Colombian soil texture: Building a spatial ensemble model

Viviana Marcela Varón-Ramírez[1], Gustavo Alfonso Araujo-Carrillo[1], and Mario Guevara[2-4]

[1]Corporación Colombiana de Investigación Agropecuaria - AGROSAVIA, Mosquera-Cundinamarca, Colombia
[2]Centro de Geociencias - Universidad Nacional Autónoma de México Campus Juriquilla, Qro. MX.
[3]University of California, Riverside, Department of Environmental Sciences, Riverside CA. 92507, USA
[4]United States Department of Agriculture, U.S. Soil Salinity National Laboratory, Riverside CA. 92507, USA

Correspondence: Viviana Marcela Varón-Ramírez (vvaron@agrosavia.co)

Abstract. Texture is a fundamental soil property for multiple applications in environmental and earth sciences. Knowing its spatial distribution allows a better understanding of the response of soil conditions to changes in the environment, such as land use. This paper describes the technical development of Colombia´s first texture maps, obtained via a spatial ensemble of national and global digital soil mapping products. This work compiles a new database with 4,203 soil profiles, which were harmonized at five standard depths (0-5, 5-15, 15-30, 30-60, and 60-100 cm) and standardized with additive log-ratio (ALR) transformation. A compilation of 83 covariates was developed and harmonized at 1 square kilometer of spatial resolution. Ensemble Machine Learning (EML) algorithms (MACHISPLIN and landmap) were trained to predict the distribution of soil particle-size fractions (PSF) (clay, sand, and silt), and a comparison with SoilGrids (SG) products was performed. Finally, a spatial ensemble function was created to identify the smallest prediction errors between EML and SG. Our results are the first effort to build a national texture map (clay, sand, and silt fractions) based on digital soil mapping in Colombia. The results of EML algorithms showed that their accuracies were very similar at each standard depth, and were more accurate than SG. The largest improvement with the spatial ensemble was found at the first layer (0-5 cm). EML predictions were frequently selected for each PSF and depth in the total area; however, SG predictions were better when increasing soil depth in some specific regions. The final error distribution in the study area showed that sand presented higher absolute error values than clay and silt fractions, specifically in eastern Colombia. The spatial distribution of soil texture in Colombia is a potential tool to provide information for water-related applications, ecosystem services, and agricultural and crop modeling. However, future efforts need to improve aspects such as treating abrupt changes in the texture between depths and unbalanced data. Our results and the compiled database (Varón-Ramírez and Araujo-Carrillo, 2022; Varón-Ramírez et al., 2022) provide new insights to solve some of the aforementioned issues.

Keywords: Soil Particle Fractions, Ensemble Machine Learning, Compositional Data, Soil Database.

1 Introduction

Soil texture is defined by the proportion of particle-size fractions (PSF), called clay, silt, and sand (Richer-de Forges et al., 2022). Soil texture is important to understand soil processes related to agriculture and the environment from the field to the continental scale (Radočaj et al., 2020; Malone et al., 2021; Bönecke et al., 2021; Caubet et al., 2019). For example, soil

texture is a fundamental soil property for characterizing soil productivity and soil fertility (Patel et al., 2021; Soropa et al., 2021). Soil texture plays a fundamental role in quantifying the capacity of soils to store carbon and retain the water required for plants to grow (Dharumarajan and Hegde, 2020; Zhang and Hartemink, 2021). Additionally, the soil texture study must include two principal statements: first, it is compositional data, which means that PSF sum to 100% (%clay+%sand+%silt), and this statement must be satisfied at each location (Amirian-Chakan et al., 2019); second, the proportion of PSF could variate between horizons depending on soil forming factors interactions (Orton et al., 2016; Poggio and Gimona, 2017).

Spatial predictions of soil properties (e.g., particle-size fractions proportion) or classes (e.g., soil textural class) across areas where no soil data exist is the primary motivation of digital soil mapping (or pedometric mapping) (McBratney et al., 2003). In digital soil mapping (DSM), soil properties (continuous or categorical) for a specific soil depth and a given location in the geographical space can be predicted as a function (e.g., empirical function) of the soil forming environment (climate, organisms, topography, geology, ecology, atmosphere, and human interventions to soils) (Grunwald et al., 2011). These environmental prediction factors are commonly acquired from four primary sources: remote sensing, digital terrain analysis, climate, and thematic maps (e.g., soil type, rock type). The use of prediction algorithms or models that can account for the spatial variability of soil distribution is the basis of DSM (Wadoux et al., 2021a; Khaledian and Miller, 2020).

Predictions of quantitative soil properties (e.g., percentages of clay, silt, and sand) (Liu et al., 2020; Li et al., 2020) or the probability of presence/absence of a soil class (e.g., a soil textural class) (Ramcharan et al., 2018; Kaya and Başayiğit, 2022) are represented on digital soil maps for a given soil depth and a specific period. These predictions or probability estimates are derived from the use of supervised statistical learning (in the presence of training data for a response variable) or unsupervised statistical learning (in the absence of a response variable) (James et al., 2013). Statistical learning methods for supervised learning (e.g., for upscaling soil texture data using digital elevation models) can be applied to categorical (e.g., to solve classification problems) or numerical (to solve prediction problems) datasets (Bischl et al., 2016). There are hundreds (if not thousands) of modeling approaches for solving regression and classification problems. We could classify these methods into two modeling cultures: one assumes that a given stochastic data model generates the data, and the another uses algorithmic models and treats the data mechanism as unknown (Breiman, 2001). However, it is not easy to classify the immense diversity of modeling approaches and their possible combinations. Therefore, the problem of model or algorithm selection to perform predictions in DSM is an emergent research question.

Geostatistics and machine learning are the two principal forms of statistical learning in DSM (Hengl and MacMillan, 2019). Geostatistics is a branch of statistics that deals with the values associated with spatial or spatial-temporal datasets (Webster and Oliver, 2007). In contrast, machine learning is a computer-assisted branch of statistics that uses algorithms developed to solve prediction problems (Witten et al., 2011). Machine learning models are commonly parameterized (selection of multiple modeling parameters) using multiple resampling techniques, such as cross-validation or bootstrapping (Brenning, 2012). These resampling techniques allow the algorithm to 'learn' from the data using the capacity of computers to store results from multiple data configurations following the same statistical treatment. This computer-assisted learning allows machine learning algorithms to reproduce the relationship between the response and the prediction factors in the statistical space and can be applied to soil datasets to generate digital soil maps. Machine learning algorithms can be roughly divided into four main

groups: a) conventional machine learning based on trees, kernels, linear based, or probabilistic algorithms, b) reinforcement learning algorithms, c) deep learning algorithms, and d) ensemble learning algorithms. The latter extracts information from multiple modeling approaches and combines them to create a better solution for a given prediction problem (Yang, 2017).

Recent developments in ensemble learning efforts have demonstrated great potential for improving the accuracy and spatial detail of current estimates of soil functional properties across scales (Hengl et al. (2021); Wadoux et al. (2020); Llamas et al. (2020)). Several experiences with the mapping of soil texture at different depths have been done: France (Mulder et al., 2016), Scotland (Poggio and Gimona, 2017), Hungary (Laborczi et al., 2019), or China (Liu et al., 2020) are some examples. However, few of them have used spatial ensemble techniques. One representative case was developed by Hengl et al. (2021) for the continent of Africa at three depths (0, 20, and 50 cm) and 30 m spatial resolution. They produced predictions using two scale 3D ensemble machine learning (EML) framework and 122,200 training samples (approximately); their study utilized an improved predictive mapping framework: spatially-adjusted EML, that better accounts for spatial clustering of points. The spatial cross-validation methodology was a special point of their work, obtaining RMSE values of 9.6%, 13.7%, and 8.9% for clay, sand, and silt, respectively. Their results proved to be more accurate than previous works, which was attributable to the addition of higher resolution remote sensing images and Digital Terrain Parameters (DTM), the adoption of methodological improvements in hyper-parameter tuning, selection of features, and implementation of ensemble models (Hengl et al., 2021).

Colombia has produced maps, either PSF or textural classes, at a national scale through conventional mapping and regional scale using DSM. On a national scale, these maps use a series of delineations based on qualitative soil characteristics, called cartographic soil units (CSU). This studies have been produced in different periods: Cortés et al. (1982) (scale 1:5.000.000), IGAC (2003) (scale 1:500.000), and IGAC (2015) (scale 1:500.000). The map carried out by IGAC (2015) represented the PSF through four textural groups of soils (very fine, fine, medium, and coarse) in a layer from 0 to 50 cm; however, this methodology ignores the spatial variability inside the CSU. On a regional scale, Araujo-Carrillo et al. (2021) used machine learning algorithms to show the spatial distribution of clay (%) and its prediction error. However, they ignore the statement of compositional data of the soil texture, and their study just included the surface layer of the soil (0-20cm).

In this work, first, we compared and tested two ensemble machine learning approaches applied to predict soil texture at national scales in Colombia. Second, we compared our results with the global product SoilGrids (SG). Third, we built an en-semble map developing a pixel-wise solution to identify the method with lower prediction error. We hypothesized that multiple prediction algorithms could capture the spatial variability of soil texture differently because they treat the data in different ways to solve prediction problems (e.g., using decision boundaries or probability thresholds or hypothesis of the empirical relation-ship between the response and the prediction factors). Understanding which prediction algorithms and approaches yield lower error levels at the pixel level could benefit model selection efforts in DSM.

2   Methodology

Our workflow contains five significant steps: harmonization and transformation of soil data, selection of covariates, spatial prediction with different algorithms, validation, and spatial ensemble. These sections will be discussed in detail below.

2.1 Dataset

A total of 4,203 georeferenced (EPSG:4326) soil profiles were collected from Sistema de Información de Suelos de Latinoamérica y el Caribe - SISLAC, a soil information system developed by FAO (FAO, 2020), that all contained information about soil PSF. These PSF are classified according to the United States Department of Agriculture (USDA) system: clay (particles smaller than 0.002 mm in diameter), silt (particles sizes from 0.002 to 0.05 mm in diameter), and sand (particles sizes from 0.05 to 2.00 mm in diameter). The soil data covered five natural regions (geographic division made based on climatic, vegetation, relief, and soil class conditions) and 31 districts of the continental area of Colombia (Figure 1) (Rangel-Ch and Aguilar, 1995). The regions were: Caribbean in the north, Pacific in the west, Andean in the center (corresponding to the Andes Mountains), Orinoquia in the east, and Amazon in the south.
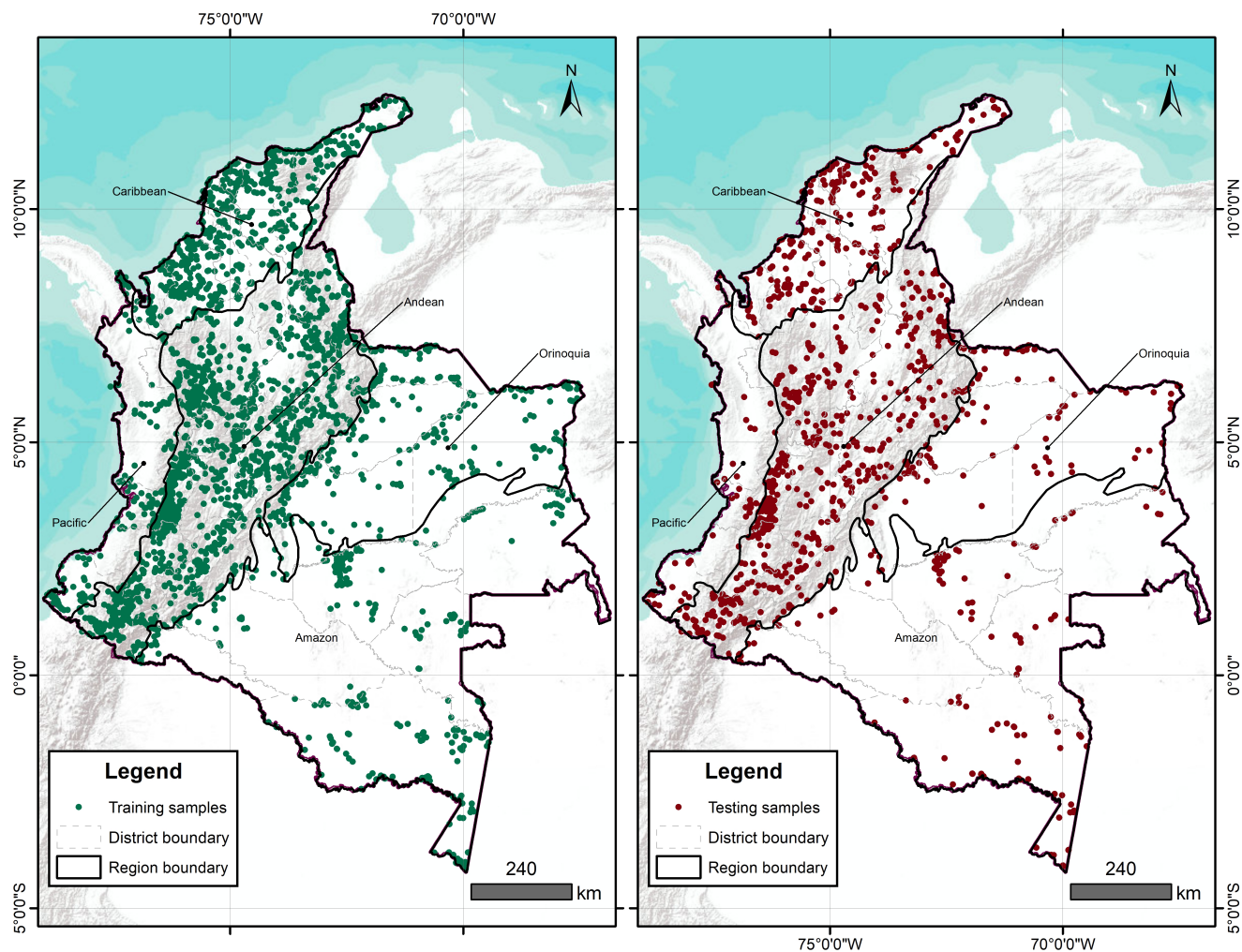
Figure 1. Soil-sample points distribution at 0-5 cm depth (Left: training samples. Right: testing samples)

2.2 Data harmonization and transformation

Dataset quality was ensured by: i) sum of PSF equals 100 % and ii) no overlapping sampling depth. PSF were harmonized to five standard depths (0-5, 5-15, 15-30, 30-60, and 60-100 cm), following the vertical discretization in the GlobalSoilMap specifications (Arrouays et al., 2014). The distribution of soil profiles by depth was: 4,203 at 0-5 cm, 4,201 at 5-15 cm, 4,153 at 15-30 cm, 3,974 at 30-60 cm, and 3,597 at 60-100 cm. The soil information for each depth was obtained using a quadratic function of depth with equal areas (spline) (Bishop et al., 1999), through the `mpspline` function of the `aqp` package (Beaudette et al., 2013) of R version 4.0.3.

