Revista Brasileira de Ciência do Solo

Division - Soil in Space and Time | Commission - Pedometrics

Optimized data-driven pipeline for digital mapping of quantitative and categorical properties of soils in Colombia

Alejandro Coca-Castro[(1)], **Joan Sebastián Gutierrez-Díaz**[(2)], **Victoria Camacho**[(1)], **Andrés Felipe López**[(1)], **Patricia Escudero**[(1)], **Pedro Karin Serrato**[(1)], **Yesenia Vargas**[(1)], **Ricardo Devia**[(2)], **Juan Camilo García**[(2)], **Carlos Franco**[(1)*] **and Janeth González**[(2)]

[(1)] Instituto Geográfico Agustín Codazzi, Centro de Investigación y Desarrollo en Información Geográfica, Bogotá, Cundinamarca, Colombia.
[(2)] Instituto Geográfico Agustín Codazzi, Sub-Directorate of Agrology, Bogotá, Cundinamarca, Colombia.

ABSTRACT: Soil maps provide a method for graphically communicating what is known about the spatial distribution of soil properties in nature. We proposed an optimized pipeline, named dino-soil toolbox, programmed in the R software for mapping quantitative and categorical properties of legacy soil data. The pipeline, composed of four main modules (data preprocessing, covariates selection, exploratory data analysis and modeling), was tested across a study area of 14,537 km^2 located between the departments of Cesar and Magdalena, Colombia. We assessed the feasibility of the toolbox to model three soil properties: pH at two depth intervals (0.00-0.30 and 0.30-1.00 m), soil taxonomy (great group) and taxonomic family by particle-size, according to a set of 25 environmental factors derived from auxiliary layers of climate, land cover and terrain. As a result, we successfully deployed the proposed semi-automatic and sequential pipeline, yielding rapid digital soil mapping (DSM) outputs across the study area. By providing multiple outputs such as tables, charts, maps, and geospatial data in four main modules, the pipeline offers considerable robustness to support outcomes and analysis of a DSM project. Future studies might be interesting to expand on further machine learning frameworks for predictive modeling of soil properties such as ensembles and deep learning models, which have shown a high performance for DSM.

Keywords: soil prediction, soil databases, machine learning, uncertainty, toolbox.

*** Corresponding author:**
E-mail: carlos.franco@igac.gov.co

Received: July 24, 2021

Approved: October 05, 2021

How to cite: Coca-Castro A, Gutierrez-Díaz JS, Camacho V, López AF, Escudero P, Serrato PK, Vargas Y, Devia R, García JC, Franco C, González J. Optimized data-driven pipeline for digital mapping of quantitative and categorical properties of soils in Colombia. Rev Bras Cienc Solo. 2021;45:e0210084.
https://doi.org/10.36783/18069657rbcs20210084

Editors: José Miguel Reichert and José Coelho de Araújo Filho.

https://doi.org/10.36783/18069657rbcs20210084

INTRODUCTION

Soil classification is a method for organizing and communicating knowledge and perceptions about soil properties. Soil maps provide a method for graphically communicating what is known about the soil properties spatial distribution in nature. Among the different approaches to generate those maps, there is a wide adoption of the digital soil mapping (DSM) framework. Incepted by McBratney et al. (2003), the framework has been widely applied to produce and analyze spatio-temporal patterns of soil properties according to environmental covariates such as climate, terrain, vegetation and land use.

The existing methods in DSM can be grouped into two main modeling types, conventional (statistical and geostatistical) and machine learning (ML). In the former type, a soil property is modeled as a linear relationship between the property and state factors, accounting for the deterministic portion of the total variation, and a spatially dependent stochastic portion by using kriging methods (Keskin and Grunwald, 2018). While scholars have proposed multiple geostatistical for soil mapping, they are computationally demanding if the sample size and/or the number of prediction locations are large (Cressie and Johannesson, 2008). Moreover, modeling non-linear relationships between soil properties and numerous cross-correlated environmental covariates are not straightforward and introduces additional challenges, e.g., estimating many parameters (Wadoux et al., 2020a). Unlike geostatistical methods in which the transformation of the original observations is often required to satisfy assumptions, ML algorithms do not assume the observations' distribution. In addition, they are more suitable for large area predictions and designed to handle non-linear relations and complexity found in soil data (Padarian et al., 2020).

Besides the contribution to soil research, ML-based methods have accelerated the production of first versions of digital soil maps at the national (Odgers et al., 2012; Akpa et al., 2014; Mulder et al., 2016; Padarian et al., 2017), continental (Hengl et al., 2015; Ballabio et al., 2016) and global (Hengl et al., 2017) extent. These maps are aligned to global efforts in storing, coding and harmonizing legacy soil data, especially soil profiles with soil analytical data (Arrouays et al., 2017). In Colombia, the Geographic Institute Agustin Codazzi (IGAC), through the Sub-Directorate of Agrology leads the inventory, study, analysis and monitoring of the country's soils for their management to support national land planning. As part of the Global Soil Partnership, IGAC has recently generated the first version of the national soil organic carbon map (Bolívar Gamboa et al., 2021) according to FAO's cookbook (FAO, 2018). Agrosavia recently launched the IRAKA platform, the first Colombian soil information system providing spatial data of multiple soil properties at the surface layer, from 0.00 to 0.20 m depth (so-called topsoil) across the Cundiboyacense high plateau (Araujo-Carrillo et al., 2021). This development roots from the concept of hybrid soil information system (SIS) by using both the techniques and data of traditional soil studies and DSM models and information technologies such as database management, ML-assisted modeling, and geographic web services.

To contribute to the optimization of SIS's for generating soil-related information in digital formats, this study aimed to test the hypothesis that a semi-automatic and sequential pipeline, compiling a variety of statistical and ML algorithms commonly used in soil mapping research, can offer robustness and rapid experimentation required for a DSM project. This involves the deployment of a pipeline in the software R, named dinoSoil-toolbox, composed of four sequential modules (data preprocessing, covariates selection, exploratory data analysis and modeling) across a large heterogeneous geographic area. To achieve this objective, we 1) tested the proposed toolbox for a rapid generation of spatial data of three soil properties: pH at two soil layers (0.00-0.30 m and 0.30-1.00 m), soil taxonomy (great group) and taxonomic family by particle-size from soil databases gathered and curated by IGAC in a mountainous valley terrain of the Momposine depression located between the departments of Cesar and Magdalena, Colombia; 2) informed the multiple formats (charts, tabular and georeferenced layers) of soil-related information outputs provided by the toolbox for their inspection and analysis; and 3) made a shareable and reproducible toolbox following open science principles with readable documentation, sample data, and code available in a public GitHub repository.

MATERIALS AND METHODS

Study area

The area of interest (AOI) encompasses a surface of 14,537 km^2 located between the departments of Cesar and Magdalena, Colombia, between 8° 56’ 11” and 10° 52’ 4” north latitude and 73° 4’ 56” and -73° 57’ 39” west longitude (Figure 1). This area is part of the research program of land policy of IGAC’s Sub-Directorate of Agrology. It has been subject to multiple soil surveys, most of them compiled by this program. With elevation values between -31 and 5364 m and an average air temperature of 18 °C, the dominant landscapes are mountains and floodplains.

