Revista Brasileira de Ciência do Solo

**Division – Soil In Space and Time** | Commission – Pedometrics

# Sample design effects on soil unit prediction with machine: randomness, uncertainty, and majority map

**Waldir de Carvalho Junior[(1)]*, Nilson Rendeiro Pereira[(1)], Elpidio Inacio Fernandes Filho[(2)], Braz Calderano Filho[(1)], Helena Saraiva Koenow Pinheiro[(3)], Cesar da Silva Chagas[(1)], Silvio Barge Bhering[(1)], Vinicius Rendeiro Pereira[(3)] and Sara Lawall[(3)]**

[(1)] Empresa Brasileira de Pesquisa Agropecuária, Embrapa Solos, Rio de Janeiro, Rio de Janeiro, Brazil.
[(2)] Universidade Federal de Viçosa, Departamento de Solos, Viçosa, Minas Gerais, Brazil.
[(3)] Universidade Federal Rural do Rio de Janeiro, Departamento de Engenharia Agrícola, Seropédica, Rio de Janeiro, Brazil.

**ABSTRACT:** Notwithstanding the importance of soil surveys, advances in digital soil mapping have mainly focused on mapping soil attributes or properties rather than developing digital maps of soil units or soil classes. The purpose of this research was to develop digital soil unit maps based on primary soil data collection in areas without previously collected soil information. The covariate variability, the random effect across the data subset and the map outputs were the focuses of this study. We used five datasets with four models (Random Forest - RF, Gradient Boosted Machine - GBM, C5.0, and multinomial log-linear model - MLR). The covariates were grouped into five datasets, where four were grouped by Region Of Interest per Class (ROIC) and one was not grouped by ROIC. To evaluate the random effect to split the dataset, we ran each model 50 times and observed the overall accuracy (OA) and kappa index, and uncertainty, majority and variety maps. The OA of Dataset01 to 04 was lower than to Dataset05 accuracy. However, map outputs of RF and GBM for Dataset01 and Dataset05 had the same majority prediction. It seems that RF and GBM produce consistent results in map outputs according to this methodology and pedologist expertise. To evaluate the uncertainty and the consistency of soil unit prediction, we used the majority maps process. Random Forest, similar to GBM, presented the best results. The increase in the number of covariates was not a guarantee of improvement in the OA or in the quality of the map output. Geographic position and distance raster did not improve the map output according to expert evaluation. Because the variance between the ROICs, when the training and validation datasets were split based on it, the subsets are quite different in relation to the covariates, and this is the reason for the worse results of this model, comparing with the Dataset05. On the other hand, when considering one complete dataset not based on ROICs, the variance of training and validation subsets is lower and produced more accurate parameters of quality.

**Keywords:** tree learners models, hillslope areas, random forest.

*** Corresponding author:**
E-mail: waldir.carvalho@embrapa.br

**Received:** September 25, 2019
**Approved:** March 16, 2020

**How to cite:** Carvalho Junior W, Pereira NR, Fernandes Filho EI, Calderano Filho B, Pinheiro HSK, Chagas CS, Bhering SB, Pereira VR, Lawall S. Sample design effects on soil unit prediction with machine: randomness, uncertainty, and majority map. Rev Bras Cienc Solo. 2020;44:e0190120.
https://doi.org/10.36783/18069657rbcs20190120

## INTRODUCTION

Soil surveys are an important tool to understand the environment and make better decisions on soil management. The methods used to produce soil class maps differ between conventional and digital approaches. Digital soil mapping (DSM) has become popular for producing maps of soil classes and properties based on spatial data from soil inventories and auxiliary landscape spatial data (McBratney et al., 2003), bridging gaps between discrete soil maps and the continuous nature of soil cover (Burrough et al., 1997). Digital soil mapping techniques, in which both soil classification and mapping are handled numerically, can represent a formalized alternative to conventional soil mapping.

To develop knowledge regarding the detailed spatial distribution of soils, the employment of DSM techniques to add value to traditional soil maps is increasing (McBratney et al., 2003). This development is based on advances in geographic information systems, computer data processing, and available global landscape data. Digital soil mapping techniques are analogous to conventional methods modeling the relationship between soil properties or classes and environmental variables or covariables (auxiliary landscape data) by spatial statistics or geostatistics approaches (Camera et al., 2017). However, pedological tacit knowledge remains a key factor in building models that generate both statistically and pedologically sound outputs (Kempen et al., 2009) and is included in almost all steps of digital soil mapping.

Terrain attributes influence soil genesis and, consequently, soil type distribution (Gruber et al., 2017). Furthermore, the set of environmental covariates can be enriched by using remotely sensed data from spectral sensors in the DSM approach. The correlation between soil information and environmental variables provides the basis for constructing the DSM dataset for soil property or soil class studies.

Accordingly, a search in the journals Geoderma, Regional Geoderma, Catena, and Pedosphere for the keywords 'digital soil mapping' in the last five years shows 229 responses; 78 % of the articles are about mapping soil properties or attributes, while only 22 % are concerned with soil class or soil unit mapping. The majority of articles concerned with soil class mapping used secondary data from preexisting soil surveys or disaggregation of map unit polygons. None of the articles discuss creating a soil unit map from primary data in areas where there is no high-resolution soil information, despite the importance of such maps to environmental knowledge and conservation. The lack of this type of study supports the idea that more effort must be made to perform digital soil class mapping to fill this gap.

Data-driven models require the use of machine learning, which is a datamining process, to optimize pattern recognition from large datasets using training models. The process of 'training' a model is described as a type of 'learning', where 'machine learning' can be defined as the process of identifying the relationships between predictor and response variables (the training dataset) by using computer-based statistical approaches (Witten and Frank, 2005; Hastie et al., 2009). Digital soil mapping techniques use 'machine learning' processes to identify the relationships between soil information and environmental variables.

Camera et al. (2017) studied soil class distribution in Cyprus based on a conventional preexisting soil survey database, investigating the optimal number of training points by using Random Forest (RF) and Multinomial Logistic Regression (MLR) models, and they concluded that RF had better performance in topographically and pedologically complex regions when compared with regression models. Heung et al. (2016) used a conventional soil survey to collect training and validation samples to evaluate a variety of datamining models and concluded that different algorithms resulted in drastically different outputs; however, RF usage in DSM appears to be more promising than other techniques.

Teske et al. (2015), by using data from preexisting soil surveys, inferred that the sample design and the method of accuracy evaluation can affect the model predictor selection. Heung et al. (2017) used a training dataset derived from soil pits and soil survey polygons to compare two types of machine learners, concluding that the RF model performed the best overall accuracy. Pásztor et al. (2018) used legacy soil information to compile a new soil map in Hungary by sequential classification methods (segmentation, classification trees, RF, and artificial neural networks) and obtained an overall accuracy of 70 %.

Adhikari et al. (2014) constructed a soil map in Denmark by applying a decision tree-based model, and more than 1170 soil profiles and 17 environmental covariables were used as input covariates. A total of 20 % of the dataset was used for validation, and the overall accuracy ranged from 60 to 76 % when considering the prediction accuracy of similar groups.

The main goals of this study were to evaluate the random effect of data splitting in developing digital soil unit maps of hillslope areas based on primary soil data collection and to propose a default procedure to reduce the random effect of data splitting. Specific objectives of this study were (i) to analyze the covariate variability between the Region Of Interest per Class (ROIC) around the ground control points (GCPs) used as the input dataset; (ii) to investigate the effect of randomly selecting ROIC samples and single samples used as training and validation soil data; and (iii) to evaluate five different input dataset combinations based on terrain attributes, geographic position, and distance rasters (spatial dependency), as well as four different models of machine learning (RF, Gradient Boosted Machine - GBM, C5.0, and MLR) applied in the digital mapping of soil units.

## MATERIALS AND METHODS

### Study area

Bonfim Watershed is located in Rio de Janeiro State, between 22.4–22.5° S and 43.4–43.8° W (Figure 1). The total area of the watershed is 3,030 hectares. The watershed is in the west side of a wide mountain chain regionally known as Serra do Mar. The altitude ranges from 675 to 2,260 m, and the mean elevation is 1,324 m. The average slope is 62 %, with a 33 % standard deviation, and the lithology comprises mainly granite and gneiss rocks (Silva and Cunha, 2001; Leite et al., 2004). The production of vegetables to supply nearby consumer centers is the main agricultural activity. The topographic conditions of the watershed make it difficult to access all areas to sample soils.

