Catena 196 (2021) 104940

Contents lists available at ScienceDirect

Catena

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

IRAKA: The first Colombian soil information system with digital soil mapping products

Gustavo A. Araujo-Carrillo[a,*], Viviana Marcela Varón-Ramírez[a], Camilo Ignacio Jaramillo-Barrios[b], Jhon M. Estupiñan-Casallas[a], Elías Alexander Silva-Arero[a], Douglas A. Gómez-Latorre[a], Fabio E. Martínez-Maldonado[a]

[a] Corporación Colombiana de Investigación Agropecuaria, Centro de Investigación Tibaitatá, Mosquera 250047, Cundinamarca, Colombia
[b] Corporación Colombiana de Investigación Agropecuaria, Centro de Investigación Nataima, Espinal 733529, Tolima, Colombia

ARTICLE INFO

Keywords:
Soil database
Machine learning techniques
Cundiboyacense high plateau
Geographic web services

ABSTRACT

Advances in science and technology combined with the need to handle and manage natural resources have resulted in a gradual increase in soil data availability. This has allowed characterizing, describing, or proposing projects to address issues such as soil degradation, agricultural productivity, threats, and ecosystem services. Despite the availability of soil data, users are increasingly looking for additional, timely, and reliable information in order to respond to current and future situations, such as food security and climate change. IRAKA is a soil information system for the Colombian Cundiboyacense high plateau and the first to provide additional information from available official sources. It has a modular design based on the organization of a soil database whose structure incorporated 299 profiles, 1432 samples, and 55 properties and the integration, harmonization, and standardization of 8 soil surveys conducted in the area. The structured information enabled modeling and validating physical and chemical properties through digital soil mapping and comparing different machine learning techniques, including random forest, ranger, support vector machine, and ensemble models of these techniques. The interpolations of 11 quantitative properties and 1 qualitative property resulted in acceptable coefficients of goodness of fit due to different factors: property variability, representativeness, distribution in the study area, and property description based on environmental covariates. All the information generated has been made available for widespread use across different geographic web services. Thus IRAKA contributes to the management of information and knowledge through techniques and tools that allow users to visualize new information, use the information, and draw attention to new studies and improve databases in which the confidence level is insufficient and does not allow decision making.

1. Introduction

Maps are one of the simplest methods for expanding soil knowledge. They have been drawn since prehistoric times, evolving from scientific thematic maps at the beginning of the 19th century (Hartemink et al., 2013) to the first world soil map in 1906 by K.D. Glinka (Karavaeva and Gerasimova, 2005). In the 20th century, the advances were exponential, and soil maps led to the need for increasingly powerful information systems. Since the first decade of the 21st century, techniques such as digital soil mapping (DSM) (McBratney et al., 2003; Lagacherie et al., 2006; Cambule et al., 2013; Arrouays et al., 2017a; Malone et al., 2017; Guevara et al., 2018) have been incorporated, backed by computer-assisted procedures that operate structured databases and are supported by geographic information systems (GIS) to collect, organize, store, and disseminate information (Olaya, 2011).

Advances in soil maps have been made possible by the numerous studies published by the scientific community, which were initially used to understand soil-landscape relationships (Hendriks et al., 2019). The identification of the constraints and potentialities of soil has generated a series of almost standardized processes that include soil description (FAO, 2006), classification (Soil Survey Staff, 2014; IUSS Working Group WRB, 2015), and mapping (Soil Science Division Staff, 2017; FAO, 2018). The results are generally presented in information systems that have a functional perspective, namely, a technological

* Corresponding author.
E-mail addresses: garaujo@agrosavia.co (G.A. Araujo-Carrillo), vvaron@agrosavia.co (V.M. Varón-Ramírez), cijaramillo@agrosavia.co (C.I. Jaramillo-Barrios), jmestupinan@agrosavia.co (J.M. Estupiñan-Casallas), esilva@agrosavia.co (E.A. Silva-Arero), dagomez@agrosavia.co (D.A. Gómez-Latorre), femartinez@agrosavia.co (F.E. Martínez-Maldonado).

https://doi.org/10.1016/j.catena.2020.104940
Received 26 February 2020; Received in revised form 5 September 2020; Accepted 26 September 2020
Available online 22 October 2020

Table 1
Soil surveys at various scales in the study area.

No	Name of the study	Scale	Year
1	General study of soils and land zoning of the Department of Cundinamarca	100,000	2000
2	General study of soils and land zoning of the Department of Boyacá	100,000	2004
3	Detailed soil survey in the flat areas of 14 municipalities in the savanna of Bogotá	10,000	2012
4	Detailed soil survey in flat areas of the municipalities of Cogua, El Rosal, Nemocón, Subachoque, Suesca, Zipacón and Zipaquirá	10,000	2013
5	Semidetailed soil survey in areas influenced by moors in Colombia – District of Páramos Altiplano[1]	25,000	2016
6	Semidetailed soil survey in areas influenced by moors in Colombia – District of Páramos Cundinamarca[1]	25,000	2016
7	Semidetailed soil survey in areas influenced by moors in Colombia – District of Páramos Boyacá[1]	25,000	2016
8	Semidetailed soil survey in the areas influenced by wetlands in Colombia - Andean region[1]	25,000	2016

[1] Study conducted by the IGAC in conjunction with the Instituto de Investigación de Recursos Biológicos Alexander von Humboldt – IAvH.

means for recording, storage, and inference, and a structural perspective, which is the result of an interrelated set of processes, data, models, and technologies that have a specific purpose, in order to support decision-making (Dwivedi, 2017).

The soil information system (SIS) concept was created in the Netherlands in 1975 by the Commission of the International Society of Soil Science (Dwivedi, 2017). Since then, various systems have been created, such as the Latin American SIS (SISLAC), developed by FAO. Among the soil mapping and soil property prediction initiatives have been GlobalSoilMap.net (Sánchez et al., 2009; Arrouays et al., 2014, 2017b), and SoilGrids (Hengl et al., 2014, 2017), a global gridded soil information promoted by the International Soil Reference Information Centre (ISRIC).

SISs have different components, such as hardware, software, data, networks, and human resources; however, from the perspective of the user, databases are the most important component (Dwivedi, 2017). A database is a structured collection of data that has information relevant to an organization (Silberschatz et al., 2002). Databases can be classified, according to their structure, as hierarchical, network, relational, and object-oriented (Camps et al., 2005; Ramakrishnan and Gehrke, 2007). Currently, object-oriented databases are the most-used ones in GIS (Olaya, 2011). Object-oriented models are variations of relational models combined with object-oriented programming languages that offer abstraction mechanisms to control the construction of complex systems (Camps et al., 2005).

Among the notable soil databases are the Global and National Soil and Terrain Databases (SOTER) (van Engelen and Dijkshoorn, 2013), active since 1986, e-SOTER-GEOSS, which integrates developments in remote sensing into SOTER under the framework of the Global Earth Observation System of Systems (GEOSS) (Pourabdollah et al., 2012); the ISRIC World Inventory of Soil Emission Potential (WISE) (Batjes, 2009), which is a large global repository of soil profiles; and the Harmonized World Soil Database (HWSD) (FAO/IIASA/ISRIC/ISS-CAS/JRC, 2012). In Latin America and the Caribbean, there is SOTERLAC, a 1:5 M scale database (Dijkshoorn et al., 2005), which is based on a project by ISRIC, FAO, and UNEP. In Colombia, the Instituto Geográfico Agustín Codazzí (IGAC) has databases within a national soil information service called SIGA SIG, which provides mapping information in soil mapping units and their representative profiles on a 1:100,000 scale. Among developments at a more local level, the results reported by Rubiano et al. (2005), in a case study of a georeferenced soil quality system (GEOSOIL) in the municipality of Puerto López, Department of Meta, stand out.

Colombia has made progress in soil resource censuses, especially those intended for departmental land management (IGAC, 2015); however, strategies are still emerging regarding increasing use, encouraging precise application, and understanding needs in terms of data for situations such as climate change and food security (FAO, 2013). In addition, users increasingly require more quantitative information provided in digital formats and with geo-positioning accuracy (Gray et al., 2009). One of the responses to these challenges is hybrid SISs, which incorporate both the techniques and data of traditional soil studies as well as DSM models and information supported by an infrastructure of information technology capable of providing it (Lilburne et al., 2012).

In this regard, the objective of this study was to build a SIS, called IRAKA, oriented toward soil information management in a specific area of Colombia based on 3 elements: organization of a spatial soil database; spatial modeling and analysis of soil properties, including pH, organic matter (OM), electrical conductivity (EC), effective cation exchange capacity (ECEC), Al, Na, Ca, Mg, P, bulk density (Bd), clay (Ar), and textural classes; and availability to users through geographic web services.

2. Methodology

2.1. Study area

The Cundiboyacense high plateau encompasses an area of 16,102 km^2 located in the Andean region of Colombia, in the departments of Cundinamarca and Boyacá, between 4°2′2″ and 6°5′58″ north latitude and 72°41′5″ and 74°32′58″ west longitude. This high plateau is the largest and the most diverse in Colombia, a bio-geographical unit that includes an expanse of the paramo zone, xerophytic enclaves, and forest formations (Rivera et al., 2004). This area has been the subject of the greatest number of soil surveys in the country, carried out at different scales (Table 1). Its altitude ranges from 460 to 4247 m.a.s.l. It has 16 hydrographic subzones and 124 agroecological zones, and from a climate standpoint, its multi-year average rainfall is 1026 ± 325 mm, the annual average air temperature is 12.5 ± 2.4 °C, the multi-year relative humidity is 81.7 ± 2.6%, and the annual sunlight is 1642 ± 163 h. According to the geomorphologic interpretation (IGAC, 2015), the study area has two principal landscapes: mountains and floodplains. Fig. 1 shows the study area.