Input data

Soil surveys

IGAC’s Sub-Directorate of Agrology provided field and laboratory observational soil databases (so-called legacy soil data). Both datasets describe soil properties at a different level of detail but at the same unit of observation (profile). The former database contains 1858 profiles and refers to soil morphological and taxonomical data from soil-survey campaigns conducted in 2020. These campaigns were conducted for defining soil cartographic units across the study area. After defining these units, more detailed soil analytical data were collected and analyzed through laboratory methods. As for laboratory database, this one contains information on pH(H_2O) (1:2.5), soil texture (sand, silt and clay percentage, Bouyoucos method), organic carbon (%, Walkley-Black chromic acid wet oxidation method), and bulk density (Mg m^{-3}, core method). Along with these databases, an additional database containing the location of soil profiles was also provided.

Auxiliary data

A total of 25 environmental covariates were constructed and harmonized from auxiliary layers provided by IGAC’s Sub-Directorate of Agrology and external sources representing

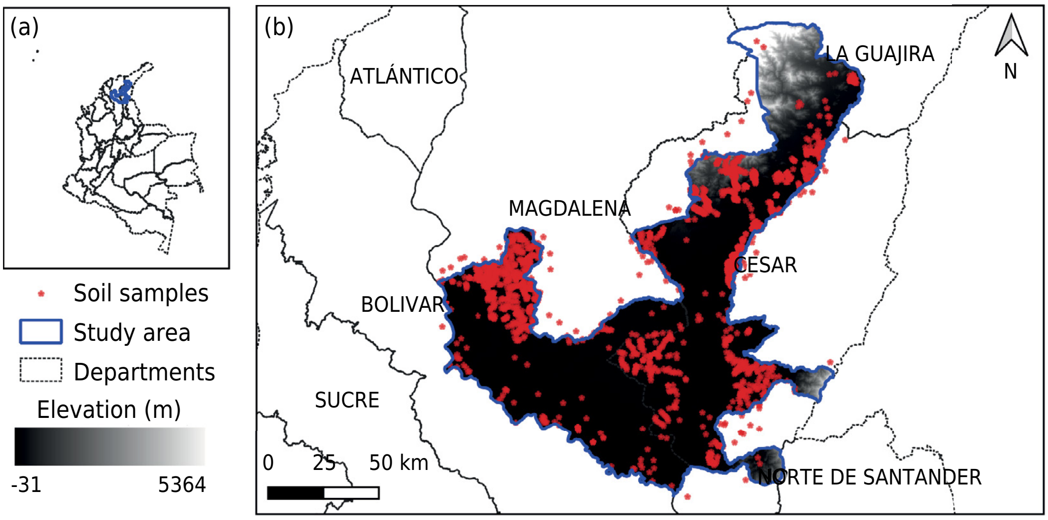

Figure 1. Location of the study area at national (a) and local (b) geographical extents. The internal divisions refer to administrative boundaries (Departments).

three soil forming factors: land cover, geomorphology, and climate (Table 1). Most of these covariates, 21 out of 25, were derived from a digital elevation model (DEM), which also defined a mapping resolution of 250 m. These multiple terrain parameters were used to represent the role of geomorphology in the spatial variability of soil properties and classes. A multiband raster stack containing all environmental information was generated, resampled, and harmonized on a 250 × 250 m pixel size grid over the AOI. Note a more detailed study using the proposed pipeline was conducted over the same study area but due to restricted access of the DEM used of 30 m, only the results with the open-access DEM are reported. While the community in general has access to 30 m DEM data, they often require post-processing according to the amount of missing data, i.e., holes in the study area. For this investigation, the restricted 30 m version was void-filled by DEM post-processing experts.

Proposed pipeline

The proposed semi-automatic pipeline comprises four main modules: data preprocessing, covariates selection, exploratory data analysis and modeling. These modules were built upon previous developments (code) provided by IGAC's Sub-Directorate of Agrology and adjusted according to the input data described above. Making decision processes are needed at some points along the workflow, mainly in the data preprocessing and covariates selection; therefore, we brought statistical criteria that, in conjunction with expert knowledge, support these decisions.

Data preprocessing

This step involves constructing data matrices containing soil profiles with each target property and related environmental layers. This procedure was conducted by extracting, i.e., intersecting the covariate values of each sample point. Then, each matrix is randomly divided using 70 % for model training and 30 % for test purposes (Guevara et al., 2018).

Before the above data extraction and partition, some data preparation procedures were conducted according to the type of the target property. An additional data preparation step is required for quantitative properties, such as pH, before the extraction procedure. This preparation refers to interpolate values at user-defined depths. It mainly requires reshaping field and/or laboratory databases from horizontal to long format. For the study area, pH was interpolated at two layers, 0.00-0.30 and 0.30-1.00 m. Interpolation is a common procedure in DSM to make the depth of the input soil dataset uniform. In this regard, a quadratic function of depth with equal areas (splines) method (Bishop et al., 1999) was implemented as part of the tool and used to interpolate pH at the target depths. For great group and family by particle-size, only those classes with at least five observations were retained following FAO (2018)´s good practices in DSM.

Table 1. Auxiliary layers for modeling soil properties in the study area

Forming factor	Number of covariates	Description	Source
Land cover	1	Median values of the Normalized Difference Vegetation Index (NDVI) from January 2019 to August 2020, 30 m resampled to 250 m	Landsat 8 Collection 1 Surface Reflectance acquired from the Google Earth Engine platform (Gorelick et al., 2017)
Topography	23	Digital elevation model (derived layers), 250 m Physiographic landscape and topography, vector	SRTM (CGIAR-CSI, 2018) Soil map scale 1:100.000 (IGAC, 2012)
Weather	1	Climate zones according to Holdridge classification, vector	IDEAM (2015)

After the above data preparation, two final procedures are considered for generating the input matrices per target property for modeling. The first procedure refers to automatically removing covariates with zero or close to zero variance. The second consists of detecting extreme values, i.e., outliers (values below the 5th percentile, or above the 95th percentile) per covariate. Users can handle observed outliers by different methods such as removal or interpolation with mean or median. For the input data of this study, we use the removal option.

Covariates selection

Recursive feature elimination with Random Forest (RFE-RF) was used to identify the optimal subset of covariates useful for prediction. After removing covariates with zero or near-zero variance, this algorithm iteratively eliminates the least remaining promising predictors based on an initial measurement of variable importance in a reference model, in this case, calibrated using the Random Forest algorithm (Guyon et al., 2002). The RFE-RF was set with a 5–fold repeated cross-validation, and variable importance was assessed by the mean decrease of Gini impurity (Guevara et al., 2018).