Soil variability in the region is exceptionally high due to the influence of geomorphology (valleys, steep slopes, and high elevation) and parental material (granite and gneiss). Rock outcrops are common and were identified from a previously available land use land cover (LULC) map. The soil taxonomic classification was based on the SiBCS - Brazilian Soil Classification System (Santos et al., 2018) and applied to the ground control points (GCPs). In the study area, we observed *Cambissolos* (Inceptisols - moderately developed soils), *Latossolos Amarelos* and *Vermelho-Amarelos* (*Oxisols* - soils with high content of kaolinite and oxides), *Neossolos Litólicos* (Lithicsols - soils that are thin or with many coarse fragments) and rocks outcrops (RckO). The soil units are composed of these soil classes.

### Soil dataset

We used primary soil information from soil profiles of 75 ground control points (GCPs) defined by conditioned Latin hypercube sampling (Minasny and McBratney, 2007) taking account the elevation, total insolation, slope, and general curvature. These 75 ground control points (GCPs) were observed and/or sampled. This dataset allowed the definition

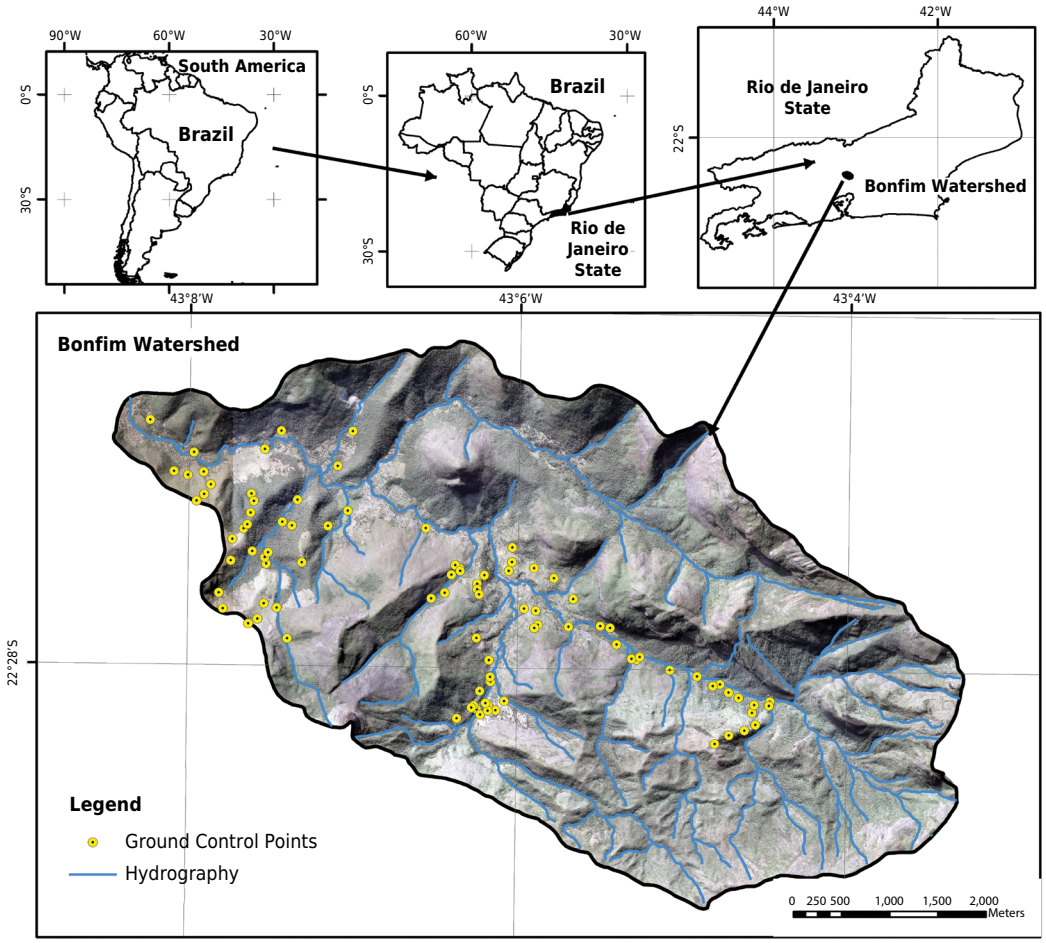



**Figure 1.** The location of Bonfim watershed and Ground Control Points over an orthophoto and its location in South America, Brazil, and Rio de Janeiro State.

of eight map units based on the second level of SiBCS (Santos et al., 2018) and the complexity of each unit. The soil units defined are CXbd (*Cambissolos Háplicos Tb Distróficos* - Inceptisols), LAd (*Latossolo Amarelo Distrófico* - Oxisols), LVAd [*Latossolo Vermelho-Amarelo Distrófico*, which corresponds to a Rhodic Ferralsol/Oxisols)], RLd (*Neossolo Litólico Distrófico* - Lithicsols) and the units that represent associations between RLd and RckO (RLd+RcKO) and RLd, CXbd and RcKO (RLd+CXbd+RckO). The units RckO and Urban Areas (Urb) were identified from previously available LULC maps and were superimposed on the maps resulting from the applied computational methods.

Simultaneously, in the fieldwork, using the pedologist expertise, were defined 75 irregular polygons (areas) around the GCPs that represent the observed soil unit. These polygons, called ROICs, help to improve and enlarge the soil dataset based on 75 profiles. The ROICs are regions chosen to represent a class. These 75 ROICs allowed us to obtain a total of surrounding 1,844 individual samples to train and validate the models based on a 10 m spatial resolution grid. Table 1 shows the distribution of the ROICs and individual samples per soil unit. To avoid an imbalance of samples, approximately 300 samples were established per soil unit.

The ROIC definitions are pedologically dependent and defined through fieldwork, and it is very important step to begin the data-driven process. It was done by the team of pedologists who worked in the area. The implementation of the ROICs was through geoprocessing in SagaGIS (Conrad et al., 2015) with the help of the orthophoto and covariates (elevation, slope, and curvature) observed around the GCPs and defining the region that represents each GCP (Figure 2) to compose the dataset.

**Table 1.** Distribution of ROICs and samples by soil unit

| Soil Unit | Number of ROICs[1] | Number of single samples |
|---|---|---|
| CXbd (Cambisols) | 15 | 317 |
| LAd (Xanthic Ferralsol) | 10 | 307 |
| LVAd (Rhodic Ferralsol) | 23 | 300 |
| RLd (Dystric Leptosols) | 09 | 312 |
| RLd+RckO[2] | 10 | 303 |
| RLd+CXbd+RckO | 08 | 305 |

