Revista Brasileira de Ciência do Solo

Division - Soil in Space and Time | Commission - Pedometry

Multinomial Logistic Regression and Random Forest Classifiers in Digital Mapping of Soil Classes in Western Haiti

Wesly Jeune[(1)], **Márcio Rocha Francelino**[(2)], **Eliana de Souza**[(3)*], **Elpídio Inácio Fernandes Filho**[(2)] **and Genelício Crusoé Rocha**[(2)]

[(1)] Université Quisqueya, Faculté des Sciences de l'Agriculture et de l'Environnement, Port-au-Prince, Ouest, Haiti.
[(2)] Universidade Federal de Viçosa, Departamento de Solos, Viçosa, Minas Gerais, Brasil.
[(3)] Universidade Federal de Viçosa, Departamento de solos, Programa de Pós-Graduação em Solos e Nutrição de Plantas, Viçosa, Minas Gerais, Brasil.

ABSTRACT: Digital soil mapping (DSM) has been increasingly used to provide quick and accurate spatial information to support decision-makers in agricultural and environmental planning programs. In this study, we used a DSM approach to map soils in western Haiti and compare the performance of the Multinomial Logistic Regression (MLR) with Random Forest (RF) to classify the soils. The study area of 4,300 km^2 is mostly composed of diverse limestone rocks, alluvial deposits, and, to a lesser extent, basalt. A soil survey was conducted whereby soils were described and classified at 258 sites. Soil samples were collected and subjected to physical and chemical analyses. Recursive Feature Elimination (RFE) was used to select the most important covariates from auxiliary data, such as climate, lithology, and morphometric properties to describe the soil-landscape relationship. Mapping performance was assessed by the Kappa index and overall accuracy derived from a confusion matrix generated using a 5-fold cross validation process. In addition, an external mapping validation was carried out using an independent soil dataset. Accordingly, the soil dataset was split into 80 % and 20 % for training and validation of the models, respectively. No significant statistical difference (Z = 0.56< |1.96|) was found between maps generated with both classifiers (Kappa index 0.45 for MLR and 0.42 for RF). Based on the Kappa values, the classification performance can be characterized as moderate for both algorithms. Surprisingly, the RF classifier outperformed MLR in the validation process (Kappa values of 0.55 and 0.33, respectively). These results suggest a higher generalization ability of RF. However, no significant statistical difference (Z = 1.83< |1.96|) was observed. The soil map derived from RF indicated the occurrence of Leptosols (48.5 %), Gleysols (19.6 %), Chernozems (8 %), and Fluvisols (6.6 %) in most of the study area. The DSM approaches proved suitable for mapping soils in western Haiti and could be used in other parts of the country, thereby closing information gaps with regard to Haitian soils.

Keywords: auxiliary data, digital soil mapping, soil survey, data-mining.

*** Corresponding author:**
E-mail: elianadsouza@yahoo.com.br

Received: June 3, 2017

Approved: January 23, 2018

How to cite: Jeune W, Francelino MR, Souza E, Fernandes Filho EI, Rocha GC. Multinomial logistic regression and random forest classifiers in digital mapping of soil classes in western Haiti. Rev Bras Cienc Solo. 2018;42:e0170133.
https://doi.org/10.1590/18069657rbcs20170133

INTRODUCTION

In the last decades, soil mapping has become increasingly automated by means of computational tools, owing to the great advances in computational technology, the development of geographic information systems (GIS), and the availability of more extensive geographic data. This development has been in some ways fostered by the current demands for soil data to address environmental issues and climate change (Grunwald, 2010; Minasny and McBratney, 2016; Muñoz-Rojas et al., 2017), and has contributed to the creation of a sub-discipline of soil science known as digital soil mapping - DSM (Minasny and McBratney, 2016). The framework for DSM was conceptualized by McBratney et al. (2003) and is based on the Scorpan model. As a spatial information system, DSM was created from numerical models to produce digital information that explains temporal and spatial variation in soil classes and properties based on correlated environmental variables. Comprehensive reviews of methods and techniques used for DSM were published by McBratney et al. (2003) and Scull et al. (2003).

In comparison to conventional soil mapping, the uncertainty and accuracy associated to the entire process of mapping can be quantified by DSM approaches. They provide fast and increasingly accurate information by using modern GIS techniques, spatial data from digital elevation models, and remote sensing imagery together with auxiliary data (Bacon et al., 2010; Moonjun et al., 2010).

In several countries, the development and use of predictive models using DSM techniques has constantly evolved, producing more precise information (Grunwald et al., 2011; Arrouays et al., 2017). Yet, there is a lack of application of DSM in some regions of the world, such e.g., Haiti, where the only soil map, representing the whole country, was produced at a scale of 1:250,000 (Libohova et al., 2017). Soil resources in Haiti are not well known and the few studies that have been carried out are largely exploratory. Notable examples of Haitian soil studies are the studies of Sweet (1926) in the Artibonite area, Haspil and Butterlin (1955), Colmet-Daage and Lagache (1965), and Guthrie and Shannon (2004) in west and southwest Haiti. The most recent studies are those of Chaves et al. (2010) on the Mapou region (Southeast of the country), of Hylkema (2011) on soil fertility in five pilot watersheds, and of Libohova et al. (2017), who developed a detailed soil map at a 1:24,000 scale for an area of 3,000 ha located in the Cul-de-Sac depression.

When performing soil mapping, the decision on which predictive model to use and how to select the best set of predictive covariates is important. Many algorithms for categorical classification, as of soil classes, are implemented in a variety of available software programs (McBratney et al., 2000). Among the most popular classifiers are Neural Networks, Fuzzy logic, tree-based model (Cart and Random Forest), and Logistic Regression.

The large amount of spatial data available for use as soil predictors has intensified the need to use data-mining technique to establish a more efficient relationship between variables and predictors and to optimize the selection of a promising predictor set (Hastie et al., 2009; Kuhn and Johnson, 2013; Heung et al., 2014). Some studies addressed the comparison of data mining approaches for predicting soil classes by different predictive models. Nevertheless, despite the broad success of DSM techniques using tree-based model and logistic regression, few studies have compared the performance of promising classifiers for soil class mapping, such as logistic regression and Random Forest (Hengl et al., 2007; Collard et al., 2014; Taghizadeh-Mehrjardi et al., 2015; Pahlavan-Rad et al., 2016; Camera et al., 2017; Heung et al., 2017). Under similar conditions of western Haiti, this kind of studies is rarer.

In this context, the main objective of this research was to assess the use of DSM techniques for mapping soils in western Haiti, specifically by assessing the importance of environmental covariates in soil mapping and comparing the performance of Multinomial Logistic Regression (MLR) with Random Forest (RF) classifiers.

MATERIALS AND METHODS

Study area

The study area (between 71° 42′ 39″ W and 73° 4′ 22″ W longitude and 18° 15′ 30″ N and 18° 58′ 25″ N latitude) is located in the western region of Haiti and covers approximately 4,300 km^2 (Figure 1). According to Köppen classification system, the regional climate is tropical (Aw), with low temperature variation in plains and hills during the year; while at higher elevations (>1,500 m) the climate is subtropical highland (Cwb), with dry winters and hot and humid summers.

The original vegetation of the study area consisted of highland forests, pine forests, xerophilous forest/grassland, savanna, and mangrove forest (Robar, 1984). Currently, due to the demographic pressure and deforestation, the natural vegetation has been incredibly reduced, being transformed into small fragments of isolated plants and species (Koohafkan and Lilin, 1989; Churches et al., 2014).

Western Haiti is geologically complex, composed of various sedimentary and igneous rocks folded, faulted, and fractured by tectonic activities in the Caribbean region (Potter et al., 2004). It is crossed by the Cul-de-Sac depression, a tectonic depression where the hypersaline Lake Azueï or Etang Saumâtre lies. Aside from basaltic rock, the lithology is markedly formed by sedimentary formations composed of diverse limestone rocks, in addition to alluvial sediments, dejection cones, and mud banks. Tertiary carbonate formations, such as marine limestone and limestone marlstone, represent a significant percentage of Haitian territory (Woodring et al., 1924). It is worth mentioning that carbonate formations account for more than 65 % of the lithology of the area.

The region is geomorphologically dominated by mountainous landforms and deep valleys, as is representative of the general geomorphological features in other parts of Haiti. The landform system was shaped by the regional endogenic forces and erosional processes. The elevation ranges from sea level to 2,666 m (Morne La Selle), the highest point in the country. The relief is rugged with slopes between 0 and 62 degrees.