Exploratory data analysis

Conducted over the training input data matrices, these analyses are complementary since they determine possible relationships, interactions, or dependencies between the covariates. For the descriptive analysis, measures of centralization and dispersion are reported. Regarding the statistical analysis, normality tests and non-parametric tests are carried out. According to a consensus of the amount of legacy soil data for DSM (n >50), the tests of Lilliefors and Jarque Bera are suitable to assess the assumptions of normality in quantitative variables. The normality test should be ignored when using ML algorithms to handle non-linear relationships, e.g., random forest, support vector machine. Kruskal-Wallis (for multiple comparisons) or Wilcoxon-Mann–Whitney U (for single comparison) with the Bonferroni correction tests were used to statistically determine differences in the distribution of each covariate among the target soil properties.

Modeling

Model training: five algorithms were selected from different classes to model the target soil properties. Table 2 shows the list of ML algorithms considered with their family, the name stated in the R library, and the variables modeled. Only Random Forest, Extreme Gradient Boosting and Generalized Linear Models predict both quantitative and categorical variables. The support vector machine with radial kernel and cubist algorithms are specific to model pH as multilayer perception and multinomial logistic regression are categorical variables.

Table 2. List of ML models, family, corresponding name in the R library, and variables modeled

MODEL	CLASS	Name in R	Variable		
			pH	Great group	Texture classes
Random Forest	Decision trees	*ranger*	X	X	X
Extreme Gradient Boosting	Gradient Boosting	*xgbTree*	X	X	X
Generalized linear models	Linear models	*glmnet*	X	X	X
Support vector machines with linear kernel	Support vector machines	*svmRadial*	X		
Cubist	Cubist	*cubist*	X		
Multilayer Perceptron	Neural network	*mlp*		X	X
Multinomial logistic regression	Regression	*multinom*		X	X

Model evaluation: For selecting the best quantitative and categorical variables model, 10-fold repeated cross-validation was used from the training partition (Hengl et al., 2017). It is worth mentioning, each of these models is calibrated using a random grid search with a size of 20 for tuning its main parameters. The value of grid search might change according to the user's knowledge in the study area or modeled soil property. A larger number implies longer model training times, in particular for ML algorithms such as SVM. It is then suggested to start with a reasonable number as we set in this research. Once the best model with its calibrated parameters is identified per target variable, the external performance is measured using the test dataset. The main metrics for assessing model performance in both stages were the root mean square error (RMSE) and coefficient of determination (R^2) for pH, and overall accuracy (OA) and kappa coefficient for the categorical variables. In addition to these metrics, other complementary ones were considered for the external validation according to the variable type. For quantitative variables, the metrics included the R-square (R2), coefficient of efficiency (COE), index of agreement (IOA), amount of variance explained (AVE) as reported by Araujo-Carrillo et al. (2021). For the categorical variables, the F1-score was incorporated in the assessment.

Model uncertainty: An essential aspect of DSM involves the measurement of the uncertainties of the ML-based predictions. The methods are different according to the variable type. We propose to derive estimates of uncertainty for quantitative variables by fitting a quantile regression forest model (QRF). As Vaysse and Lagacherie (2017) suggested, this method allows interpolating the response of the best model's residuals for each unobserved location. We use parallel computing at the pixel level to compute the standard deviation of predictions made with QRF. As this is a pixel-wise process, we can obtain a continuous surface of model uncertainty. Regarding the categorical variables, the uncertainty is measured by the scaled Shannon Entropy Index (H_s), which is calculated by using the probability maps of the target classes (Equation 1) in both great groups and taxonomic family by particle-size.

$$H_s(x) = \sum_{k=1}^{K} p_k(x) \times \log_{K(p_K(x))} \qquad \text{Eq. 1}$$

in which K is the number of possible classes; $\log_K$ is the logarithm to base K and p_k is the probability of class k. Ranging from 0–1, 0 indicates no ambiguity, and 1 indicates maximum confusion. This metric should not be confused with classification accuracy assessment as H_s is an internal accuracy measure derived from the model and not based on a comparison of predictions with validation data, such as the OA and kappa metrics (Hengl et al., 2017).

Model use and predictions: it mainly consists in predicting over unobserved locations by using the best model fitted in the previous steps. The multi-band raster stack of covariates was used for predicting each target variable across the whole surface of the AOI.

Software and implementation

All the steps of the pipeline presented here were implemented in R software version 4.0.2. SAGA version 7.8.2 was used to generate the derivatives of DEM. The main R-libraries used according to the key steps of the pipeline are presented in table 3. We deployed these steps using the sample data on a machine with a 4-core 2.5 GHz Intel Core i5 CPU, 16 GB RAM and macOS.

RESULTS

Data preprocessing

Input data matrices were generated separately by the target variable. Table 4 presents the amount of data, reported with the number of soil profiles after preprocessing, i.e., outliers' removal, and the corresponding sizes of the training and test partitions. The

Table 3. Lists of key R-libraries used by the pipeline per step

Step	R libraries and versions
Preprocessing	read excel files (readxl 1.3.1), geospatial data (raster 3.13-3, rgeos 0.5-5, rgdal 1.5-18, sf 0.9-16, smoothr 0.1.2, gdalUtilities 1.1.1, rgee 1.0.6), interpolation of soil properties (GSIF 0.5-15, aqp 1.19), machine learning (caret 6.0-86), utilities (tidyr 1.1.2, dplyr 1.02, magrittr 1.5, stringr1.4.0)
Covariables selection	machine learning (caret 6.0-86, e1071 1.7-4, randomForest 4.6-14, Boruta 7.0.0), parallel processing (doParallel 1.0.15, snow 0.4-3), utilities (dplyr 1.02, stringr1.4.0)
Exploratory analysis	statistical (psych 2.0.9, nortest 1.0-4, tseries 0.10-48, rstatix 0.6.0, PMCMR 4.3, rcompanion 2.3.26), machine learning (caret 6.0-86), plots (ggplot2 3.3.2, PerformanceAnalytics 2.0.4), utilities (tidyr 1.1.2, purr 0.3.4, multcompView 0.1-8, stringr1.4.0)
Modeling	training (caret 6.0-86), handle geospatial data (raster 3.13-3, sf 0.9-16), model evaluation (Metrics 0.1.4, hydroGOF 0.4-0), model uncertainty (quantregForest 1.3-7), parallel processing (doParallel 1.0.15, snow 0.4-3), plots (ggplot2 3.3.2, ggspatial 1.1.4, viridis 0.5.1), utilities (tidyr 1.1.2, dplyr 1.02, purr 0.3.4, stringr1.4.0)

Table 4. Number of soil profiles per target variable after preprocessing and related task type. The decrease is based on an initial 1858 profiles

Variable	Profiles (% decrease)	Training (70 %)	Test (30 %)	Task
pH (0.00-0.30 m)	1537 (17 %)	1077	460	Regression
pH (0.30-1.00 m)	1537 (17 %)	1078	459	Regression
Great groups	1069 (42 %)	756	468	Classification (20 classes)
Taxonomical family by particle-size	762 (59 %)	536	226	Classification (15 classes)

number of profiles decreased from 17 % (pH) to 59 % (taxonomic family by particle-size). The modeled classes for the taxonomic family by particle-size and USDA great group were 15 and 20, respectively.