[1] ROICs: regions of interest per class. [2] RckO: rock outcrops

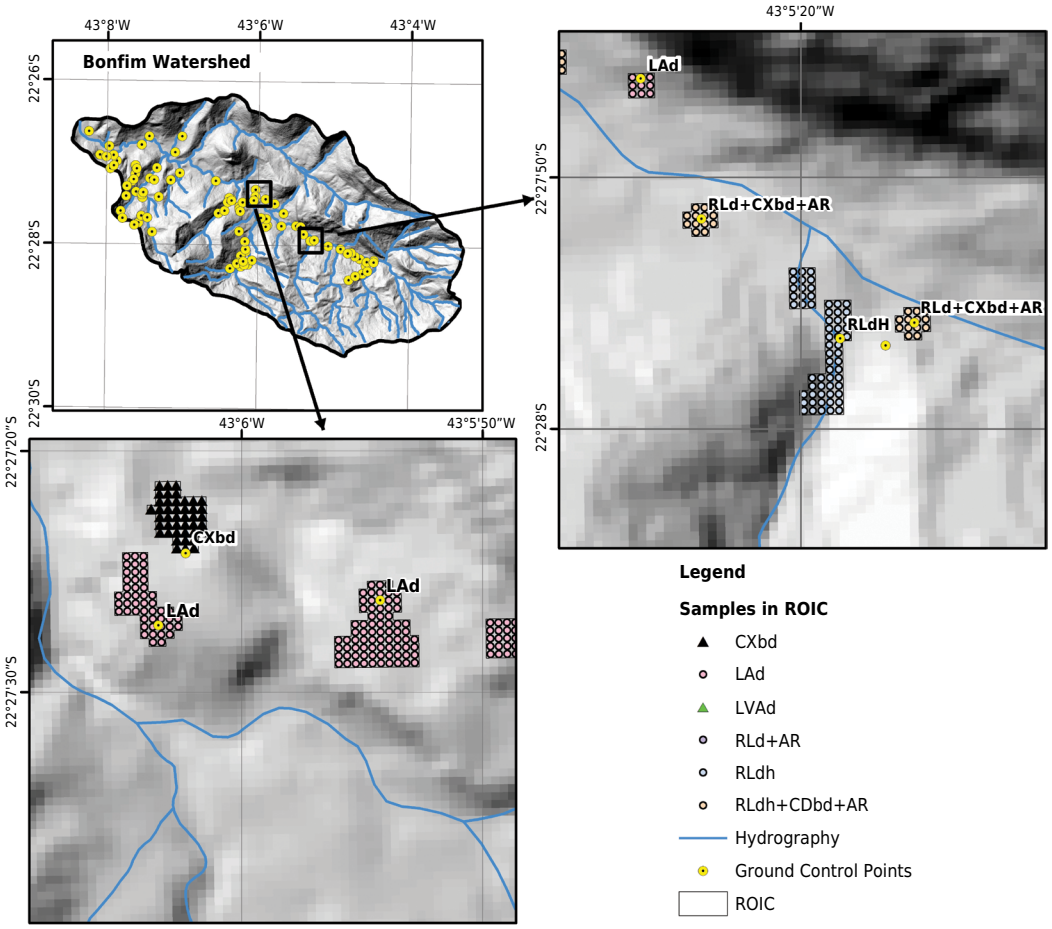



**Figure 2.** Spatial distribution of point samples in ROICs around the GCPs over the DEM hillshade, in which CXbd: Inceptisols; LAd and LVAd: Oxisols; RLd and RLdh: Lithicsols.

## Environmental covariates

Thirty-one environmental variables (Table 2) were derived from an acquired 10 m spatial resolution digital elevation model (DEM) to create the dataset using SagaGIS software (Conrad et al., 2015). The covariates derived from the DEM are those most commonly used by DSM users (Calderano Filho et al., 2014; Carvalho Junior et al., 2014; Bhering et al., 2016; Chagas et al., 2016; Camera et al., 2017; Gruber et al., 2017; Heung et al., 2017).

Covariates also included were point coordinates X and Y (in meters), as well as 16 rasters of distance calculated from 16 different points, spread regularly over the study area. These 16 rasters were named 'dist_xn', where 'n' varies from 1 to 16 (Table 2) and, together with the X and Y positions, represent a test to verify whether there were spatial dependencies in the soil dataset.

**Table 2.** Covariates derived from a 10 m DEM, from the distances calculation and the geographic position and its representation

| Covariate | Representation |
|---|---|
| Elevation; aspect; slope; plan curvature; profile curvature; curvature classification; general curvature; maximal curvature; minimal curvature; standardized height; tangential curvature; total curvature; cross sectional curvature; longitudinal curvature | Local scale morphometry |
| Multi-resolution ridge top flatness index; multi-resolution valley bottom flatness index; mid slope position; normalized height; slope height; valley depth; euclidian distance to rivers; topographic position index | Landscape scale morphometry |
| slope length factor; flow accumulation; flow direction; flow line curvature; topographic wetness index; terrain ruggedness index | Hydrologic characteristics |
| Diffuse insolation; total insolation; direct insolation | Landscape exposure |
| dist_x1 to dist_x16 | Spatial dependence |
| UTM coordinates X and Y | Geographic position |

Thus, the covariates were organized into five different datasets combining the DEM derived, the position and distance covariates, with the 1,844 samples composed of i) Dataset01 – the 27 DEM-derived attributes grouped by ROICs; ii) Dataset02 – Dataset01 plus the Y coordinate with 28 covariates grouped by ROICs; iii) Dataset03 – Dataset01 plus 11 distance rasters with 38 covariates grouped by ROICs; iv) Dataset04 – all the covariates with 39 covariates grouped by ROICs; and v) Dataset05 – Dataset01 with a single dataset for each soil unit (the default in the majority of studies), aggregating all the ROICs with 1,844 samples, but without grouping by ROIC with the 27 DEM-derived attributes (Table 1 in the single samples).

## Computational operation and methods

The five datasets were run through models in R (www.r-project.org) to apply the computational procedures of machine learning. The packages and models used were randomForest (Liaw and Wiener, 2018), generalized boosted model – 'gbm' (Ridgeway, 2017), C5.0 (Kuhn, 2017), and nnet (Ripley and Venables, 2016) to a multinomial log-linear model called 'multinom', and these are referred to as RF, GBM, C50, and MLR, respectively. The first three models are based on decision trees, and all models were tested for all datasets.

The RF is a classifier consisting of a collection of tree structured classifiers in which the random vector is independent identically distributed and each tree casts a unit vote for the most popular class at input x. In addition, it is very user-friendly in the sense that it has only two parameters (the number of variables in the random subset at each node and the number of trees in the forest), and is usually not very sensitive to their values (Liaw and Wiener, 2002).

Friedman (2001) developed a new approach called gradient boosting (GBM). The performance of such method is improved, and the overfitting is reduced by the introduction of randomness and by stochastic gradient boosting, and each decision tree is constructed by taking a random subsample of the training dataset (Friedman, 2002). The aim of the GBM is to improve the model performance by combining a large number of simple trees, i.e., the final outcome is a collection of weak learners. The model fit over different trees is improved by considering the previous learners and by emphasizing those observations incorrectly classified.

C5.0 is a decision tree algorithm, and it is the improved version of the C4.5 algorithm. The boosting feature of C5.0 helps improve the accuracy of the model (Emre et al., 2019). The C50 is a large decision tree and gives the acknowledge of noise and missing data. The C50 algorithm solves the problem of over fitting and error pruning. In the classification technique, the C50 classifier can anticipate which attributes are relevant and which are not relevant in classification (Pandya and Pandya, 2015).

The MLR is a multinomial log-linear model that was an improved version of logistic regression that incorporates an artificial neural network approach for parameter optimization (Sim et al., 2018).

Covariates with high correlation (Pearson correlation > +0.95 or < -0.95) were excluded from the dataset to decrease multicollinearity between them. Between two high correlated covariates, we used to exclude the one with bigger sum of Pearson correlations. These constraints reduced the number of covariates and facilitated the routine in R.

The operational methods included splitting the samples between training (60 %) and validation (40 %) subsets grouping by ROIC identifiers, where all the samples from one ROIC could be used for validation or training purposes with Dataset01, Dataset02, Dataset03, and Dataset04. This allowed us to analyze the variance between and within the ROICs to verify whether the variance was greater. Considering that the ROICs do not have the same area and consequently different numbers of cells, the validation and training datasets varied the number of total samples when a random routine session was performed. On the other hand, in Dataset05, the default method without grouping by ROICs (a single dataset) the amount of samples for training and validation (60 and 40 %, respectively) was constant, respectively 1,106 and 738 samples. To evaluate the random effect of splitting the data into validation and training, and calculate the uncertainty, the models were performed 50 times with 50 random subsets for training and validation, and the statistical results and output maps were recorded.

The evaluation of the models was based on the mean overall accuracy (OA) and mean kappa index (kappa) (Rossiter, 2008; Rossiter et al., 2017) of the 50 repetitions. The OA was obtained from the confusion matrix and represents the total success classification of the model when applied to the validation subset. Kappa is an association measure used to describe the concordance level of the map unit prediction (Wolski et al., 2017) over the validation subset.

