very steep slopes and the predominantly acidic nature of the parent material, the dominant soils in the INP are the Regosols and Cambic Umbrisols. Under the forest coverage, but with higher slopes, some intermediate developed soils such as Cambisols and Folic Umbrisols occur, with the Regosols occupying the strongly sloping and steep areas. In the higher elevation areas, with altitude fields vegetation and rock exposures, shallow soils such as Histosols and Leptosols with a histic horizon are dominant. In the lower elevations of INP, under forest coverage and located in flat and gently slopes, more weathered and deeper soils such as Rhodic Acrisols and Ferrasols predominate.

In summary, the INP plateau is relatively complex in terms of soil variation and their attributes and most soil properties predictions were produced with admissible modeling diagnostics and uncertainty ranges for LOO-CV (Table 4), when the limitation due to the number of soil samples and their spatial distribution is taken in account. For the remote areas of the INP plateau, the models tend to extrapolate the results for soil carbon content and CEC in deeper layers, especially the MLR in relation to GAM. The same results were observed by Kidd et al. (2015), suggesting that maps should be created with continuous improvements, from the input of newly collected data. Prediction uncertainty can help to choose supplemental sampling to improve the DSM (Li et al., 2016).

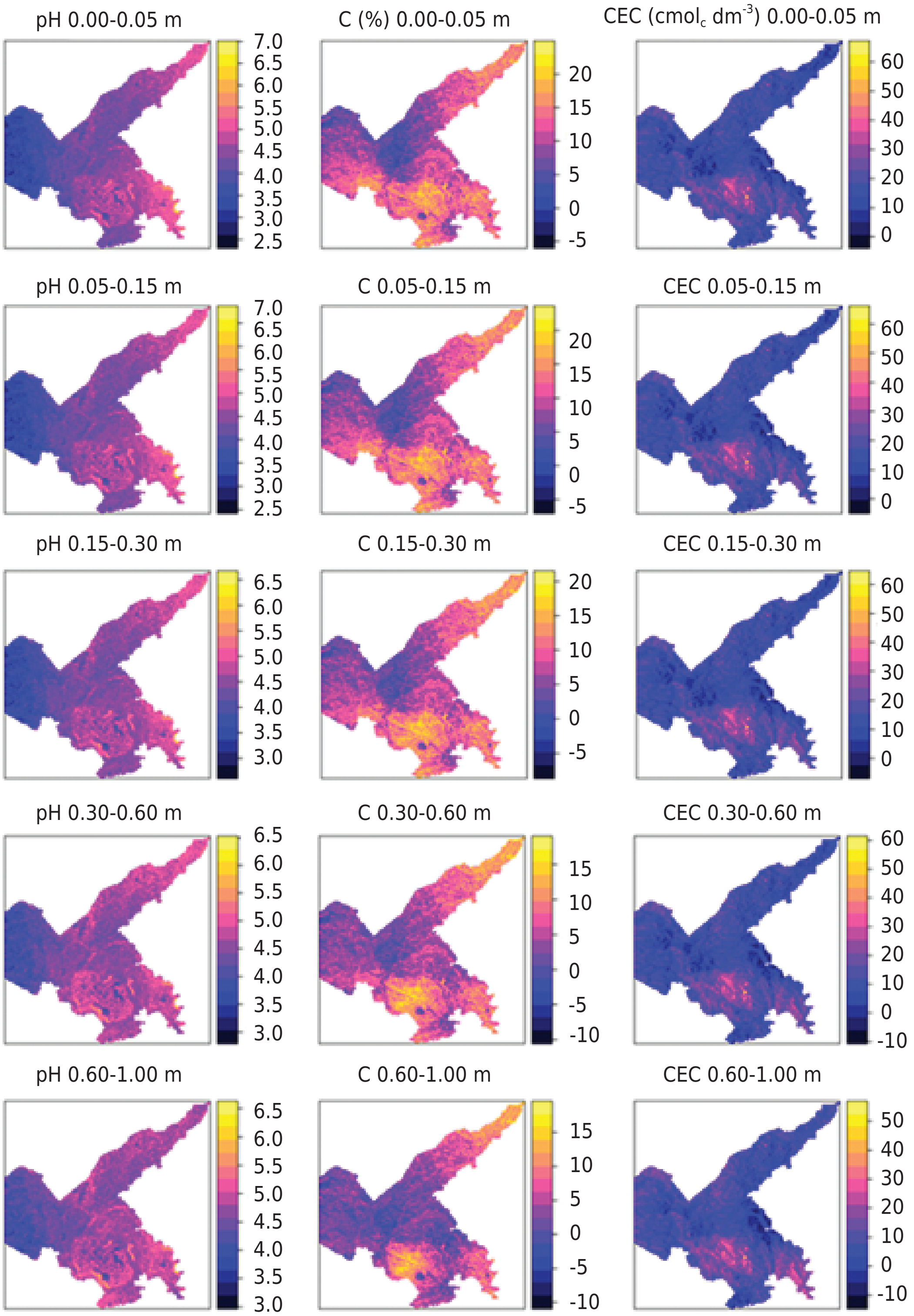


Figure 7. Maps of the soil properties at five layerss (pH left, carbon content (%), CEC (cmol$_c$ dm^{-3}) right). They were predicted with models evaluated with cross-validation. The five layers are 0.00-0.05, 0.05-0.15, 0.15-0.30, 0.30-0.60 and 0.60-1.00 m.

CONCLUSIONS

In general, the GAM model had superior performance MLR. The approach based on soil-forming factors showed to be a simple and viable method for covariates selection in the GAM model, especially considering limitations regarding degrees of freedom due to the limited number of soil samples.

The elevation, parental material, and covariates from the RapidEye sensor were the factors that most influenced the soil properties of the Itatiaia National Park plateau.