Covariates selection

The number of optimal covariates variated by target soil property. Figure 2 shows an example of a diagram obtained using the RFE-RF algorithm for pH 0.00-0.30 m. Eight covariates are suggested as the optimal number of covariates (see the dot marker). For the remaining variables, the optimal covariates were 5, 3 and 7 for pH 0.30-1.00 m, great group, and taxonomical family by particle size, respectively. The following section provides a list of the covariates selected.

Exploratory data analysis

According to the statistical analysis for pH, the Lilliefors and Jarque-Bera tests indicate that this variable is not normally distributed at two layers. The Kruskal-Wallis reveals significant differences ($p<0.0001$) in the distribution between pH 0.00-0.30 m and two covariates (climate and terrain) (Table 5).

The post-hoc test using Wilcoxon-Mann–Whitney U with the Bonferroni correction indicated the climate levels in which pH 0.00-0.30 m resulted significantly different. They were between humid warm and dry warm ($p<0.001$) as well as very dry warm and dry warm ($p<0.05$).

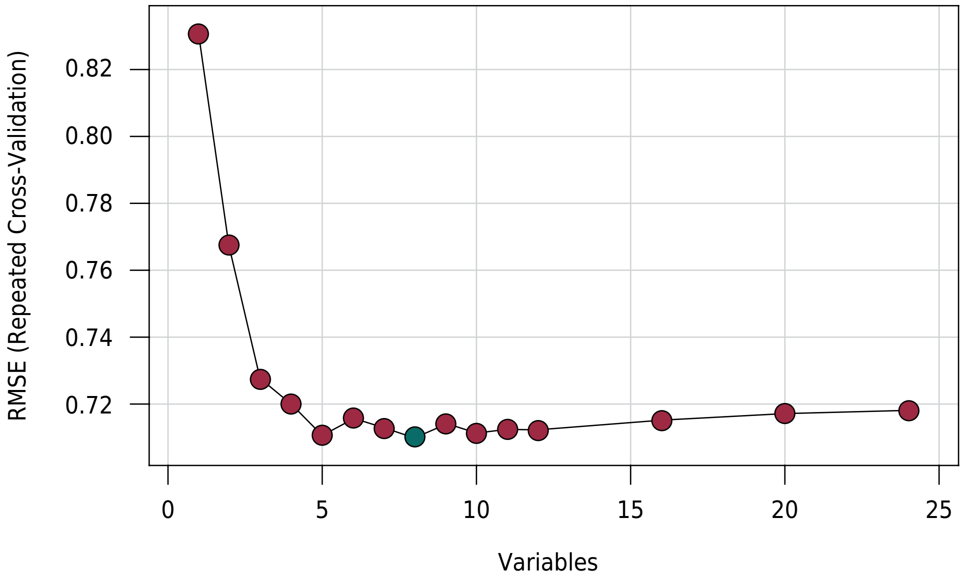

Figure 2. Diagram of the RFE-RF curve for pH 0.00-0.30 m showing the number of covariates (x-axis) and changes in RMSE values (y-axis).

Table 5. Spearman's (S) rho test to determine the association between pH at two layers and their optimal continuous covariates. Kruskall Wallis (KW) is also provided when comparing the means of pH among groups from optimal categorical variables

pH depth	Covariate	Source (layer)	Test	Statistic	p-value	rho
0.00-0.30 m	Channel Network Base Level	DEM and its derivatives	S	178346514.49	0.00 ****	0.14
	DEM		S	176884870.09	0.00 ****	0.15
	MRRTF		S	193342679.57	0.019 *	0.07
	MRVBF		S	190954081.66	0.006 **	0.08
	Valley Depth		S	228246417.32	0.001 **	-0.10
	SkyViewFactor		S	222052147.18	0.029 *	-0.07
	Terrain	Topography	KW	149.78	0.00 ****	N/A
	Climate	Weather	KW	53.76	0.00 ****	N/A
0.30-1.00 m	Channel Network Base Level	DEM and its derivatives	S	170484456.77	0.00 ****	0.18
	DEM		S	170231364.19	0.00 ****	0.18
	MRRTF		S	188564995.90	0.001 **	0.10
	MRVBF		S	202336351.40	0.311 NS	0.03
	ValleyDepth		S	233699592.63	0.00 ****	-0.12

NS: not significant; *: $p<0.05$; **: $p<0.01$; ***: $p<0.001$; ****: $p<0.0001$. N/A: not applicable.

For the categorical variables, the distribution of great groups was significantly different ($p<0.0001$) to DEM as taxonomic family by particle-size classes were against DEM ($p<0.01$) and terrain ($p<0.0001$) (Table 6).

Modeling

Overall, Random Forest had the best performance among the five algorithms proposed to predict pH at two layers, great group and taxonomic family by particle-size classes. Figure 3 illustrates boxplots with the performance of the predictions models of pH 0.00-0.30 m and great groups using five trained algorithms (with different configurations, i.e., parameters).

From the different configurations of trained Random Forest algorithms, the one that performed best in cross-validation with the training partition was used for

Table 6. Kruskall Wallis (KW) test compares means of optimal continuous covariates among groups of target categorical variables. Chi-square (CS) test when investigating relationships between categorical variables

Variable	Covariate	Source (layer)	Test	Statistic	p-value
Great groups	Channel Network Base Level	DEM and its derivatives	KW	742.51	0.37
	DEM		KW	315.79	0.00 ****
	Climate	Weather	CS	358.28	0.00 ****
Taxonomical family by particle-size	Channel Network Base Level	DEM and its derivatives	KW	527.38	0.43
	DEM		KW	258.93	0.00 **
	MRVBF		KW	534.91	0.38
	ValleyDepth		KW	534.91	0.35
	SkyViewFactor		KW	522.68	0.48
	PositiveOpenness		KW	534.91	0.36
	Terrain	Topography	CS	358.28	0.00 ****