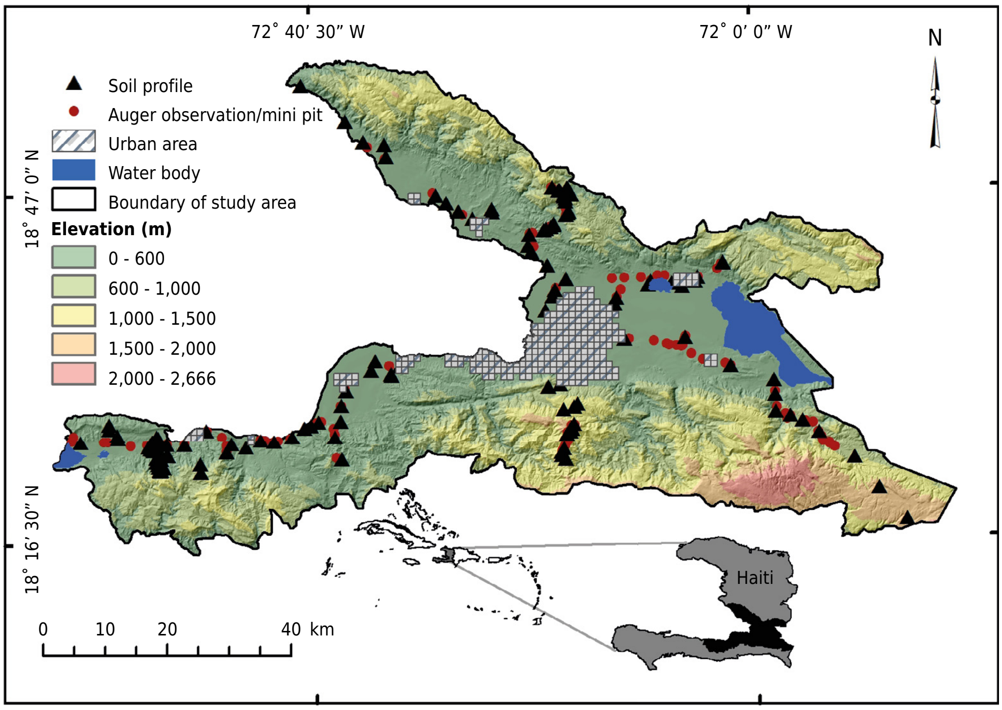

Figure 1. Spatial distribution of the soil pits described and sampled in western Haiti, Caribbean.

Agricultural land, forestland (mixed deciduous, coniferous, and bushland), mixed urban or built-up land, and bare exposed rock are the most common categories of land use in the region. This land use pattern has accelerated the erosional processes that are considered a particularly severe problem in Haiti (Bayard et al., 2006).

The area of this study encompasses the Haitian capital, Port-au-Prince, along with its metropolitan region including the municipalities of Gressier, Leogâne, Grand-Goâve, Petit-Goâve, Cabaret, Arcahaie, Thomazeau, Ganthier, Fond-Verette, Cornillon, and Kenscoff. The area has a population of approximately 4 million people, corresponding to approximately 36 % of the country population (IHSI, 2015).

Soil survey and classification

A pedological survey was carried out in two phases: the first based on a conventional survey approach, in which the sampling was undertaken in transects; and in the second phase, a randomly stratified sample was established by selecting the locations for soil sampling based on the sampling density of the first phase. During these surveys, soils at 258 sites were described and sampled, accounting for 140 complete soil profiles and, 118 auger observations and mini pits (Figure 1). The soil samples collected were subjected to physical, chemical, and mineralogical analyses, according to the guidelines of Claessen (1997).

The soils were classified according to the Brazilian Soil Classification System - SiBCS (Santos et al., 2006) at the suborder level (Table 1). Eight soil groups were observed in the study area: Cambisols, Chernozems, Fluvisols, Gleysols, Leptosols, Luvisols, Nitisols, and Vertisols. The corresponding reference soil groups in the World Reference Base (WRB, 2015) were used to delineate the soil map units in the mapping process.

Due to the low level of detail which characterizes the soil survey scale (1:50,000), coupled with the spatial resolution of most covariates (30 m), nine soil map units were delineated, of which four represented a single component (Chernozems, Gleysols, Leptosols, and Fluvisols), for being considered major components of the study area. Likewise, soil associations were formed by two or three soil components (Table 1). The associations define the map unit by grouping soil taxonomic units that are distinctly and geographically associated and can be mapped individually in more detailed scale mapping. Among the map units formed by associations, the extension of Luvisols and Vertisols is minor. Leptosols and rocky outcrops occur as a complex and were therefore mapped within a single map unit.

Table 1. Soil map units and landscape properties

MU[(1)]	Taxonomic unit		n	Altitude			Lithology
	IUSS-WRB	SiBCS[(2)]		Min	Med	Max	
					m		
LP	Leptosols	*Neossolos Litólicos*	59	14	575	1,772	Qa, Cb, Ep, Ms, O, Qa, Qc
LP1	Leptosols + rocky outcrops	*Neossolos Litólicos + Afloramento de rocha*	19	220	868	1,674	Qa, Cb, Qa, Qc
CM	Cambisols + Fluvisols	*Cambissolos Háplicos + Neossolos Flúvicos*	26	29	180	663	Ms, O, Qa, Qc
GL	Gleysols	*Gleissolos Háplicos*	33	0	20	40	Qa, Cf
CH	Chernozems	*Chernossolos Rênzicos + Chernossolos Ebânicos*	19	22	436	733	Ep, Qa, Cb, O, Qc
NT	Nitisols + Leptosols	*Nitossolos Vermelho + Neossolos Litólicos*	17	1,300	1,571	1,758	Qc, Cb
FL	Fluvisols	*Neossolos Flúvicos*	67	15	50	455	Ep, Cf, Cb, O, Qa
LV	Luvisols + Chernozems	*Luvissolos Háplico + Chernossolos Rênzicos + Chernossolos Ebânicos*	6	29	155	356	Ep, Qa, Cb
VR	Vertisols + Chernozems	*Vertissolos Háplicos + Chernossolos Háplicos*	12	32	538	666	Bm, Qa, Qc

[(1)] Map units. [(2)] Brazilian Soil Classification System. n = number of soil data; Bm = basalts; Cb = volcanic sedimentary rocks; Cf = flysch and, sandstone, and limestone; Ep = marlstone and limestone; Ms = sandstone and marl; O = gypsum and limestone; Qa = alluvial deposits and detrital particles; Qc = hard limestone.

Soil covariates

A set of 14 environmental covariates, including lithology, indices based on satellite imagery, climatic maps, and primary and secondary terrain properties derived from a digital elevation model (DEM), was used as predictor variable set in the soil mapping (Table 2). The DEM was obtained from images from a Shuttle Radar Topography Mission (SRTM) with a spatial resolution of 30 m. Before calculating the topographic properties, the DEM was preprocessed by removing the sinks. The following properties were derived from the DEM: elevation above sea level, aspect, slope, topographic wetness index, flow direction, curvature, channel network base level, terrain surface texture, vertical distance to channel network, and relative slope position. In addition to the terrain properties, mean precipitation and lithology were incorporated as variables, representing the climate and parent material soil formation factors, respectively.

The vegetation covariate, a normalized difference vegetation index (NDVI) calculated by equation 1, was derived from Landsat 8 satellite images, sensor OLI and the clay mineral index (CMI), according to Boettinger et al. (2008) (Equation 2).

$$\text{NDVI} = \frac{NIR\ \text{-}\ R}{NIR + R} \qquad \text{Eq. 1}$$

$$\text{CMI} = \frac{(SWIR1)}{(SWIR2)} \qquad \text{Eq. 2}$$

In which NIR is the near infrared band (0.85-0.88 µm), R the red band (0.64-0.67 µm), and SWIR1 and SWIR2 are short infrared bands 6 (1.57-1.65 µm) and band 7 (2.11-2.29 µm), respectively, of Landsat-8 imagery.

The maps of covariates were generated using ArcGIS v.10.1, SAGA 2.1.0, and R software. The references used for generating the covariates and their relationship with soil formation factors are presented in table 2. Since the topographic properties are 1 arc-second (30 m) resolution data, the lithological map was rasterized (1:500,000) at 30 m. Likewise, the low-resolution precipitation map (≈ 30 arc-second or 1 km) was resampled at 30 m (Table 2).

Training and validation datasets

For the classification training and validation, the data were partitioned with “Data Splitting functions” using R and the “caret” package (Kuhn, 2017) at 80 % for model training and 20 % for independent validation. This random stratified splitting tries to preserve the overall distribution of the data within each class. The soil data distribution in classes is shown in figure 2.

Table 2. Covariates evaluated as soil predictors

Covariate	Soil-forming factor[(1)]	Spatial resolution/Scale	Reference
Elevation	R	30 m	Burrough (1986)
Aspect	R	30 m	Burrough and McDonnell (1998)
Slope	R	30 m	Burrough (1986)
Topographic wetness Index (TWI)	R	30 m	Beven and Kirkby (1979)
Flow direction	R	30 m	Jenson and Domingue (1988)
Curvature	R	30 m	Moore et al. (1991)
Channel network base level (CNW)	R	30 m	Conrad et al. (2015)
Terrain surface texture	R	30 m	Iwahashi and Pike (2007)
Vertical distance to channel network (VDCN)	R	30 m	Conrad et al. (2015)
Relative slope position (RSP)	R	30 m	Conrad et al. (2015)
Lithology	P	1:250,000	BME (2005)
Clay Mineral Index (CMI)	P	30 m	Boettinger et al. (2008)
NDVI	O	30 m	Rouse (1974)
Precipitation	Cl	1 km	Hijmans et al. (2005)

[(1)] Variables related to the Jenny equation (Jenny, 1941) - (Clorpt): R = relief; P = parent material; O = organisms; Cl = climate.