The Cundiboyacense high plateau is the zone that surrounds Bogotá, the largest city in Colombia, and these areas provide a part of its energy and food requirements. Cold weather vegetable crops (e.g. onion, carrot, and lettuce), potato, and corn are common (Agronet, 2020).

2.2. Data collection

Soil data were collected from three sources: a). Soil surveys conducted by the IGAC (Table 1). b). Characterization study due to climate variability phenomena such as “La Niña” (results of agreement 1723/2011 between the Colombian Ministry of Agriculture and the Corporación Colombiana de Investigación Agropecuaria - AGROSAVIA). c). Characterization study from the Secretary of Agriculture of the Department of Cundinamarca (results of agreement 1847/2014 between the Secretary and AGROSAVIA). All soil data included geographical coordinates and were projected onto the official Colombian coordinate reference system MAGNA-SIRGAS, with the Bogotá (EPGS: 3116) area as the origin.

Table 1 shows that soil data were collected in different years, between 2000 and 2016. That situation causes problems, mainly in the

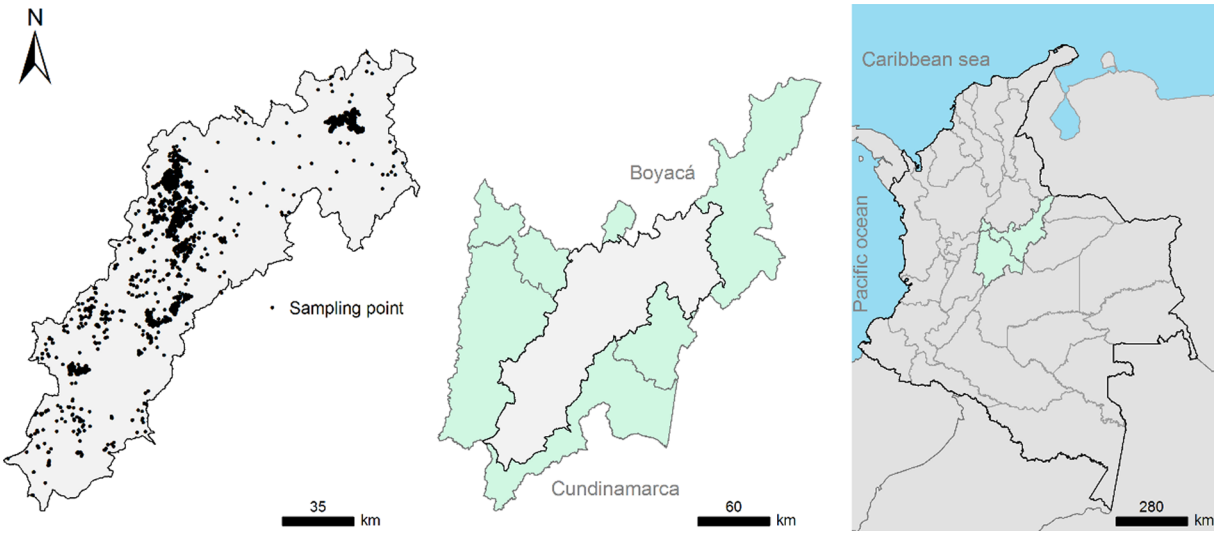

Fig. 1. Study area within the Cundiboyacense high plateau (Colombia).

analysis of the chemical properties, because they easily change with agronomic practices or fertilization. However, the soil data were the legacy data in the study area.

2.3. Spatial soil database

The spatial soil database (SSDB) contained information provided in the surveys described in Section 2.2. Thus 299 soil profiles were incorporated along with their respective horizons from the studies described in Table 1. The additional sampling covered 1432 points, and they were topsoil observations from the characterization studies b and c described in Section 2.2. For the SSDB, the object-relational database management system PostgreSQL 9.5 was used with its pgAdmin 4.13 administration tool and 2.4.3 PostGIS module.

The object-relational model created for the IRAKA system included 16 objects: 6 spatial (S) and 10 alphanumeric (AN) (Table 2). Because most of the data were from the IGAC, the objects and their attributes were adjusted according to the codebook of agrologic surveys handled by that entity (IGAC, 2014).

Fig. 2 shows the structure and attributes of the objects Profile, Soil_Sample, Property, and Soil_Sample_Property.

The SSDB design considered 55 soil properties; however, none of the samples or profiles contained complete information.

2.4. Spatial modeling

The density of points in the study area was 9.3 observations per km^2 (16,102 km^2 with 1731 observations), a result that, associated with the intensity level, was defined as a low- or semi-detailed conventional soil survey (Avery, 1987). Despite this, an approach that included DSM techniques with an evaluation method was proposed (Malone et al., 2017; Guevara et al., 2018). Using the SSDB, the variables that could be spatially represented were identified, particularly those with more observations (> 850 points).

Table 2
Objects and classes established in the SSDB of the IRAKA system.

Object name	Class	Object name	Class
Study_Area	S	Weather	AN
Mapping_Unit	S	Study_Cuelo	AN
Characteristic_Mapping_Unit	AN	Department	S
Characteristic	AN	Municipality	S
Characteristic_Type	AN	Profile	S
Mapping_Unit_Profile	AN	Profile_Characteristic	AN
Property	AN	Soil_Sample_Property	AN
Soil_Sample	S	Nomenclature	AN

S: spatial. AN: alphanumeric.

2.4.1. Soil properties analyzed

Table 3 shows the laboratory methods and units for each of the mapped properties. The texture class was determined by the fraction of sand, silt, and clay for each sampling point; this procedure was performed using the TT.points.in.classes function of the soiltexture package of R version 3.5.3.

Spatial modeling was performed only for the surface layer of the soil, from 0 to 20 cm deep (called "topsoil" in this study). To make the depth of study of the data set uniform, a standard horizon was constructed using a quadratic function of depth with equal areas (spline) (Bishop et al., 1999; Malone et al., 2009). Its main feature is the conservation of the values measured for each horizon (mass conservation), which can be calculated by integrating a continuous function (Malone et al., 2017). This procedure was carried out using the mpspline function of the aqp (*Algorithms for Quantitative Pedology*) package.

2.4.2. Exploratory data analysis

A descriptive and exploratory analysis was performed with all properties studied to estimate measures of central tendency and dispersion (Barbat et al., 2009; Rodríguez-Garay et al., 2016). Subsequently, an exploratory analysis was performed through visual diagnostic tests using box plots and histograms. For mild outlier and extreme outlier data (Dawson, 2011), each property was reviewed in order to determine the observations outside the range of the variable; these values were also compared with other correlated soil properties, and with this method, it was decided whether the sample actually corresponded to an outlier datum or not.

2.4.3. Environmental covariates

A total of 128 covariates for the region of study were constructed, 21 of which were binary class variables (Samuel-Rosa et al., 2015). The 128 layers were used to generate a stack adjusted to a pixel size of 125 m, according to the inspection density and cartographic rules proposed by Hengl (2006). The pixel size was 152 m in fine resolution by inspection of the density, but it was adjusted to half the size of the pixel defined by cartographic rules, at a scale 1:100.000 (125 m). Environmental covariates were selected to broadly reflect the factors that form the soil, as described by Jenny (1941) (Eq. (1)):

$$S_p = f(cl, o, r, p, t) \quad (1)$$

Profile		
PK	Profile_Id	Varchar
FK	Mplty_Id	Varchar
	Profile_Depth_Effect	Int
FK	Char_Id	Varchar
	Profile_CodOrg	Varchar
	Profile_geom	Geometry
	Profile_Altitude	Int
	Profile_Location	Varchar
	Profile_Date	Date
	Profile_Taxonomy	Varchar
	Profile_MatPar	Varchar
	Profile_Veget	Varchar
	Profile_LimUse	Varchar

Soil_Sample		
PK	Soil_Sample_Id	Varchar
	Soil_Sample_geom	Geometry
	Soil_Sample_InitialDepth	Int
	Soil_Sample_FinalDepth	Int
	Soil_Sample_Description	Varchar
FK	Profile_Id	Varchar
FK	Nomenclature_Id	Varchar

Property		
PK	Property_Id	Varchar
	Property_Description	Varchar
	Property_Units	Varchar
	Property_Type	Varchar

Soil_Sample_Property		
PK	Soil_Sample_Prop_Id	Varchar
FK	Property_Id	Varchar
FK	Soil_Sample_Id	Varchar
	Soil_Sample_Prop_Value	Varchar

Fig. 2. Structure and attributes of the objects Profile, Soil_Sample, Property and Soil_Sample_Property (PK and FK correspond to the primary key and foreign key, respectively).

where a soil property (S_p) is a function of the weather (*cl*), biological organisms or land cover (*o*), relief (*r*), parent material (*p*), and time (*t*) (Table 4).

In stacks of covariates, there are surfaces with little variability, or high correlation, between them; these aspects could lead to multicollinearity effects that decrease the predictive capacity of the models (Kuhn et al., 2018). To analyze the stack, 3 sequential processes were carried out: first, a correlation analysis between properties and covariates was performed; then covariates with zero or close to zero variance were eliminated through the nearZeroVar function of the caret package (Kuhn et al., 2018); and lastly, the Spearman correlation coefficients (r) were calculated between environmental covariates, and those with a coefficient higher than 0.9 were eliminated (Zeraatpisheh et al., 2019).

A regression matrix was built with the selected covariates, and this allowed extraction of the covariate values at the coordinates of each sampling point (Massawe et al., 2018). The dataset was stratified into eight strata according to the source of the studies described in Section 2.2. In each stratum, each property was randomly divided using 25% for validation purposes and 75% for model training (Vaysse and Lagacherie, 2017; Guevara et al., 2018).