**: p<0.01; ****: p<0.0001.

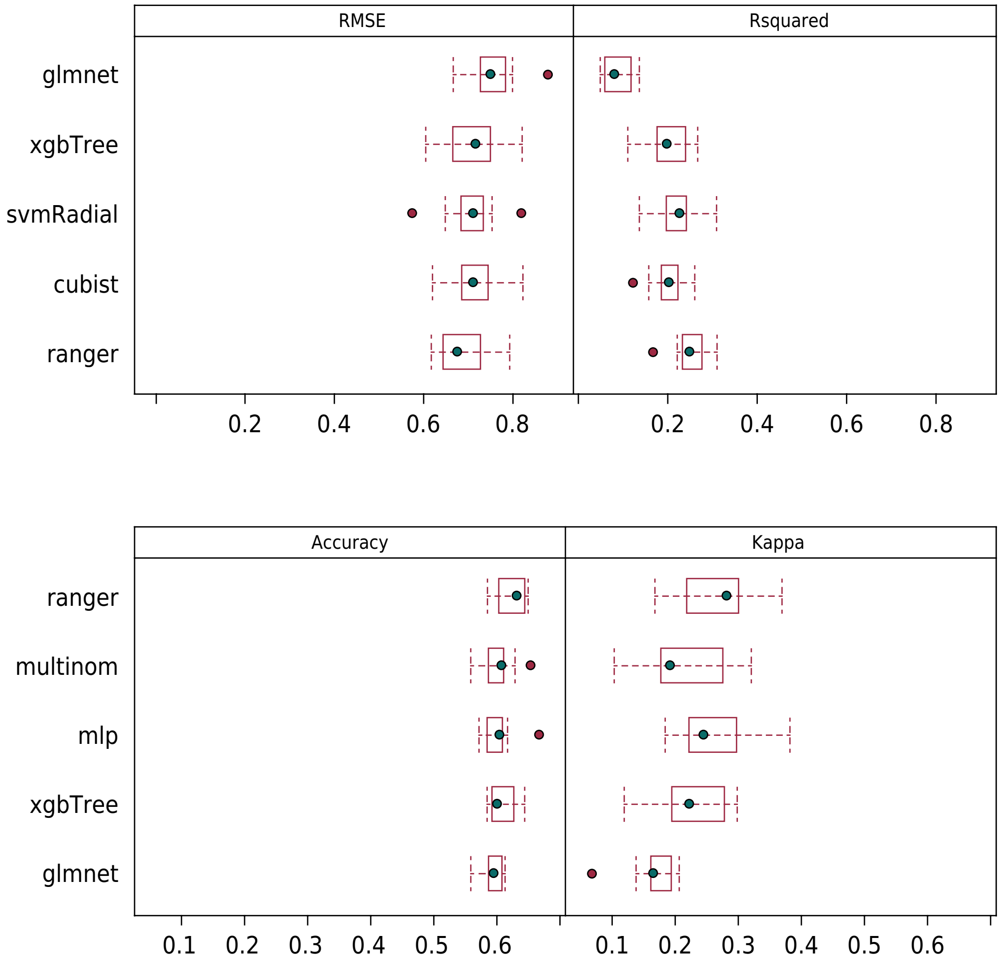

Figure 3. Performance of the trained ML models for predicting pH 0.00-0.30 m (a) and great groups (b). The random forest (ranger) obtained the best performance according to the main metrics, RMSE and Accuracy, used for quantitative and categorical variables, respectively.

independent evaluation with the test set. This assessment can be inspected through scatterplots and confusion matrices for quantitative and categorical variables, respectively (Figure 4).

From the above outputs, a set of evaluation metrics were computed for the target variables. According to the main metrics, RMSE and Overall Accuracy for quantitative

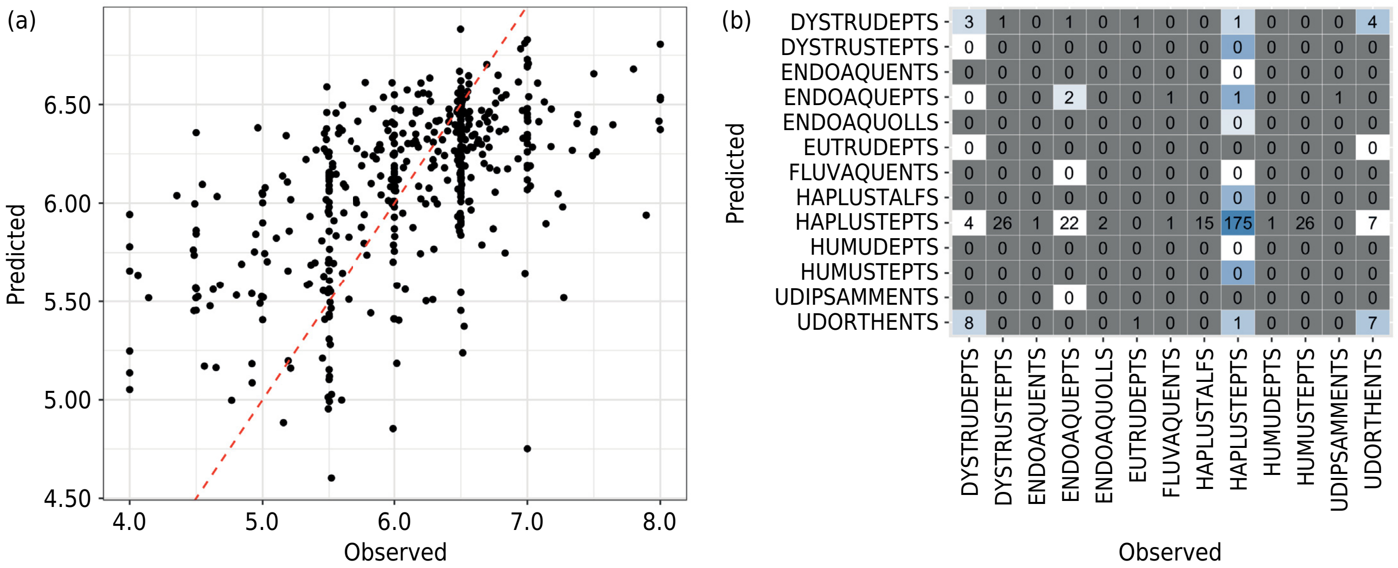

Figure 4. Scatter plot (a) and confusion matrix (b) charts comparing predictions vs observations in the test set according to the best model for predicting pH 0.00-0.30 m and great groups, respectively.