A D-part composition is an element where all its components are strictly positive real numbers, they stock relative information, and these components must sum to 100% (Amirian-Chakan et al., 2019). In this way, soil texture is a 3-part composition, which means that PSF sum to 100% (%clay+%sand+%silt), and this statement must be satisfied at each location. In order to address this statement, PSF at each profile in standard depth were transformed based on additive log-ratio (ALR) transformation (Aitchison, 1986). The properties of the transformations when applied to regionalized compositions were discussed by Pawlowsky-Glahn and Olea (2004). ALR is commonly used for mapping soil PSF (Odeh et al., 2003; Poggio and Gimona, 2017; Wang et al., 2020; Li et al., 2020), preserving information about spatial correlation, showing a distribution more likely to be closer to a normal distribution (Li et al., 2020), and maintaining the compositional aspect of the variables (Lark and Bishop, 2007).

Let z_i, i=1, 2, 3 (D) represent the clay, sand, and silt fractions, where $D = 3$ is the number of soil particle-size categories and $D - 1 = 2$ is the number of transformations. ALR transformation is defined in equation 1, and the inverse transformation to obtain the original values of clay, sand, and silt is defined in equation 2:

$$Trans\_i = ln\left(\frac{z_i}{z_D}\right), \quad i = 1, 2, \ldots, D-1 \tag{1}$$

$$z_i = \begin{cases} \frac{exp\,(Trans\_i)}{1+\sum_{j=1}^{D-1} exp\,(Trans\_j)}, & i = 1, 2, \ldots, D-1 \\ \frac{1}{1+\sum_{j=1}^{D-1} \exp\,(Trans\_j)}, & i = D \end{cases} \tag{2}$$

Where $Trans\_i$ is the transformed value, and z_i is the original value. According to Poggio and Gimona (2017) and after verifying the possible selection of denominators (normality), this study used clay as the denominator variable. In this way, $Trans\_1 = ln\left(\frac{sand}{clay}\right)$ and $Trans\_2 = ln\left(\frac{silt}{clay}\right)$. The ALR transformation was implemented using the `alr` function in `Compositional` package (Tsagris et al., 2022). The predicted results were back-transformed to the original PSF values (clay, sand, and silt) using `alrinv` function.

2.3 Soil covariates

Using ESRI®ArcGIS™version 10.3, a total of 83 environmental covariates were selected to broadly reflect soil forming factors, as described by McBratney et al. (2003):

$$S_a = f(s, c, o, r, p, a, n) \tag{3}$$

where a soil attribute (S_a) is a function of other properties of the soil at a point (s), the climate (c), organisms (o), relief (r), parent material (p), age (a), and space (s) (Table 1). The pixel size of the environmental covariates was adjusted to 1 square kilometer using two methods: nearest neighbor and bilinear interpolation. Then, a stack of covariates (collection of rasters) was compiled for Colombia.

A recursive feature elimination (RFE) was run for each depth and transformation, using the function `rfe` of the `caret` package (Kuhn et al., 2020). The RFE is an algorithm that implements a backward selection of covariates based on predictor importance ranking (Kuhn et al., 2020). The goal was to find a subset of covariates used to produce the most accurate model possible. A regression matrix for each depth and transformation was built with the selected covariates, and this allowed extraction of the covariate values at the coordinates of each soil sample. With the regression matrix the dataset was divided using the function `createDataPartition` of the `caret` package (Kuhn et al., 2020). This function generates a stratified random split of the data and aims to create balanced splits of the data: a part for model training (75 %) (training samples in Figure 1) and another independent part for validation purposes (25 %) (testing samples in Figure 1) (Guevara et al., 2018).

2.4 Prediction models

The spatial distribution of the PSF at each of the five standard depths was modeled through Ensemble Machine Learning (EML) algorithms in two R packages: `MACHISPLIN` (Brown, 2021) and `landmap` (Hengl, 2021). EML consists of various approaches based on different methodologies, including stacking methods, averaging methods, bagging, and boosting approaches (Zounemat-Kermani et al., 2021).

MACHISPLIN evaluates different combinations to predict the input data, weighing and evaluating the fit. MACHISPLIN algorithm interpolates multivariate data through EML using six algorithms: boosted regression trees (BRT), neural networks (NN), generalized additive model (GAM), multivariate adaptive regression splines (MARS), support vector machines (SVM), and random forest (RF). This approach evaluates (via k-fold cross validation, where k=10) a method's ability to predict the input data and ensembles of all combinations of the six algorithms weighting each from 0 to 1. The best model will have the lowest Akaike information criterion with a correction for small sizes. After the best model is determined, the function runs the ensemble on the full dataset. Then, residuals are calculated and interpolated using a thin-plate-smoothing spline; this creates a continuous error surface and is used to correct most of the residual errors in the final ensemble model (Brown, 2021).

The landmap algorithm applies the stacking ensemble type. Stacking (sometimes called stacked generalization or committee machine approach) learns in parallel and fits a meta-model to predict ensemble estimates (Zhang and Ma, 2012). The 'meta-model' is an additional model that combines all individual or 'base learners' (Hengl, 2021). The landmap approach ensembles

No	Factor	Covariate	Source	Scale or resolution	Type	Unit
1		Clay ratio	Google Earth Engine		Cont	Dimensionless
2	Soil	Grain Size Index - GSI	(2020)	30 m	Cont	Dimensionless
	(s)					
3-22		Alofana-Kaolinite, Kaolinite, Kaolinite-Alofana, Kaolinite-Gibsite, Kaolinite-Integrated, Kaolinite-Micas, Kaolinite-Montmorillonite, Kaolinite-Quartz, Montmorillonite-Vermiculite, Montmorillonite-Kaolinite, Altered-Quartz, Feldspars-Altered, Feldspars-Amphiboles, Feldspars-Quartz, Feldspars-VolcanicGlass, Quartz-Amphiboles, Quartz-Altered, Quartz-Feldspars, Quartz-Micas, Quartz.	Soil map. IGAC (2015)	1:100,000	Bin	-
23-28		Udic, Ustic, Udic - Aquic, Aquic - Udic, Aquic, Aridic	Soil map. IGAC (2015)	1:100,000	Bin	-
29	Climate	Precipitation	Climatological database		Cont	Millimeters
30		Relative humidity	1980 - 2011.	1,000 m	Cont	Percentage
31	(c)	Mean air temperature	IDEAM (2015)		Cont	Celsius degree
32	Organisms	Pastures, Heterogeneous agriculture, Shrubs, Forests, Permanent crops	Corine Land Cover Classification 2010–2012. IDEAM (2014)	1:100,000	Bin	-
37	(o)	NDVI - index	Google Earth Engine		Cont	Dimensionless
38-39		Averages Band 6 and Band 7 - Landsat 8	(2020)	30 m	Cont	Dimensionless
40-54		Saladares, Mesas, Hills, Glacis, Terraces, Filas - vigas, Slopes, Fans, Crest, Vallecitos, Hogback, Flood Plan, Dunes, Rocky Glacis, Flattening.	Soil map. IGAC (2015)	1:100,000	Bin	-
55		Elevation			Cont	Meter
56	Relief	Aspect			Cont	Degree
57	(r)	Curvature			Cont	Dimensionless
58		Drainage Distance	SRTM (2000)	90 m	Cont	Meter
59		Slope			Cont	Degree
60		Terrain Ruggedness Index - TRI			Cont	Dimensionless
61		Topographic Wetness Index - TWI			Cont	Dimensionless
62		Valley Depth			Cont	Meter
63-68	Parental material (p)	Alluvial Deposits, Colluvial Deposits, Volcanic Ash, Igneous Rocks, Sedimentary Rocks, Metamorphic Rocks	Soil map. IGAC (2015)	1:100,000	Bin	-
69-78	Age (a)	Andisols, Alfisols, Entisols, Mollisols, Inceptisols, Oxisols, Ultisols, Histosols, Vertisols, Aridisols.	Soil map. IGAC (2015)	1:100,000	Bin	-
79-83	Space (n)	Oblique geographic coords 0.0π, 0.2π, 0.4π, 0.6π, 0.8π	Møller et al. (2020)	1,000 m	Bin	Degree

Table 1. Environmental covariates by soil forming factor. Cont: Continuous; Bin: Binary

different machine learning algorithms, a fast implementation of RF (ranger), extreme gradient boosting (xgboost), support vector machines (ksvm), neural networks (nnet), and generalized linear models (GLM) (with Lasso or Elastic Net Regularization). The landmap package extends the functionality of the deprecated mlr 'meta-package' (now mlr3) (Lang et al., 2019) and is based on super learner. It is a prediction method designed to find the optimal combination of a collection of prediction algorithms, and its framework is built on the theory of cross-validation and allows for a general class of prediction algorithms to be considered for the ensemble (Polley and Van der Laan, 2010).

The main difference between these two algorithms is the cross-validation. MACHISPLIN randomly makes cross-validation, while landmap makes spatial cross-validation. In the random process, the testing and training dataset would not be independent in this scenario, with the consequence that cross-validation fails to detect a possible overfitting (Lovelace et al., 2019). That situation is performed by landmap, due it blocks some training points based on spatial dependence (it makes a semivariogram model) to prevent producing biased estimations predictions (Hengl, 2021). However, Wadoux et al. (2021b) indicate that spatial cross-validation methods may provide biased estimates of map accuracy, but standard cross-validation is deficient in the case of clustered data. Additionally, MACHISPLIN constructs the best linear model, systematically assigning a weight for each algorithm and evaluating the fit of the ensemble algorithm. In contrast, landmap constructs the meta-model with the predictions of the cross-validation (indicated in the method of ensemble).

2.5 Validation

The validation process had three steps: i) inclusion of SG layers, ii) calculation of map quality measures, and iii) identification of the spatial distribution of the prediction error.

In the first step, one aim of the work was to compare the spatial prediction of techniques mentioned in section 2.4 against the products from SG Version 2.0 (Poggio et al., 2021). In this step, SG layers (250 m pixel size) of the PSF at each standard depth were downloaded.

In the second step, the prediction errors, which is the difference between the predicted value and observed value (Brus et al., 2011), were calculated for each testing sample. After that, were calculated some quantitative statistics: mean error (ME), root mean square error (RMSE), amount of variance explained (AVE), and concordance correlation coefficient (CCC). ME measures bias in the prediction and is defined as the population mean of the prediction errors (Yigini et al., 2018), values close to 0 indicate that the predictions are unbiased. RMSE is a measure of prediction accuracy, and a perfect model would have a value ≈ 0 (Kempen et al., 2012). The AVE measures the fraction of the overall dispersion of the observed values that the model explains, and this measure has an optimal value of 1 (Samuel-Rosa et al., 2015). Finally, the CCC measures the agreement of predicted values with observed values (relationship 1:1), where 1 is a perfect concordance and 0 is no correlation (Lawrence and Lin, 1989).

In the third step, the prediction errors were interpolated using ordinary kriging (OK), a widely used geostatistical technique that assumes intrinsic stationarity (Webster and Oliver, 2007). This process was done for prediction errors of the three approaches (MACHISPLIN, landmap, and SG) in the five standard depths and each PSF. The layers obtained in this step were harmonized according to this study's resolution and extent framework.