The greater uncertainty of the maps was associated with the low accessibility areas, which had low sampling density and/or noises in the covariates. The 2- and 3-D soil properties modeling with the correspondent uncertainty propagation can be used for INP management of ecosystems.

The high resolution of soil attributes and uncertainties produced for INP in the 3-D space are an important step in developing a comprehensive soil database, carrying quantitative soil-information on a scale adequate to the INP demands.

The sampling planning using cLHS provided a convenient subset of the area representation, increasing the possible combination of ranges from the pool of covariates. This approach enhances the potential of the best model (GAM scorpan) to produce maps, by accounting with the uncertainty, confining sampling points to the absolutely necessary, especially in areas with limited access such as the INP. This strategy contributes to minimize costs when balancing challenges and sampling requirements.

ACKNOWLEDGMENTS

We are grateful to INP administrators for providing access and support to data collection as well as the cartographic base, and the training and laboratory funding from Graduate Course in Agronomy – Soil Science, UFRRJ, Brazil. Thanks to The James Hutton Institute, Aberdeen, Scotland, for providing physical structure and specialized knowledge that aided the research, particularly the supervision of Dr. Laura Poggio. The work was supported by scholarships from the Ministry of Science, Technology, Innovations and Communications through CNPq (Process 141391/2015-4), the Ministry of Education through CAPES Foundation (Process 88881.135776/2016-01), and grants from FAPERJ Foundation (Process E_34/2014).

AUTHOR CONTRIBUTIONS

Conceptualization: Elias Mendes Costa (lead), Helena Saraiva Koenow Pinheiro (supporting), and Lúcia Helena Cunha dos Anjos (supporting).

Formal analysis: Elias Mendes Costa (lead), Robson Altiellys Tosta Marcondes (supporting), and Yuri Andrei Gelsleichter (supporting).

Methodology: Elias Mendes Costa (lead) and Helena Saraiva Koenow Pinheiro (supporting).

Software: Elias Mendes Costa (lead) and Yuri Andrei Gelsleichter (supporting).

Writing – original draft: Elias Mendes Costa (lead).

Writing – review & editing: Elias Mendes Costa (equal), Helena Saraiva Koenow Pinheiro (equal), Lúcia Helena Cunha dos Anjos (equal), and Yuri Andrei Gelsleichter (equal).