2.4.4. Modeling and validation

Due the fact that the study area does not have a history of the use of DSM techniques, the spatial distribution of the properties analyzed was modeled through machine learning algorithms, which are routinely used in soil property mapping (Malone et al., 2017). Random forest (randomForest) (Liaw and Wiener, 2002), fast implementation of random forest (ranger) (Wright and Ziegler, 2017), and support vector machine (svmLinear) (Karatzoglou et al., 2006) models were used. In addition, the ranger and svmLinear models were incorporated into a generalized linear model (glm) (Hengl et al., 2017). This was done using the caretEnsemble package, where the weight of each proposed model was assigned according to the root mean square error (RMSE) obtained individually in a cross validation of 5 repeats in the generation of each model with 5 subsets of different data (Guevara et al., 2018).

After the surfaces were interpolated with the models mentioned above, external validation was performed in order to assess the quality based on the prediction error. The quantitative statistics used included the coefficient of efficiency (COE), index of agreement (IOA), amount of variance explained (AVE), and RMSE. The COE, defined by Legates and McCabe (1990), is the complement ratio of the mean squared error to the variance of the observed data. Presents a range between $-\infty$ and 1, where values close to 1 indicate high efficiency, negative values indicate the level of inefficacy or uselessness of the model, and zero

Table 3
Method and units for each property mapped.

Property	Method	Unit	Number of samples for mapping
Al	KCl 1 N	cmol (+) kg^{-1}	1730
OM	Walkley and Black	%	1727
Mg	Ammonium acetate 1 N pH 7	cmol (+) kg^{-1}	1713
Ca	Ammonium acetate 1 N pH 7	cmol (+) kg^{-1}	1711
Na	Ammonium acetate 1 N pH 7	cmol (+) kg^{-1}	1709
P (available)	Bray II	mg kg^{-1}	1703
ECEC	Sum Na, Mg, Ca, K and Al	cmol (+) kg^{-1}	1702
pH	Potentiometer soil:water 1:1		1060
EC	Conductivity meter soil:water 1:5[1]	dS m^{-1}	954
Bd	Known-volume ring	Mg m^{-3}	894
Clay	Bouyoucos and pipette	%	853

[1] 301 soil samples measured in saturated paste were transformed into a soil:water 1:5 ratio extract with the Kargas et al. (2018) function.

Table 4
Environmental covariates used in the predictive model.

Formation factor	Number of covariates	Description	Source
Weather	71	1980–2011 monthly average values of precipitation, relative humidity, sunlight, minimum temperature, and average temperature.	Climatological database, 1980 – 2011. IDEAM, 2015.
		Surface temperature of the earth, global evapotranspiration, and radiation.	Time series of land surface temperature (LST), evapotranspiration (ET), and MODIS radiation. Reuter and Hengl, 2012.
Biological organisms or land cover	11	Corine land cover classification categories (binary type).	Corine Land Cover Classification 1:100.000 2010–2012. IDEAM, 2014.
		Vegetation indices.	Time series of the enhanced vegetation index (EVI) and MODIS LAI (leaf area index). Reuter and Hengl, 2012.
Topography	28	Digital elevation model (derived parameters).	SRTM mission of 2000, at 30 m.
		Physiographic landscape and topography.	Soil map, scale 1:100.000. IGAC, 2012.
Parent material	10	Lithology with correction for source materials reported in soil studies.	Soil map, scale 1:100.000. IGAC, 2012.
Time	8	Order of soils as a relation to the progression of time.	Soil map, scale 1:100.000. IGAC, 2012.

indicates that the model cannot predict values different from the observed mean. The revised IOA proposed by Willmott et al. (2012) indicates the relationship between the sum of the error between the predicted and observed variables and the sum of the deviation of the observed variable from its mean value. This index can take values between −1 and 1, where a value of 1 means a perfect model, in which the sum of the errors is 0. The AVE measures the fraction of the overall dispersion of the observed values that is explained by the model, with an optimal value of 1 (Samuel-Rosa et al., 2015). RMSE is a measure of prediction accuracy, since it has the same unit of measurement as the mapped property and can, therefore, more easily be compared to it (Ramos et al., 2017). A perfect model would have a RMSE of ~0. For the evaluation of the properties with the best results, the RMSE-observations standard deviation ratio (RSR) was used, a model evaluation statistic developed by Moriasi et al. (2007). RSR varies from the optimal value of 0, which indicates zero RMSE or residual variation and therefore perfect model simulation, to a large positive value.

One of the graphical methods used was the Taylor diagram, which summarizes multiple aspects of model performance, such as the agreement and variance between observed and predicted values (Taylor, 2001). The diagram provides a way of showing how three complementary model performance statistics vary simultaneously. These statistics are the Pearson correlation coefficient (r), the centered pattern RMS error, and the standard deviation (sigma). The correlation coefficient is shown by the angle (azimuth), from perfect (0°) to none (90°). The higher the correlation, the closer the predictions match the original values. The centered RMS error in the predicted pattern is proportional to the distance from the point on the x-axis (standard deviation of the actual pattern) marked with a purple dot labelled “observed”. The further from the that point, the higher the centered RMS. The standard deviation of the predicted pattern is proportional to the radial distance from the origin; ideally, this would be on the arc from the standard deviation of the reference set. If the points are inside this arc, the prediction under-represents the variability of the original data. The ideal model will be coincident with the point on the x-axis representing the “original” dataset. So the closer a point is to that, the better it is. For the qualitative variables, a confusion or error matrix was generated in order to calculate, based on the generated categories, the user's, producer's, and overall accuracies (Congalton, 1991).

Once the best model was selected, uncertainty prediction was estimated with the validation data set (25% of the total data). The prediction error was calculated at each point as the difference between the observed and the predicted value (Brus et al., 2011). Due to the fact that the selection of that set did not respond to a probabilistic sampling, nor was it based on the spatial variability of the covariates used, the comparison of the methods was not applied to all the soils in the study area, but only to those represented by the set of validation sampling.

Subsequently, the independent residual model was calculated (where the predicted error was the target variable) (Guevara et al., 2018). This model was built with the quantile regression method (a kind of random forest for estimation of conditional quantiles) within the quantregForest function (Meinshausen, 2017). Through this function, the full conditional response of these residuals was estimated for each pixel. Finally, the mean of the errors was calculated at each pixel and plot in order to show the approximated error trend of the estimations of each study variable.

2.5. Geographical web services

Because the SIS was designed for widespread use by decision makers, part of the SSDB and all the properties analyzed were obtained from geographical web services, following the specifications defined by the Open Geospatial Consortium (OGC). The services included a web mapping service (WMS) for all layers; a web feature service (WFS) for vectoral layers; and a web coverage service (WCS) for the layers of the properties analyzed. The generation of the services was carried out using the open data server GeoServer v. 2.15.0.

3. Results and discussion

3.1. Cundiboyacense high plateau and its soils

The geographic location of the high plateau and its socio-ecological relationship with Bogota have caused the agricultural sector to require more and better information on biophysical resources such as soil, in order to analyze topics such as food security and climate change. These analyses would allow stakeholders and the academic community to improve the environmental and territorial management of this region.

Eight soil orders defined by USDA Soil Taxonomy (Survey Soil Staff, 2014) were identified in the high plateau: Alfisols, Andisols, Entisols, Histosols, Inceptisols, Mollisols, Ultisols, and Vertisols. Inceptisols and Andisols are predominant in the study area. Inceptisols are the most representative order and are characterized by their incipient evolution; in the study area, these soils originated from sedimentary rocks and alluvial and colluvial-alluvial deposits. These soils are found in mountains, hills, valleys, and floodplains and have Andic and Pachic intergrades with ustic, udic, and aquic moisture regimes. Andisols have a higher level of evolution and originated from volcanic ash deposits on the floodplains of the study area; these soils are characterized by an acidic pH and low bulk density (IGAC, 2015).

Table 5
Descriptive statistics of soil properties in the analyzed topsoil of the SSDB.

Soil parameter	Mean	Min	1st Qu	Median	3rd Qu	Max	SD	Skewness	Kurtosis
Clay (%)	33.80	2.20	21.00	33.80	48.18	87.18	17.04	0.19	−0.72
Bd (Mg m^{-3})	0.90	0.10	0.64	0.89	1.15	1.81	0.34	0.15	−0.74
pH	5.19	3.00	4.60	5.10	5.60	7.90	0.70	0.77	0.70
OM (%)	11.88	0.17	5.10	8.30	15.70	68.40	9.79	1.71	3.29
P (ppm)	32.39	0.32	6.35	12.70	36.82	682.94	54.88	4.77	32.51
Ca (cmol (+) kg^{-1})	7.22	0.01	2.40	4.94	9.50	68.80	7.52	2.81	11.66
Mg (cmol (+) kg^{-1})	2.16	0.02	0.71	1.42	3.02	24.80	2.15	2.61	12.65
Na (cmol (+) kg^{-1})	0.80	0.01	0.07	0.19	0.58	17.27	1.86	4.68	26.36
Al (cmol (+) kg^{-1})	1.39	0.00	0.00	0.60	1.80	31.00	2.39	4.82	39.16
ECEC (cmol (+) kg^{-1})	12.43	1.34	5.77	9.60	15.67	92.50	9.57	2.38	8.96
EC (dS m^{-1})	0.32	0.04	0.16	0.23	0.36	4.51	0.37	6.12	50.69

Bd, bulk density; OM, organic matter; ECEC, effective cation exchange capacity; EC, electric conductivity.