Each random repetition of the models generated a map. The 50 maps for each model and dataset were layer stacked, and cell statistics were calculated using ArcGIS (ESRI). To examine the map results from each model, the 'Composite Bands' tool was used to perform the layer stack of the maps, followed by a local analysis tool 'Cell Statistics' to obtain the variety (how many classes were predicted by pixel) and majority (the prediction that occurred most often by pixel). We calculated the frequency of the majority class prediction ('Equal Frequency' tool) and the uncertainty of the prediction by the formula in 'Raster Calculator':

$$U = 1 - MAJFREQ / n$$

in which U is the uncertainty; MAJFREQ is the frequency of the prediction that occurs most often; and n is the number of repetitions.

The uncertainty ranges from 0 to 1 and represents a synthesis of the variability of cell prediction, with higher values identify points that require more sampling and the prediction is less reliable. The uncertainty was classified as low (≤0.2), medium (0.2< uncertainty ≤0.4), high (0.4< uncertainty ≤0.6), and very high (>0.6) and was quantified by the number of cells. To improve the evaluation of the output maps, we used the mean value of the uncertainty for all areas. The majority map led to the creation of a digital soil unit

map that represents the most frequently predicted soil unit distribution according to the 50 repetitions of the models.

The maps of the majority soil unit distribution were used to represent the soil distribution of the area (for each model and dataset) and were evaluated statistically and by pedologist expertise.

### Workflow

Pedological expertise remains a key factor in building DSM models (Kempen et al., 2009). The workflow diagram (Figure 3) highlights the importance of pedologists in conducting the analysis. Only the machine learning model step does not require pedological expertise. The entire process, except the machine learning process described above, is dependent on pedological expertise. We can verify that part of the process is data-driven (machine learning) and part is knowledge-driven (dependent on pedological expertise).

Pedology expertise is needed to assemble legacy data. Planning the sample design is the first step, followed by fieldwork to collect primary soil information and define the ROICs. After the analytical results and final soil classification of the GCPs, the soil map units were defined, and the ROIC database was created in the GIS environment through polygons/regions around the GCPs that represent each soil map unit. The covariates created for the entire area were extracted to the ROICs, and a database was built and transformed into a data frame in the R environment to apply the machine learning models.

The machine learning step is a routine that does not depend on pedologist expertise, wherein all the procedures are automatically performed by a script that performs the random selection of training and validation subsets, executes the training over the subset to four models, applies the models over the validation subset, and finally calculates the OA, kappa, and the predicted map at every loop (50 iterations).

The evaluation of statistics and output maps considers the mean OA and kappa of the models, the predicted map statistics of variety, the majority and uncertainty, and pedological expertise.

## RESULTS

The urban areas and rock outcrops (units Urb and RckO, respectively) obtained from the previously available LULC map were superimposed on the soil unit maps achieved from the applied computational methods and represented 1.8 and 33 % of the total area, respectively.

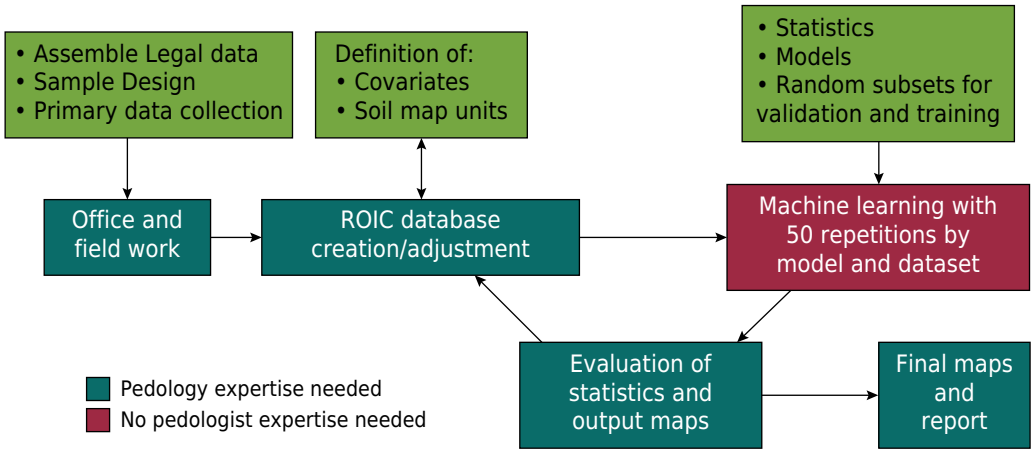



**Figure 3.** Flow chart summarizing the methodology used in the study.

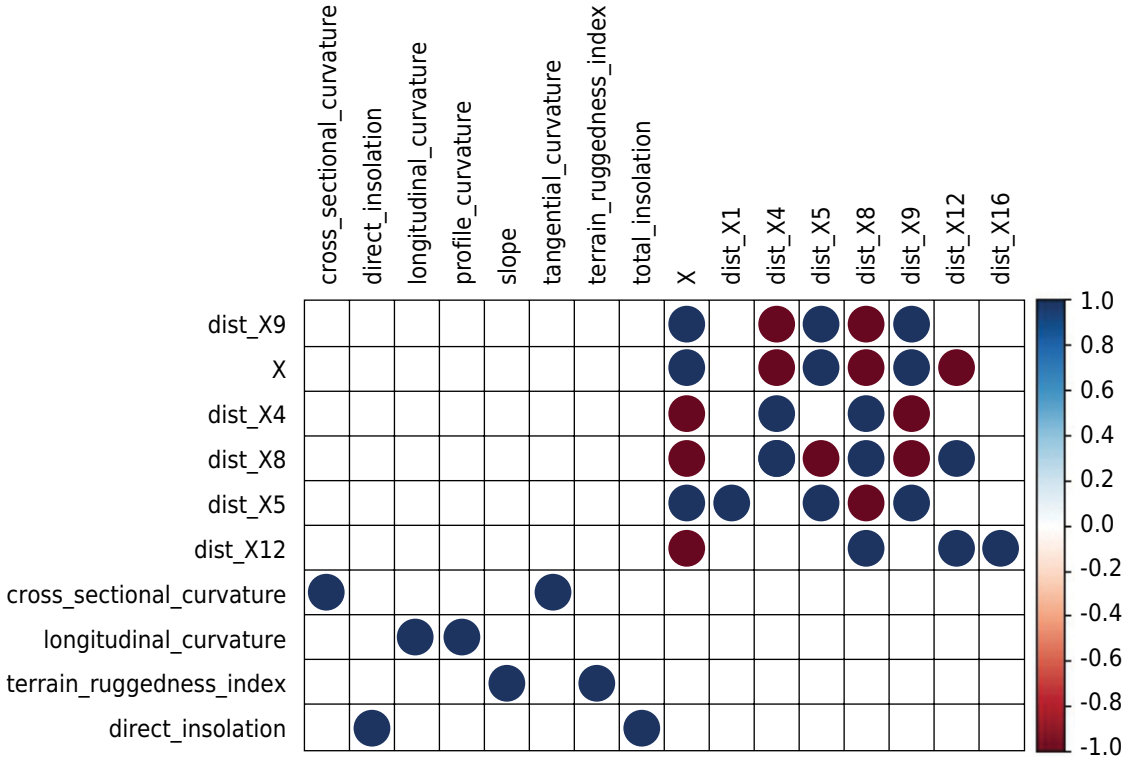



**Figure 4.** The high correlation covariates, considering all covariates. The Y-axis shows the eliminated covariates.

In the process of reducing the covariates with high correlation, ten covariates were discarded remaining in the final dataset 39 covariates. Figure 4 shows these correlations and the eliminated covariates. Dataset01 and Dataset05 used 27 covariates, Dataset02 used 28, Dataset03 used 38, and Dataset04 remains with 39 covariates.

The input Dataset01 to Dataset04 were split into training (60 %) and validation (40 %) sets grouped by ROICs because we realized that in each ROIC, the variance of the covariate was lower than between ROICs. In figure 5, the boxplot of covariates 'aspect' and 'elevation' to soil units CXbd, RLd+RckO, and RLd+CXbd+RckO shows the differences between ROICs and the single dataset (Dataset05). It is possible to verify the variability of the covariates to the same soil unit in the function of the ROICs.