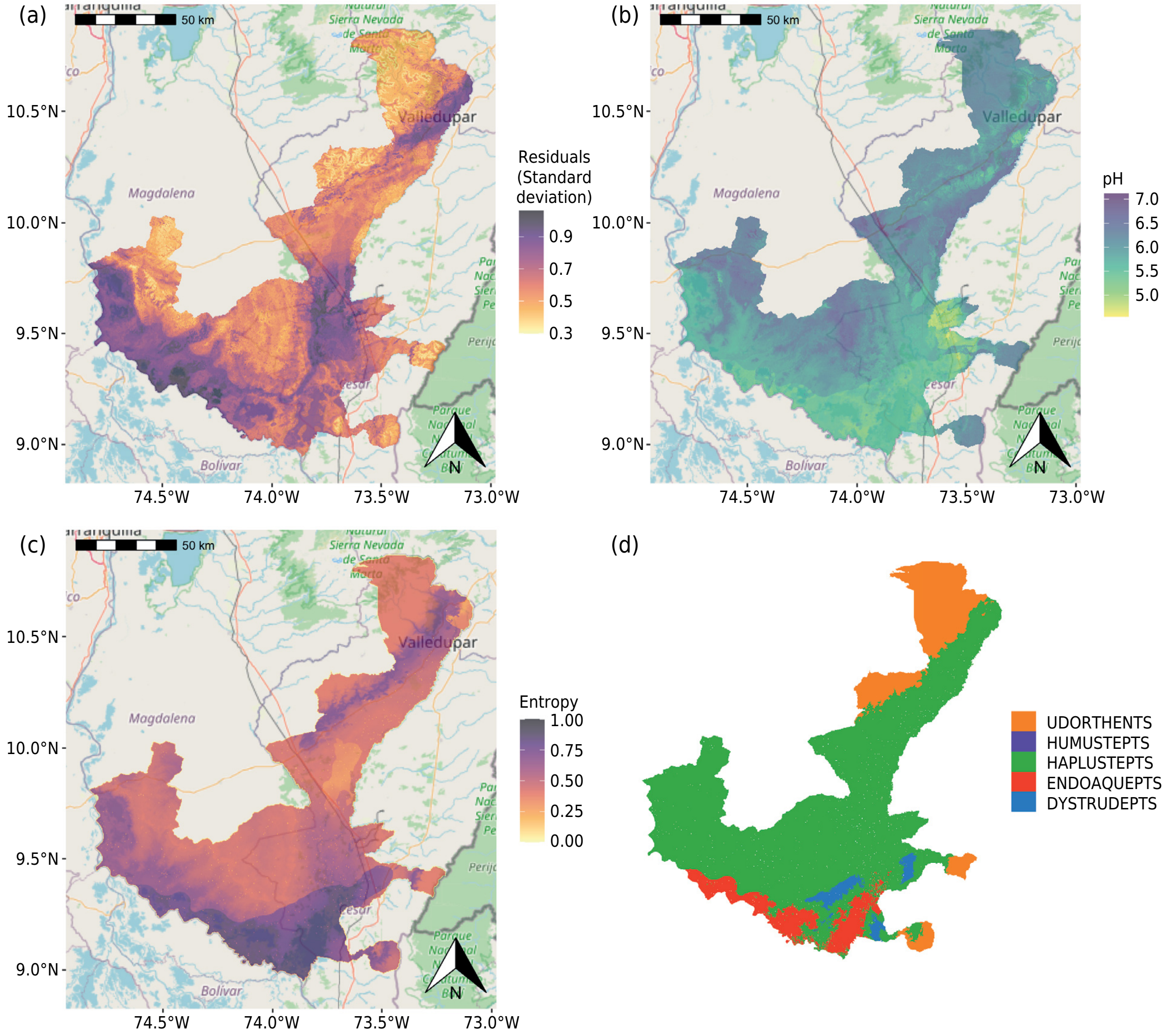

Figure 5. Uncertainty (left) and prediction (right) maps of pH 0.00-0.30 m (a, b) and great groups (c, d). Open Street Map is used as base map.

and categorical variables, respectively, the best model has error values of 0.63 and 0.68 for pH 0.00-0.30 m and 0.30-1.00 m, respectively; and accuracy values of 60 and 67 % for great groups and taxonomic family by particle-size, correspondingly.

To report the predictions' error, the uncertainty estimation was a key to highlight areas in which the best models tend to have a lower certainty. For instance, figure 5 shows maps of uncertainty and prediction of pH 0.00-0.30 m and great groups. For pH, the dark areas indicate a higher standard deviation of the residuals, which can be associated with a high error. For great groups, the interpretation remains different as the dark areas indicate a higher mix of the mapped classes. These mixed pixels can be related to the mapping unit, in this case, 250 m, which might be too coarse to discriminate against a dominant great group.

DISCUSSION

The proposed pipeline, tested over a considerably large area of 14,537 km^2, shows its feasibility to handle multiple data-driven methodologies (from preprocessing to modeling) to generate information related to the spatial distribution of soil properties. According to a review on 150 studies about mapping soil properties using ML algorithms, the use of legacy samples as they were tested in the pipeline, is predominant for local and regional scale areas (about 10^4 km^2) (Wadoux et al., 2020a). For the deployment of the pipeline, 1858 soil profiles or 7.8 units per km^2 were considered. However, the initial density was reduced between 17 and 59 % after the preprocessing step per target variable. This number is still relatively high if compared with that found by Wadoux et al. (2020a), who reported an average sampling of 0.24 units per km^2.

The recursive feature elimination algorithm, in this case, RFE-RF widely used in previous DSM studies (Shi et al., 2018; Gomes et al., 2019), recursively removes the least important covariates from the initial pool, with little or no decrease in model prediction accuracy. While the optimal subsets of environmental covariates yielded by the RFE-RF provided reasonable results for each target variable according to the mapping resolution of 250 m, further covariates representing multi-scale or temporal variation should be incorporated. The covariates have to represent only soil-forming factors and, where possible, a physical explanation to ensure an unbiased ML-DSM soil knowledge generation (Wadoux et al., 2020b).

After the selection of covariates, a complementary component of the pipeline is the exploratory data analysis. The statistical analysis is an added value to this component as it complements the interpretation of predictions obtained by the ML algorithms. For instance, the statistical analysis supported a thematic validation conducted over maps generated for the same soil properties but with the restricted access DEM. For either modeling exercises using open or restricted DEM, it was evident that pH 0.00-0.30 m has a significant response according to the climate levels. This evidence corroborates previous studies, which indicate soils from different climates have distinct soil. Specifically, climate can affect the process of soil chemical reaction and consequently influence soil pH (Zhang et al., 2019).

Regarding the algorithms assessed, Random Forest yielded the best performance consistently for the target soil properties. This particular algorithm is the most popular for modeling quantitative and categorical variables in ML-DSM (Wadoux et al., 2020b). It is worth mentioning, the random grid search for parameter tuning allows maximizing the performance of all trained models. For instance, the main parameters tuned for the RF algorithm were the min node size, number of predictors per division and partition rule. Additionally, to facilitate the interpretability of the best model, the pipeline returns charts of the covariate importance. For the RF algorithm, the most essential covariates are identified using the mean decrease in the variance of the response (Wright and Ziegler, 2017).

Table 7. Processing times for pH 0.00-0.30 m with(out) parallel processing

Step	Processing time (min)	
	Single core	Four-core (parallel)
Covariate selection	14.6	6.5
Exploratory analysis	0.2	0.2
Modeling	14.0	11.7
Total	28.8	18.4

After selecting the best model, the assessment with the test partition indicates an acceptable accuracy according to studies in similar environmental conditions, i.e., tropics and input data. For instance, using legacy soil data collected across the Cundiboyacense high plateau, Colombia, Araujo-Carrillo et al. (2021) reported a RMSE of 0.57 for pH at the topsoil (0.00-0.20 m) and overall accuracy of 48 % for the taxonomic family by particle-size classes (9 groups) modeled at a spatial resolution of 125 m using the exact Random Forest implementation in R (ranger). While there were no studies conducted at similar conditions to compare results of the second depth (0.30-1.00 m), the RMSE is close to the first depth. Wadoux et al. (2020a) indicate most of the existing ML-DSM studies (around 70 %) predicted a soil property or class for a single depth (topsoil). The existing studies at multiple depths mostly focus on modeling soil properties in temperate climates (Lacoste et al., 2014; Viscarra Rossel et al., 2015).