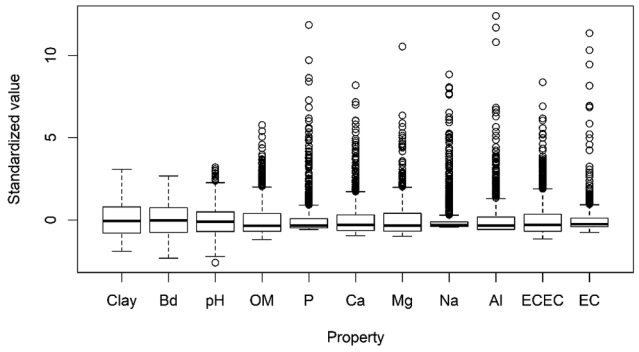

Fig. 3. Boxplot of soil properties in the analyzed topsoil of the SSDB The thick line indicates the median, the box indicates Q1 and Q3, and the whiskers Q1 and Q3 plus multiplicate 1.5 the interquartile range.

3.2. Exploratory data analysis

Of the 55 properties considered in the SSDB, only 11 were analyzed in the topsoil and are described in Table 5. Fig. 3 shows the variability and the mild and extreme outliers in the boxplots of the properties. They were previously standardized. The physical properties (Bd and clay) did not show outliers and had a lower skewness than the chemical properties. Al, EC, Mg, and P exhibited the highest outlier values.

According to the data analyzed, most soils of the high plateau belong to the fine-loamy textural family (18% to 35% clay) with very low Bd (< 1 Mg m^{-3}), pH between strong and extremely acidic (≤ 5.5), high OM ($\geq 10\%$), non-saline (EC $\leq$ 2 dS m^{-1}), medium ECEC, with high P and Ca contents, medium Mg and Na contents, and low Al content (ICA, 1992).

3.3. Environmental covariates

The correlation coefficients between the studied soil properties and the environmental covariates showed that no property had correlations < -0.5 or > 0.5. The highest and most significant correlations (p-value < 0.05) were found between the maximum and minimum temperatures with chemical properties such as Al (-0.40), Ca (0.35), OM (-0.34), and Mg (0.33). Covariates derived from the digital elevation model (DEM), such as the topographic wetness index (TWI), surface texture (TEX), slope (SL), and relative slope position (RSP) showed significant correlations (p-value < 0.05) with properties such as Ca (0.35 TWI, −0.38 TEX, −0.32 SL, −0.34 RSP), Mg (0.44 TWI, −0.48 TEX, −0.39 SL, −0.36 RSP), Na (0.30 TWI, −0.36 TEX, −0.25 SL, −0.21 RSP), and Al (−0.27 TWI, 0.24 TEX, 0.31 SL, 0.44 RSP). By contrast, the properties that exhibited the lowest correlations with the covariates used were clay, pH, and P.

The near-zero variance diagnosis indicated that the covariates with less variation in the study area were those related to the presence or absence of deposits of very fine materials and mixtures with organic deposits; landforms such as glacial valleys, valleys, ridges, and hills; areas covered with permanent and annual crops; and the presence or absence of Alfisols, Mollisols, Ultisols, and Entisols. In addition, the high correlations among the continuous predictors ($r > 0.9$) were those associated with weather, such as the maximum and minimum temperature for all months of the year, sunshine for 5 months, and relative humidity for 10 months a year, and some covariates derived from the DEM, such as slope, the RSP, and distance to drainage. With the above procedures, the number of covariates was reduced from 128 to 57.

3.4. Modeling and spatial validation

For each of the analyzed properties, three models were constructed (randomForest, ranger, and svmLinear), and the ensemble model was built from them. Table 6 shows the results of the selected indices.

The model with the highest COE was the randomForest model for Bd (0.67), and the model with the lowest COE was randomForest for EC (0.13). In general, the models generated using svmLinear had the lowest COE values, which indicates that this technique is less efficient for the properties in the study area (Legates and McCabe, 2013). The IOA had values with a range from 0.84 for Bd with the randomForest model to 0.56 for EC with the svmLinear model, which indicates that in both cases the sum of the errors is less than half the sum of the deviations of the observed data or of the deviation from the perfect model (Willmott et al., 2012). For the AVE, the highest value was also obtained for Bd with the randomForest and ensemble models (0.86), while the lowest (0.22) was for P and CE with the ranger and randomForest models, respectively. The AVEs obtained for the variables clay and ECEC were very similar to those recorded by Samuel-Rosa et al. (2015), and in general they have models with acceptable to poor performance; however, Bd stands out, because its AVE was higher than that reported by Hengl et al. (2014, 2017). The RMSE had values ranging from 0.43 (CE with svmLinear model) to 34.07 (P with svmLinear model). For some properties, for example pH, the RMSE for randomForest model was 0.57, less than that found in France (0.78) by Vaysse and Lagacherie (2015) and in Portugal (0.65) by Ramos et al. (2017).

The Taylor diagrams (Fig. 4) show goodness of fit. In general, it was observed that the svmLinear model predicted values farther from the original data set. Importantly, the deviations of the predicted data sets for each model for each property in all cases showed a standard deviation lower than that of the original data set, with the values concentrated around the mean prediction of each property.

According to the RSR, the property with the best index was Bd (0.37). Following this were P (0.44), Clay (0.51), Mg and Na (0.61), OM (0.63), ECEC (0.77), pH (0.81), Ca (0.84), and Al (0.86). The worst-modeled property was EC (1.16). Bd most closely matched the pattern between the observed and modeled data. Its good results were the product of a low variation in its data. It did not show outliers, and it had

Table 6
Indices for the models tested for the modeled properties.

Property	Model	COE	IOA	AVE	RMSE	Property	Model	COE	IOA	AVE	RMSE
Clay	randomForest	0.57	0.79	0.72	8.61	Mg	randomForest	0.38	0.69	0.49	1.35
	ranger	0.46	0.73	0.60	10.31		ranger	0.44	0.72	0.51	1.32
	svmLinear	0.34	0.67	0.41	12.53		svmLinear	0.40	0.70	0.45	1.39
	ensemble	0.55	0.77	0.71	8.82		ensemble	0.39	0.69	0.49	1.34
Bd	randomForest	0.67	0.84	0.86	0.13	Na	randomForest	0.51	0.76	0.68	1.13
	ranger	0.66	0.83	0.85	0.13		ranger	0.57	0.78	0.64	1.20
	svmLinear	0.42	0.71	0.58	0.22		svmLinear	0.52	0.76	0.53	1.36
	ensemble	0.67	0.84	0.86	0.13		ensemble	0.51	0.75	0.67	1.15
pH	randomForest	0.29	0.64	0.46	0.57	Al	randomForest	0.26	0.63	0.37	2.05
	ranger	0.28	0.64	0.45	0.57		ranger	0.25	0.63	0.35	2.07
	svmLinear	0.20	0.60	0.31	0.64		svmLinear	0.29	0.64	0.28	2.18
	ensemble	0.29	0.65	0.46	0.57		ensemble	0.27	0.63	0.35	2.07
OM	randomForest	0.40	0.70	0.57	6.19	ECEC	randomForest	0.30	0.65	0.38	7.47
	ranger	0.42	0.71	0.52	6.60		ranger	0.37	0.68	0.39	7.40
	svmLinear	0.36	0.68	0.42	7.23		svmLinear	0.42	0.71	0.32	9.76
	ensemble	0.42	0.71	0.57	6.20		ensemble	0.32	0.66	0.38	7.43
P	randomForest	0.30	0.65	0.39	24.38	EC	randomForest	0.13	0.56	0.22	0.45
	ranger	0.32	0.66	0.22	27.54		ranger	0.16	0.58	0.26	0.44
	svmLinear	0.28	0.64	0.25	34.07		svmLinear	0.28	0.64	0.27	0.43
	ensemble	0.32	0.66	0.39	24.38		ensemble	0.16	0.58	0.25	0.44
Ca	randomForest	0.32	0.66	0.40	6.31						
	ranger	0.35	0.68	0.35	6.58						
	svmLinear	0.33	0.67	0.30	6.83						
	ensemble	0.33	0.67	0.39	6.35						

a good fit with the evaluated models. However, in the case of EC, the results were poor, as a consequence of the highest outlier values and the harmonization of the laboratory method of part of the data. In general, the results obtained after the analysis showed that the efficiency of the models is completely acceptable.

According to the goodness of fit metric and Taylor diagrams, only one model was selected for each property: clay – randomForest (RMSE = 8.61), Bd – randomForest (RMSE = 0.13), pH – ensemble (RMSE = 0.57), OM – ensemble (RMSE = 6.20), P – randomForest (RMSE = 24.38), Ca – randomForest (RMSE = 6.31), Mg – ranger (RMSE = 1.32), Na – randomForest (RMSE = 1.13), Al – randomForest (RMSE = 2.05), ECEC – ranger (RMSE = 7.40), and EC – ranger (RMSE = 0.43). For the texture class, 3 models were tested: randomForest, ranger, and svmLinear, with an overall accuracy of 0.86, 0.48, and 0.43, respectively. The model selected was randomForest, due to its better performance. Table 7 shows the producer's and user's accuracies for the randomForest model.

Although according to the Taylor diagrams there were properties with very similar results among the models (e.g. Ca, Mg, and EC), the selection was based on the models with the easiest interpretation (randomForest and ranger) versus the more complex ones (svmLinear and ensemble) (Heung et al., 2016).

Of the 12 properties, in 10 randomForest (including ranger) was the selected method, according to that proposed by several studies (Vaysse and Lagacherie, 2015; Heung et al., 2016; Nussbaum et al., 2018). The randomForest effectiveness came from recording training data in the range of values of the covariates and from better representing complex nonlinear relationships (Hengl et al., 2017). However, the performance of the techniques on the study area was highly variable from one property to another, similar to that reported by Malone et al. (2009) and Vaysse and Lagacherie (2015). In future research on DSM, the differences found should be deepened, emphasizing other techniques and their selection (Heung et al., 2016).