Supervision: Helena Saraiva Koenow Pinheiro (supporting) and Lúcia Helena Cunha dos Anjos(lead).

REFERENCES

Adhikari K, Hartemink AE. Linking soils to ecosystem services - a global review. Geoderma. 2016;262:101-11. https://doi.org/10.1016/j.geoderma.2015.08.009

Amirian Chakan A, Taghizadeh-Mehrjardi R, Kerry R, Kumar S, Khordehbin S, Yusefi Khanghah S. Spatial 3D distribution of soil organic carbon under different land use types. Environ Monit Assess. 2017;189:131. https://doi.org/10.1007/s10661-017-5830-9

Arrouays D, Grundy MG, Hartemink AE, Hempel JW, Heuvelink GBM, Hong SY, Lagacherie P, Lelyk G, McBratney AB, McKenzie NJ, Mendonca-Santos ML, Minasny B, Montanarella L, Odeh IOA, Sanchez PA, Thompson JA, Zhang G-L. GlobalSoilMap: toward a fine-resolution global grid of soil properties. In: Sparks DL, editor. Advances in agronomy. Newark: Academic Press; 2014. v. 157. p. 93-134.

Barreto CG, Campos JB, Roberto DM, Roberto DM, Schwarzstei NT, Alves GSG, Coelho W. Encarte 3: Análise da unidade de conservação. In: Plano de manejo: Parque Nacional do Itatiaia. Brasília: Instituto Chico Mendes de Conservação da Biodiversidade; 2013. Available from: http://www.icmbio.gov.br/portal/images/stories/docs-planos-de-manejo/pm_parna_itatiaia_enc3.pdf

Beguin J, Fuglstad G-A, Mansuy N, Paré D. Predicting soil properties in the Canadian boreal forest with limited data: Comparison of spatial and non-spatial statistical approaches. Geoderma. 2017;306:195-205. https://doi.org/10.1016/j.geoderma.2017.06.016

Brus DJ, Kempen B, Heuvelink GBM. Sampling for validation of digital soil maps. Eur J Soil Sci. 2011;62:394-407. https://doi.org/10.1111/j.1365-2389.2011.01364.x

Cambule AH, Rossiter DG, Stoorvogel JJ. A methodology for digital soil mapping in poorly-accessible areas. Geoderma. 2013;192:341-53. https://doi.org/10.1016/j.geoderma.2012.08.020

Cambule AH, Rossiter DG, Stoorvogel JJ, Smaling EMA. Soil organic carbon stocks in the Limpopo National Park, Mozambique: Amount, spatial distribution and uncertainty. Geoderma. 2014;213:46-56. https://doi.org/10.1016/j.geoderma.2013.07.015

Carvalho Junior W, Calderano Filho B, Chagas CS, Bhering SB, Pereira NR, Pinheiro HSK. Regressão linear múltipla e modelo Random Forest para estimar a densidade do solo em áreas montanhosas. Pesq Agropec Bras. 2016;51:1428-37. https://doi.org/10.1590/S0100-204X2016000900041

Carvalho Júnior W, Chagas CS, Muselli A, Pinheiro HSK, Pereira NR, Bhering SB. Método do hipercubo latino condicionado para a amostragem de solos na presença de covariáveis ambientais visando o mapeamento digital de solos. Rev Bras Cienc Solo. 2014;38:386-96. https://doi.org/10.1590/S0100-06832014000200003

Chagas CS, Carvalho Junior W, Bhering SB, Calderano Filho B. Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions. Catena. 2016;139:232-40. https://doi.org/10.1016/j.catena.2016.01.001

Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference and prediction. 2nd ed. New York: Springer-Verlag; 2009.