The OA presents mean values between 0.34 and 0.62 when the dataset is grouped by ROIC and between 0.69 and 0.97 for Dataset05 (Figure 6a). Although the OA values are greater when using Dataset05, we can compare its majority map output against the majority map produced by Dataset01 because the covariates are the same; however, the method of splitting into training and validation subsets is different.

To investigate the effect of selected sample subsets used as training and validation, we ran 50 repetitions across the datasets and models and recorded the values of OA and kappa as well as the map output from each loop. The general mean results of OA are shown in figure 6a.

Considering Dataset04, we can see in figure 6b the OA variation for each loop and model. The OA standard deviation for these models was 0.09, which represents 15 % of the mean value. In general, for Dataset04, the OA values range from 0.29 (MLR) to 0.82 (RF). For the RF model with this dataset, the values of OA range from 0.40 to 0.82.

The results of OA to Dataset05 (Figure 6c) are greater than the other datasets and varies from 0.65 to MLR and 0.99 to C5.0 and GBM.

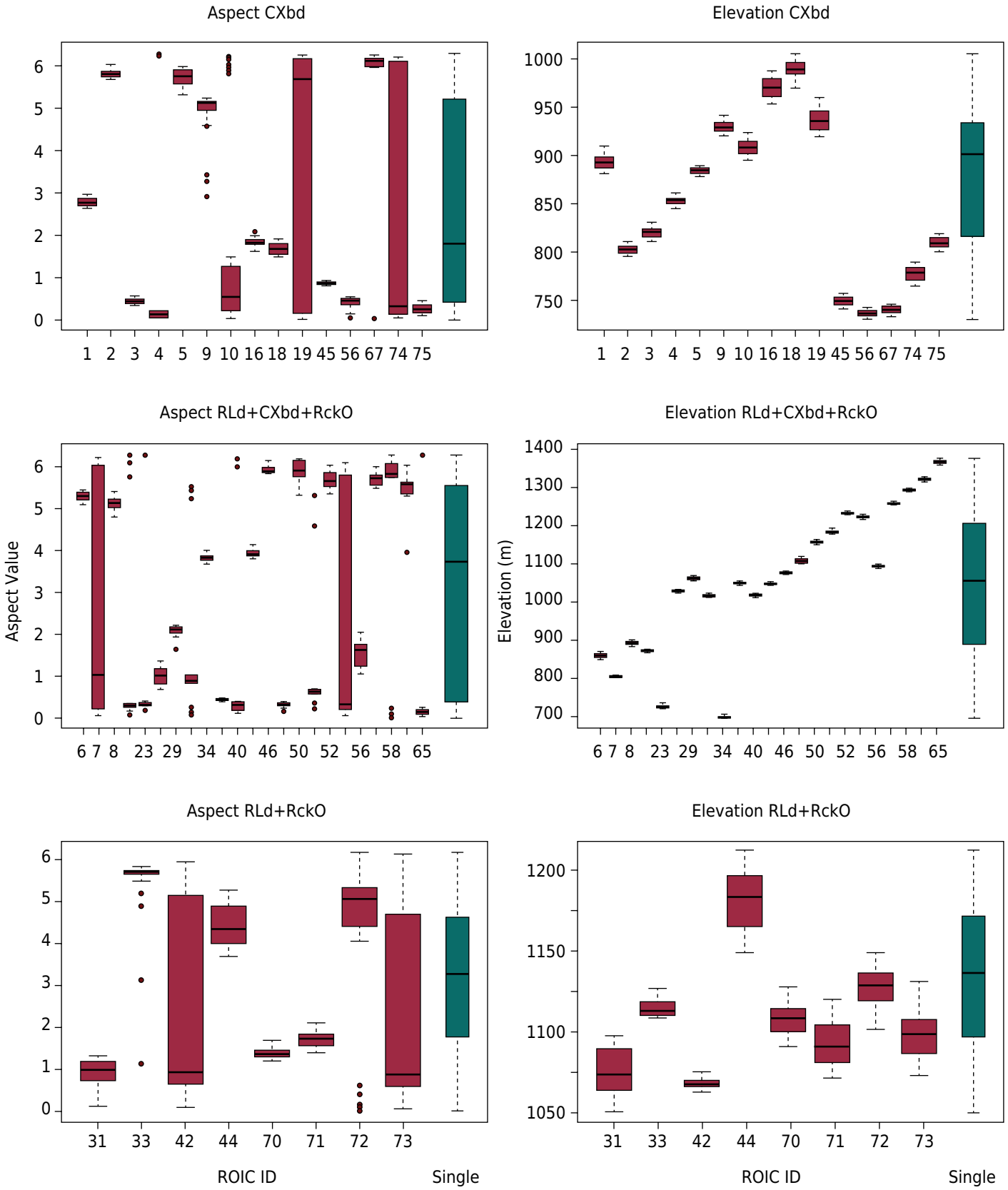
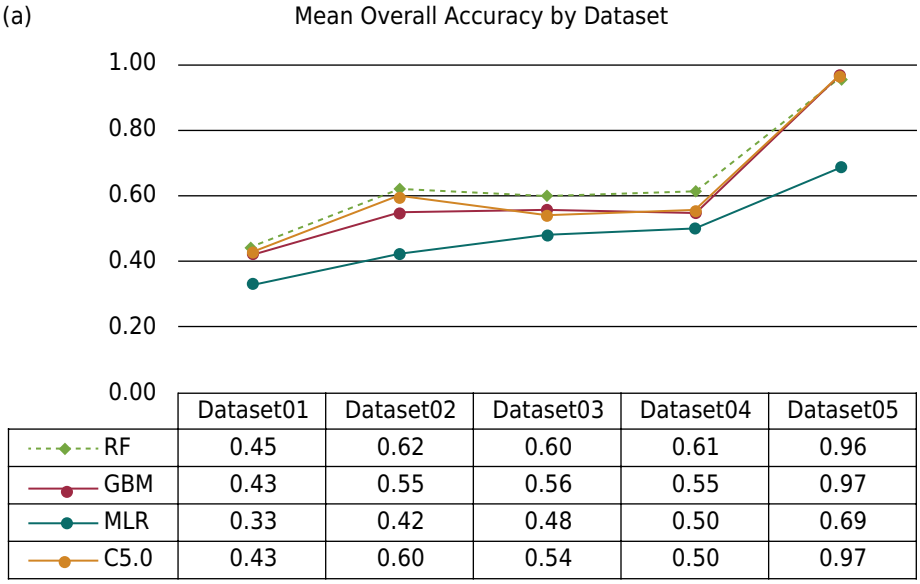


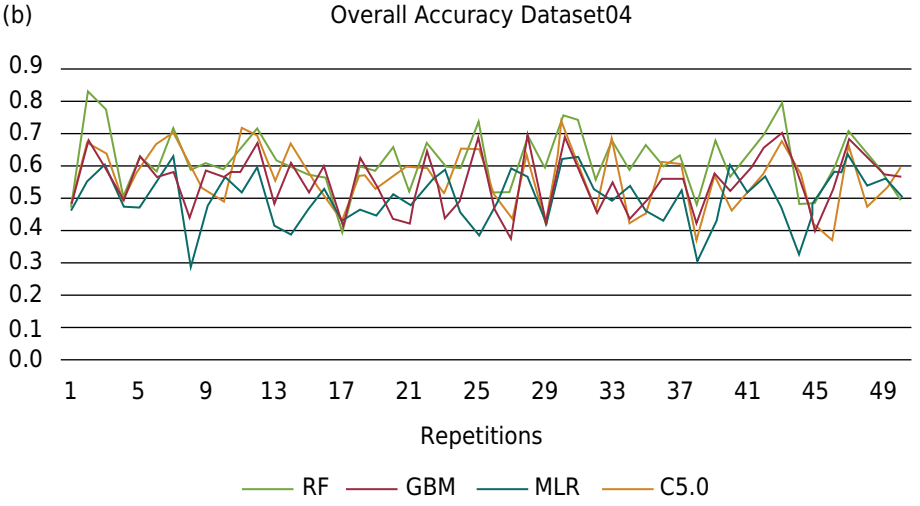

**Figure 5.** Boxplot of Aspect and Elevation ROICs and the single dataset (green color) of three soil units.

(a)