The clay content prediction (Fig. 5a) ranged from 5.8% to 66.7%, and the lowest content was seen in moor areas, where alterites from sedimentary rock (sandstone) were reported as parent material (IGAC, 2004). The highest clay contents were found in areas that provide predominantly fine alluvial materials.

Clay content in the soil is correlated with the retention of water and nutrients in the soil (Koch et al., 2016; Lipiec et al., 2018); therefore, sites with high clay content also have a high ECEC (Fig. 5s). However, sites with a high ECEC can also correspond to areas with high exchangeable Ca content (Fig. 5k), most likely because the sedimentary rock that produced the soil is carbonated (IGAC, 2000). Areas with thick textures but with high OM content tend to have a high ECEC. This can be explained by the fact that organic colloids have a higher CEC than do mineral colloids such as clays (Jaramillo, 2002). With regard to exchangeable Ca content, values between 0.1 and 34.5 cmol (+) kg^{-1} were mapped, and most had low to medium content of this cation (ICA, 1992). High Ca contents are associated with high pH values (Behera and Shukla, 2015; Pérez de los Reyes et al., 2015); this occurs in soils of the study area because sites with a high Ca content also had a pH close to 7 (Fig. 5e), and it is in these areas where growers correct acidic pH with lime-based chemicals, which have high Ca and Mg contents.

As for OM content in the soil, values between 1% and 54% were mapped (Fig. 5g), with greater percentages seen in the soils on the periphery of the study area, which are places at higher elevation that correspond to moor ecosystems, where the OM mineralization rate is lower (IGAC, 2015). Additionally, most of these places correspond to sites with an acidic pH (Fig. 5e) due to the conservation of humic acids in the soil (IGAC, 2015). The exchangeable Mg content ranged from 0 to 9.7 cmol (+) kg^{-1} (Fig. 5m), with higher values in farming areas and where human intervention has changed the natural soil conditions due to practices such as fertilization and liming. In Fig. 5o, the same condition is seen for exchangeable Na content in the Bogotá savanna.

The development of the structure of the soil is supported by the OM content (Yang et al., 2016), and it is therefore expected that areas with high OM content will have an adequate soil Bd (from 1.2 to 1.4 Mg m^{-3}) (IGAC, 2015). Moreover, those sites in Fig. 5c where Bd values are close to 0.23 g cm^{-3} have soils with organic horizons from the moor ecosystem or from a lacustrine origin mainly composed of plant materials in different states of decomposition. In addition, the locations with higher Bd values correspond to Andisols with vertic groups (with sliding crack surfaces).

Regarding the EC of the soil, values between 0 and 2.8 ds m^{-1} were mapped (Fig. 5u), which demonstrates the presence of soils that are saline to strongly saline, although to a lesser extent. High EC values of the soil are found in growing regions where excessive application of fertilizers (Fig. 5k and m) can generate salinity problems, which leads to soil degradation, crop loss, water quality deterioration, and

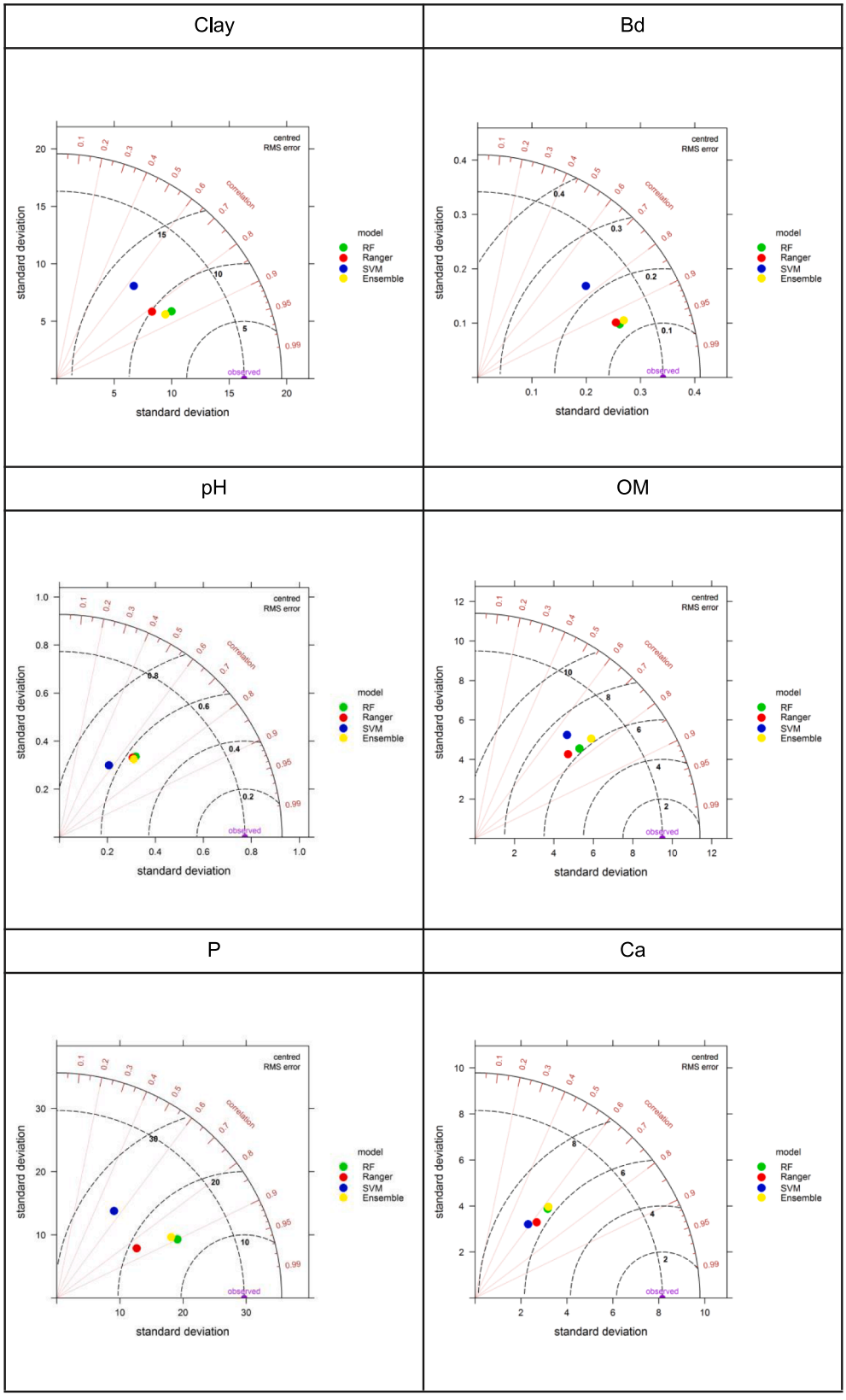

Fig. 4. Taylor diagrams of the high plateau soil properties analyzed (RF – randomForest, SVM – svmLinear).

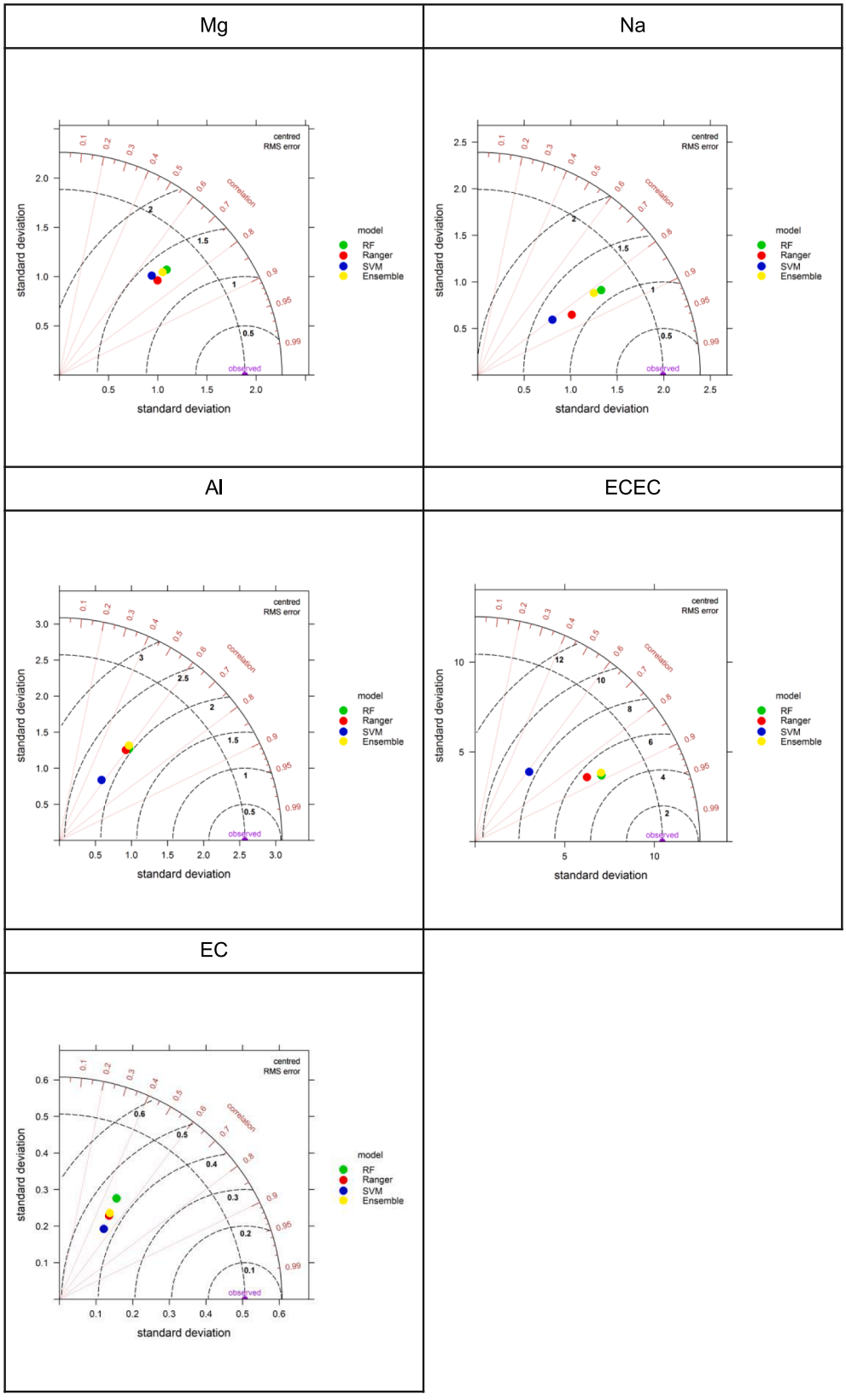

Fig. 4. (*continued*)

Table 7
Producer's and user's accuracies for the texture class.