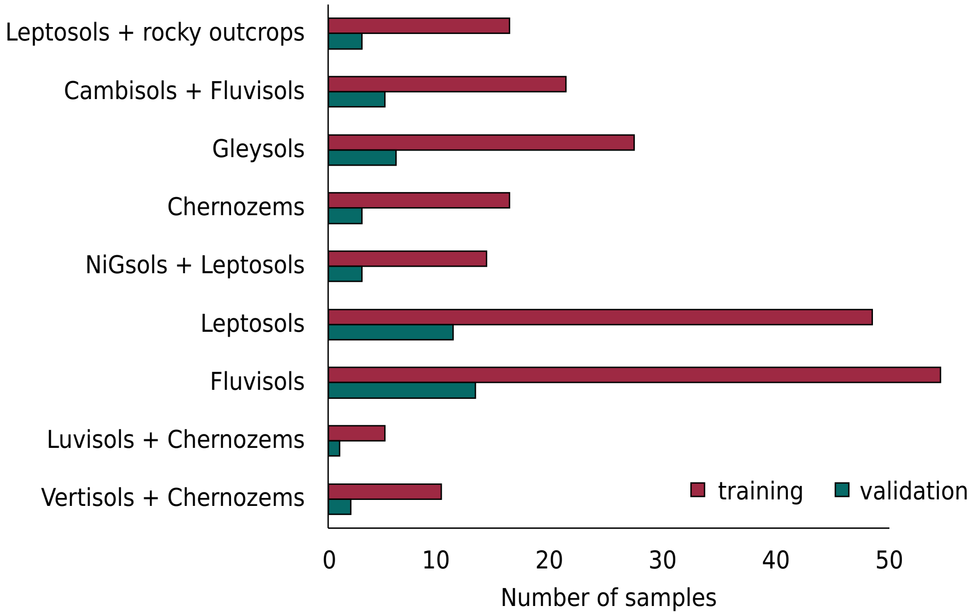

Figure 2. Distribution of soil classes in the training and validation dataset (n = 258).

Covariate selection

To identify an optimal subset of covariates from the set of all available covariates (Table 2), we used Recursive Feature Elimination (RFE), which is a backward selection algorithm that iteratively eliminates the least promising predictors from the model based on an initial predictor importance measure (Kuhn and Johnson, 2013). It is implemented within the caret package (Kuhn, 2017) and available for model training within a variety of models, such as the support vector machine and linear model. We chose to run RFE with Random Forest, as it can deal with both numeric and categorical variables and capture the non-linear relationship between predictor and response variables as well as the interactions between predictors (Heung et al., 2017). The methodological efficiency of variable selection using RFE for soil class mapping was demonstrated by Brungard et al. (2015) and for soil nutrient prediction by Jeong et al. (2017).

Classifiers for soil mapping

In the process of digital soil mapping, the target variable is represented by the soil classes described in the area and organized in soil map units (Table 1), and the explanatory variables are represented by environmental covariates correlated with the soils, according to the soil-environment relationship conceptualized by McBratney et al. (2003) in the Scorpan Model.

For classification by Random Forest (RF) and Multinomial Logistic Regression (MLR) classifiers, we used the packages Random Forest (Liaw and Wiener, 2015), nnet (Ripley and Venables, 2016), and caret (Kuhn, 2017) in R software, version 3.3.2 (R Development Core Team, 2016). The models were run in parallel using default tuning for selecting the best model, considering the “mtry” for RF and “decay” value for MLR. In the following, we give a brief description of the RF and MLR classifiers.

Random Forest (RF)

Random Forest was developed by Breiman (2001) to perform regression and classification. The model is based on the construction of a large set of random trees during the model training, leading to a single prediction. Each tree of the forest is constructed based on a bootstrap sample (sample with replacement) of the original training data with each bootstrap set, disregarding about one-third of the observations (Breiman, 2002; Hastie et al., 2009). The best split of each node of the tree is only searched among a randomly selected subset of the total number of predictors and the final prediction in the regression case is the average of the individual tree value, whereas for classification, the correct soil class is determined by a majority vote of the trees (as was the case in this study) (Liaw and Wiener, 2015).

Random Forest uses the out-of-bag (oob) samples, i.e., training observations not included in the bootstrap, to estimate the classification error. The oob samples are also used to construct a different variable importance measure (mean square errors, Gini index) to express the strength of each variable in the prediction (Hastie et al., 2009). Here, for categorical classification, the Gini index determines the importance of each predictive covariate in discriminating the soils: the higher the value of the index for a particular variable, the greater is its contribution to discriminating the class.

The model has two parameters to be tuned “mtry” (number of variables randomly sampled as candidates at each split) and “ntree” (number of trees to grow). To determine the best value of “mtry”, Breiman (2002) suggests starting with the default value (square root of the number of predictors for classification) and trying a default value twice as high and half as low to search for the optimal value. The number of “ntree”, on the other hand, does not require specific judgment for the setting of the parameter.

As a tree-based model, RF has advantages compared to linear models such as Multinomial Logistic Regression. It is able to model non-linear relationships between predictors and the response variable to handle noise data (observations with missing covariate data) and other situations in which a small dataset is associated with a large number of covariates (Collard et al., 2014). Although RF has shown better performance for soil class mapping when compared to a set of other classifiers (Pahlavan-Rad et al., 2016; Heung et al., 2017), the studies performed by Collard et al. (2014), Taghizadeh-Mehrjardi et al. (2015), and Camera et al. (2017) showed that MLR performed better than RF.

Multinomial Logistic Regression (MLR)

The MLR classifier is part of the family of generalized linear models and is used when the response variable has more than two categories (Hosmer and Lemeshow, 1989). This helps predict the probability of occurrence of each soil class in the landscape studied. The MLR applies a non-linear log transformation that allows to calculate the probability of occurrence of any number of classes of a dependent variable (in this case, soil class “y”) based on explanatory variables. Due to the sigmoidal behavior of its curve, the MLR differs from linear regression models.

The probability of occurrence of y_1 is π_1 and that of y_2 is $\pi_2 = 1 - \pi_1$. Logistic regression relates probability π_1 to a set of predictors using the logit link function (Equation 3):

$$\log \frac{\pi_j\,(x)}{\pi_J\,(x)} = \alpha_j + \beta'_j X,\ j = 1, J\text{-}1 \qquad \text{Eq. 3}$$

in which α is a constant, x a vector of predictive variables, and j is the model vector of coefficients. The model is analogous to a logistic regression model, except that the probability distribution of the response variable is multinomial instead of binomial and there are $J - 1$ equations instead of one, so that: $\pi_j\,(x) = P\,(Y = j \mid x)$, where x explanatory variables, with $x = 1$, and $\sum_j \pi_j(x) = 1$. Hence, the counts at J categories of Y can be treated as multinomial with probabilities of $\{\pi_1(x), \pi_2(x), \ldots..\ \pi_j(x)\}$.

Finally, the equation 4 expresses multinomial logistic models in terms of probabilities of answers (in this case, x soil class):

$$\pi_j\,(x) = \frac{\exp\,(\alpha_j + \beta'_j x)}{1 + \sum_{h-1}^{j-1} \exp\,(\alpha_h + \beta'_h x)} \qquad \text{Eq. 4}$$

Contrary to linear regression models that use the *least squares* as estimator, MLR coefficients are typically estimated using maximum likelihood. A more extensive description is found in the literature, notably of Hosmer and Lemeshow (1989).

Evaluation of MLR for soil mapping has been reported in a variety of situations regarding soil data acquisition (legacy polygon map, soil pits, and soil data derived from spectral reflectance spectroscopy), landscape, and level of soil classification (i.e. order, group, family and soil map unit component). Due to differences in data handling, the performance of MLR has been compared to a set of other classifiers for soil mapping purposes (Hengl et al., 2007; Collard et al., 2014; Brungard et al., 2015; Pahlavan-Rad et al., 2016; Heung et al., 2017) with some differences in accuracy depending on the sampling scheme, covariate scale and resolution, landscape pattern, and use and level of soil classification.

Soil classes were predicted at the order level by Vasques et al. (2015) using MLR with a coupled model combining field data with visible-near infrared (vis-NIR) spectral data. The use of soil data extracted from a legacy soil polygon map is more common since field surveys and laboratory analysis are the most expensive aspects of soil mapping. This approach has become widespread and was assessed by Giasson et al. (2008), Kempen et al. (2009), and ten Caten et al. (2011), who each used a conventional map at a 1:50,000 scale to extract training data for the soil classes. Similarly, data extracted from a 1:80,000 scale soil map were used by Figueiredo et al. (2008) and Collard et al. (2014) used soil data extracted from a polygon map to update a reconnaissance soil map (1:250,000) and compare MLR to other classifiers, observing a satisfactory performance of MLR for most soil classes.

An assessment of MLR, which compared the use of legacy data with ground data from soil pit data was performed by Heung et al. (2017), resulting in a better MLR performance when using a MLR-bagging than a single MLR model.