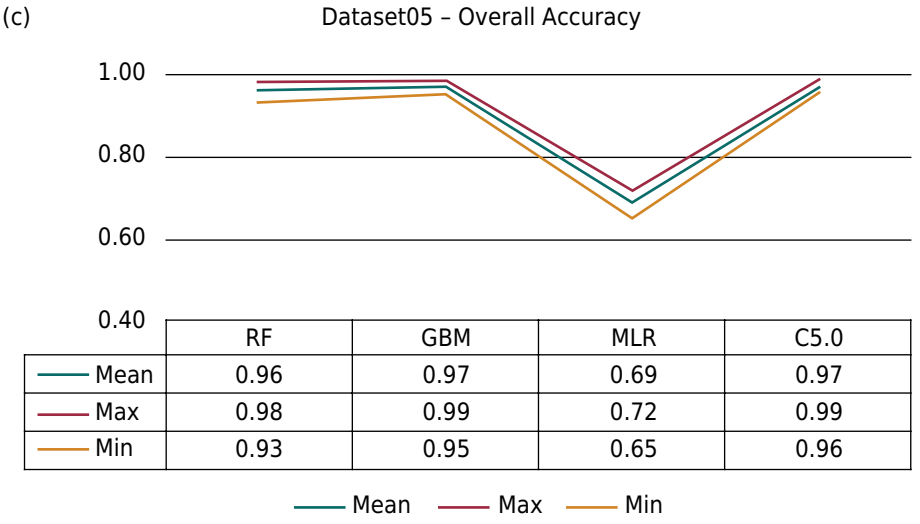



| | Dataset01 | Dataset02 | Dataset03 | Dataset04 | Dataset05 |
|---|---|---|---|---|---|
| RF | 0.45 | 0.62 | 0.60 | 0.61 | 0.96 |
| GBM | 0.43 | 0.55 | 0.56 | 0.55 | 0.97 |
| MLR | 0.33 | 0.42 | 0.48 | 0.50 | 0.69 |
| C5.0 | 0.43 | 0.60 | 0.54 | 0.50 | 0.97 |

(b)




(c)




| | RF | GBM | MLR | C5.0 |
|---|---|---|---|---|
| Mean | 0.96 | 0.97 | 0.69 | 0.97 |
| Max | 0.98 | 0.99 | 0.72 | 0.99 |
| Min | 0.93 | 0.95 | 0.65 | 0.96 |

**Figure 6.** Mean overall accuracy considering four models and five datasets (a); the OA behavior of the random repetitions of the models to Dataset04 (b); and mean, maximum, and minimum OA of the models with Dataset05 (c).

According to the workflow approach (Figure 3), the output maps need to be evaluated by a pedologist and by statistical parameters. First, we took all 50 maps generated by running the models and accounted for the variety (how many classes were predicted by pixel) and the majority maps, followed by an uncertainty calculation (Table 3) and statistics (maximum, mean, minimum, and standard deviation).

Table 3 shows the variety in % of the area of each model and dataset, and a variety value of 1 means that all 50 random repetitions of the model predicted the same soil unit in the cells and that the uncertainty is zero. On the other hand, a variety of 6 identifies the cells that showed a prediction of all six soil units in 50 random repetitions and the uncertainty is bigger. Considering low values of variety of 1 and 2, the GBM, RF, and C50 models, in that order, showed the best performance in Dataset01. The MLR model showed a worse result compared to the others.

Considering the mean value of uncertainty, RF had the lowest uncertainty value for all the datasets, ranging from 0.091 to 0.286. Uncertainty values for GBM and C50 ranged from 0.105 to 0.363. The MLR showed a larger variation range from 0.096 to 0.516.

The uncertainty of the maps results was also classified and computed based on the number of cells for each defined class (Figure 7). It can be observed to Dataset01 that low uncertainty class is bigger to C5.0, GBM, and RF, and RLM is the lower and significantly different. To Dataset05, all the models present a greater area with low uncertainty, showing

**Table 3.** Variety in % of area and Uncertainty basic statistics to the models and datasets

| Model | Variety (% of area) | | | | | | Uncertainty | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | MIN | MEAN | MAX | SD |
| | | | | | Dataset01 | | | | | |
| C50 | 7.45 | 12.02 | 26.40 | 36.34 | 17.27 | 0.52 | 0 | 0.317 | 0.78 | 0.188 |
| RF | 3.68 | 18.60 | 32.36 | 30.89 | 14.02 | 0.46 | 0 | 0.286 | 0.78 | 0.180 |
| GBM | 4.35 | 21.41 | 32.78 | 28.10 | 12.52 | 0.84 | 0 | 0.297 | 0.78 | 0.179 |
| MLR | 0.16 | 2.79 | 42.17 | 28.04 | 21.86 | 4.99 | 0 | 0.435 | 0.78 | 0.151 |
| | | | | | Dataset02 | | | | | |
| C50 | 4.76 | 14.18 | 29.47 | 40.65 | 10.54 | 0.40 | 0 | 0.317 | 0.74 | 0.177 |
| RF | 7.02 | 24.04 | 33.71 | 28.06 | 7.07 | 0.10 | 0 | 0.237 | 0.76 | 0.186 |
| GBM | 2.23 | 20.37 | 34.34 | 33.89 | 8.62 | 0.56 | 0 | 0.328 | 0.78 | 0.177 |
| MLR | 0.15 | 0.81 | 7.76 | 57.35 | 27.88 | 6.05 | 0 | 0.515 | 0.80 | 0.120 |
| | | | | | Dataset03 | | | | | |
| C50 | 2.46 | 16.79 | 27.15 | 33.66 | 18.35 | 1.57 | 0 | 0.363 | 0.74 | 0.169 |
| RF | 8.29 | 36.61 | 33.35 | 18.72 | 2.92 | 0.10 | 0 | 0.217 | 0.74 | 0.181 |
| GBM | 3.01 | 24.57 | 37.52 | 24.76 | 9.84 | 0.29 | 0 | 0.299 | 0.76 | 0.189 |
| MLR | 0.07 | 0.28 | 3.27 | 11.41 | 24.32 | 60.65 | 0 | 0.496 | 0.80 | 0.154 |
| | | | | | Dataset04 | | | | | |
| C50 | 2.34 | 13.47 | 28.82 | 35.14 | 17.42 | 2.80 | 0 | 0.363 | 0.74 | 0.172 |
| RF | 5.39 | 36.05 | 37.69 | 17.61 | 3.26 | 0.00 | 0 | 0.229 | 0.76 | 0.175 |
| GBM | 2.61 | 19.15 | 44.37 | 24.27 | 9.50 | 0.11 | 0 | 0.295 | 0.76 | 0.177 |
| MLR | 0.09 | 0.31 | 1.64 | 7.30 | 32.10 | 58.56 | 0 | 0.516 | 0.80 | 0.135 |
| | | | | | Dataset05 | | | | | |
| C50 | 35.23 | 33.04 | 23.17 | 6.72 | 1.84 | 0.01 | 0 | 0.152 | 0.76 | 0.177 |
| RF | 44.23 | 39.92 | 13.67 | 2.02 | 0.15 | 0.00 | 0 | 0.091 | 0.74 | 0.141 |
| GBM | 49.28 | 34.47 | 12.69 | 3.22 | 0.33 | 0.01 | 0 | 0.105 | 0.74 | 0.155 |
| MLR | 46.56 | 36.94 | 14.06 | 2.37 | 0.07 | 0.00 | 0 | 0.096 | 0.72 | 0.147 |

the significative difference in using or not the ROICs to split the data into validation and training datasets.

The evaluation of the output maps used the RF and GBM models for all datasets, considering their better values for OA, variety, and uncertainty. For these two models, an expert evaluation regarding the soil unit distribution (visual interpretation) in the majority map was performed. Figure 8 shows the majority of soil map units to these models for Dataset01 to 04.