2.6 Spatial ensemble

The spatial ensemble maps and their ensemble errors were generated for each depth and PSF (Figure 2). First, with the error prediction maps for each PSF, a mean absolute error (MAE) was calculated for each EML and SG. After, a spatial ensemble function was created to perform a conditional evaluation using the prediction maps and their respective MAE; at each pixel, the function identified which model (EML or SG) had the minimum MAE and selected the predictions for clay, sand, and silt of it. At last, an ensemble prediction error map was built for each PSF doing a mask by model selected and assigning their respective prediction error (calculated in section 2.5). After, the map quality measures statistics were newly calculated.
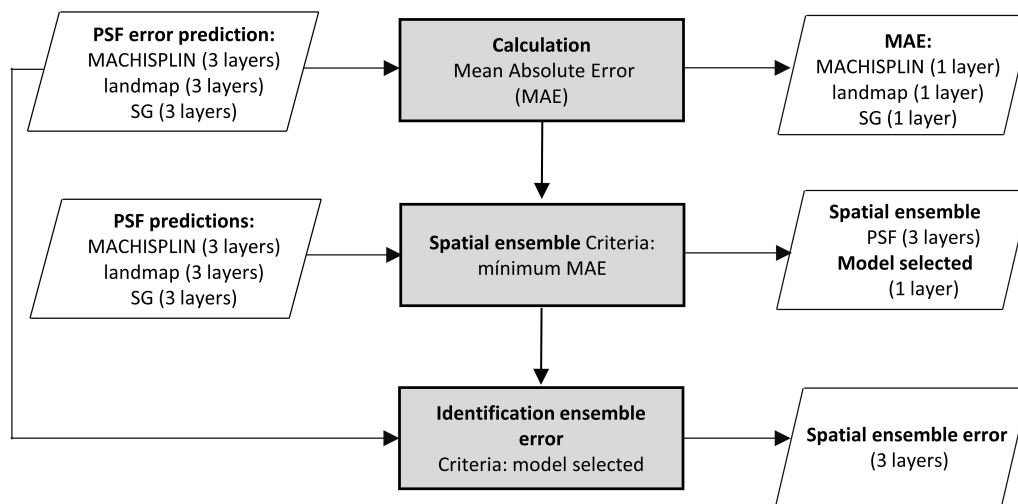
Figure 2. Framework used for generating ensemble maps. PSF means: inputs or outputs for clay, sand, and silt

3 Results

This study represents the first effort to provide a national map of soil texture using a digital soil mapping framework in Colombia. This work used EML algorithms to improve the accuracy of national soil texture predictions, with a fully independent dataset, concerning the global product (SG). Also, it provided new insights for assessing the quality and accuracy of global soil texture predictions. The main results are going to be shown in the following subsections.

3.1 Soil texture characterization and dataset

The textural classes for each sample point at five standard depths are shown in Figure 3. In most USDA textural classes, the dataset has soil samples. Sandy loam, loam, clay loam, sandy clay loam, and clay were the most frequent textural classes in all

standard depths, and silty and silty loam textural classes were less frequent. Also, it is easy to identify that, in the dataset, there
210 are not many soil samples with high content of clay and silt, while extreme contents of sand fraction are frequent.

Descriptive statistics are shown in Table 2 for PSF and its transformations (Trans_1 and Trans_2). The PSF covers the entire range of measures (0 to 100 %), which is expected for national-scale analysis. The mean and median for sand fraction was higher than clay and silt fractions in all standard depths, indicating that sand is the dominant PSF in Colombian soils. The standard deviation grew for all fractions in the deepest layers; sand content was the fraction with the highest variation in all
215 depths. These results suggest that sand fraction has the highest variability in the PSF for Colombian soils, which rises with increasing depth. The skewness coefficients were positive and less than 1 for all textural fraction dataset, and the sand fraction showed less deviation from the normal distribution, except for 60-100 cm layer. Regarding transformations, the ranges took values from negatives to positives with means around zero; this signifies that ALR transformation improved the sand and silt distribution for all depths, except the sand in the first three layers.
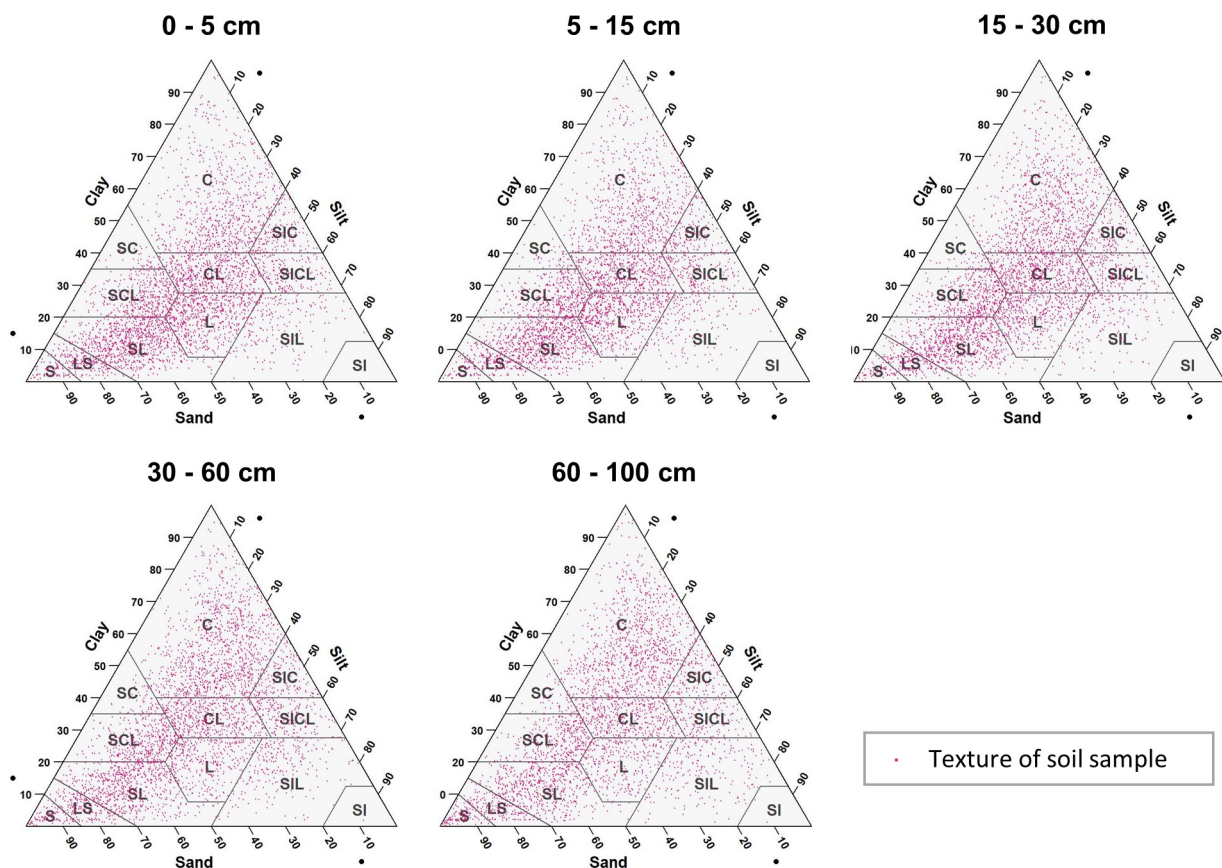
Figure 3. Particle-size soil samples representation in a textural diagram for each standard depth. C: Clay, SC: Sandy clay, SCL: Sandy clay loam, CL: Clay loam, SIC: Silt clay, SICL: Silt clay loam, L: Loam, SIL: Silt loam, SI: Silt, SL: Sandy loam, LS: Loamy sand, and S: Sand.

Standard depth (cm)	PSF and transformations	Min	Max	Mean	Median	SD	Kurtosis	Skewness
5	Clay	0.02	95.07	27.88	25.17	17.23	0.43	0.82
	Sand	0.13	99.19	42.69	42.00	22.45	-0.83	0.13
	Silt	0.10	83.65	29.43	28.00	13.95	0.00	0.46
	Trans_1	-6.06	7.68	0.44	0.50	1.44	1.20	-0.36
	Trans_2	-6.33	7.87	0.14	0.15	0.89	4.19	-0.22
15	Clay	0.44	94.50	28.64	26.07	17.07	0.28	0.76
	Sand	0.12	98.00	42.02	41.53	22.18	-0.83	0.17
	Silt	1.00	81.74	29.33	28.00	13.61	0.00	0.47
	Trans_1	-6.42	4.63	0.38	0.43	1.40	1.01	-0.40
	Trans_2	-3.45	0.11	0.12	4.82	0.82	1.51	-0.26
30	Clay	0.29	94.78	30.47	28.40	17.75	-0.15	0.58
	Sand	0.04	98.00	40.48	38.53	22.62	-0.82	0.27
	Silt	0.36	76.76	29.05	27.34	13.75	0.05	0.55
	Trans_1	-6.83	5.83	0.27	0.27	1.45	1.05	-0.32
	Trans_2	-3.84	3.33	0.04	0.05	0.83	1.00	-0.07
60	Clay	0.01	94.50	32.32	30.10	19.16	-0.47	0.45
	Sand	0.03	99.86	39.10	36.08	23.67	-0.79	0.38
	Silt	0.05	90.13	28.58	26.13	14.61	0.15	0.66
	Trans_1	-7.47	9.14	0.16	0.14	1.60	1.53	-0.20
	Trans_2	-4.67	6.78	-0.03	-0.05	0.93	1.84	0.21
100	Clay	0.06	97.07	32.76	30.71	19.70	-0.55	0.41
	Sand	0.01	99.80	38.57	34.76	24.38	-0.72	0.46
	Silt	0.14	87.50	28.67	26.24	15.40	0.33	0.71
	Trans_1	-8.43	7.35	0.13	0.08	1.64	1.22	-0.16
	Trans_2	-4.25	4.19	-0.05	-0.07	0.94	1.04	0.14

Table 2. Descriptive statistics of PSF and its transformations for each standard depth. Min: minimum; Max: maximum; SD: standard deviation

3.2 Covariate selection

Ten recursive feature-elimination models were obtained, and individual covariate stacks were built for each transformation in all standard depths. For each layer, covariates were selected for Trans_1 and Trans_2, respectively: 44 and 83 (0-5 cm), 54 and 54 (5-15 cm), 59 and 83 (15-30 cm), 56 and 58 (30-60 cm), and 56 and 83 (60-100 cm) (Table 3). The top 5 covariates (Table 3) included soil forming factors: climatic (temperature, relative humidity, and precipitation), topographic (altitude, slope, and presence of flood planes), parent material (presence of alluvial deposits), activity of organisms (bands 6 and 7 of Landsat 8)

Standard depth (cm)	Variable	Covs selected	RMSE	Top 5 Covs selected
5	Trans_1	44	1.21	TEM, RH, PPT, Altitude, L8 b7
	Trans_2	83	0.76	TEM, RH, PPT, Altitude, Alluvial
15	Trans_1	54	1.17	TEM, Altitude, RH, PPT, Clay ratio
	Trans_2	54	0.74	TEM, RH, PPT, GSI, Altitude
30	Trans_1	59	1.24	TEM, RH, Altitude, Flood plane, PPT
	Trans_2	83	0.73	Alluvial, PPT, RH, TEM, Altitude
60	Trans_1	56	1.34	RH, TEM, PPT, Altitude, Slope
	Trans_2	58	0.82	L8 b6, PPT, Alluvial, RH, Clay ratio
100	Trans_1	56	1.34	PPT, RH, TEM, Altitude, L8 b6
	Trans_2	83	0.84	L8 b6, PPT, GSI, Alluvial, Clay ratio

Table 3. Top 5 covariates selection for each transformation and standard depth. TEM: temperature, RH: relative humidity, PPT: precipitation, L8 b7: Landsat 8 band 7, L8 b6: Landsat 8 band 6, GSI: Grain Size Index.

and previous soil index information (Clay ratio and Grain Size Index). It is important to highlight that the covariates selection had only two binary covariates (alluvial and flood plains).

3.3 Soil texture predictions and SG products validation

Map quality measures are given in Table 4. Referring to ME, silt fraction was overestimated (positive ME), and clay and sand fractions were underestimated (negative ME); this happened for all depths and EML algorithms, except clay in 60-100 cm layer. For sand fraction, the bias values of landmap predictions were fewer than MACHISPLIN predictions. The ME values closets to zero were found for clay in the deepest layer, suggesting unbiased predictions for both MACHISPLIN and landmap algorithms. Comparing MACHISPLIN and landmap, the RMSE values were similar for each PSF and depth. However, comparing RMSE values between PSF, RMSE values for sand were always higher than clay and silt.

The AVE values were under 0.35. For all standard depths and algorithms, sand fraction had higher AVE values than clay and silt, except in the 60-100 cm layer, where AVE values for clay were higher. In general, for MACHISPLIN and landmap, the capacity of each model to explain the variance decreased when increasing depth; however, it is important to highlight the AVE values for silt fraction in 0-5 and 5-15 cm layers, which were so close to zero. On the other hand, the CCC values were from 0.32 to 0.54. The CCC values for sand were higher than clay and silt in all cases. Also, the lowest CCC value between standard depths was found in the deepest layer (60-100 cm), where the data set had fewer sample points than the superficial layers. These results suggest that, for sand, the predicted and observed values agree more than the clay and silt fraction.

An evaluation of SG products with the dataset validation (the same used for EML validation) is shown in Table 4. About ME, the highest bias values were found for sand fraction, followed by clay and silt. For sand fraction, the estimations were underestimated (negative ME) in all standard depths; in contrast, for clay and silt fractions, the estimations were overestimated

245 (positive ME), except by silt in 30-60 cm layer. These ME values were higher than EML predictions. Concerning RMSE values, the sand fraction had higher values than the clay and silt fraction, and the RMSE increased with increasing depth. For AVE, negative and close to zero values were found in all standard depths and fractions (-0.27 to -0.08). Similarly, the CCC values were close to zero (0.04 to 0.16), and the highest values were obtained for sand and silt fraction in the three most superficial layers.

250 In Figure 4 there is a visual comparison of our results and SG products. In this representation, we can see some aspects: first, the ranges of the predicted values are wider for MACHISPLIN and landmap than SG (low contrast of colors). Second, the general pattern of PSF distribution is different between our results and SG products. Principally, in the Andean and Caribbean regions, our results suggest that these areas have higher values of sand (blue colors) than clay and silt fractions, while SG suggests soils with more content of clay and silt (orange and green colors); also, in Orinoquia region, our results show soils
255 with high content of silt (green colors) and SG displays soils with more content of clay and sand (purple colors). Third, in the southern and eastern areas, the SG products do not have artifacts generated for the prediction of the algorithms; this is an advantage of SG products.