The accuracy of MLR for soil class mapping proved moderately accurate compared to other classifiers. For mapping soils in the Rio Doce Basin, Brazil, Souza et al. (2014) used MLR and found a kappa value of 0.35. A similar accuracy was found by Hengl et al. (2007), with a kappa of 0.36 for mapping WRB soil groups. Mapping soils in a northern province of Iran for taxonomic levels (great group, subgroup, and series), Pahlavan-Rad et al. (2016) reported a difference of around 10 % when comparing MLR to RF model in soil mapping on Northern of Iran. Nevertheless, soil mapping approaches used in those studies were different from the one proposed in this study.

Classifier evaluation

Classification is seen as a statistical and probabilistic process that tries to bring digital mapping as close to reality as possible. Accuracy is usually expressed in terms of indexes computed from a confusion matrix, which establishes the agreement between the classified map and the set of reference ground data (Foody, 2004). The confusion matrix compares, class by class, the relationship between the reference data from the field and the corresponding results obtained by classification (Congalton and Green, 1999).

Thus, to assess the model fit, we used a 5-fold cross validation, and compared the Kappa index of the classification with the kappa value of the external validation. This validation was performed by standard procedure, in which the model adjusted with the training dataset was applied to the independent dataset. The agreement between the predicted classes and the observations in the independent dataset was then measured by means of the confusion matrix technique, as described by Congalton and Green (1999).

The accuracy of the maps produced by the RF and MLR classifiers was evaluated by the confusion matrix indexes, expressed by the Kappa index and Overall Accuracy. The quality of the classification associated with the Kappa statistics, as proposed by Landis and Koch (1977), can be classified as: 0.00-0.20 = slight; 0.21-0.40 = fair; 0.41-0.60 = moderate; 0.61-0.80 = substantial; and 0.81-1.00 = almost perfect.

The soil maps predicted using RF and MLR classifiers were validated using an independent dataset, and the metrics of the classification (Kappa and variance) were compared by means of the Z test (Congalton and Green, 1999), at significance level of 0.05. The Z test is calculated as presented on equation 5:

$$Z = \frac{|\widehat{K}_1 - \widehat{K}_2|}{\sqrt{v\hat{a}r\ (\widehat{K}_1) + v\hat{a}r\ (\widehat{K}_2)}} \qquad \text{Eq. 5}$$

where $\widehat{K}_1$ and $\widehat{K}_2$ stand for the Kappa values of map 1 and map 2, respectively; and "*VÂR*" stands for the variance of the respective map.

RESULTS AND DISCUSSION

Covariates and selection of optimal dataset

In the exploratory data analysis of the covariates by hierarchical grouping (Figure 3), a high correlation was observed between elevation and channel network (cnw) (0.97). The NDVI and CMI (0.89) and relative slope position (rsp) and vertical distance to channel network (vdcn) (0.89) were also highly correlated.

To run the data-mining approach using RFE, we defined the minimal predictor number as 5 of (all) 14 covariates. The accuracy and oob error for the groups of covariates are shown in figure 4. The dataset with eight covariates resulted in the highest accuracy (0.53), equal to the accuracy of the datasets with 13 and 14 covariates. Although the kappa value was one unit lower when reducing the number of covariates, the reduction of the covariates in the model is justified by considering parsimony in the model.

The eight covariates included in the most accurate model consisted of: CNW, elevation, lithology, NDVI, precipitation, slope, terrain surface texture, and VDCN. In soil mapping, we proceeded with these selected covariates. Although CNW presents more than 97 % of correlation with elevation, both covariates were included in the set of best predictors by the data mining approach applied through the RFE algorithm. This selection showed diversity of the sources of selected covariates, including the categorical map (lithology), maps from satellite imagery (NDVI), and topographic maps.

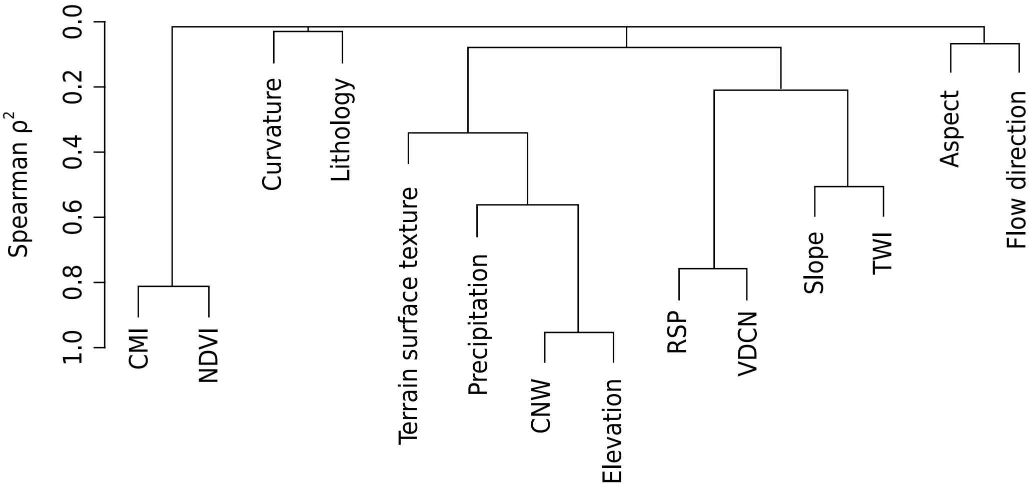

Figure 3. Dendogram of Spearman's correlation among covariates. CMI = clay mineral index; CNW = channel network base level, correlation among covariates; RSP = relative slope position; VDCN = vertical distance to channel network; TWI = topographic wetness index; NDVI = normalized difference vegetation index.

Classification by Random Forest

The RF model was trained using 5-fold cross validation repeated 5 times, using the predictors covariates selected with the RFE. It was verified that the best model parameter was the one with “mtry” = 2, oob (out-of-bag) error of 48 %, overall accuracy of 52 % (Table 3). This oob error criteria is lower than the one reported by Taghizadeh-Mehrjardi et al. (2015) who obtained 78 %. Using the Gini index, irrelevant covariates with weak correlation with the soils were eliminated by a data-mining approach using the RFE algorithm. The variables channel network base level (CNW), elevation, terrain surface texture, precipitation, NDVI, VDCN, slope, lithology, and curvature were the most important predictive covariates (Figure 5). These results appear to be consistent with those reported in the literature, where elevation is regarded as the most common variable used in digital soil mapping (McBratney et al., 2003). On the other hand, lithology and curvature contributed least to improve the model performance, as their relevance in terms of explaining the spatial distribution of the soils was ranked low (Figure 4). With regard to lithology, its poor capability to discriminate the soils in the study area may be related to the low spatial variability, since limestone occurs in more than 65 % of the area.

It is worth mentioning that the single soil components (LP, GL, FL) and soil associations NV and VX showed a fair contribution of the covariates CNW and elevation, indicating

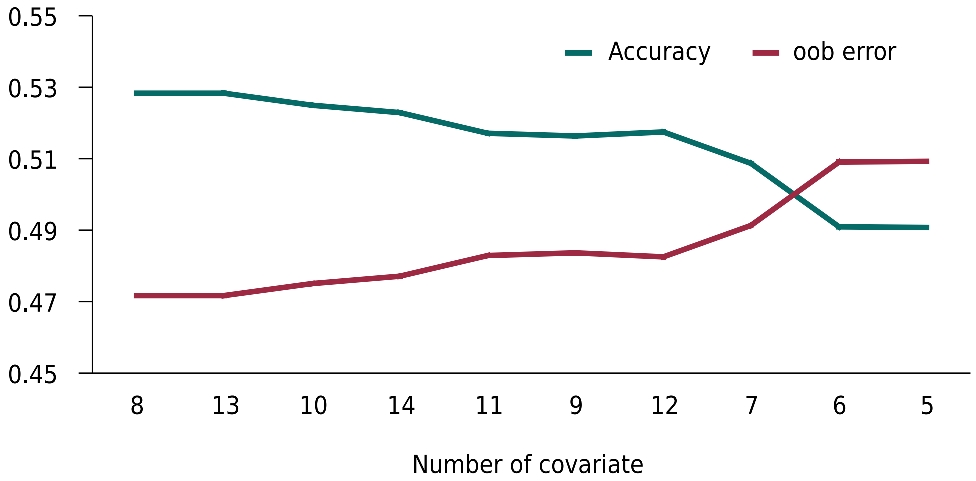

Figure 4. Overall accuracy and oob error per group of predictor covariates selected using RFE.

Table 3. Confusion matrix of soil classification using Random Forest

	LP	LP1	CM	GL	CH	NT	FL	LV	VR	Total	User’s Ac
											%
LP	25	9	5	0	6	3	3	1	0	52	48
LP1	2	2	0	0	0	2	0	0	1	7	29
CM	4	0	7	1	1	0	3	0	0	16	44
GL	0	0	0	12	0	0	8	0	0	20	60
CH	5	0	0	0	9	0	0	1	0	15	60
NT	6	4	0	0	0	9	0	0	0	19	47
FL	4	0	8	14	0	0	39	3	2	70	56
LV	1	0	0	0	0	0	1	0	0	2	0
VR	1	1	1	0	0	0	0	0	7	10	70
Total	48	16	21	27	16	14	54	5	10	211	
Prod. Ac. (%)	52	13	33	44	56	64	72	0	70		
					Kappa = 0.42 Overall accuracy = 0.52						

LP = Leptosols; LP1 = Leptosols + rocky outcrops; CM = Cambisols + Fluvisols; GL = Gleysols; CH = Chernozems; NT = Nitisols + Leptosols; FL = Fluvisols; LV = Luvisols + Chernozems; VR = Vertisols + Chernozems; User Ac. = user’s accuracy; Prod. Ac. = producer’s accuracy.

a well-defined relationship of these soils with the landscape features (Figure 6). Slope was inefficient in explaining the spatial distribution of the map unit NV, since this soil association depicts large slope ranges. On the other hand, the soils of this map unit were found to be adequately correlated with precipitation, due to occurring in the most humid part of the study area.