It is worth mentioning the observed performances of the best predictive models are considered poor. One of the potential explanations for poor performance is the quality and robustness of datasets which are of greater importance than the classifier itself (Meir et al., 2018). For reproducibility purposes, we intentionally deployed the tool using coarse spatial datasets, harmonized on a 250 × 250 m pixel size according to the spatial resolution of the input DEM. Cavazzi et al. (2012) claim the landscape complexity of the study area can play a pivotal role in choosing the optimal resolution for modeling soil properties. The authors found varied morphological areas with abrupt changes in topography, like the Momposine depression in Colombia, can yield better prediction results with fine resolutions (30 m). We, therefore, expect the predictive models can be improved with a higher resolution DEM. In addition to the spatial datasets, the quality of the legacy soil data might impact models' performance. In figure 4, the scatter plot between observations and predictions of pH at 0.00-0.30 m shows a poor performance in predicting extreme pH values and minor classes for great groups. Taking the imbalance distribution into account would improve the performance of the predictive models. In this regard, some scholars have proposed approaches to tackling imbalanced datasets from low-quality legacy soil data, few of them tested across tropical environments (Hounkpatin et al., 2018).

Additional to the feature of modeling at multiple depths, the pipeline includes the estimation of uncertainty according to the variable type. As suggested by Hengl et al. (2017), maps of uncertainty could be potentially very useful for planning new soil surveys. Finally, in terms of computing performance, table 7 reports the processing times of the pipeline for modeling pH 0.00-0.30 m from the covariate selection step with and without parallel processing. This parallel processing feature provides substantial gains for both covariates selection and modeling steps.

CONCLUSIONS

This investigation successfully deployed a semi-automatic and sequential pipeline, programmed in the R software, named dinoSoil-toolbox, to generate soil-related information in digital formats. As it was shown using legacy soil data from a mountainous valley terrain in Colombia, the proposed pipeline facilitates a rapid mapping of quantitative and categorical soil properties. By providing multiple outputs such as tables, charts, maps, and geospatial data in four main steps, the pipeline offers considerable robustness to support outcomes and analysis of a DSM project. These components are aligned to the recommendations by Wadoux et al. (2020a) of plausibility, interpretability, and explainability in ML-DSM developments that enable soil scientists to couple model prediction with pedological explanation and understanding of the underlying soil processes. Furthermore, we released the pipeline on a public GitHub repository (https://github.com/acocac/dinoSOIL-toolbox) with readable documentation and facilitating reproducibility by making available the input data presented in this research.

Future studies might be interesting to implement further ML frameworks such as ensembles and deep learning models, which have shown a high performance for DSM. Moreover, while the pipeline was designed under an open science framework for users with relative knowledge in R, it would be helpful to design a GUI that facilitates its use for non-programming experts.

SUPPLEMENTARY DATA

Supplementary data to this article can be found online at https://www.rbcsjournal.org/wp-content/uploads/articles_xml/1806-9657-rbcs-45-e0210084/1806-9657-rbcs-45-e0210084-suppl01.pdf

ACKNOWLEDGEMENTS

The authors thank Diana Galindo and Napoleón Ordoñez, head of CIAF and the Sub-Directorate of Agrology at IGAC, respectively, for their support and technical feedback in the development of this work.

AUTHOR CONTRIBUTIONS

Conceptualization: Alejandro Coca-Castro (lead), Andrés Felipe López (equal), Joan Sebastián Gutierrez-Díaz (supporting), Juan Camilo García (equal), Patricia Escudero (equal), Pedro Karin Serrato (equal), Ricardo Devia (equal), Victoria Camacho (equal) and Yesenia Vargas (equal).

Data curation: Joan Sebastián Gutierrez-Díaz (supporting), Juan Camilo García (equal) and Ricardo Devia (equal).

Formal analysis: Alejandro Coca-Castro (equal), Andrés Felipe López (equal), Joan Sebastián Gutierrez-Díaz (supporting), Patricia Escudero (equal), Pedro Karin Serrato (equal), Victoria Camacho (equal) and Yesenia Vargas (equal).

Funding acquisition: Carlos Franco (lead) and Janeth González (equal).

Investigation: Alejandro Coca-Castro (lead), Andrés Felipe López (equal), Joan Sebastián Gutierrez-Díaz (supporting), Juan Camilo García (supporting), Patricia Escudero (equal), Pedro Karin Serrato (equal), Ricardo Devia (supporting), Victoria Camacho (equal) and Yesenia Vargas (equal).

Methodology: Alejandro Coca-Castro (lead), Andrés Felipe López (equal), Joan Sebastián Gutierrez-Díaz (equal), Juan Camilo García (supporting), Patricia Escudero (equal), Pedro Karin Serrato (equal), Ricardo Devia (supporting), Victoria Camacho (equal) and Yesenia Vargas (equal).

Project administration: Carlos Franco (equal) and Janeth González (equal).

Resources: Carlos Franco (equal) and Janeth González (equal).

Software: Alejandro Coca-Castro (lead), Andrés Felipe López (supporting), Carlos Franco (equal), Joan Sebastián Gutierrez-Díaz (supporting), Patricia Escudero (supporting) and Victoria Camacho (supporting).

Supervision: Carlos Franco (lead) and Janeth González (lead).

Validation: Andrés Felipe López (equal), Juan Camilo García (equal), Patricia Escudero (equal), Pedro Karin Serrato (equal), Ricardo Devia (equal), Victoria Camacho (equal) and Yesenia Vargas (equal).

Visualization: Alejandro Coca-Castro (lead).

Writing – original draft: Alejandro Coca-Castro (lead).

Writing – review & editing: Alejandro Coca-Castro (lead), Carlos Franco (supporting), Joan Sebastián Gutierrez-Díaz (equal) and Pedro Karin Serrato (equal).

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

Araujo-Carrillo GA, Varón-Ramírez VM, Jaramillo-Barrios CI, Estupiñan-Casallas JM, Silva-Arero EA, Gómez-Latorre DA, Martínez-Maldonado FE. IRAKA: The first Colombian soil information system with digital soil mapping products. Catena. 2021;196:104940. https://doi.org/10.1016/j.catena.2020.104940

Arrouays D, Lagacherie P, Hartemink AE. Digital soil mapping across the globe. Geoderma Reg. 2017;9:1-4. https://doi.org/10.1016/j.geodrs.2017.03.002

Ballabio C, Panagos P, Monatanarella L. Mapping topsoil physical properties at European scale using the LUCAS database. Geoderma. 2016;261:110-23. https://doi.org/10.1016/j.geoderma.2015.07.006

Bishop TFA, McBratney AB, Laslett GM. Modelling soil attribute depth functions with equal-area quadratic smoothing splines. Geoderma. 1999;91:27-45. https://doi.org/10.1016/S0016-7061(99)00003-8

Cavazzi S, Corstanje R, Mayr T, Hannam J, Fealy R. Are fine resolution digital elevation models always the best choice in digital soil mapping? Geoderma. 2013;195-196:111-21. https://doi.org/10.1016/j.geoderma.2012.11.020

CGIAR-CSI Consortium for Spatial Information. SRTM 90m Digital Elevation Database v4.1. Rome: CGIAR-CSI; 2018 [cited 2020 Nov 10]. Available from: https://cgiarcsi.community/data/srtm-90m-digital-elevation-database-v4-1/.