3.4 Spatial ensemble

Figure 5 to 9 display the final spatial ensemble maps for each PSF, which contain their final error and the model selected
260 for each pixel at each standard depth. The spatial ensemble, which, as described above, is a collection of best-fit data from 3 different algorithms (MACHISPLIN, landmap, and SG), contained common elements/features in most standard depths.

In all standard depths, predictions of MACHISPLIN (yellow color) and landmap (green color) represented the PSF distribution more than SG (gray color). The SG selection increased with depth, especially in 60-100 cm layer. The most significant areas where SG had the fewest prediction errors were in Orinoquia and Amazon regions. Similar to SG, the landmap predic-
265 tions were frequently selected in the deepest layers overall in the southern areas of Colombia. MACHISPLIN predictions were most chosen in the areas with the lowest sample density (Orinoquia and Amazon regions).

Concerning PSF distribution, sand fraction had the highest variation in the geographical space in all depths. The highest mountains in the central areas, the Sierra Nevada in the northern, and the hills in the eastern areas had a significant content of sand; in contrast, the finest fractions (clay and silt) were found mainly in valley landscapes between mountain chains in central
270 areas of Colombia and hill landscape in the eastern areas.

The external validation for the spatial ensemble maps showed an improvement in their metrics vs. the maps using EML algorithm or SG product (Table 5). The ME values were closer to zero, showing an improvement in the prediction; however, in this ensemble model, predictions of silt fraction had the highest bias, which is a different behavior of EML algorithms (MACHISPLIN and landmap), where sand fraction had the most biased predictions. RMSE values decreased for all PSF
275 and standard depths, which means a raising in the precision of the map. On the other hand, AVE values increased with the spatial ensemble model; interesting that the AVE values for silt fraction at 0-5 and 5-15 cm were higher compared to landmap predictions (Table 4); also the sand fraction had the most betterment respect to clay and silt fraction. Concern to CCC values, these are still equal to or higher than individual algorithms.

Depht (cm)	Method	Size Fraction	MACHISPLIN			landmap			SoilGrids		
			Clay	Sand	Silt	Clay	Sand	Silt	Clay	Sand	Silt
0 - 5	ME		-0.96	-0.72	1.68	-1.88	-0.06	1.94	2.62	-8.86	2.45
	RMSE		15.00	18.57	11.96	15.38	19.09	13.94	18.48	24.54	14.92
	AVE		0.28	0.35	0.28	0.24	0.31	0.02	-0.10	-0.14	-0.13
	CCC		0.42	0.51	0.43	0.40	0.54	0.36	0.08	0.14	0.12
5 - 15	ME		-0.34	-0.71	1.05	-0.90	-0.47	1.37	2.16	-8.35	1.73
	RMSE		14.24	17.82	11.30	14.77	19.01	13.62	18.41	23.93	14.55
	AVE		0.32	0.36	0.30	0.27	0.27	-0.01	-0.14	-0.15	-0.16
	CCC		0.46	0.53	0.46	0.43	0.54	0.36	0.06	0.13	0.12
15 - 30	ME		-0.48	-1.03	1.50	-0.76	-0.47	1.23	2.50	-7.89	1.54
	RMSE		15.75	19.02	11.76	15.83	18.93	11.75	18.64	24.43	14.38
	AVE		0.23	0.32	0.29	0.22	0.32	0.29	-0.08	-0.13	-0.06
	CCC		0.37	0.48	0.44	0.38	0.51	0.46	0.09	0.12	0.16
30 - 60	ME		-0.04	-0.78	0.82	-0.08	-0.76	0.84	3.55	-6.98	-0.01
	RMSE		17.11	20.68	13.13	17.08	20.87	13.30	20.54	25.15	15.70
	AVE		0.23	0.26	0.25	0.23	0.24	0.23	-0.11	-0.10	-0.08
	CCC		0.36	0.42	0.39	0.38	0.43	0.39	0.07	0.09	0.09
60 - 100	ME		0.01	-3.16	3.15	0.05	-2.87	2.82	3.17	-9.83	1.12
	RMSE		17.29	21.89	14.09	17.53	22.36	14.24	21.44	27.67	16.39
	AVE		0.24	0.20	0.14	0.21	0.17	0.12	-0.17	-0.27	-0.16
	CCC		0.36	0.38	0.31	0.37	0.38	0.34	0.07	0.03	0.08

Table 4. Map quality measures of each algorithm for PSF in five standard depths. ME: Mean Error; RMSE: Root Square Mean Error; AVE: Amount of variance Explained; CCC: Concordance Correlation Coefficient. These map quality measures are based on the validation dataset.

Depth (cm)	0 - 5			5 - 15			15 - 30			30 - 60			60 - 100		
Size Fraction	Clay	Sand	Silt	Clay	Sand	Silt	Clay	Sand	Silt	Clay	Sand	Silt	Clay	Sand	Silt
ME	-0.92	-0.98	1.81	-0.31	-0.61	0.89	-0.44	-0.85	1.26	0.00	-0.85	0.82	0.03	-2.11	1.99
RMSE	14.30	17.14	12.61	13.87	16.94	11.45	15.54	18.55	11.60	16.88	20.32	13.10	17.26	21.74	13.86
AVE	0.34	0.44	0.19	0.35	0.42	0.28	0.25	0.35	0.31	0.25	0.28	0.25	0.24	0.21	0.17
CCC	0.47	0.60	0.42	0.49	0.58	0.45	0.39	0.51	0.46	0.38	0.44	0.40	0.37	0.38	0.31

Table 5. Summary of map quality measures for spatial ensemble model for PSF in five standard depths. ME: Mean Error; RMSE: Root Mean Square Error; AVE:Amount of variance Explained; CCC: Concordance Correlation Coefficient. These map quality measures are based on the validation dataset

Texture fractions at 0-5 cm

Texture fractions at 5-15 cm
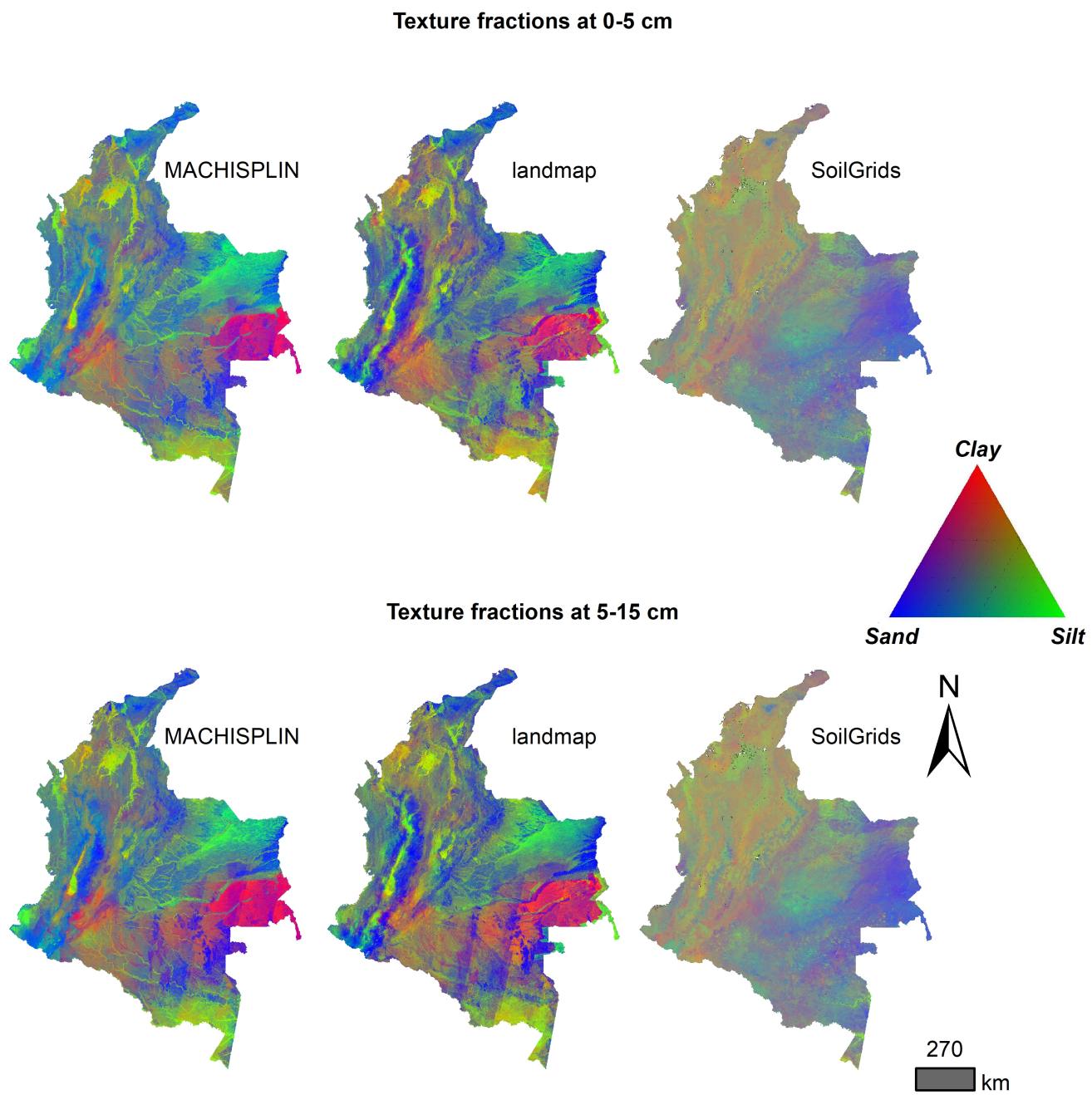
Figure 4. Color composite map of soil texture fractions predictions at two depths: 0-5 and 5-15 cm
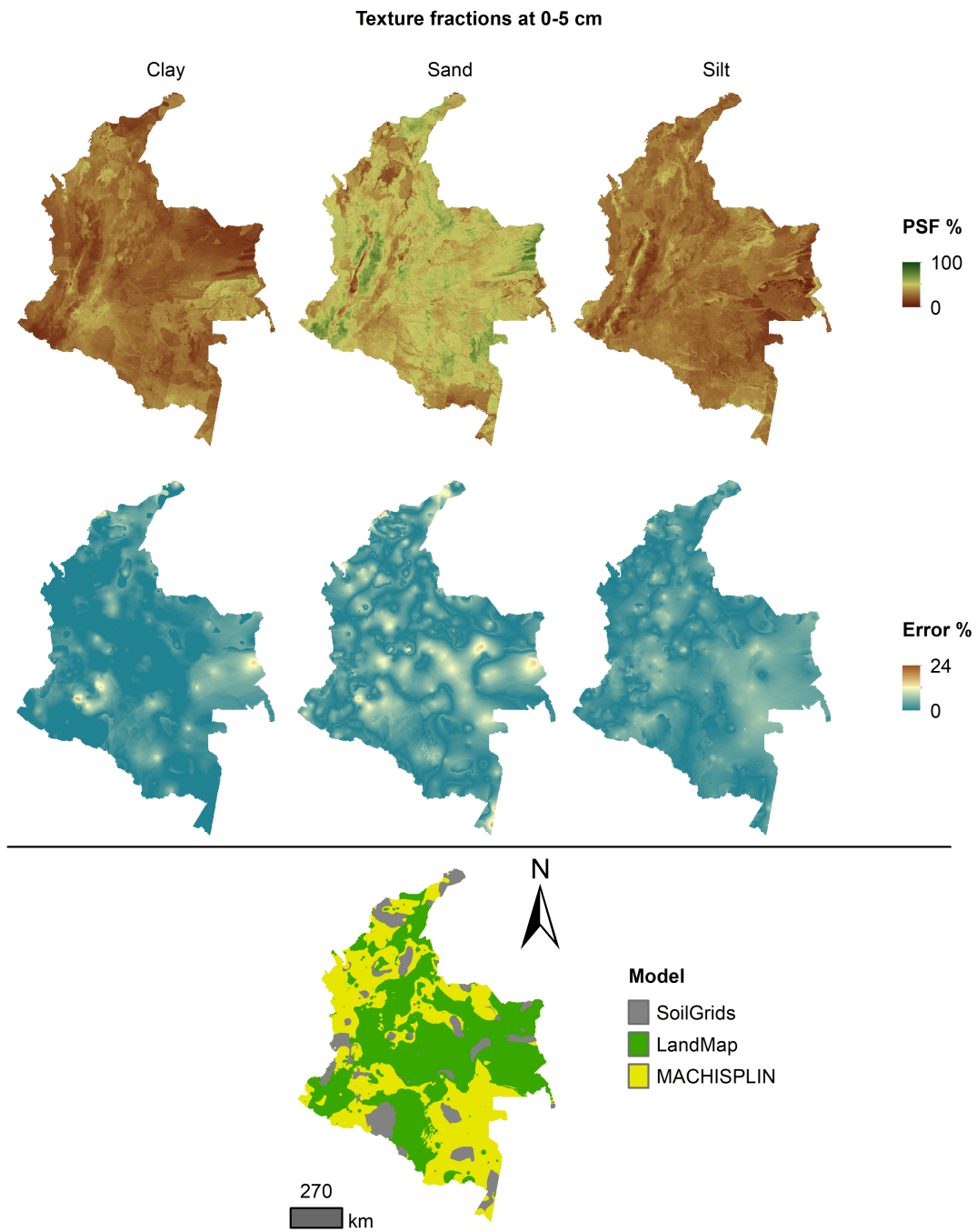
Figure 5. Ensemble model, percentage-error distribution, and best model selected at 0-5 cm