Map unit LP1 was stronger correlated with slope and terrain surface texture, while CM distribution was reasonably explained by slope, precipitation, terrain surface texture, and lithology. Map unit LV was best explained by CNW, precipitation, NDVI, and terrain surface texture, although these correlations were relatively weak.

The soil map produced by the RF classifier is shown in figure 7. Leptosols and Rocky outcrops (LP1) were found mostly in the uplands in the northern part of the study area, corresponding to the steepest and most eroded slopes. Correspondingly, Leptosols (LP) occurred extensively in the study area and were found in both the northern and southern

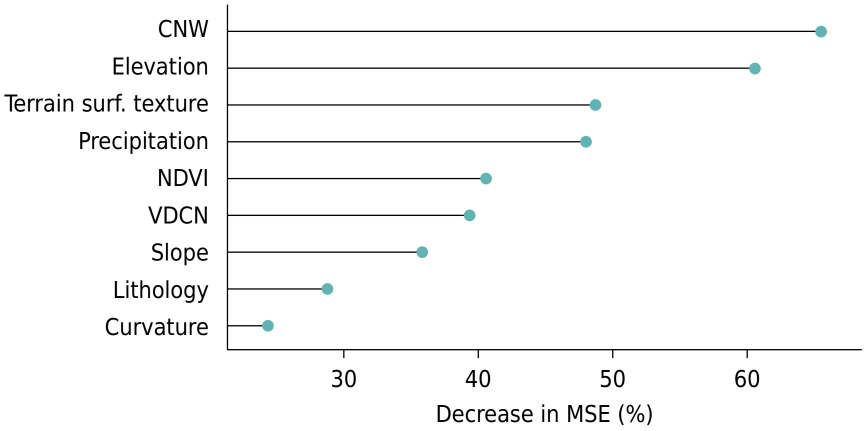

Figure 5. Importance of the covariates for soil mapping using Random Forest classifier.

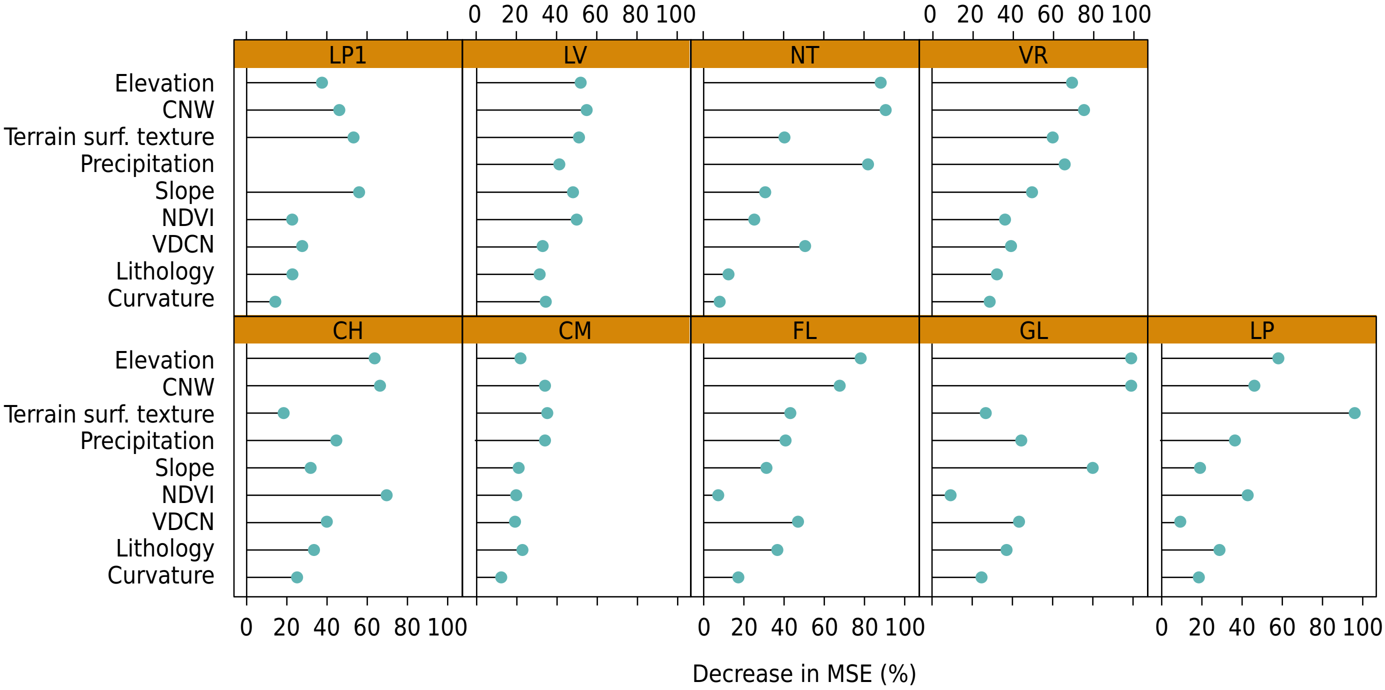

Figure 6. Importance of the covariates in mapping soils by the Random Forest classifier. LP1 = Leptosols + rocky outcrops; LV = Luvisols + Chernozems; NT = Nitisols + Leptosols; VR = Vertisols + Chernozems; CH = Chernozems; CM = Cambisols + Fluvisols; FL = Fluvisols; GL = Gleysols; LP = Leptosols.

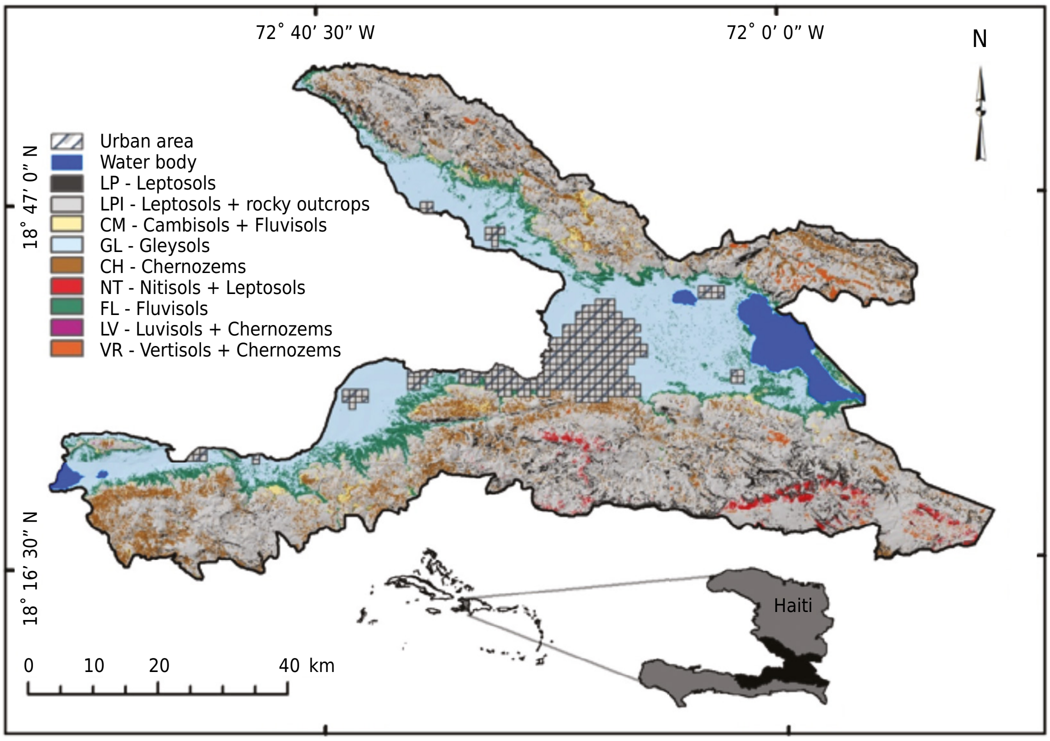

Figure 7. Soil map generated with classifier Random Forest - western Haiti, Caribbean.

part, on diverse parental materials, elevations, and slope ranges. These soils occurred mostly in dissected areas.

The map unit CM (Cambisols + Fluvisols) occupied primarily footslopes and lowlands. It is constantly affected by gully erosion and sediment depositions carried by rivers. These soils are intensively used as crop lands (banana/plantain, corn, common bean, and sugarcane).

Gleysols (GL) were commonly found in the central part, particularly in concave terrains where drainage is inhibited. Also, they occupied coastal areas near mangroves and other wetlands.