Cressie N, Johannesson G. Fixed rank kriging for very large spatial data sets. J R Statist Soc B. 2008;70:209-26. https://doi.org/10.1111/j.1467-9868.2007.00633.x

Food and Agriculture Organization of the United Nations FAO. Soil organic carbon mapping cookbook. 2nd ed. Rome: FAO; 2018.

Gamboa AC, Hilarión CAC, Delgado NO, Díaz JG, Lucero GA, Santamaría MG, Olivera C, Olmedo G, Bunning S, Vargas R. Estimación de carbono orgánico del suelo en Colombia, una herramienta de gestión del territorio. Ecosistemas. 2021;30:2019. https://doi.org/10.7818/ECOS.2019

Gomes LC, Faria RM, Souza E, Veloso GV, Schaefer CEGR, Fernandes Filho EI. Modelling and mapping soil organic carbon stocks in Brazil. Geoderma. 2019;340:337-50. https://doi.org/10.1016/j.geoderma.2019.01.007

Gorelick N, Hancher M, Dixon M, Ilyushchenko S, Thau D, Moore R. Google earth engine: planetary-scale geospatial analysis for everyone. Remote Sens Environ. 2017;202:18-27. https://doi.org/10.1016/j.rse.2017.06.031

Guevara M, Olmedo GF, Stell E, Yigini Y, Duarte YA, Hernández CA, Arévalo GE, Arroyo-Cruz CE, Bolivar A, Bunning S, Cañas NB, Cruz-Gaistardo CO, Davila F, Dell Acqua M, Encina A, Tacona HF, Fontes F, Herrera JAH, Navarro ARI, Loayza V, Manueles AM, Jara FM, Olivera C, Hermosilla RO, Pereira G, Prieto P, Ramos IA, Brina JCR, Rivera R, Rodríguez-Rodríguez J, Roopnarine R, Ibarra AR, Riveiro KAR, Schulz GA, Spence A, Vasques GM, Vargas RR, Vargas R. No silver bullet for digital soil mapping: country-specific soil organic carbon estimates across Latin America. Soil. 2018;4:173-93. https://doi.org/10.5194/soil-4-173-2018

Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn. 2002;46:389-422. https://doi.org/10.1023/A:1012487302797

Hengl T, Heuvelink GBM, Kempen B, Leenaars JGB, Walsh MG, Shepherd KD, Sila A, MacMillan RA, Jesus JM, Tamene L, Tondoh JE. 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

Hengl T, Jesus JM, Heuvelink GBM, Gonzalez MR, Kilibarda M, Blagotić A, Shangguan W, Wright MN, Geng X, Bauer-Marschallinger B, Guevara MA, Vargas R, MacMillan RA, Batjes NH, Leenaars JGB, Ribeiro E, Wheeler I, Mantel S, Kempen B. SoilGrids250m: Global gridded soil information based on machine learning. PLoS One. 2017;12:e0169748. https://doi.org/10.1371/journal.pone.0169748

Hounkpatin KOL, Schmidt K, Stumpf F, Forkuor G, Behrens T, Scholten T, Amelung W, Welp G. Predicting reference soil groups using legacy data: A data pruning and Random Forest approach for tropical environment (Dano catchment, Burkina Faso). Sci Rep. 2018;8:9959. https://doi.org/10.1038/s41598-018-28244-w

Instituto de Hidrología, Meteorología y Estudios Ambientales - IDEAM. Atlas climatológico de Colombia [internet]. Bogota: IDEAM; 2005 [cited 2020 Nov 30]. Available from: http://atlas.ideam.gov.co/presentacion/

Instituto Geográfico Agustín Codazzi - IGAC. Conflictos de uso del territorio colombiano. Escala 1:100.000 [internet]. Bogota: Gobierno de Colombia; 2012 [cited 2020 Nov 20]. Available from: https://geoportal.igac.gov.co/contenido/datos-abiertos-agrologia/

Keskin H, Grunwald S. Regression kriging as a workhorse in the digital soil mapper's toolbox. Geoderma. 2018;326:22-41. https://doi.org/10.1016/j.geoderma.2018.04.004

Lacoste M, Minasny B, McBratney A, Michot D, Viaud V, Walter C. High resolution 3D mapping of soil organic carbon in a heterogeneous agricultural landscape. Geoderma. 2014;213:296-311. https://doi.org/10.1016/j.geoderma.2013.07.002

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

Meier M, de Souza E, Francelino M, Fernandes-Filho E, Schaefer C. Digital Soil Mapping Using machine learning algorithms in a tropical mountainous area. Rev Bras Cienc Solo. 2018;42:1-22. https://doi.org/10.1590/18069657rbcs20170421

Mulder VL, Lacoste M, Richer-de-Forges AC, Arrouays D. GlobalSoilMap France: High-resolution spatial modelling the soils of France up to two meter depth. Sci Total Environ. 2016;573:1352-69. https://doi.org/10.1016/j.scitotenv.2016.07.066

Odgers NP, Libohova Z, Thompson JA. Equal-area spline functions applied to a legacy soil database to create weighted-means maps of soil organic carbon at a continental scale. Geoderma. 2012;189-90:153-63. https://doi.org/10.1016/j.geoderma.2012.05.026

Padarian J, Minasny B, McBratney AB. Machine learning and soil sciences: a review aided by machine learning tools. Soil. 2020;6:35-52. https://doi.org/10.5194/soil-6-35-2020

Padarian J, Minasny B, McBratney AB. Chile and the Chilean soil grid: A contribution to GlobalSoilMap. Geoderma Reg. 2017;9:17-28. https://doi.org/10.1016/j.geodrs.2016.12.001

Shi J, Yang L, Zhu A-X, Qin C, Liang P, Zeng C, Pei T. Machine-learning variables at different scales vs. knowledge-based variables for mapping multiple soil properties. Soil Sci Soc Am J. 2018;82:645-56. https://doi.org/10.2136/sssaj2017.11.0392

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

Rossel RAV, Chen C, Grundy MJ, Searle R, Clifford D, Campbell PH. The Australian three-dimensional soil grid: Australia's contribution to the GlobalSoilMap project. Soil Res. 2015;53:845-64. https://doi.org/10.1071/SR14366

Wadoux AMJ-C, Minasny B, McBratney AB. Machine learning for digital soil mapping: Applications, challenges and suggested solutions. Earth-Science Rev. 2020a;210:103359. https://doi.org/10.1016/j.earscirev.2020.103359

Wadoux AMJ-C, Samuel-Rosa A, Poggio L, Mulder VL. A note on knowledge discovery and machine learning in digital soil mapping. Eur J Soil Sci. 2020b;71:133-6. https://doi.org/10.1111/ejss.12909

Wright MN, Ziegler A. Ranger : A fast implementation of random forests for high dimensional data in C++ and R. J Stat Softw. 2017;77:1-17. https://doi.org/10.18637/jss.v077.i01

Zhang YY, Wu W, Liu H. Factors affecting variations of soil pH in different horizons in hilly regions. PLoS One. 2019;14:e0218563. https://doi.org/10.1371/journal.pone.0218563