Texture fractions at 5-15 cm
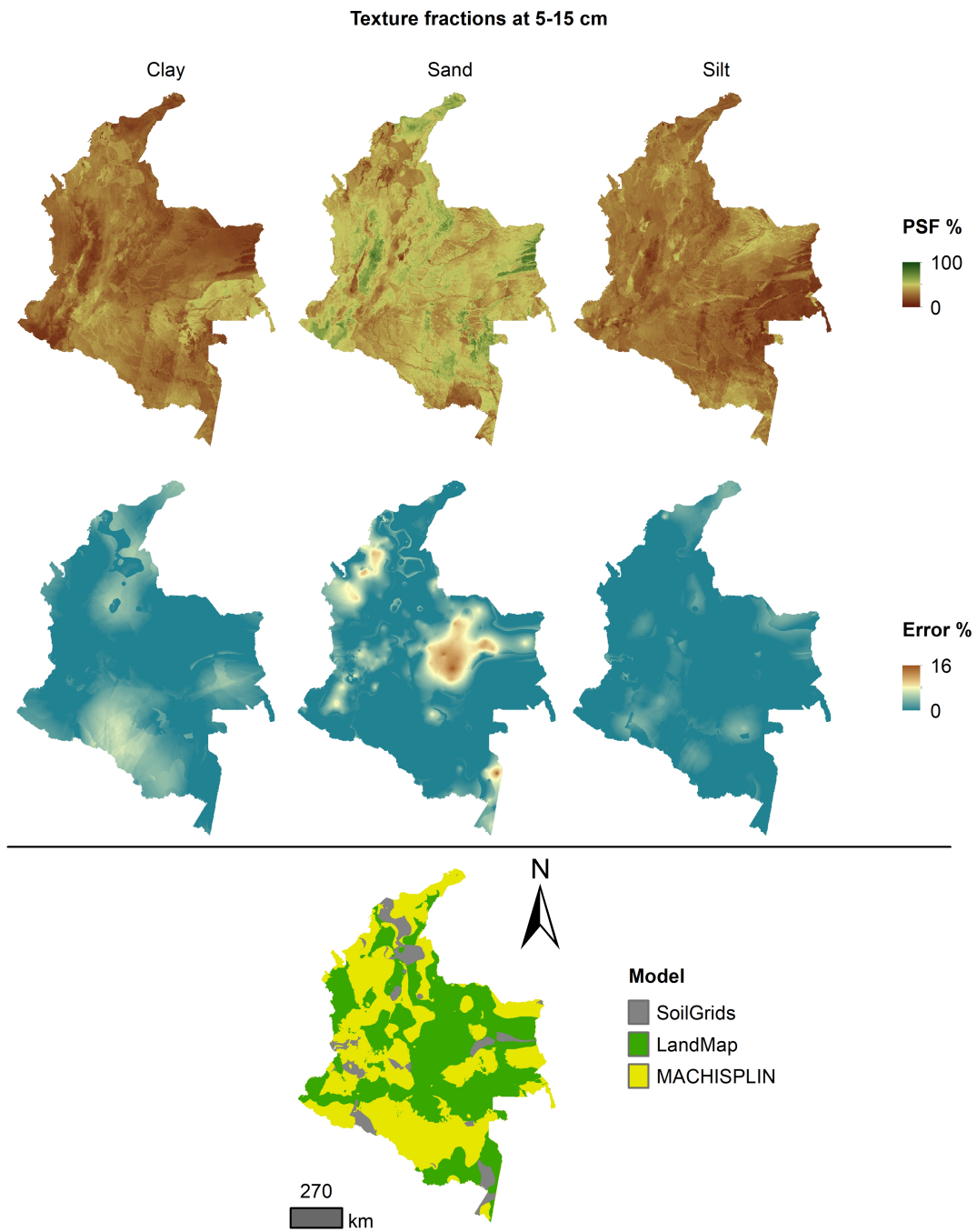
Figure 6. Ensemble model, percentage-error distribution, and best model selected at 5-15 cm

Texture fractions at 15-30 cm
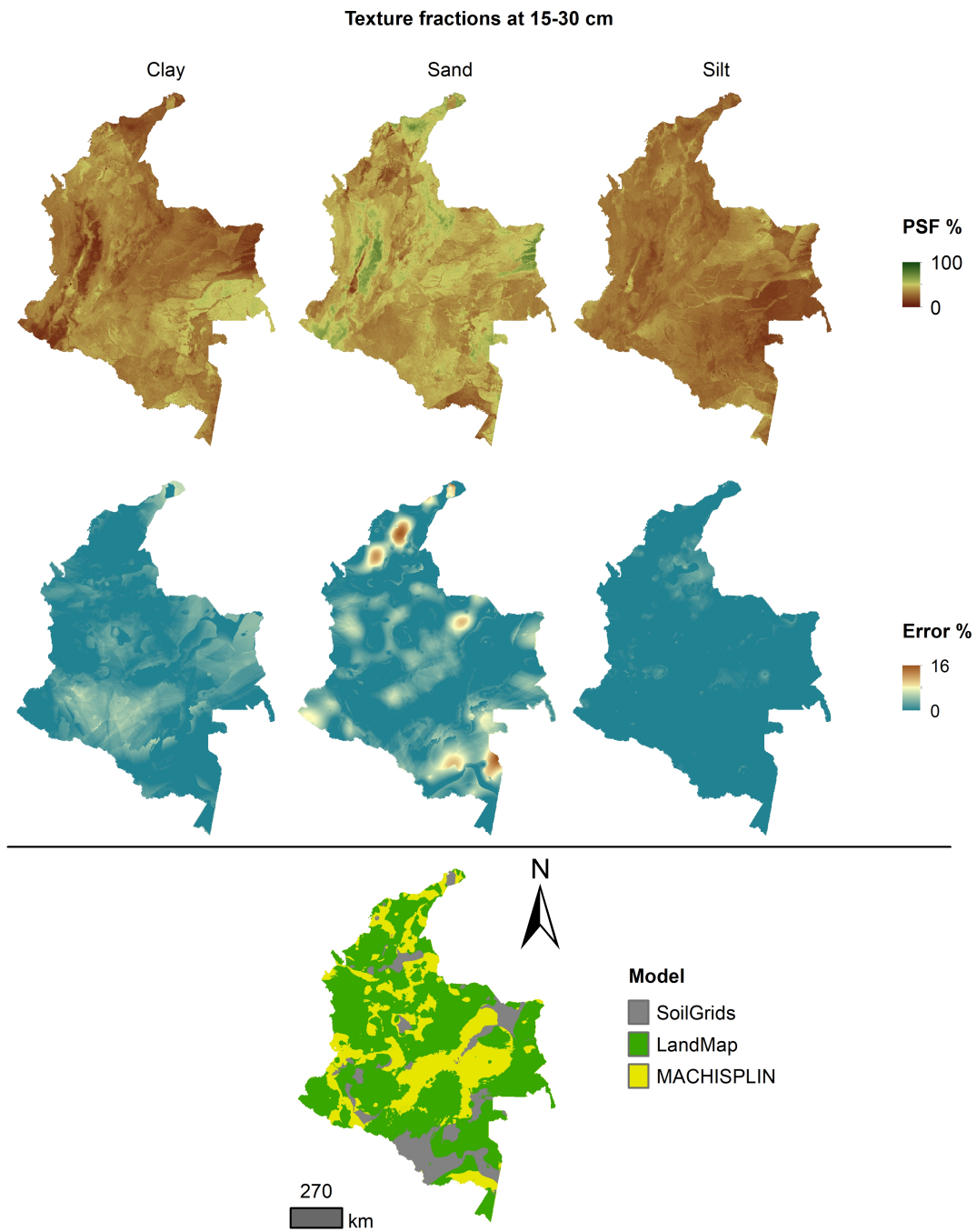
Figure 7. Ensemble model, percentage-error distribution, and best model selected at 15-30 cm

Texture fractions at 30-60 cm
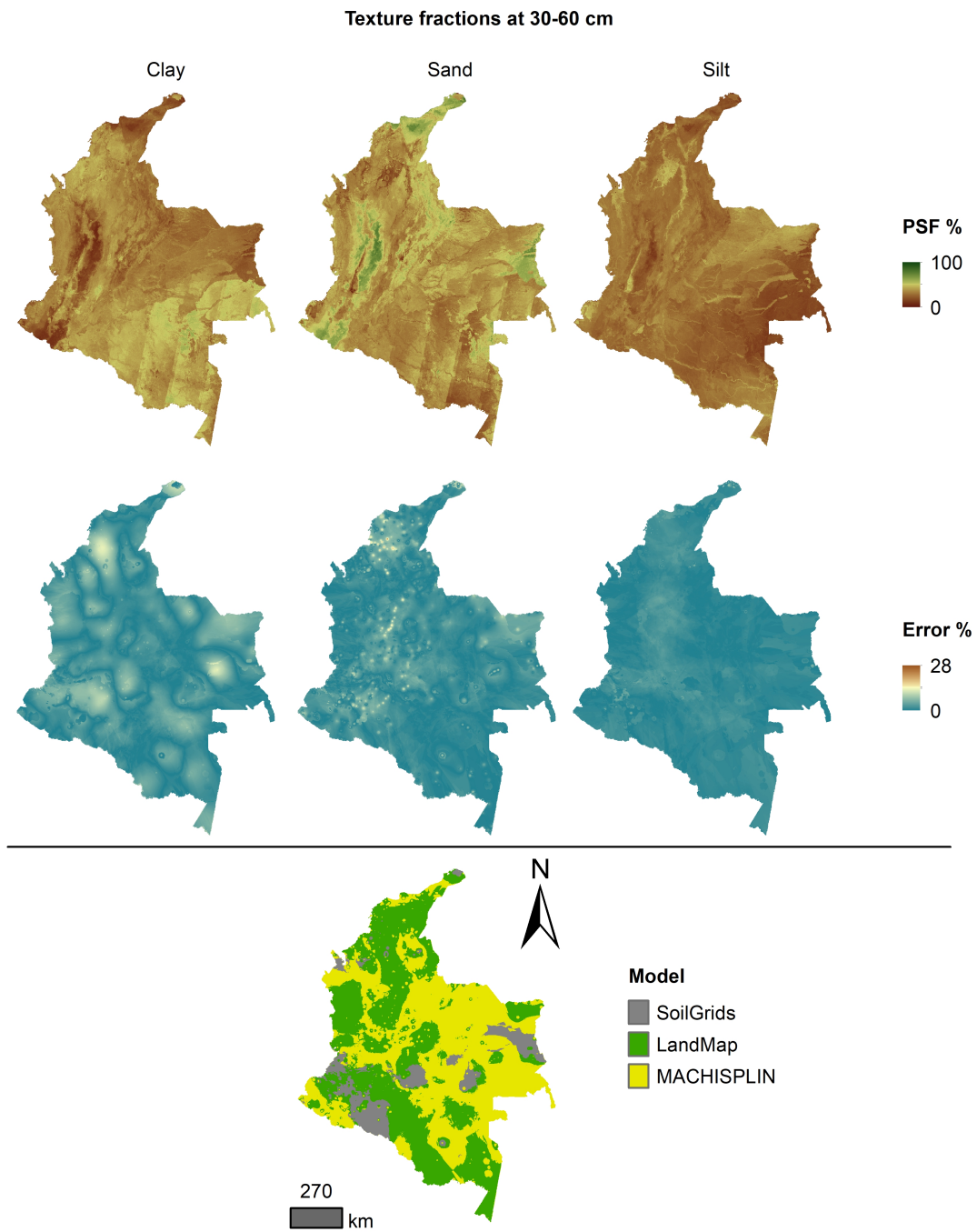
Figure 8. Ensemble model, percentage-error distribution, and best model selected at 30-60 cm