The occurrence of Chernozems (CH) was associated with hills, foothills, and usually calcareous material. These soils develop on gentle slopes and can be found from lower heights to about 1,000 m above the sea level, where rainfall is not abundant (≈ 762-1,194 mm yr^{-1}) in the study area. Chernozems have been used for diverse cultivation systems such as common bean-corn rotation, cassava, and sweet potatoes.

Nitosols + Leptosols (NT) are found on the uplands, particularly in the moistest part of the region, at higher elevations (≈ 1,524-2,683 m). This association corresponds to the most weathered soils in the area and are mainly used for vegetable cultivation. Moreover, a typical pine forest (*Pinus occidentalis)* occurs on those well-drained soils, on hill tops.

Fluvisols (FL) occupy the flatlands on recent alluvial deposits, in the central part of the area. They are poorly drained and frequently flooded. Nevertheless, they are intensively cultivated, as they are seen as very productive lands.

The association Luvisols + Chernozems (LV) are the least representative map unit, which occurs mostly in the southwestern part of the area and is located mainly on mid slopes. Similarly, the association Vertisols + Chernozems (VR) was found to be irrelevant in terms of geographical extent, occurring on lowlands as well as hills, and developed on low slopes. It occupies mostly lowlands of the southwestern part, at elevations between 274 and 396 m.

Classification by Multinomial Logistic Regression

Curvature showed no contribution to the soil mapping, while lithology contributed to 3 %. The variable importance of the other covariates (15 to 100 %), indicated that elevation and cnw (channel network base level) were the most important predictive variables. Precipitation presented a contribution of 49 % and some morphometric properties, e.g., terrain surface texture, vdcn (vertical distance to channel network), and slope, reached up to 40 % on the global scale (Figure 8).

The set of covariates showed different rankings of importance, compared to those presented in the RF model, particularly for the terrain surface texture and slope. In MLR, the slope was considered the fourth most important variable, whereas in RF, it was placed in seventh position. Terrain surface texture which ranked third in RF, ranked sixth in MRL. Also, a difference was observed for the variable curvature: it contributed to about a 10 % decrease in the MSE in RF while in MLR, it has no contribution.

The confusion matrix for the classification using MLR is displayed in table 4. It shows a Kappa of 0.45 and overall accuracy of 0.55. Out of nine soil classes, MRL failed to classify one (LV) and it classified CM with an accuracy of 19 % while the other soil classes were classified with an accuracy level, ranging from 39 to 79 %.

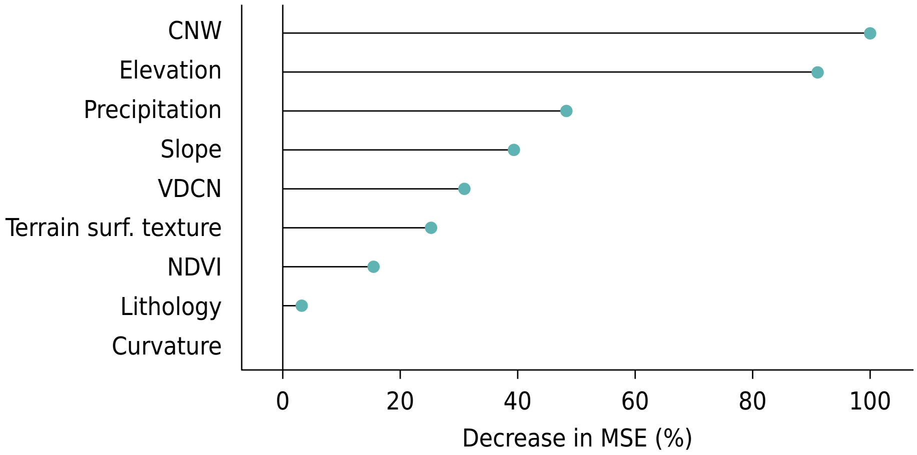

Figure 8. Importance of covariates in soil mapping by Multinomial Logistic Regression (MLR) classifier.

Table 4. Confusion matrix of soil classification using Multinomial Logistic Regression

	LP	LP1	CM	GL	CH	NT	FL	LV	VR	Total	User's Ac.
											%
LP	28	7	5	0	7	1	3	1	4	56	50
LP1	4	7	0	0	0	2	0	0	0	13	54
CM	2	0	4	0	2	0	2	0	0	10	40
GL	0	0	2	16	0	0	8	0	0	26	62
CH	3	0	1	0	6	0	1	1	0	12	50
NT	4	1	0	0	0	11	0	0	0	16	69
FL	5	0	8	11	0	0	40	3	2	69	58
LV	1	0	0	0	0	0	0	0	0	1	0
VR	1	1	1	0	1	0	0	0	4	8	50
Total	48	16	21	27	16	14	54	5	10	211	
Prod. Ac. (%)	58	44	19	59	38	79	74	0	40		
					Kappa = 0.45		Overall accuracy = 0.55				

LP = Leptosols; LP1 = Leptosols + rocky outcrops; CM = Cambisols + Fluvisols; GL = Gleysols; CH = Chernozems; NT = Nitisols + Leptosols; FL = Fluvisols; LV = Luvisols + Chernozems; VR = Vertisols + Chernozems; User Ac. = user's accuracy; Prod. Ac. = producer's accuracy.

Soil mapping validation

The independent validation of the soil mapping classification, assessed with 20 % of the dataset (47 soil data) by means of the confusion matrix, showed a Kappa index of 0.55 and overall accuracy of 0.64 for RF, and of 0.33 and 0.47, respectively, for MLR (Table 5).

Although the difference between the kappa values derived from RF and MLR was about 22 %, the Z test (1.83<|Z=1.96|) (Table 5) showed no significant difference between the two classifiers.

Classifiers comparison and soil class representativeness

The accuracy of classification with regard to both classifiers is presented on the confusion matrix and reflects the performance of MLR and RF in predicting the soils in the target area (Tables 3 and 4). In the model training, MLR performed slightly better, with an overall accuracy of 0.55 and Kappa of 0.45, while the overall accuracy and Kappa of RF were 0.52 and 0.42, respectively.

Similar to the findings of this study, Taghizadeh-Mehrjardi et al. (2015) mapped soil classified at the order level and also found a better performance by Logistic Regression (0.84) than by

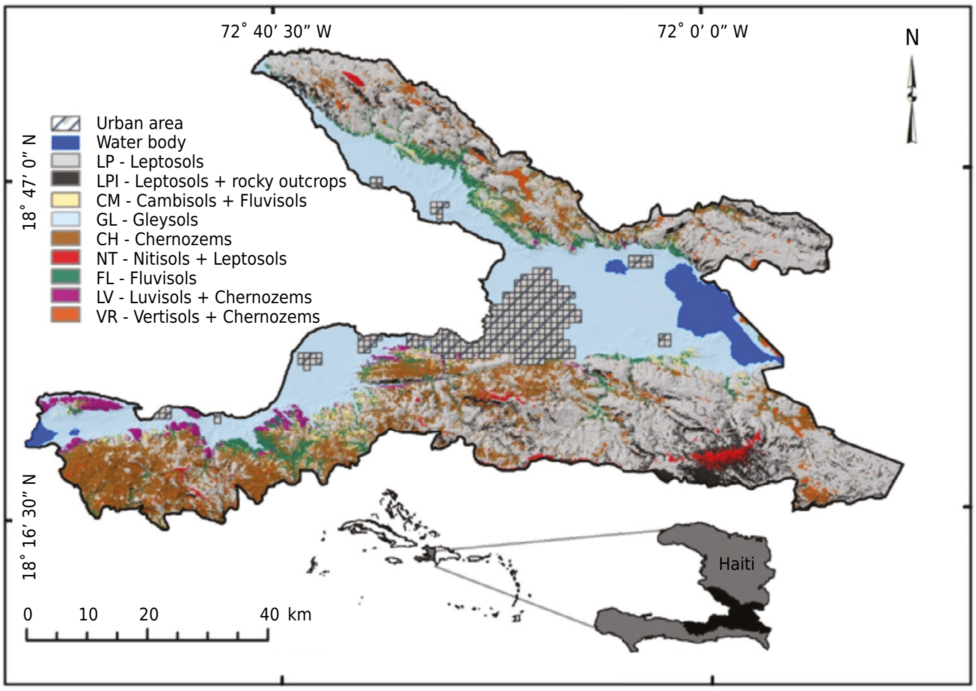

Figure 9. Soil map generated with the classifier Multinomial Logistic Regression - western Haiti, Caribbean.

Table 5. Comparison of classification performance by Random Forest and Multinomial Logistic Regression for external validation

Classifier	Overall accuracy	Kappa index	Kappa variance
Random Forest	0.64	0.55	0.0071416
Logistic regression	0.47	0.33	0.0073965
	Z test of the Kappa index		
	Random Forest		Logistic Regression
Random Forest	0.00		1.83[1]
Logistic Regression			0.00

[1] Z value calculated, not significant at 95 %.

RF (0.78). Consequently, for lower levels of soil classification (great group and family), the authors reported that RF outperformed Logistic Regression. On the other hand, contrary to what was found in this study, Camera et al. (2017) observed a better performance of RF than MLR for predicting soil groups in Cyprus, and Taghizadeh-Mehrjardi et al. (2015) showed an about 10 % higher accuracy of MLR than RF, when mapping soils in northern Iran.