Instituto Brasileiro de Geografia e Estatística - IBGE. Manual técnico de pedologia. 3. ed. Rio de Janeiro: IBGE; 2015. (Manuais técnicos em geociências, n. 4). Available from: https://biblioteca.ibge.gov.br/visualizacao/livros/liv95017.pdf

Jeong G, Oeverdieck H, Park SJ, Huwe B, Ließ M. Spatial soil nutrients prediction using three supervised learning methods for assessment of land potentials in complex terrain. Catena. 2017;154:73-84. https://doi.org/10.1016/j.catena.2017.02.006

Kempen B, Brus DJ, Heuvelink GBM, Stoorvogel JJ. Updating the 1:50,000 Dutch soil map using legacy soil data: a multinomial logistic regression approach. Geoderma. 2009;151:311-26. https://doi.org/10.1016/j.geoderma.2009.04.023

Kempen B, Brus DJ, Stoorvogel JJ. Three-dimensional mapping of soil organic matter content using soil type-specific depth functions. Geoderma. 2011;162:107-23. https://doi.org/10.1016/j.geoderma.2011.01.010

Kidd D, Webb M, Malone B, Minasny B, McBratney A. Eighty-metre resolution 3D soil-attribute maps for Tasmania, Australia. Soil Res. 2015;53:932-55. https://doi.org/10.1071/SR14268

Kuhn M. caret: classification and regression training. R package version 6.0.84; 2017. Available from: http://topepo.github.io/caret/index.html

Li Y, Zhu A-X, Shi Z, Liu J, Du F. Supplemental sampling for digital soil mapping based on prediction uncertainty from both the feature domain and the spatial domain. Geoderma. 2016;284:73-84. https://doi.org/10.1016/j.geoderma.2016.08.013

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

Menezes MD, Silva SHG, Mello CR, Owens PR, Curi N. Knowledge-based digital soil mapping for predicting soil properties in two representative watersheds. Sci Agric. 2018;75:144-53. https://doi.org/10.1590/1678-992x-2016-0097

Menezes MD, Silva SHG, Mello CR, Owens PR, Curi N. Solum depth spatial prediction comparing conventional with knowledge-based digital soil mapping approaches. Sci Agric. 2014;71:316-23. https://doi.org/10.1590/0103-9016-2013-0416

Minasny B, McBratney AB. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Comput Geosci. 2006;32:1378-88. https://doi.org/10.1016/j.cageo.2005.12.009

Mulder VL, Lacoste M, Richer-de-Forges AC, Martin MP, Arrouays D. National versus global modelling the 3D distribution of soil organic carbon in mainland France. Geoderma. 2016;263:16-34. https://doi.org/10.1016/j.geoderma.2015.08.035

Nussbaum M, Spiess K, Baltensweiler A, Grob U, Keller A, Greiner L, Schaepman ME, Papritz AJ. Evaluation of digital soil mapping approaches with large sets of environmental covariates. Soil. 2018;4:1-22. https://doi.org/10.3929/ethz-b-000228435

Nussbaum M, Walthert L, Fraefel M, Greiner L, Papritz A. Mapping of soil properties at high resolution in Switzerland using boosted geoadditive models. Soil Discuss. 2017;53:1-32. https://doi.org/10.5194/soil-2017-13

Poggio L, Gimona A. 3D mapping of soil texture in Scotland. Geoderma Reg. 2017a;9:5-16. https://doi.org/10.1016/j.geodrs.2016.11.003

Poggio L, Gimona A. Assimilation of optical and radar remote sensing data in 3D mapping of soil properties over large areas. Sci Total Environ. 2017b;579:1094-110. https://doi.org/10.1016/j.scitotenv.2016.11.078

Poggio L, Gimona A, Brewer MJ. Regional scale mapping of soil properties and their uncertainty with a large number of satellite-derived covariates. Geoderma. 2013;209-210:1-14. https://doi.org/10.1016/j.geoderma.2013.05.029

R Development Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2018. Available at: http://www.R-project.org.

RapidEye. RapidEye^{TM} Mosaic product specifications. Campinas: Embrapa Monitoramento por Satélite; 2011.

Roudier P, Hewitt AE, Beaudette DE. A conditioned Latin hypercube sampling algorithm incorporating operational constraints. In: Minasny B, Malone BP, McBratney AB. Proceedings of the 5th Global Workshop on Digital Soil Mapping. Sydney: Digital soil assessments and beyond; 2012. p. 227-31.