Accuracy	LoSa	Cl	SiCl	Lo	SaLo	ClLo	SaClLo	SiClLo	SiLo
Producer's	100	94	100	83	77	80	82	80	82
User's	100	82	100	88	89	85	100	67	100

LoSa, loamy sand; Cl, clay; SiCl, silty clay; Lo, loam; SaLo, sandy loam; ClLo, clay loam; SaClLo, sandy clay loam; SiClLo, silty clay loam; SiLo, silt loam.

environmental degradation in general (Kitamura et al., 2006). The available P content in the study region ranged from 2 to 130 mg kg^{-1} (Fig. 5i). These higher P contents are seen primarily in locations with Andisols or Inceptisols with andic intergrades, characterized by their high phosphate retention and ability to be used to sow crops such as potato, where constant applications of P are carried out and can lead to an increase in the content of this element in the soil.

In regard to texture, sandy loam (SaLo) was the most representative texture class, which is related to the parent material that predominates in the study area and corresponds to sedimentary rocks (sandstone) (Fig. 5w). In the floodplain, the predominant textures range from fine to very fine (SaClLo, ClLo, Cl), which originate from the alluvial and lacustrine deposits in the region.

Some generated maps are not exempt from artefacts in some places (round spot areas), especially the mean error prediction of some variables (e.g. Fig. 5b, l, t, and v). Slight changes in the input feature can have a big impact on the predicted outcome, which is usually not desirable (Molnar, 2019). These round spots are there due to the high sensitivity of the randomForest model to artefacts in the input data (Hengl et al., 2015). In this case, some raster input of the environmental covariates stack (e.g. relative humidity and sunlight) had several circular areas.

Other factors that explain the results obtained with the DSM models come from the data set used to make the maps. As indicated before, the inspection density was considered low, and although no spatial autocorrelation analyses were performed, the high heterogeneity of the soils was not adequately recorded in the data set. This is because a large part of the data comes from studies carried out by IGAC, which had a soil classification approach, and their sites were selected in terms of defining geopedological units than determining variability over short distances. It should be remembered that the data were also collected at different times (from 2000 to 2016). This could be solved by incorporating more soil studies from the study area or by collecting new information from an adequate sample design built from the space feature of the covariates used in the study.

3.5. IRAKA web services

The maps of the soil properties modeled and error trends were stored in the IRAKA information system and are available at https://iraka.agrosavia.co/. Maps are available in WCS and WMS format. Additionally, IRAKA has a geographic viewer that allows one to query, view, and download the modeled soil observations. The system provides a view of the soil orders of the study area according to USDA soil taxonomy and generated based on the representativeness of the modal profiles described by the mapping units. For all the maps, IRAKA generates a general overview, and for the interpolations, the models and goodness of fit units are included. This information system is also associated with the opensource cataloging tool Geonetwork for resources referenced to the geographic space. The geographic metadata (according to the ISO 19115 and the Colombian Technical Rule – NTC 4611) is in Geonetwork. This is a tool that is in line with the Colombian Spatial Data Infrastructure (ICDE). Additional information on handling the tool can be found at AGROSAVIA (2020).

Compared to other web services, IRAKA has several advantages: unlike SoilGrids, IRAKA uses local information from different sources, adjusted to the characteristics of the high plateau. In contrast to SISLAC, IRAKA has maps with information about chemical and physical properties. SIGA GIS only includes soil information in polygons, while IRAKA has raster representation from DSM techniques. In the case of GEOSOIL, whose results are for a municipality, IRAKA presents information for an entire natural subregion in Colombia. Its disadvantages are: the information is only available for the topsoil, while other systems present it at different depths (e.g. SoilGrids, SISLAC). There is very little information on physical properties, only the texture and bulk density. Otherwise, to facilitate handling, and network time server in the geographic viewer, the information was represented to 250 m, although it was built to 125 m. Though IRAKA allows multiple users to view the information, the web geographic systems are intended for users who can make use of it in a GIS, in order to carry out geospatial operations or to include it in land management, food security, or climate change analyses. For example, an important element within IRAKA is the possibility to produce maps of soil quality or suitability using the results of the mapping. In the case of crops like bulb and long green onion, the system has maps of soil quality indices. These maps were generated from the products of IRAKA and other information on the absorption of nutrients.

4. Conclusions

IRAKA is the result of building an SIS based on information available for the Cundiboyacense high plateau of Colombia. Its structure comprises 3 aspects: data from different sources stored in spatial databases, administrated through management systems that enable database creation, queries, updates or deletions; spatial modeling through digital mapping techniques, which generate error trend surfaces; and provision of information through geographic web services for online consumption or through a GIS. Its advantages over other information systems are the use of local information, the analysis of different DSM models, and distribution on raster and vector layers. Its disadvantage is that only topsoil information is presented.

The spatial modeling in this study used machine learning techniques such as random forests, support vector machines, and their incorporation into an ensemble through a generalized linear function. Random forest (including ranger technique) gave the best results, due to the fact that it was used in 10 of the 12 mapped soil properties. A special random forest technique (conditional quantiles) was also used to calculate uncertainty, and this allowed obtaining the mean of the error for each property in the study area.

IRAKA was developed to integrate available information, standardize the information according to national procedures, and analyze the information according to its characteristics. Therefore, it is a tool that easily presents updated soil information for planning and managing the edaphic resource of the Cundiboyacense high plateau. This system represents an important step towards integration with information systems that address other biophysical resources such as climate, land cover, streams, etc. Although IRAKA was based on a specific natural subregion, its focus, methodologies, and development can be extended to other areas and can have nationwide coverage.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

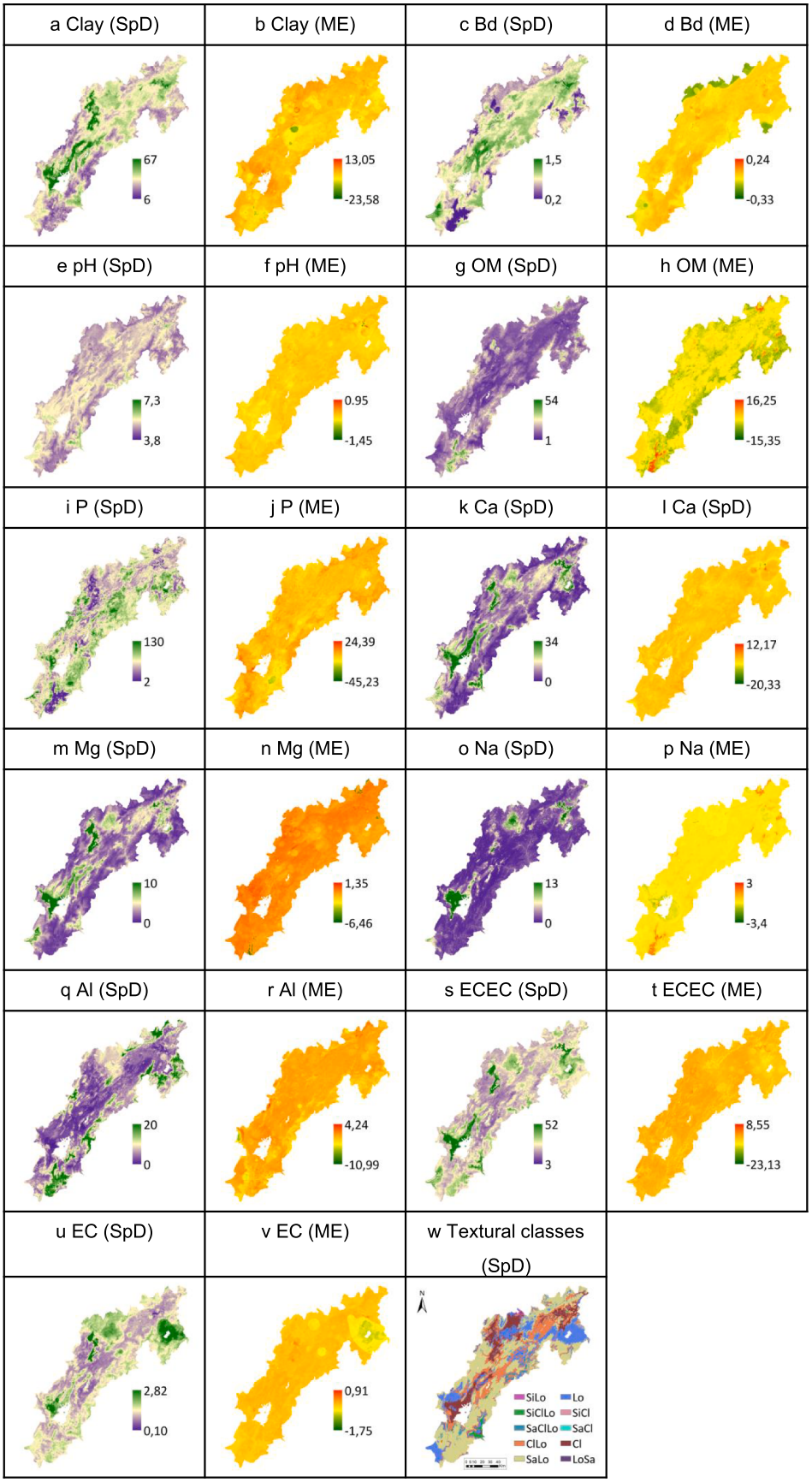

Fig. 5. Spatial distribution (SpD) and mean error (ME) of soil variables.