For all soil map units in this study, the classification accuracy by MLR was equal to or higher than for those classified by RF. The accuracy by MLR varied from 19 to 79 % among soil classes, whereas for RF, the highest value was 72 % and the lowest 13 %, while one map unit (LV) was not classified, neither by RF nor MLR (Table 3 and 4).

The low representativeness of the map unit LV (Luvisols + Chernozems), with a total of five observations during the training step, was trivial for the learning pattern and therefore misclassified. This fact can be related to the process of random recursive binary partitioning of soil observations to build a forest of trees, in which the number of samples in the training and internal validation does not always include the less representative classes (Strobl et al., 2009). The partitioning algorithm tends to favor the most representative target category, which typically leads to drastic overfitting.

The quality of the classification assessed by external validation, according to Landis and Koch (1977), can be characterized as moderate for RF, and fair for the MLR classifier. The Kappa index of 0.33 for the MLR classification was lower than what was found by Hengl et al. (2007) and by Figueiredo et al. (2008), which verified a Kappa index of 0.36 and 0.39, respectively.

Table 6. Conditional Kappa of soil mapping by Random Forest (RF) and Multinomial Logistic Regression (MLR) for training and validation

	Training		Validation	
	RF	MLR	RF	MLR
LP - Leptosols	0.33	0.35	0.53	0.39
LP1 - Leptosols + rocky outcrops	0.23	0.50	1	1
CM - Cambisols + Fluvisols	0.38	0.33	0.16	0
GL - Gleysols	0.54	0.56	0.81	0.43
CH - Chernozems	0.57	0.46	0.47	0.47
NT - Nitisols + Leptosols	0.44	0.67	0.73	0.47
FL - Fluvisols	0.40	0.44	0.51	0.27
LV -Luvisols + Chernozems	0	0	0	0
VR - Vertisols + Chernozems	0.69	0.48	0.48	0

Table 7. Area of soil classes mapped by Random Forest and Multinomial Logistic Regression (MLR)

Map unit	Random Forest		MLR	
	Area	Area	Area	Area
	km^2	%	km^2	%
LP - Leptosols	2,087	48.5	1,654	38.5
LP1 - Leptosols + rocky outcrops	223	5.2	257	6
CM - Cambisols + Fluvisols	59	1.4	85	2
GL - Gleysols	841	19.6	987	23
CH - Chernozems	344	8	540	12.6
NT - Nitisols + Leptosols	42	1	46	1.1
FL - Fluvisols	283	6.6	198	4.6
LV -Luvisols + Chernozems	2	0	71	1.6
VR - Vertisols + Chernozems	31	0.7	75	1.7
Lake	131	3	131	3
Urban area	258	6	258	6
Total	3,911	100	3,911	100

The soil unit LP (48.5 %) and LP1 (5.2 %), corresponding, respectively, to Leptosols and Leptosols-Rocky outcrops in RF mapping, account together for 53.7 % of the whole area (Table 7). This large proportion of area covered by these map units confirms once again that the soils of the region are relatively young. Additionally, this pattern reflects the strong influence of morphogenetic processes on pedogenesis. The association of Vertisols-Chernozems is not very expressive, with a total of less than 2 % of the area by both classifiers, MLR and RF.

CONCLUSIONS

Of the covariates selected in this mapping process, channel network base level and elevation had the strongest power in separating the soils and therefore stressed the close relationship between the landscape and spatial distribution of soils in Western Haiti. On the other hand, the low importance of lithology in the mapping process may be due to its low variability, since limestone occurred in more than 65 % of the study area.

The close soil-landscape relationship was highlighted by the high contribution of topographic covariates to satisfactorily explain the soil distribution in the western region of Haiti.

The two evaluated classifiers proved capable of satisfactorily mapping the soils of the region, although RF outperformed MLR, reflecting its ability for learning and generalizing soil data of larger geographic areas.

Digital sol mapping techniques using MLR and RF can be used to produce soil maps for other parts of Haiti, as they were efficient in mapping the soils of the study area (overall accuracy of 0.47 and 0.64, respectively). Moreover, they can provide information for environmental planning and response programs of the country in which soil degradation has reached an advanced degree.

ACKNOWLEDGEMENT

The authors are indebted to the Brazilian National Council for Scientific and Technological Development (CNPq) for the scholarship of the first author to carry out this study in the framework of the PEC-PG program, to the Inter-American Institute for Cooperation on Agriculture (IICA-Haiti) for funding the data collection and, to the Program PNPD/Capes and the 'Appui à la Recherche et au Développement Agricole - AREA' (USAID) for their support. The authors also wish to thank the two anonymous reviewers for their helpful comments.

REFERENCES

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

Bacon SN, McDonald EV, Dalldorf GK, Baker SE, Sabol DE, Minor TB, Bassett SD, MacCabe SR, Bullard TF. Predictive soil maps based on geomorphic mapping, remote sensing, and soil databases in the desert Southwest. In: Boettinger JL, Howell DW, Moore AC, Hartemink AE, Kienast-Brown S, editors. Digital soil mapping: bridging research, environmental application, and operation. Berlin: Springer; 2010. p. 411-21.

Bayard B, Jolly CM, Shannon DA. The adoption and management of soil conservation practices in Haiti: the case of rock walls. Agricultural Economics Review. 2006;7:28-39.

Beven KJ, Kirkby MJ. A physically based, variable contributing area model of basin hydrology. Hydrol Sci B. 1979;24:43-69. https://doi.org/10.1080/02626667909491834

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

Breiman L. Manual on setting up, using, and understanding random forests. Berkeley: Statistics Department University of California Berkeley; 2002.

Breiman L. Random forests. Mach Learn. 2001;45:5-32. https://doi.org/10.1023/A:1010933404324

Brungard CW, Boettinger JL, Duniway MC, Wills SA, Edwards Jr TC. Machine learning for predicting soil classes in three semi-arid landscapes. Geoderma. 2015;239-240:68-83. https://doi.org/10.1016/j.geoderma.2014.09.019

Bureau des Mines et de l'Énergie - BME. Haiti geology (géologie), BME [08.2005] - polygon. Haiti: Port-au-Prince; 2005 [accessed on 2016 May 10]. Available at: http://www.haitidata.org/layers/geonode:hti_geology_geology_polygon_082005#map=611.49622628141/-8131084.26/2161122.51/0

Burrough PA, McDonnell RA. Principles of geographical information systems. 2nd ed. New York: Oxford University Press; 1998.

Burrough PA. Principles of geographic information systems for land resources assessment. Oxford: Oxford University Press; 1986.

Camera C, Zomeni Z, Noller JS, Zissimos AM, Christoforou IC, Bruggeman A. A high resolution map of soil types and physical properties for Cyprus: a digital soil mapping optimization. Geoderma. 2017;285:35-49. https://doi.org/10.1016/j.geoderma.2016.09.019

Chaves DA. Solos e aptidão agrícola das terras nas seções comunais do Mapou, Collines dês Chaines e Pichon - Haiti [dissertação]. Seropédica: Universidade Federal Rural do Rio de Janeiro; 2010.

Churches CE, Wampler PJ, Sun W, Smith AJ. Evaluation of forest cover estimates for Haiti using supervised classification of Landsat data. Int J Appl Earth Obs. 2014;30:203-16. https://doi.org/10.1016/j.jag.2014.01.020

Claessen MEC, organizador. Manual de métodos de análise de solo. 2. ed. Rio de Janeiro: Centro Nacional de Pesquisa de Solos; 1997.

Collard F, Kempen B, Heuvelink GBM, Saby NPA, Forges ACR, Lehmann S, Nehlig P, Arrouays D. Refining a reconnaissance soil map by calibrating regression models with data from the same map (Normandy, France). Geoderma Regional. 2014;1:21-30. https://doi.org/10.1016/j.geodrs.2014.07.001

Colmet-Daage F, Lagache P. Caractéristiques de quelques groupes de sols dérivés de roches volcaniques aux Antilles françaises. Cahiers ORSTOM. Série pédologie. 1965;8:91-121.

Congalton RG, Green K. Assessing the accuracy of remotely sensed data: principles and practices. New York: Lewis Publishers; 1999.

Conrad O, Bechtel B, Bock M, Dietrich H, Fischer E, Gerlitz L, Wehberg J, Wichmann V, Böhner J. System for Automated Geoscientific Analyses (SAGA) v.2.1.4. Geosci Model Dev. 2015;8:1991-2007. https://doi.org/10.5194/gmd-8-1991-2015

Figueiredo SR, Giasson E, Tornquist CG, Nascimento PC. Uso de regressões logísticas múltiplas para mapeamento digital de solos no Planalto Médio do RS. Rev Bras Cienc Solo. 2008;32:2779-85. https://doi.org/10.1590/S0100-06832008000700023

Foody GM. Thematic map comparison: evaluating the statistical significance of differences in classification accuracy. Photogramm Eng Rem S. 2004;70:627-33. https://doi.org/10.14358/PERS.70.5.627

Giasson E, Figueiredo SR, Tornquist CG, Clarke RT. Digital soil mapping using logistic regression on terrain parameters for several ecological regions in Southern Brazil. In: Hartemink AE, McBratney A, Mendonça-Santos ML, editors. Digital soil mapping with limited data. Dordrecht: Springer; 2008. p. 225-32. https://doi.org/10.1007/978-1-4020-8592-5_19

Grunwald S, Thompson JA, Boettinger JL. Digital soil mapping and modeling at continental scales: finding solutions for global issues. Soil Sci Soc Am J. 2011;75:1201-13. https://doi.org/10.2136/sssaj2011.0025

Grunwald S. Current state of digital soil mapping and what is next. In: Boettinger JL, Howell DW, Moore AC, Hartemink AE, Kienast-Brown S, editors. Digital soil mapping: bridging research, environmental application, and operation. Berlin: Springer; 2010. p. 3-12.