Ruppert D, Wand MP, Carroll RJ. Semiparametric regression. Cambridge: Cambridge University Press; 2003.

Samuel-Rosa A, Heuvelink GBM, Vasques GM, Anjos LHC. Do more detailed environmental covariates deliver more accurate soil maps? Geoderma. 2015;243-244:214-27. https://doi.org/10.1016/j.geoderma.2014.12.017

Santos RF, Pires Neto AG, Csordas SM. O Parque Nacional do Itatiaia. Fundação Bras. para o Desenvolv. Sustentável 2000;1:9-19.

Sindayihebura A, Ottoy S, Dondeyne S, Van Meirvenne M, Van Orshoven J. Comparing digital soil mapping techniques for organic carbon and clay content: Case study in Burundi's central plateaus. Catena. 2017;156:161-75. https://doi.org/10.1016/j.catena.2017.04.003

Soares PFC, Anjos LHC, Pereira MG, Pessenda LCR. Histosols in an Upper Montane Environment in the Itatiaia Plateau. Rev Bras Cienc Solo. 2016;40:e0160176. https://doi.org/10.1590/18069657rbcs20160176

Somarathna PDSN, Minasny B, Malone BP. More data or a better model? Figuring out what matters most for the spatial prediction of soil carbon. Soil Sci Soc Am J. 2017;81:1413-26. https://doi.org/10.2136/sssaj2016.11.0376

Stumpf F, Schmidt K, Behrens T, Schönbrodt-Stitt S, Buzzo G, Dumperth C, Wadoux A, Xiang W, Scholten T. Incorporating limited field operability and legacy soil samples in a hypercube sampling design for digital soil mapping. J Plant Nutr Soil Sci. 2016;179:499-509. https://doi.org/10.1002/jpln.201500313

Taghizadeh-Mehrjardi R, Nabiollahi K, Kerry R. Digital mapping of soil organic carbon at multiple depths using different data mining techniques in Baneh region, Iran. Geoderma. 2016;266:98-110. https://doi.org/10.1016/j.geoderma.2015.12.003

ten Caten A, Dalmolin RSD, Pedron FA, Mendonça-Santos ML. Estatística multivariada aplicada à diminuição do número de preditores no mapeamento digital do solo. Pesq Agropec Bras. 2011;46:554-62. https://doi.org/10.1590/S0100-204X2011000500014

Tomzhinski GW, Ribeiro KT, Fernandes MC. Análise geoecológica dos incêndios florestais do Parque Nacional do Itatiaia. Boletim de Pesquisa do Parque Nacional do Itatiaia. 2012;15:1-158.

Vašát R, Kodešová R, Borůvka L, Jakšík O, Klement A, Brodský L. Combining reflectance spectroscopy and the digital elevation model for soil oxidizable carbon estimation. Geoderma. 2017;303:133-42. https://doi.org/10.1016/j.geoderma.2017.05.018

Vaysse K, Lagacherie P. Using quantile regression forest to estimate uncertainty of digital soil mapping products. Geoderma. 2017;291:55-64. https://doi.org/10.1016/j.geoderma.2016.12.017

Vermeulen D, Niekerk AV. Machine learning performance for predicting soil salinity using different combinations of geomorphometric covariates. Geoderma. 2017;299:1-12. https://doi.org/10.1016/j.geoderma.2017.03.013

Vermote EF, Herman M, Morcrette J. Second simulation of the satellite signal in the solar spectrum, 6S: an overview. IEEE T Geosci Remote. 1997;35:675-86. https://doi.org/10.1109/36.581987

Veronesi F, Corstanje R, Mayr T. Landscape scale estimation of soil carbon stock using 3D modelling. Sci Total Environ. 2014;487:578-86. https://doi.org/10.1016/j.scitotenv.2014.02.061

Wood SN. Generalized additive models: an introduction with R. 2nd ed. Boca Raton: CRC Press; 2017.

Zhang G-l, Liu F, Song X-d. Recent progress and future prospect of digital soil mapping: a review. J Integr Agr. 2017;16:2871-85. https://doi.org/10.1016/S2095-3119(17)61762-3