For Dataset02, the geographic position was included in the covariates, a pedologist identified straight lines separating the LAd (yellow color) soil unit in both models. The pedologist noted the artefact effect produced by the geographic position and deal

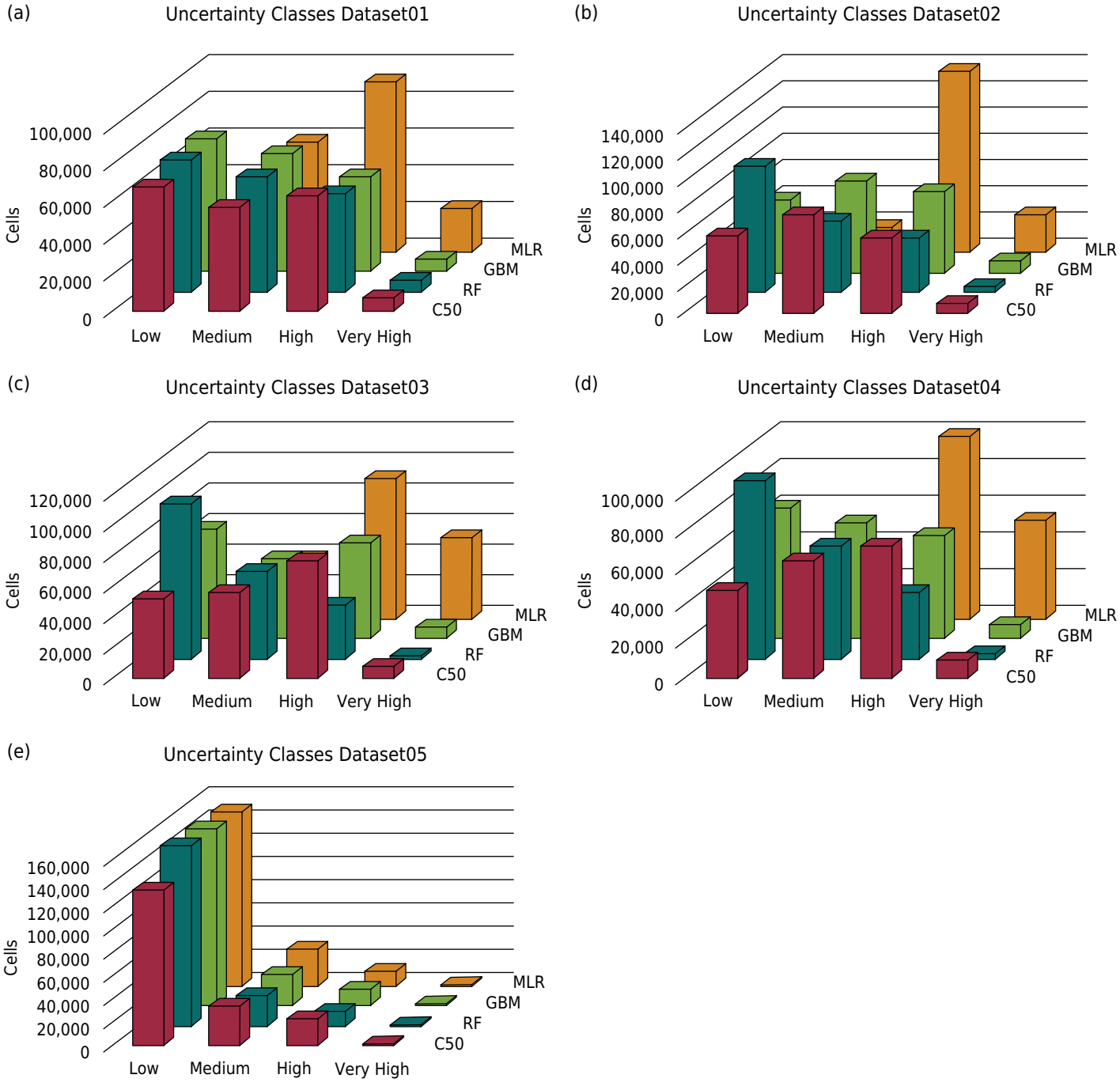



**Figure 7.** Graphical distribution of uncertainty classes by dataset and model, where Low (≤0.2); Medium (>0.2 to ≤0.4); High (>0.4 to ≤0.6) and Very High (>0.6).