Texture fractions at 60-100 cm
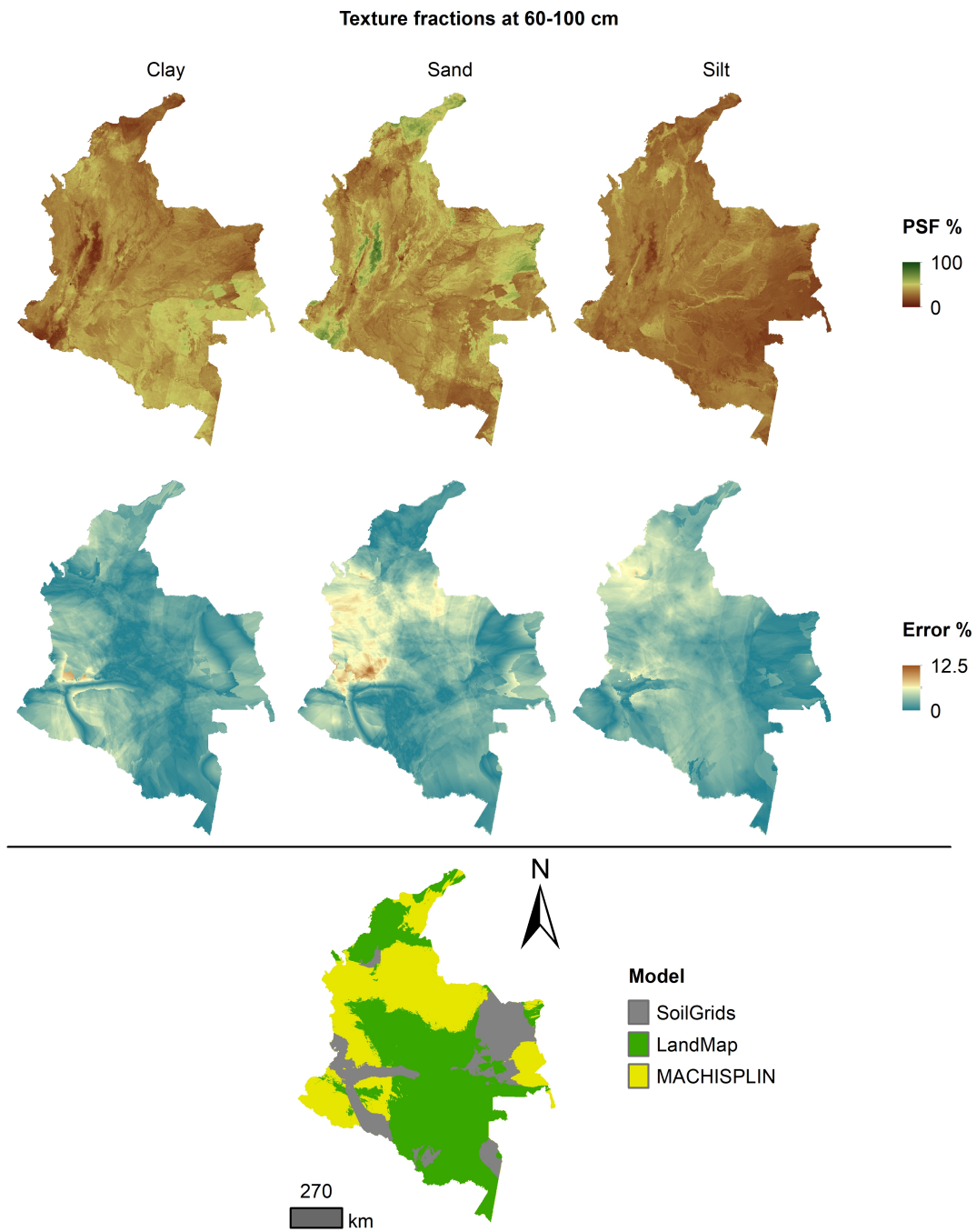
Figure 9. Ensemble model, percentage-error distribution, and best model selected at 60-100 cm

4 Discussion

In this paper, we developed a new digital soil texture dataset that contains legacy soil data, environmental covariates, and the first digital soil texture maps across Colombia. Colombia's literature on machine learning applied to soil texture mapping is limited. We improved the accuracy and spatial resolution of previous conventional maps. While many studies focus on mapping soil properties such as pH and organic matter, fewer studies focus on comparing and testing global approaches, such as SoilGrids, for maximizing accuracy. Our results contribute to a national benchmark of the reliability of global predictions compared to national predictions. We first discuss the general geography of soil texture across the country and compare and discuss our findings with previous work.

Colombia has a great diversity of soils, and their properties change with depth. In the five standard depths, soil texture in Colombia has representation in all textural classes defined by Soil Survey Staff (2014). As depth increases, the soil texture is finer, and the proportion of clay and silt rises. On the other hand, coarse soils are in central and northern areas, and these soil textures hold with increasing depth. This high diversity of soil texture is due to the high number of interactions between soil forming factors, particularly the great diversity of parent materials within Colombia (IGAC, 2015; Araujo et al., 2017).

Some topography and parent material covariates were the principal drivers in texture modeling. The focal areas with fine and medium textures are found in the northwest (floodplain and land depressions), in central areas (Magdalena River valley), in the west (Cauca River valley), in the south (Amazon region), and the east (Orinoquia region). All these regions have specific soil forming factors, such as alluvial parent material deposited by one or many rivers, which are soil fine-size fractions driver (Flórez, 2003)). On the other hand, medium and coarse textures are principally in the hillsides of mountain landscapes in central, southern, and southwestern regions. Mainly, these coarse soil textures are due to the presence of sandstone, conglomerate sandstone, granites, and gneisses, among others, that have siliceous and quartz rocks (Catoni et al., 2016), volcanic materials, and glacial clast (IGAC, 2015) that are in these areas. Despite the relationship between soil texture distribution, relief, and parent material covariates, only altitude (quantitative), slope (quantitative), alluvial (binary), and flood plane (binary) covariates were present in the top 5 predictors for each standard depth.

Although parental material is critical in the soil texture spatial distribution, the covariates selection identified that the climatic covariates were more important (i.e. temperature, relative humidity, and precipitation). The covariates used to describe the parental material were binary class variables; maybe the following exercises should include quantitative variables to identify this soil forming factor, for example, using radar remote sensing (Niang et al., 2014) or based on the spectral response in the visible and near-infrared spectrum (Vis-NIR), medium infrared (MIR), and Vis-NIR-MIR (Campbell et al., 2019). In the PSF predictions (in specific, the ALR components), the importance of the climatic covariates did not have apparent changes with depth. The country's climate conditions have led to relatively strong physical weathering in the soil forming process (Osman, 2013). Due to the country's location, it is influenced climatologically by the atmospheric circulation of the Caribbean Sea, Pacific Ocean, the Amazon basin, and the orographic barrier of the three branches of the Andes Mountain (Poveda, 2004). Furthermore, in this study, the variables were chosen to maximize the predictive power of the models, not their explanatory capabilities.

Colombia has not produced maps on PSF at a national scale with DSM products, but Colombia has developed soil surveys through conventional mapping (IGAC, 2015); then, the textural soil distribution of Colombia presented in this study is not

315 directly comparable with previous national textural soil maps. Due to the methodology used in IGAC (2015), the depth studied is different, the polygons delineated (CSU) have a unique value for an entire area, and CSUs are not an uncertainty value associated. These last two reasons are the primary use limitations in traditional soil surveys (Angelini et al., 2016). Despite that, the maps produced by this study and those of the IGAC project show two significant areas with similar attributes. In the northwest (Caribbean region) and southern (Amazon region), the IGAC study presents a fine group texture (clay between 40

320 and 60%), and this current result shows that levels of clay percentages in that clay range. However, there is a principal region in the western (Orinoquia region), where the two results are very different. The previous result shows these areas with a coarse textural group, and this current result displays low percentages of sand fractions for 5-15, 15-30, and 30-60 cm depths. These differences are due to the low soil sampling density, where there is just one observation, and in this current study, its nearest predictions are driven by soil data.

325 Regarding map quality measures, RMSE had the highest values for the sand fraction. This is the same behavior found by other studies that implemented different algorithms (geoestatistics and machine learning) (Poggio and Gimona, 2017; Laborczi et al., 2019; Liu et al., 2020); however, the joint statement in those studies was that, for sand fraction, the ranges were wider and the SD was higher compared with clay and silt fraction. The CCC values for sand were higher than clay and silt in all cases, and this is the same behavior found by Mulder et al. (2016) in France. Additionally, for all PSF, they found that the predictions

330 were less reliable for deepest layers; this is the same statement found in this work. Their map quality measures are better than ours; however, the soil samples used in that work were between 28.000 and 3000, decreasing with increasing depth.

The qualitative evaluation for SG at a global scale showed that coarse-scale patterns are well reproduced (Poggio et al., 2021). Nevertheless, in a quantitative evaluation with Colombian soils, SG products cannot explain the variance (AVE values negative), their predictions are not according with our validation dataset (CCC is not close to zero), and their RMSE values

335 are significantly higher than ours. Liu et al. (2020) built a national map of silt in China, and they compared their results with SG products through RMSE values. They found that their RMSE values were higher than the RMSE of SG, and in many specific areas, SG did not represent the local behavior of the PSF. In this way, we suggest, for applications that need textural soil information at a national scale, to use our results obtained with individual algorithms (MACHISPLIN and landmap) and the ensemble maps. However, it is important to stand out that in some areas of Orinoquia and Amazon region, the SG had the

340 fewest prediction error; these regions had in common that there was a low soil sampling density.

This work could identify the PSF's better models, error trends, and prediction layers. However, in many areas, depths, and textural fractions, the map quality measurements are low; for example, we desire to increase AVE and CCC values. The causes can be many: the relations between some soil properties and landscape attributes are nonlinear, complex, or unknown, a concept defined by Minasny and McBratney (2010). Linked to the aforementioned is the distribution of the soil samples.

345 The study had an unbalanced representation and spatial clustering; for example, the central zone (Andean region) was the most represented (bias towards potentially productive areas), while the east and southeast zones were the least represented, so many predictions were largely controlled by point data, then, some large artifacts (e.g. lines and blocky outputs) are shown in

these areas, a similar case to that reported by Hengl et al. (2014). These artifacts are derived mainly from covariates related to satellite images, such as bands 6 and 7 from Landsat 8, and its derivatives (Clay ratio and Grain Size Index). These four covariates were obtained from Google Earth Engine with seam carving and had a significant importance in the RFE model and their score importance (overall in RFE model) are between 10 and 14% respect to the best subset of covariates (Appendix A1). Additionally, for agricultural studies, the use of this results will be straitened to the agricultural limit defined in Colombia, places where the results do not have artifacts (Appendix A2).

In Colombia, DSM has great challenges to attend map-user's requirements, such as soil texture predictions with uncertainty improvements and soil maps with better spatial resolution. There are three principal strategies to improve predictions: treatment of unbalanced soil data, management of PSF transformations, and incorporation of new environmental covariates related to soil texture drivers. Attending the first strategy is necessary to raise the soil database with available soil information from other sources such as detailed soil surveys, soil degradation, and soil management studies made by national and governmental institutions (e.g. IGAC, IDEAM, or UPRA); or obtaining the amount of each fraction from other kinds of soil analysis, such as Visible Near InfraRed-Short of soil minerals (Lagacherie et al., 2020). Also, model-building processes by soil group (Kempen et al., 2009) or homosoil (Mallavan et al., 2010; Angelini et al., 2020; Malone et al., 2016) have been used to get pedologically-plausible predictions in areas without high soil-sampling density. Other log-ratio transformations could be applied as a second strategy to improve ALR transformation issues. For example, Wang and Shi (2017) indicated that in some datasets, the changes in the denominator selection in Additive log-ratio transformation could represent different predictions and decrease the accuracy of the estimates; then, using centered log-ratio transformation, this issue could be avoided (Amirian-Chakan et al., 2019); also, data sets with zero values must be threatened with symmetry and isometric log-ratio transformation (Li et al., 2020). Finally, as a third strategy, some qualitative and quantitative environmental covariates could buttress the predictors' stack, such as depth to bedrock and soil horizons designations and thickness; also, to improve the visual quality of the results, a previous covariate analysis could be used, such as principal components (Hengl et al., 2014), or a smoothed strategy.

5   Conclusions

We provided the first comparison of the PSF across Colombia between EML models (MACHISPLIN and landmap) and SG's existing soil texture maps. The study shows that the spatial distribution of soil texture prediction with national datasets was, on average, 17% better (in terms of RMSE) using EML models than the SG products. Between MACHISPLIN and landmap, there was no better EML model because the quantitative statistics were very similar. In function of the PSF, the spatial distributions did not exhibit a fraction with better results. However, layers of 0-5, 5-15, and 15-30 cm obtained the best results, which indicate the effectiveness in the depths closest to the soil surface.

Another valuable contribution developed in this study was the implementation of the spatial ensemble of soil texture fractions on a national scale and at different depths. This implementation identified the best result for each depth and each pixel. Although the SG products had the worst quantitative statistics, in some areas of the country, these products performed well, mainly in the south. However, with the spatial ensembled, the best composition of the models was possible.

The spatial distribution of soil particle-size fractions can provide soil information for water-related applications, ecosystem services, and agricultural and crop modeling. However, the results had limitations, especially with some artifacts in the southern and eastern areas. Treatment of unbalanced soil data and incorporation of more appropriate environmental covariates are crucial to improving accuracy in the future.

385 6 Code and data availability

Dataset are available at: https://portal.edirepository.org/nis/mapbrowse?packageid=edi.746.2. This repository contains the data set for each standard depth. For each sample point are shown PSF and ALR transformations (Trans_1 and Trans_2) (Varón-Ramírez and Araujo-Carrillo, 2022).

Textural soil maps are available at: https://portal.edirepository.org/nis/mapbrowse?packageid=edi.972.3. In this repository the users are going to find 9 raster stacks: PSF obtained with landmap and MACHISPLIN algorithms (2 stacks); PSF obtained from SG (1 stack); residual of the PSF predictions for landmap and MACHISPLIN algorithms and SG (3 stacks); and finally PSF predictions obtained through spatial ensemble technique (3 stacks). All stacks contain information at 5 standard depths (Varón-Ramírez et al., 2022)

Rproject scripts to reproduce the spatial ensemble procedure and models validation area are available at: https://github.com/VimiVaron/Textural-maps-Colombia.git

Appendix A

A1   Importance of covariates in recursive feature elimination for Trans_1 and Trans_2 prediction

Depth (cm)	Variable	Covariates selected	Sum of total score	GSI, Clay_ratio, L8_b6, L8_b7	
				Score	Proportion (%)
0-5	Trans_1	44	501.1	67.8	13.5
	Trans_2	83	451.0	58.3	12.9
5-15	Trans_1	54	549.8	66.5	12.1
	Trans_2	54	494.7	68.8	13.9
15-30	Trans_1	59	704.1	67.9	9.6
	Trans_2	83	692.0	75.9	11.0
30-60	Trans_1	56	629.7	64.8	10.3
	Trans_2	58	651.2	78.2	12.0
60-100	Trans_1	56	564.5	60.0	10.6
	Trans_2	83	593.0	70.6	11.9

Table A1. Representation of importance scores for satellite-derived covariates. GSI: Grain Size Index, L8 b7: Landsat 8 band 7, L8 b6: Landsat 8 band 6.