Guthrie RL, Shannon DA. Soil profile descriptions for steeplands research sites in Haiti. Alabama: Auburn University; 2004. (Technical Bulletin No. 2004-01).

Haspil A, Butterlin J. Les principaux types de sols de la République d'Haïti et leur répartition géographique; 1955. (Bulletin Agric Départ Agric Damien, 4).

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

Hengl T, Toomanian N, Reuter HI, Malakouti MJ. Methods to interpolate soil categorical variables from profile observations: lessons from Iran. Geoderma. 2007;140:417-27. https://doi.org/10.1016/j.geoderma.2007.04.022

Heung B, Ho HC, Zhang J, Knudby A, Bulmer CE, Schmidt MG. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. Geoderma. 2016;265:62-77. https://doi.org/10.1016/j.geoderma.2015.11.014

Heung B, Hodúl M, Schmidt MG. Comparing the use of training data derived from legacy soil pits and soil survey polygons for mapping soil classes. Geoderma. 2017;290:51-68. https://doi.org/10.1016/j.geoderma.2016.12.001

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A. Very high resolution interpolated climate surfaces for global land areas. Int J Climatol. 2005;25:1965-78. https://doi.org/10.1002/joc.1276

Hosmer DW, Lemeshow S. Applied logistic regression. New York: John Wiley & Sons; 1989.

Hylkema AL. Haiti soil fertility analysis and crop interpretations for principal crops in the five WINNER watershed zones of intervention [dissertation]. Florida: University of Florida; 2011.

Institut Haïtien de Statistique et D'informatique - IHSI. Population totale, de 18 ans e plus: Ménages et densités estimés en 2015. Port-au-Prince: Ministère de l'Économie et des Finances (MEF); 2015 [accessed on 2016 May 10]. Available at: http://www.ihsi.ht/

IUSS Working Group WRB. World reference base for soil resources 2014, update 2015: International soil classification system for naming soils and creating legends for soil maps. Rome: Food and Agriculture Organization of the United Nations; 2015. (World Soil Resources Reports, 106).

Iwahashi J, Pike RJ. Automated classifications of topography from DEMs by an unsupervised nested-means algorithm and a three-part geometric signature. Geomorphology. 2007;86:409-40. https://doi.org/10.1016/j.geomorph.2006.09.012

Jenny H. Factors of soil formation: a system of quantitative pedology. New York: McGraw-Hill; 1941.

Jenson SK, Domingue JO. Extracting topographic structure from digital elevation data for geographic information system analysis. Photogramm Eng Remote Sensing. 1988;54:1593-600.

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

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

Koohafkan AP, Lilin CH. Arbres et arbustes d'Haïti: utilisation des espèces ligneuses en conservation des sols et en aménagement des bassins versants. Port-au-Prince: Ministère de l'Agriculture des Ressources Naturelles et du Développment Rural; 1989.

Kuhn M. Classification and regression training. R package version 6.0-76; 2017 [accessed on 2017 May 3]. Available at: https://CRAN.R-project.org/package=caret.

Kuhn M, Johnson K. Applied predictive modeling. New York: Springer; 2013.

Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159-74. https://doi.org/10.2307/2529310

Liaw A, Wiener M. Classification and regression with random forest. R package version 4.6-12; 2015 [accessed on 2017 Fev 24]. Available at: https://CRAN.R-project.org/package=randomForest.

Libohova Z, Wysocki D, Schoeneberger P, Reinsch T, Kome C, Rolfes T, Jones N, Monteith S, Matos M. Soils and climate of Cul de Sac Valley, Haiti: a soil water and geomorphology perspective. J Soil Water Conserv. 2017;72:91-101. https://doi.org/10.2489/jswc.72.2.91

McBratney AB, Odeh IOA, Bishop TFA, Dunbar MS, Shatar TM. An overview of pedometric techniques for use in soil survey. Geoderma. 2000;97:293-327. https://doi.org/10.1016/S0016-7061(00)00043-4

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

Minasny B, McBratney AB. Digital soil mapping: a brief history and some lessons. Geoderma. 2016;264:301-11. https://doi.org/10.1016/j.geoderma.2015.07.017

Moonjun R, Farshad A, Shrestha DP, Vaiphasa C. Artificial neural network and decision trees in predictive soil mapping of Hoi Num Rin sub-watershed, Thailand. In: Boettinger JL, Howell DW, Moore AC, Hartemink AE, Kienast-Brown S, editors. Digital soil mapping: bridging research, environmental application, and operation. Berlin: Springer; 2010. p. 151-64.

Moore ID, Grayson RB, Ladson AR. Digital terrain modelling: a review of hydrological, geomorphological, and biological applications. Hydrol Process. 1991;5:3-30. https://doi.org/10.1002/hyp.3360050103

Muñoz-Rojas M, Pereira P, Brevik E, Cerda A, Jordan A. Soil mapping and processes models to support climate change mitigation and adaptation strategies: a review. In: 19th EGU General Assembly, EGU2017-10677, proceedings from the conference held 23-28 April, 2017. Vienna, Austria: Geophysical Research Abstracts; 2017.

Pahlavan-Rad MR, Khormali F, Toomanian N, Brungard CW, Kiani F, Komaki CB, Bogaert P. Legacy soil maps as a covariate in digital soil mapping: a case study from Northern Iran. Geoderma. 2016;279:141-8. https://doi.org/10.1016/j.geoderma.2016.05.014

Potter RB, Barker D, Conway D, Klak T. The contemporary Caribbean. New York: Routledge, 2004.

R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria; 2016 [accessed on 2017 Jan 20]. Available at: http://www.R-project.org/.

Ripley B, Venables W. Feed-forward neural networks and multinomial log-linear models. R package version 7.3-12; 2016 [accessed on 2017 Fev 24]. Available at: https://CRAN.R-project.org/package=nnet.

Robar G. Végétation de la République d'Haïti [thèse]. France: L'Université Scientifique et Médicale de Grenoble; 1984.

Rouse JW, Haas RH, Schell JA, Deering DW, Harlan JC. Monitoring the vernal advancement of retrogradation (greenwave effect) of natural vegetation. Greenbelt: Texas A & M University; 1974.

Santos HG, Jacomine PKT, Anjos LHC, Oliveira VA, Oliveira JB, Coelho MR, Lumbreras JF, Cunha TJF. Sistema brasileiro de classificação de solos. 2. ed. Rio de Janeiro: Embrapa Solos; 2006.

Scull P, Franklin J, Chadwick OA, McArthur D. Predictive soil mapping: a review. Prog Phys Geog. 2003;27:171-97. https://doi.org/10.1191/0309133303pp366ra

Souza E, Hengl T, Kempen B, Heuvelink GBM, Fernandes Filho EI, Schaefer CEGR. Comparing spatial prediction methods for soil property mapping in Brazil: a case study for the Rio Doce Basin. In: Arrouays D, McKenzie N, Hempel J, Forges ACR, McBratney A, editors. GlobalSoilMap: basis of the global spatial soil information system. London: CRC Press; 2014. p. 267-71.

Strobl C, Malley J, Tutz G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychol Methods. 2009;14:323-48. https://doi.org/10.1037/a0016973

Sweet AT. The soils of Haiti. Soil Sci Soc Am J. 1926;7:75-86. http://dx.doi.org/10.2136/sssaj1926.036159950OB700010012x

Taghizadeh-Mehrjardi R, Nabiollahi K, Minasny B, Triantafilis J. Comparing data mining classifiers to predict spatial distribution of USDA-family soil groups in Baneh region, Iran. Geoderma. 2015;253-254:67-77. https://doi.org/10.1016/j.geoderma.2015.04.008

ten Caten A, Dalmolin RSD, Pedron FA, Mendonça-Santos ML. Regressões logísticas múltiplas: fatores que influenciam sua aplicação na predição de classes de solos. Rev Bras Cienc Solo. 2011;35:53-62. https://doi.org/10.1590/S0100-06832011000100005

Vasques GM, Demattê JAM, Rossel RAV, López LR, Terra FS, Rizzo R, Souza Filho CR. Integrating geospatial and multi-depth laboratory spectral data for mapping soil classes in a geologically complex area in southeastern Brazil. Eur J Soil Sci. 2015;66:767-79. https://doi.org/10.1111/ejss.12255

Woodring WP, Brown JS, Burbank WS. Geology of the Republic of Haiti. Port-Au-Prince: Republic of Haiti, Department of Public Works; 1924.