Acknowledgments

The authors wish to thank the Instituo Geográfico Agustín Codazzi – IGAC, especially the Agrology Branch, for providing the data. This study was carried out under the framework of the project "Implementation of a Soil Information System for long green and bulb onion crops in the Cundiboyacense High Plateau - Colombia", developed by the Corporación Colombiana de Investigación Agropecuaria - AGROSAVIA and funded by KoLFACI (Korean – Latin American Food & Agriculture Cooperation Initiative), agreement M490-1 of 2017. We are thankful to the editor and the referees for their numerous helpful suggestions, which improved the manuscript.

References

Agronet, 2020. Anuario Estadístico del sector Agropecuario – EVA. https://www.agronet.gov.co/estadistica/Paginas/home.aspx?cod=59 (accessed 08 May 2020).

AGROSAVIA, 2020. Sistema de información de suelos del altiplano cundiboyacense – IRAKA. https://iraka.agrosavia.co/Content/archivos/Manual_usuario_IRAKA_v2.pdf (accessed 29 January 2020).

Arrouays, D., McKenzie, N.J., Hempel, J.W., Richer de Forges, A.C., McBratney, A.B., 2014. GlobalSoilMap: Basis of the Global Spatial Soil Information System, first ed. Taylor & Francis Group, London.

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

Arrouays, D., Leenaars, J., Richer de Forges, A., Adhikari, K., Ballabio, C., Greve, M., et al., 2017b. Soil legacy data rescue via GlobalSoilMap and other international and national initiatives. GeoResJ. 14, 1–19. https://doi.org/10.1016/j.grj.2017.06.001.

Avery, B., 1987. Soil Survey Methods: A Review. Technical Monograph No. 18. Silsoe: Soil Survey & Land Resource Centre.

Barbat, J.M.B., Timm, L.C., Pauletto, E.A., Sousa, R.O.D., Castilhos, D.D., Ávila, C.L.D., Reckziegel, N.L., 2009. Spatial variability of the chemical, physical and biological properties in lowland cultivated with irrigated rice. Rev. Bras. Ciênc. Solo. 33 (4), 819–830. https://doi.org/10.1590/S0100-06832009000400007.

Batjes, N.H., 2009. Harmonized soil profile data for applications at global and continental scales: Updates to the WISE database. Soil Use Manage. 25, 124–127. https://doi.org/10.1111/j.1475-2743.2009.00202.x.

Behera, S.K., Shukla, A.K., 2015. Spatial distribution of surface soil acidity, electrical conductivity, soil organic carbon content and exchangeable potassium, calcium and magnesium in some cropped acid soils of India. Land Degrad. Dev. 26 (1), 71–79. https://doi.org/10.1002/ldr.2306.

Bishop, T.F.A., McBratney, A.B., Laslett, G.M., 1999. Modelling soil attribute depth functions with equal-area quadratic smoothing splines. Geoderma 91, 27–45. https://doi.org/10.1016/S0016-7061(99)00003-8.

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

Cambule, A.H., Rossiter, D.G., Stoorvogel, J.J., 2013. A methodology for digital soil mapping in poorly-accessible areas. Geoderma 192, 341–353. https://doi.org/10.1016/j.geoderma.2012.08.020.

Camps, R., Casillas, L.A., Costal, D., Ginestà, M.G., Martín, C., Pérez, O., 2005. Bases de datos. Versión 1. Fundació per a la Universitat Oberta de Catalunya. Eureca Media SL, Barcelona.

Congalton, R.G., 1991. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sens. Environ. 37 (1), 35–46. https://doi.org/10.1016/0034-4257(91)90048-B.

Dawson, R., 2011. How significant is a boxplot outlier? J. Stat. Educ. 19, 2. https://doi.org/10.1080/10691898.2011.11889610.

Dijkshoorn, J.A., Huting, J.R.M., Tempel, P., 2005. Update of the 1:5 million Soil and Terrain Database for Latin America and the Caribbean (SOTERLAC; version 2.0). Report 2005/01. ISRIC - World Soil Information, Wageningen.

Dwivedi, R.S., 2017. Soil information systems. In: Remote Sensing of Soils. Springer, Berlin Heidelberg, pp. 359–398. https://doi.org/10.1007/978-3-662-53740-4_8.

Engelen, V.M.P. van, Dijkshoorn, J.A., 2013. Global and National Soils and Terrain Databases (SOTER). Procedures Manual, Version 2.0. ISRIC – World Soil Information, Wageningen.

FAO, 2006. Guidelines for Soil Profile Description, fourth ed. FAO, Rome.

FAO, 2013. State of the Art Report on Global and Regional Soil Information: Where are we? Where to go? Global Soil Partnership. Technical report. FAO, Rome.

FAO, 2018. Soil Organic Carbon Mapping Cookbook, second ed. FAO, Rome.

FAO/IIASA/ISRIC/ISS-CAS/JRC, 2012. Harmonized World Soil Database (version 1.2). FAO, Rome.

Gray, J.M., Humphreys, G.S., Deckers, J.A., 2009. Relationships in soil distribution as revealed by a global soil database. Geoderma 150, 309–323. https://doi.org/10.1016/j.geoderma.2009.02.012.

Guevara, M., Olmedo, G.F., Stell, E., Yigini, Y., Aguilar Duarte, Y., Arellano Hernández, C., et al., 2018. No silver bullet for digital soil mapping: country-specific soil organic carbon estimates across Latin America. Soil 4 (3), 173–193. https://doi.org/10.5194/soil-4-173-2018.

Hartemink, A.E., Krasilnikov, P., Bockheim, J.G., 2013. Soil maps of the world. Geoderma 207–208, 256–267. https://doi.org/10.1016/j.geoderma.2013.05.003.

Hendriks, C.M.J., Stoorvogel, J.J., Lutz, F., Claessens, L., 2019. When can legacy soil data be used, and when should new data be collected instead? Geoderma 348, 181–188. https://doi.org/10.1016/j.geoderma.2019.04.026.

Hengl, T., 2006. Finding the right pixel size. Comput. Geosci. UK 32 (9), 1283–1298. https://doi.org/10.1016/j.cageo.2005.11.008.

Hengl, T., Mendes de Jesus, J., MacMillan, R.A., Batjes, N.H., Heuvelink, G.B.M., Ribeiro, E., et al., 2014. SoilGrids1km - global soil information based on automated mapping. PLoS ONE 9 (12), e105992. https://doi.org/10.1371/journal.pone.0105992.

Hengl, T., Heuvelink, G.B.M., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Shepherd, K.D., et al., 2015. Mapping soil properties of Africa at 250 m resolution: random forests significantly improve current predictions. PLoS ONE 10 (6), e0125814. https://doi.org/10.1371/journal.pone.0125814.

Hengl, T., Mendes de Jesus, J., Heuvelink, G.B.M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., et al., 2017. SoilGrids250m: Global gridded soil information based on machine learning. PLoS ONE 12 (2), e0169748. https://doi.org/10.1371/journal.pone.0169748.

Heung, B., Ho, H.K., Zhang, J., Knudby, A., Bulmer, C.E., Schmidt, M.G., 2016. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. Geoderma 265, 62–77. https://doi.org/10.1016/j.geoderma.2015.11.014.

ICA, 1992. Fertilización en diversos cultivos. Quinta aproximación. Manual de Asistencia Técnica No. 25. Centro de Investigación Tibaitatá. Instituto Colombiano Agropecuario, Bogotá.

IDEAM, 2014. Mapa de Coberturas de la Tierra Metodología Corine Land Cover Adaptada para Colombia Escala 1:100.000 (Período 2010 - 2012). Instituto de Hidrología, Meteorología y Estudios Ambientales, Bogotá.

IDEAM, 2015. Climatological atlas of Colombia – Interactive year 2015. accessed 30 September 2019. http://atlas.ideam.gov.co/visorAtlasClimatologico.html.

IGAC, 2000. Estudio general de suelos y zonificación de tierras del departamento de Cundinamarca. Subdirección de agrología. Instituto Geográfico Agustín Codazzi, Bogotá.

IGAC, 2004. Estudio general de suelos y zonificación de tierras del departamento de Boyacá. Subdirección de agrología. Instituto Geográfico Agustín Codazzi, Bogotá.

IGAC, 2012. Conflictos de Uso del Territorio Colombiano. Escala 1:100.000. Convenio marco de cooperación especial. Instituto Geográfico Agustín Codazzi, Bogotá.

IGAC, 2014. Códigos para los levantamientos de suelos. Grupo interno de trabajo de levantamientos agrológicos. Instituto Geográfico Agustín Codazzi, Bogotá.

IGAC, 2015. Suelos y Tierras de Colombia. Subdirección de Agrología. Instituto Geográfico Agustín Codazzi, Bogotá.

IUSS Working Group WRB, 2015. World Reference Base for Soil Resources 2014, Update 2015 International Soil Classification System for Naming Soils and Creating Legends for Soil Maps. World Soil Resources Reports No. 106. FAO, Rome.

Jaramillo, D.J., 2002. Introducción a la ciencia del suelo. Facultad de Ciencias, Universidad Nacional de Colombia, Medellín.