Appendix B

B1 Agricultural Frontier in Colombia
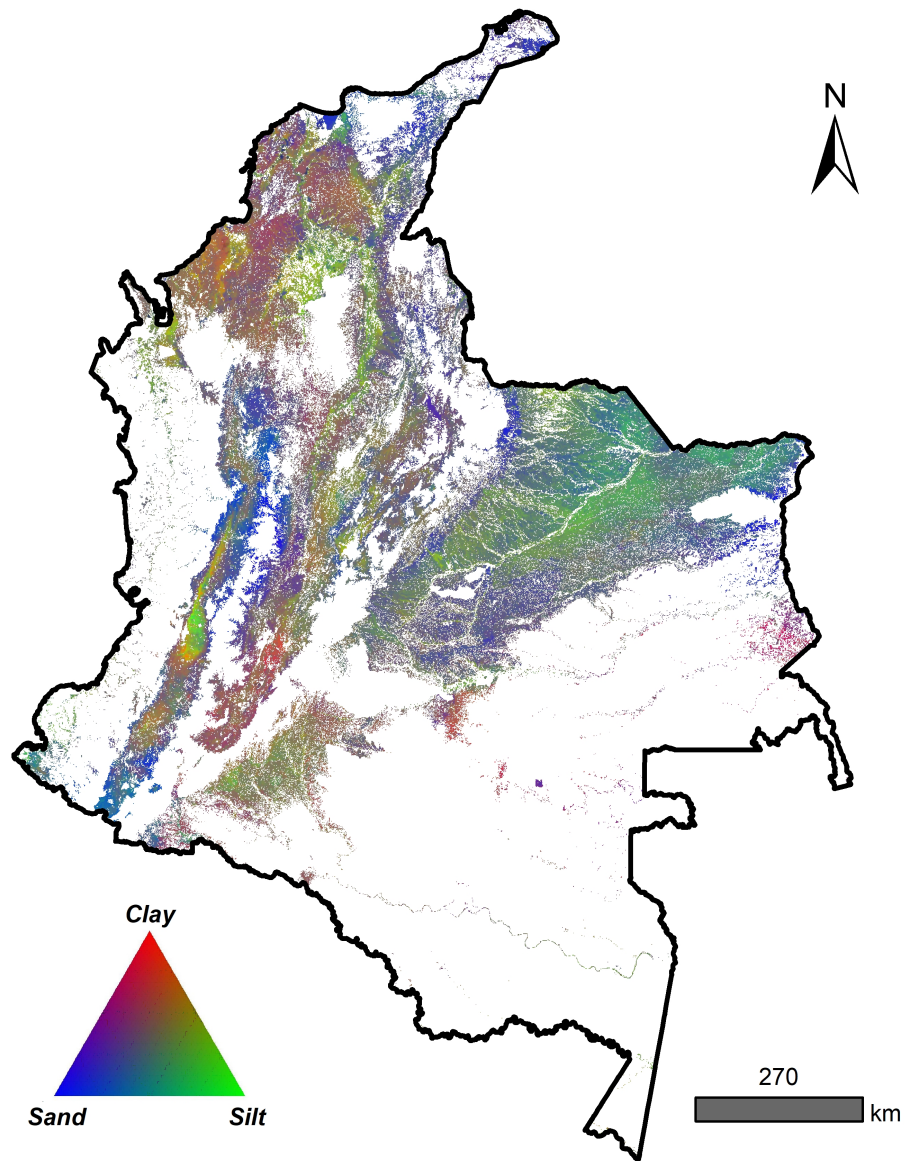
Figure B1. Compositional texture map (Clay, Sand, and Silt), integrated from 0 to 100 cm, in the agricultural frontier in Colombia.

Author contributions. Varón-Ramírez contributed with conceptualization of soil textural-data management methodology, data cleaning, harmonization of the soil dataset, writing of results and discussion, and elaboration of cartography results. Araujo-Carrillo contributed with environmental covariates construction, writing of methodology, results and discussion, and elaboration of cartography results. Guevara contributed with conceptualization of spatial ensemble models strategies and the comparison with global products and writing the introduction and discussion.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. The authors want to thank the Ministerio de Agricultura y Desarrollo Rural (MADR), the Corporación Colombiana de Investigación Agropecuaria (AGROSAVIA), and the Universidad Nacional Autónoma de México (UNAM) for supporting this study. Mario Guevara acknowledges support from grants: UNESCO-IGCP-IUGS, 2022 (#765), UNAM-PAPIIT, 2021 (#IA204522) and USDA-NIFA-AFRI, USA, 2019 (#2019-67022-29696)

References

Aitchison, J.: The statistical analysis of compositional data, Chapman and Hall, 1986.

Amirian-Chakan, A., Minasny, B., Taghizadeh-Mehrjardi, R., Akbarifazli, R., Darvishpasand, Z., and Khordehbin, S.: Some practical aspects of predicting texture data in digital soil mapping, Soil Tillage Research., 194, 104 289, https://doi.org/10.1016/j.still.2019.06.006, 2019.

Angelini, M., Kempen, B., Heuvelink, G., Temme, A., and Ransom, M.: Extrapolation of a structural equation model for digital soil mapping, Geoderma, 367, 114 226, https://doi.org/10.1016/j.geoderma.2020.114226, https://www.sciencedirect.com/science/article/pii/S0016706119325376, 2020.

Angelini, M. E., Heuvelink, G. B., Kempen, B., and Morrás, H. J.: Mapping the soils of an Argentine Pampas region using structural equation modelling, Geoderma, 281, 102–118, https://doi.org/10.1016/j.geoderma.2016.06.031, https://www.sciencedirect.com/science/article/pii/S0016706116302798, 2016.

Araujo, M. A., Zinn, Y. L., and Lal, R.: Soil parent material, texture and oxide contents have little effect on soil organic carbon retention in tropical highlands, Geoderma, 300, 1–10, https://doi.org/10.1016/j.geoderma.2017.04.006, https://www.sciencedirect.com/science/article/pii/S0016706117301076, 2017.

Araujo-Carrillo, G. A., Varón-Ramírez, V. M., Jaramillo-Barrios, C. I., Estupiñan-Casallas, J. M., Silva-Arero, E. A., Gómez-Latorre, D. A., and Martínez-Maldonado, F. E.: IRAKA: The first Colombian soil information system with digital soil mapping products, Catena, 196, 104 940, https://doi.org/10.1016/j.catena.2020.104940, 2021.

Arrouays, D., Grundy, M. G., Hartemink, A. E., Hempel, J. W., Heuvelink, G. B., Hong, S. Y., Lagacherie, P., Lelyk, G., McBratney, A. B., McKenzie, N. J., d.L. Mendonca-Santos, M., Minasny, B., Montanarella, L., Odeh, I. O., Sanchez, P. A., Thompson, J. A., and Zhang, G.-L.: Chapter Three - GlobalSoilMap: Toward a Fine-Resolution Global Grid of Soil Properties, vol. 125 of *Advances in Agronomy*, pp. 93–134, Academic Press, https://doi.org/10.1016/B978-0-12-800137-0.00003-0, https://www.sciencedirect.com/science/article/pii/B9780128001370000030, 2014.

Beaudette, D. E., Roudier, P., and O'Geen, A.: Algorithms for quantitative pedology: A toolkit for soil scientists, Computers & Geosciences, 52, 258–268, https://doi.org/10.1016/j.cageo.2012.10.020, 2013.

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., Casalicchio, G., and Jones, Z. M.: mlr: Machine Learning in R, Journal of Machine Learning Research, 17, 1–5, http://jmlr.org/papers/v17/15-066.html, 2016.

Bishop, T., McBratney, A., and Laslett, G.: 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, 1999.

Bönecke, E., Meyer, S., Vogel, S., Schröter, I., Gebbers, R., Kling, C., Kramer, E., Lück, K., Nagel, A., Philipp, G., Gerlach, F., Palme, S., Scheibe, D., Zieger, K., and Rühlmann, J.: Guidelines for precise lime management based on high-resolution soil pH, texture and SOM maps generated from proximal soil sensing data, Precis. Agric., 22, 493–523, https://doi.org/10.1007/s11119-020-09766-8, 2021.

Breiman, L.: Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author), Statist. Sci., 16, 199–231, https://doi.org/10.1214/ss/1009213726, 2001.

Brenning, A.: Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The R package sperrorest, in: 2012 IEEE International Geoscience and Remote Sensing Symposium, pp. 5372–5375, https://doi.org/10.1109/IGARSS.2012.6352393, 2012.

Brown, J.: MACHISPLIN, urlhttps://github.com/jasonleebrown/machisplin, 2021.

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

Campbell, P. M. d. M., Fernandes, E. I., Francelino, M. R., Demattê, J. A. M., Pereira, M. G., Guimarães, C. C. B., and Pinto, L. A. d. S. R.: Digital soil mapping of soil properties in the "Mar de Morros" environment using spectral data, Revista Brasileira de Ciência do Solo, 42, https://doi.org/10.1590/18069657rbcs20170413, 2019.

Catoni, M., D'Amico, M. E., Zanini, E., and Bonifacio, E.: Effect of pedogenic processes and formation factors on organic matter stabilization in alpine forest soils, Geoderma, 263, 151–160, https://doi.org/10.1016/j.geoderma.2015.09.005, https://www.sciencedirect.com/science/article/pii/S0016706115300781, 2016.

Caubet, M., Román Dobarco, M., Arrouays, D., Minasny, B., and Saby, N. P. A.: Merging country, continental and global predictions of soil texture: Lessons from ensemble modelling in France, Geoderma, 337, 99–110, https://doi.org/10.1016/j.geoderma.2018.09.007, 2019.

Cortés, A., Cortés, M., Guevara, J., and Palacino, A.: Mapas de suelos de Colombia. Memoria explicativa., Instituto Geográfico Agustín Codazzi (IGAC), Subdirección Agrológica, Bogotá, 1982.

Dharumarajan, S. and Hegde, R.: Digital mapping of soil texture classes using Random Forest classification algorithm, Soil Use Manage., n/a, https://doi.org/10.1111/sum.12668, 2020.

FAO: Sistema de Información de Suelos de Latinoamérica y el Caribe – SISLAC, urlhttp://54.229.242.119/sislac/es, 2020.

Flórez, A.: Colombia: evolución de sus relieves y modelados, Unilibros, Bogot´, 2003.

Grunwald, S., Thompson, J. A., and Boettinger, J. L.: Digital Soil Mapping and Modeling at Continental Scales: Finding Solutions for Global Issues, Soil Sci. Soc. Am. J., 75, 1201–1213, https://doi.org/10.2136/sssaj2011.0025, 2011.

Guevara, M., Olmedo, G. F., Stell, E., Yigini, Y., Aguilar Duarte, Y., Arellano Hernández, C., Arévalo, G. E., Arroyo-Cruz, C. E., Bolivar, A., Bunning, S., Bustamante Cañas, N., Cruz-Gaistardo, C. O., Davila, F., Dell Acqua, M., Encina, A., Figueredo Tacona, H., Fontes, F., Hernández Herrera, J. A., Ibelles Navarro, A. R., Loayza, V., Manueles, A. M., Mendoza Jara, F., Olivera, C., Osorio Hermosilla, R., Pereira, G., Prieto, P., Ramos, I. A., Rey Brina, J. C., Rivera, R., Rodríguez-Rodríguez, J., Roopnarine, R., Rosales Ibarra, A., Rosales Riveiro, K. A., Schulz, G. A., Spence, A., Vasques, G. M., Vargas, R. R., and Vargas, R.: No silver bullet for digital soil mapping: country-specific soil organic carbon estimates across Latin America, SOIL, 4, 173–193, https://doi.org/10.5194/soil-4-173-2018, https://soil.copernicus.org/articles/4/173/2018/, 2018.

Hengl, T.: landmap, urlhttps://github.com/envirometrix/landmap, 2021.

Hengl, T. and MacMillan, R. A.: Predictive soil mapping with R, Lulu. com, 2019.

Hengl, T., de Jesus, J. M., MacMillan, R. A., Batjes, N. H., Heuvelink, G. B., Ribeiro, E., Samuel-Rosa, A., Kempen, B., Leenaars, J. G., Walsh, M. G., et al.: SoilGrids1km—global soil information based on automated mapping, PloS one, 9, e105 992, https://doi.org/10.1371/journal.pone.0105992, 2014.

Hengl, T., Miller, M. A. E., Križan, J., Shepherd, K. D., Sila, A., Kilibarda, M., Antonijević, O., Glušica, L., Dobermann, A., Haefele, S. M., McGrath, S. P., Acquah, G. E., Collinson, J., Parente, L., Sheykhmousa, M., Saito, K., Johnson, J.-M., Chamberlin, J., Silatsa, F. B. T., Yemefack, M., Wendt, J., MacMillan, R. A., Wheeler, I., and Crouch, J.: African soil properties and nutrients mapped at 30 m spatial resolution using two-scale ensemble machine learning - Scientific Reports, Sci. Rep., 11, 1–18, https://doi.org/10.1038/s41598-021-85639-y, 2021.

IDEAM: Mapa de Coberturas de la Tierra Metodología Corine Land Cover adaptada para Colombia Escala 1:100.000 (Período 2010 - 2012), 2014.

IDEAM: Climatological atlas of Colombia – Interactive year 2015, urlhttp://atlas.ideam.gov.co/visorAtlasClimatologico.html, 2015.

IGAC: Mapa Suelos de Colombia, IGAC, Bogotá, 2003.