Jenny, H., 1941. Factors of Soil Formation: A System of Quantitative Pedology. McGraw-Hill, New York.

Karatzoglou, A., Meyer, D., Hornik, K., 2006. Support vector algorithm in R. J. Stat. Softw. 15, 1–28. https://doi.org/10.18637/jss.v015.i09.

Karavaeva, N.A., Gerasimova, M.I., 2005. World soil maps: the contribution of I.P. Gerasimov and recent advances. Eurasian. J. Soil Sci. 38 (12), 1309–1316.

Kargas, G., Chatzigiakoumis, I., Kollias, A., Spiliotis, D., Massas, I., Kerkides, P., 2018. Soil salinity assessment using saturated paste and mass soil:water 1:1 and 1:5 ratios extracts. Water 10, 1589. https://doi.org/10.3390/w10111589.

Kitamura, Y., Yano, T., Honna, T., Yamamoto, S., Inosako, K., 2006. Causes of farmland salinization and remedial measures in the Aral Sea Basin-Research on water management to prevent secondary salinization in rice-based cropping system in arid land. Agric. Water Manag. 85, 1–14. https://doi.org/10.1016/j.agwat.2006.03.007.

Koch, S., Kahle, P., Lennartz, B., 2016. Visualization of colloid transport pathways in mineral soils using titanium (IV) oxide as a tracer. J. Environ. Qual. 45, 2053–2059. https://doi.org/10.2134/jeq2016.04.0131.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., The R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C., Hunt, T., 2018. caret: Classification and Regression Training. https://CRAN.R-project.org/package=caret (accessed: 23 September 2019).

Lagacherie, P., McBratney, A.B., Voltz, M., 2006. Digital Soil Mapping. An Introductory Perspective. Developments in Soil Science, vol. 31. Elsevier, Amsterdam.

Legates, D.R., McCabe, G.J., 1990. Evaluating the use of "Goodness-of-fit" measures in hydrologic and hydroclimatic model validation. Water Resour. Res. 35 (1), 233–241.

Legates, D.R., McCabe, G.J., 2013. A refined index of model performance: e rejoinder. Int. J. Climatol. 33 (4), 1053–1056. https://doi.org/10.1002/joc.3487.

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

Lilburne, L.R., Hewitt, A.E., Webb, T.W., 2012. Soil and informatics science combine to develop S-map: A new generation soil information system for New Zealand. Geoderma 170, 232–238. https://doi.org/10.1016/j.geoderma.2011.11.012.

Lipiec, J., Czyż, E.A., Dexter, A.R., Siczek, A., 2018. Effects of soil deformation on clay dispersion in loess soil. Soil Tillage Res. 184, 203–206. https://doi.org/10.1016/j.still.2018.08.005.

Malone, B.P., McBratney, A.B., Minasny, B., Laslett, G.M., 2009. Mapping continuous depth functions of soil carbon storage and available water capacity. Geoderma 154, 138–152. https://doi.org/10.1016/j.geoderma.2009.10.007.

Malone, B.P., Minasny, B., McBratney, A.B., 2017. Using R for Digital Soil Mapping. Progress in Soil Science. Springer International Publishing Switzerland.

Massawe, B.H., Subburayalu, S.K., Kaaya, A.K., Winowiecki, L., Slater, B.K., 2018. Mapping numerically classified soil taxa in Kilombero Valley, Tanzania using machine learning. Geoderma 311, 143–148. https://doi.org/10.1016/j.geoderma.2016.11.020.

McBratney, A.B., Mendonca, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117 (1–2), 3–52. https://doi.org/10.1016/S0016-7061(03)00223-4.
Meinshausen, N., 2017. quantregForest: Quantile Regression Forests. R package. https://cran.r-project.org/web/packages/quantregForest/index.html (accessed: 23 September 2019).
Molnar, C., 2019. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, Leanpub. Creative Commons Attribution.
Moriasi, D.N., Arnold, J.G., Van Liew, M.W., Bingner, R.L., Harmel, R.D., Veith, T.L., 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Trans. ASABE 50 (3), 885–900. https://doi.org/10.13031/2013.23153.
Nussbaum, M., Spiess, K., Baltensweiler, A., Grob, U., Keller, A., Greiner, L., Schaepman, M.E., Papritz, A., 2018. Evaluation of digital soil mapping approaches with large sets of environmental covariates. Soil 4 (1), 1–22. https://doi.org/10.5194/soil-4-1-2018.
Olaya, V., 2011. Sistemas de Información Geográfica. Versión 1.0. Creative Commons Atribución.
Pérez de los Reyes, C., Ortíz Villajos, J.A., Navarro, F.G., Martín Consuega, S.B., 2015. Grapevine leaf uptake of mineral elements influenced by sugar foam amendment of an acidic soil. Vitis 52(4), 157–164, https://doi.org/10.5073/vitis.2013.52.157-164.
Pourabdollah, A., Leibovici, D.G., Simms, D.M., Tempel, P., Hallett, S.H., Jackson, M.J., 2012. Towards a standard for soil and terrain data exchange: SoTerML. Comput. Geosci. UK 45, 270–283. https://doi.org/10.1016/j.cageo.2011.11.026.
Ramakrishnan, R., Gehrke, J., 2007. Sistemas de gestión de bases de datos, tercera ed. McGraw-Hill/Interamericana de España, Madrid.
Ramos, T.B., Horta, A., Gonçalves, M.C., Pires, F.P., Duffy, D., Martins, J.C., 2017. The INFOSOLO database as a first step towards the development of a soil information system in Portugal. Catena 158, 390–412. https://doi.org/10.1016/j.catena.2017.07.020.
Reuter, H.I., Hengl, T., 2012. Worldgrids - a public repository of global soil covariates. In: B. Minasny, B., Malone, B.P., McBratney, A.B. (Eds.), Digital Soil Assessments and Beyond. Taylor & Francis Group, Sydney, pp. 287–292.
Rivera, D., Gómez, F., Goodhew, P., 2004. Altiplanos de Colombia, 1st ed. Banco de Occidente Credencial, Cali.
Rodríguez-Garay, F.A., Camacho-Tamayo, J.H., Rubiano-Sanabria, Y., 2016. Variabilidad espacial de los atributos químicos del suelo en el rendimiento y calidad de café. Cienc. Tecnol. Agropecuaria 17(2), 237–254. https://doi.org/10.21930/rcta.vol17_num2_art:493.
Rubiano, Y., Amézquita, E., Beaulieau, N., 2005. Sistema georreferenciado de indicadores de calidad de suelos para los Llanos Orientales de Colombia Estudio de caso: Municipio de Puerto López. Meta. Acta Agron. 54 (3), 1–10.
Samuel-Rosa, A., Heuvelink, G.B.M., Vasques, G.M., Anjos, L.H.C., 2015. Do more detailed environmental covariates deliver more accurate soil maps? Geoderma 243, 214–227. https://doi.org/10.1016/j.geoderma.2014.12.017.
Sánchez, P., Ahamed, S., Carré, F., Hartemink, A., Hempel, J., Huising, J., Lagacherie, P., McBratney, A., McKenzie, N., Mendonça-Santos, M., Minasmy, B., Montanarella, L., Okoth, P., Palm, C.A., Sachs, J., Shepherd, K., Vägen, T., Vanlauwe, B., Walsh, M., Winowiecki, L., Zhang, G., 2009. Digital soil map of the world. Science 325 (5941), 680–681. https://doi.org/10.1126/science.1175084.
Silberschatz, A., Korth, H.F., Sudarshan, S. 2002. Fundamentos de bases de datos, cuarta ed. McGraw-Hill Inc./Interamericana de España, Madrid.
Soil Science Division Staff, 2017. Soil survey manual. In: Ditzler, C., Scheffe, K., Monger, H.C. (Eds.), USDA Handbook 18. Government Printing Office, Washington.
Soil Survey Staff, 2014. Keys to Soil Taxonomy, 12th ed. USDA-Natural Resources Conservation Service, Washington.
Taylor, K.E., 2001. Summarizing multiple aspects of model performance in a single diagram. J. Geophys. Res. Atmos. 106 (D7), 7183–7192. https://doi.org/10.1029/2000JD900719.
Vaysse, K., Lagacherie, P., 2015. Evaluating Digital Soil Mapping approaches for mapping GlobalSoilMap soil properties from legacy data in Languedoc-Roussillon (France). Geoderma Reg. 4, 20–30. https://doi.org/10.1016/j.geodrs.2014.11.003.
Vaysse, K., Lagacherie, P., 2017. Using quantile regression forest to estimate uncertainty of digital soil mapping products. Geoderma 291, 55–64. https://doi.org/10.1016/j.geoderma.2016.12.017.
Willmott, C.J., Robeson, S.M., Matsuuraa, K., 2012. A refined index of model performance. Int. J. Climatol. 32, 2088–2094. https://doi.org/10.1002/joc.2419.
Wright, M.N., Ziegler, A., 2017. ranger: A fast implementation of random forests for high dimensional data in C++ and R. J. Stat. Softw. 77(1), https://doi.org/10.18637/jss.v077.i01.
Yang, Q.Y., Luo, W.Q., Jiang, Z.C., Li, W.J., Yuan, D.X., 2016. Improve the prediction of soil bulk density by cokriging with predicted soil water content as auxiliary. J. Soil Sediment. 16, 77–84. https://doi.org/10.1007/s11368-015-1193-4.
Zeraatpisheh, M., Ayoubi, S., Jafari, A., Tajik, S., Finke, P., 2019. Digital mapping of soil properties using multiple machine learning in a semi-arid region, central Iran. Geoderma 338, 445–452. https://doi.org/10.1016/j.geoderma.2018.09.006.