485 IGAC: Suelos y Tierras de Colombia, IGAC, Bogotá, 2015.

James, G., Witten, D., Hastie, T., and Tibshirani, R.: An Introduction to Statistical Learning: with Applications in R, Springer, https://doi.org/10.1007/978-1-0716-1418-1, https://faculty.marshall.usc.edu/gareth-james/ISL/, 2013.

Kaya, F. and Başayiğit, L.: Spatial Prediction and Digital Mapping of Soil Texture Classes in a Floodplain Using Multinomial Logistic Regression, in: Intelligent and Fuzzy Techniques for Emerging Conditions and Digital Transformation, edited by Kahraman, C., Cebi, S., Ce-

490 vik Onar, S., Oztaysi, B., Tolga, A. C., and Sari, I. U., pp. 463–473, Springer International Publishing, Cham, https://doi.org/10.1007/978-3-030-85577-2_55, 2022.

Kempen, B., Brus, D. J., Heuvelink, G. B., and Stoorvogel, J. J.: Updating the 1:50,000 Dutch soil map using legacy soil data: A multinomial logistic regression approach, Geoderma, 151, 311–326, https://doi.org/10.1016/j.geoderma.2009.04.023, https://www.sciencedirect.com/science/article/pii/S0016706109001475, 2009.

495 Kempen, B., Brus, D. J., Stoorvogel, J. J., Heuvelink, G. B., and de Vries, F.: Efficiency comparison of conventional and digital soil mapping for updating soil maps, Soil Science Society of America Journal, 76, 2097–2115, https://doi.org/10.2136/sssaj2011.0424, 2012.

Khaledian, Y. and Miller, B. A.: Selecting appropriate machine learning methods for digital soil mapping, Applied Mathematical Modelling, 81, 401–418, 2020.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., Team, R. C., et al.: Package

500 'caret', The R Journal, 2020.

Laborczi, A., Szatmári, G., Kaposi, A. D., and Pásztor, L.: Comparison of soil texture maps synthetized from standard depth layers with directly compiled products, Geoderma, 352, 360–372, https://doi.org/10.1016/j.geoderma.2018.01.020, 2019.

Lagacherie, P., Arrouays, D., Bourennane, H., Gomez, C., and Nkuba-Kasanda, L.: Analysing the impact of soil spatial sampling on the performances of Digital Soil Mapping models and their evaluation: A numerical experiment on Quantile Random Forest using clay

505 contents obtained from Vis-NIR-SWIR hyperspectral imagery, Geoderma, 375, 114 503, https://doi.org/10.1016/j.geoderma.2020.114503, https://www.sciencedirect.com/science/article/pii/S0016706119322736, 2020.

Lang, M., Binder, M., Richter, J., Schratz, P., Pfisterer, F., Coors, S., Au, Q., Casalicchio, G., Kotthoff, L., and Bischl, B.: mlr3: A modern object-oriented machine learning framework in R, Journal of Open Source Software, https://doi.org/10.21105/joss.01903, https://joss.theoj.org/papers/10.21105/joss.01903, 2019.

510 Lark, R. and Bishop, T.: Cokriging particle size fractions of the soil, European Journal of Soil Science, 58, 763–774, https://doi.org/10.1111/j.1365-2389.2006.00866.x, 2007.

Lawrence, I. and Lin, K.: A concordance correlation coefficient to evaluate reproducibility, Biometrics, pp. 255–268, https://doi.org/10.2307/2532051, 1989.

Li, J., Wan, H., and Shang, S.: Comparison of interpolation methods for mapping layered soil particle-size fractions and texture in an arid

515 oasis, CATENA, 190, 104 514, https://doi.org/10.1016/j.catena.2020.104514, 2020.

Liu, F., Zhang, G.-L., Song, X., Li, D., Zhao, Y., Yang, J., Wu, H., and Yang, F.: High-resolution and three-dimensional mapping of soil texture of China, Geoderma, 361, 114 061, https://doi.org/10.1016/j.geoderma.2019.114061, 2020.

Llamas, R. M., Guevara, M., Rorabaugh, D., Taufer, M., and Vargas, R.: Spatial Gap-Filling of ESA CCI Satellite-Derived Soil Moisture Based on Geostatistical Techniques and Multiple Regression, Remote Sens., 12, 665, https://doi.org/10.3390/rs12040665, 2020.

520 Lovelace, R., Nowosad, J., and Muenchow, J.: Geocomputation with R, CRC Press, 2019.

Mallavan, B., Minasny, B., and McBratney, A.: Homosoil, a Methodology for Quantitative Extrapolation of Soil Information Across the Globe, pp. 137–150, Springer Netherlands, Dordrecht, https://doi.org/10.1007/978-90-481-8863-5_12, 10.1007/978-90-481-8863-5_12, 2010.

Malone, B., Searle, R., Malone, B., and Searle, R.: Updating the Australian digital soil texture mapping (Part 2∗): spatial modelling of merged field and lab measurements, Soil Res., 59, 435–451, https://doi.org/10.1071/SR20284, 2021.

Malone, B. P., Jha, S. K., Minasny, B., and McBratney, A. B.: Comparing regression-based digital soil mapping and multiple-point geostatistics for the spatial extrapolation of soil data, Geoderma, 262, 243–253, https://doi.org/10.1016/j.geoderma.2015.08.037, 2016.

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

Minasny, B. and McBratney, A.: Methodologies for global soil mapping, in: Digital soil mapping, pp. 429–436, Springer, https://doi.org/10.1007/978-90-481-8863-5_34, 2010.

Møller, A. B., Beucher, A. M., Pouladi, N., and Greve, M. H.: Oblique geographic coordinates as covariates for digital soil mapping, SOIL, 6, 269–289, https://doi.org/10.5194/soil-6-269-2020, 2020.

Mulder, V. L., Lacoste, M., Richer-de Forges, A., and Arrouays, D.: GlobalSoilMap France: High-resolution spatial modelling the soils of France up to two meter depth, Science of the Total Environment, 573, 1352–1369, https://doi.org/10.1016/j.scitotenv.2016.07.066, 2016.

Niang, M. A., Nolin, M. C., Jégo, G., and Perron, I.: Digital Mapping of Soil Texture Using RADARSAT-2 Polarimetric Synthetic Aperture Radar Data, Soil Science Society of America Journal, 78, 673–684, https://doi.org/10.2136/sssaj2013.07.0307, 2014.

Odeh, I. O., Todd, A. J., and Triantafilis, J.: Spatial prediction of soil particle-size fractions as compositional data, Soil Science, 168, 501–515, https://doi.org/10.1097/01.ss.0000080335.10341.23, 2003.

Orton, T., Pringle, M., and Bishop, T.: A one-step approach for modelling and mapping soil properties based on profile data sampled over varying depth intervals, Geoderma, 262, 174–186, https://doi.org/10.1016/j.geoderma.2015.08.013, 2016.

Osman, K. T.: Soils: principles, properties and management, Dordrecht; New York: Springer, 2013.

Patel, K. F., Fansler, S. J., Campbell, T. P., Bond-Lamberty, B., Smith, A. P., RoyChowdhury, T., McCue, L. A., Varga, T., and Bailey, V. L.: Soil texture and environmental conditions influence the biogeochemical responses of soils to drought and flooding, Communications Earth & Environment, 2, 127, https://doi.org/10.1038/s43247-021-00198-4, 2021.

Pawlowsky-Glahn, V. and Olea, R. A.: Geostatistical analysis of compositional data, Oxford University Press, 2004.

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

Poggio, L., de Sousa, L. M., Batjes, N. H., Heuvelink, G. B. M., Kempen, B., Ribeiro, E., and Rossiter, D.: SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty, SOIL, 7, 217–240, https://doi.org/10.5194/soil-7-217-2021, 2021.

Polley, E. C. and Van der Laan, M. J.: Super learner in prediction, U.C. Berkeley Division of Biostatistics Working Paper Series, 266, 2010.

Poveda, G.: La hidroclimatología de Colombia: una síntesis desde la escala inter-decadal hasta la escala diurna, Rev. Acad. Colomb. Cienc, 28, 201–222, 2004.

Radočaj, D., Jurišić, M., Zebec, V., and Plaščak, I.: Delineation of Soil Texture Suitability Zones for Soybean Cultivation: A Case Study in Continental Croatia, Agronomy, 10, 823, https://doi.org/10.3390/agronomy10060823, 2020.

Ramcharan, A., Hengl, T., Nauman, T. W., Brungard, C. W., Waltman, S. W., Wills, S., and Thompson, J.: Soil Property and Class Maps of the Conterminous United States at 100-Meter Spatial Resolution, Soil Science Society of America Journal, 82, 186–201, https://doi.org/10.2136/sssaj2017.04.0122, 2018.

Rangel-Ch, J. O. and Aguilar, M.: Una aproximación sobre la diversidad climática en las regiones naturales de Colombia, Diversidad Biótica I. Instituto de Ciencias Naturales-Universidad Nacional de Colombia-Inderena, Bogotá, pp. 25–77, 1995.

Richer-de Forges, A. C., Arrouays, D., Chen, S., Dobarco, M. R., Libohova, Z., Roudier, P., Minasny, B., and Bourennane, H.: Hand-feel soil texture and particle-size distribution in central France. Relationships and implications, Catena, 213, 106 155, https://doi.org/10.1016/j.catena.2022.106155, 2022.

Samuel-Rosa, A., Heuvelink, G., Vasques, G., and Anjos, L.: Do more detailed environmental covariates deliver more accurate soil maps?, Geoderma, 243, 214–227, https://doi.org/10.1016/j.geoderma.2014.12.017, 2015.

Soil Survey Staff: Keys to Soil Taxonomy, 12th ed, USDA-Natural Resources Conservation Service, 2014.

Soropa, G., Mbisva, O. M., Nyamangara, J., Nyakatawa, E. Z., Nyapwere, N., and Lark, R. M.: Spatial variability and mapping of soil fertility status in a high-potential smallholder farming area under sub-humid conditions in Zimbabwe, SN Appl. Sci., 3, 1–19, https://doi.org/10.1007/s42452-021-04367-0, 2021.

Tsagris, M., Giorgos, A., Alenazi, A., and Adam, C.: Compositional: Compositional Data Analysis, urlhttps://cran.r-project.org/web/packages/Compositional/index.html, 2022.

Varón-Ramírez, V. and Araujo-Carrillo, G.: Textural soil data, Colombia, 0 - 100 cm, https://doi.org/10.6073/pasta/3f91778c2f6ad46c3cc70b61f02532db, 2022.

Varón-Ramírez, V., Araujo-Carrillo, G., and Guevara, M.: Textural soil maps, Colombia, 0 - 100 cm, https://doi.org/10.6073/pasta/d6c0bf5847aa40836b42dcc3e0ea874e, 2022.

Wadoux, A. M.-C., Román-Dobarco, M., and McBratney, A. B.: Perspectives on data-driven soil research, European Journal of Soil Science, 72, 1675–1689, https://doi.org/10.1016/j.apm.2019.12.016, 2021a.

Wadoux, A. M. J.-C., Minasny, B., and McBratney, A. B.: Machine learning for digital soil mapping: Applications, challenges and suggested solutions, Earth-Sci. Rev., 210, 103 359, https://doi.org/10.1016/j.earscirev.2020.103359, 2020.

Wadoux, A. M. J.-C., Heuvelink, G. B. M., de Bruin, S., and Brus, D. J.: Spatial cross-validation is not the right way to evaluate map accuracy, Ecol. Model., 457, 109 692, https://doi.org/10.1016/j.ecolmodel.2021.109692, 2021b.

Wang, Z. and Shi, W.: Mapping soil particle-size fractions: A comparison of compositional kriging and log-ratio kriging, Journal of Hydrology, 546, 526–541, https://doi.org/10.1016/j.jhydrol.2017.01.029, https://www.sciencedirect.com/science/article/pii/S0022169417300306, 2017.

Wang, Z., Shi, W., Zhou, W., Li, X., and Yue, T.: Comparison of additive and isometric log-ratio transformations combined with machine learning and regression kriging models for mapping soil particle size fractions, Geoderma, 365, 114 214, https://doi.org/10.1016/j.geoderma.2020.114214, 2020.

Webster, R. and Oliver, M. A.: Geostatistics for environmental scientists, John Wiley & Sons, https://doi.org/10.1002/9780470517277, 2007.

Witten, I., Frank, E., Hall, M., and Pal, C.: What's It All About?, Data Mining: Practical machine learning tools and techniques, pp. 3–38, https://doi.org/10.1016/C2009-0-19715-5, 2011.

Yang, Y.: Chapter 4 - Ensemble Learning, in: Temporal Data Mining Via Unsupervised Ensemble Learning, edited by Yang, Y., pp. 35–56, Elsevier, https://doi.org/10.1016/B978-0-12-811654-8.00004-X, 2017.

Yigini, Y., Olmedo, G., Reiter, S., Baritz, R., Viatkin, K., and Vargas, R.: Soil organic carbon mapping: Cookbook, 2018.

Zhang, C. and Ma, Y.: Ensemble machine learning: methods and applications, Springer, 2012.

Zhang, Y. and Hartemink, A. E.: Quantifying short-range variation of soil texture and total carbon of a 330-ha farm, Catena, 201, 105 200, https://doi.org/10.1016/j.catena.2021.105200, 2021.

Zounemat-Kermani, M., Batelaan, O., Fadaee, M., and Hinkelmann, R.: Ensemble machine learning paradigms in hydrology: A review, Journal of Hydrology, 598, 126 266, https://doi.org/10.1016/j.jhydrol.2021.126266, https://www.sciencedirect.com/science/article/pii/S0022169421003139, 2021.