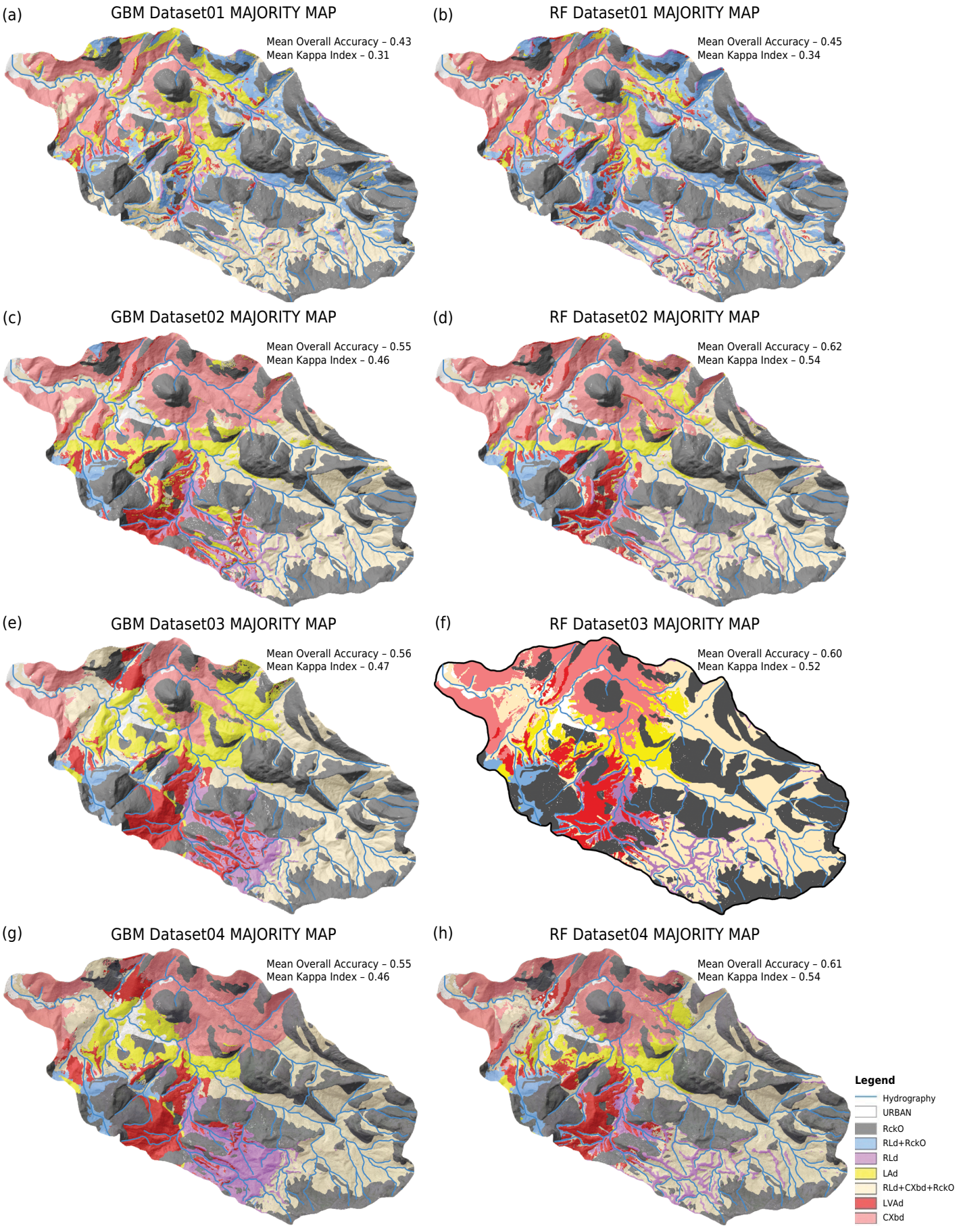



**Figure 8.** Map output by RF and GBM using the four datasets with ROICs.

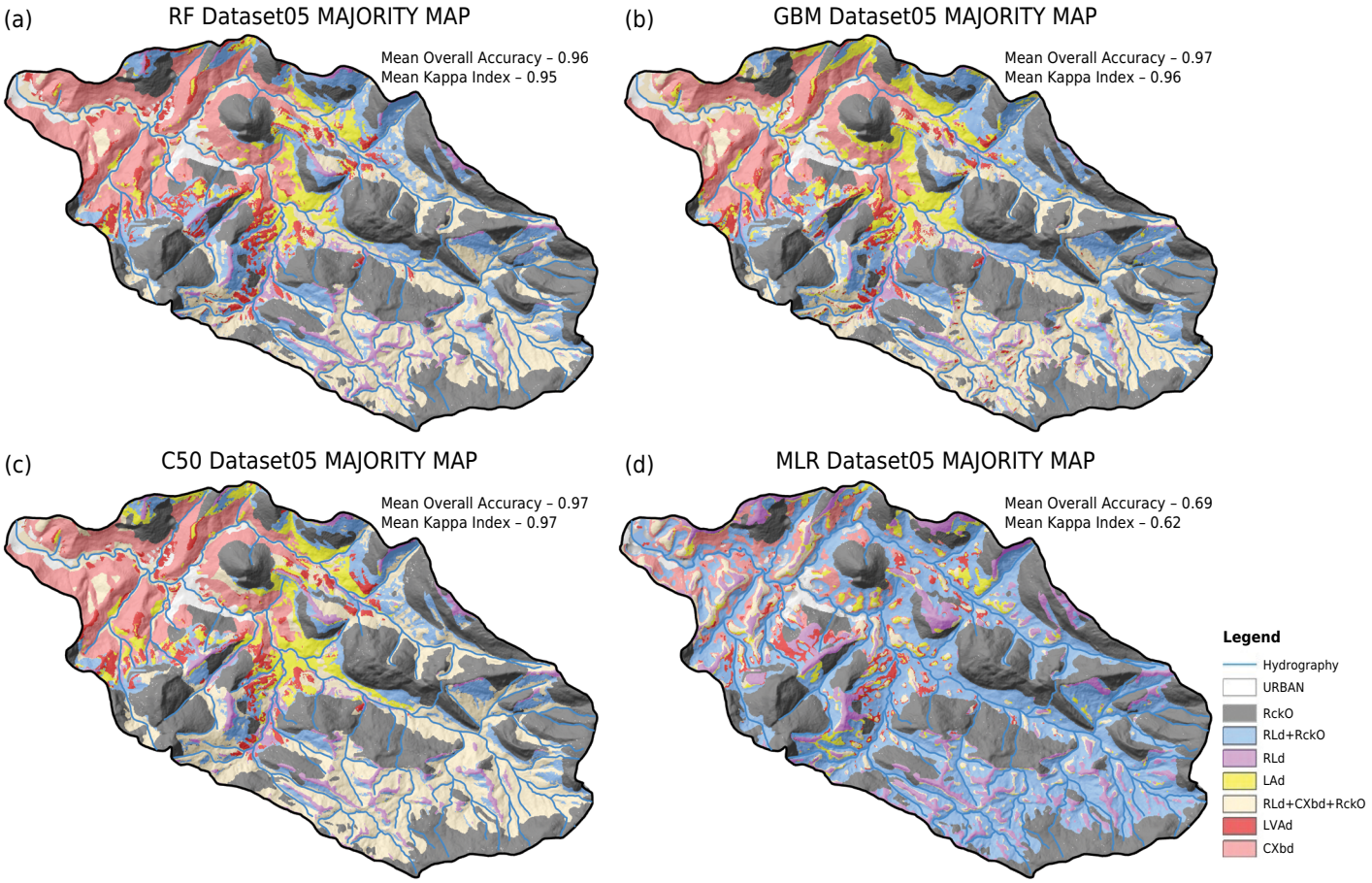



**Figure 9.** Output maps produced by Dataset05, with mean overall accuracy and kappa index.

**Table 4.** Coincidence of majority cells prediction of soil units by RF and GBM with Dataset01 and 05 in percentage of area

| | RF Dataset01 | GBM Dataset01 | RF Dataset05 | GBM Dataset05 |
|---|---|---|---|---|
| RF Dataset01 | | | | |
| GBM Dataset01 | 79.9 | | | |
| RF Dataset05 | 90.6 | 78.1 | | |
| GBM Dataset05 | 79.6 | 88.7 | 78.9 | |

with caution the maps produced by the Dataser02. Actually, this does not occur in the environment because soil unit distribution is mainly due to geomorphologic aspects, and these maps do not represent natural soil unit distribution.

The output maps produced by the models with Dataset05 are presented in figure 9. The values of mean overall accuracy (Figure 6) and kappa index were greater than those achieved by the other datasets.

A comparison between the majority maps from RF and GBM to Dataset01 and Dataset05 shows the cell coincidence prediction in percent of the area (Table 4), varying between 78.1 to 90.6 %.

## DISCUSSION

Depending on whether the dataset is grouped by ROIC or not, the results are quite different. We observed that the values of the mean, range, and standard deviation of

each ROIC for the covariates to one soil unit were very different from each other, even for the values for ROICs belonging to the same soil unit. This is probably the main aspect to explain the differences in the results from Dataset01 to Dataset04 and the Dataset05.

When the training and validation datasets were split based on ROICs, the subsets are quite different because of the variance between ROICs of the same covariate and soil unit (Figure 5). This is the reason for the worse results of these models, comparing with the Dataset05.

On the other hand, when used one complete dataset not based on ROICs, the variance of training and validation subsets to one covariate and one soil unit is lower, because the random split takes account to subdivide one ROIC into samples to both subsets. The lower variance produces more accurate parameters of models quality.

Camera et al. (2017) achieved an OOB error of 8.6 %, which is comparable to the RF results in Dataset05. The results of OA are in the same range as those obtained by Taghizadeh-Mehrjardi et al. (2012), Heung et al. (2017), Wolski et al. (2017), Pásztor et al. (2018), and Afshar et al. (2018), even though they used different approaches to build the dataset and process the data.

The RF is the model with the highest OA for datasets 01 to 04 and the same OA as GBM and C50 for Dataset05. Other studies have also identified RF as the best model for digital soil unit mapping (Camera et al., 2017; Heung et al., 2017; Mosleh et al., 2017).

The best OA result was for Dataset05, followed by Dataset02, Dataset03, and Dataset04. Dataset01 presented the worst response in the OA index compared with the others that used the ROICs, but this dataset is related to geomorphological conditions; thus, the map outputs were considered by pedological experts by a visual interpretation to evaluate the models. The OA behavior in relation to all 50 random repetitions considering the datasets showed the same trend, where MLR had the lowest OA and RF had the best performance for all datasets.

The relationship between OA and the number of covariates was not verified. The method based on the ROICs for Dataset02, 03, and 04 did not show significant differences in OA, although each dataset had a different number of covariates. The smallest datasets by number of covariates, Dataset01 and Dataset05 with 27 covariates, showed extreme responses: Dataset01 had the lowest OA, and Dataset05 had the highest.

These results show the influence of the split method grouped by ROICs (Figure 6a). Figure 6c shows the mean, maximum, and minimum OA for Dataset05, which achieved the greatest values for this index.

Dataset02, 03, and 04 showed the same trend, and the model with the best response was RF, followed by GBM, C50, and MLR. For Dataset05, the best responses were from RF and GBM, followed by C50 and MLR.

There was no significant difference between the variety results for RF, GBM, and C50 for Dataset01. For Dataset02, 03, and 04, RF showed a better performance that was significantly different from the others. For Dataset05, RF, GBM, and MLR showed the best performance in variety and lower uncertainty maps.

The MLR model showed the worse performance in all datasets, but in the Dataset05, it was improved enough to be the third best model.

In this sense, the RF and GBM models can be considered more efficient than the others based on mean uncertainty by this approach to all datasets.

Dataset02, Dataset03, and Dataset04 with low values for variety (1 or 2) showed that RF had the best performance, followed by GBM, C50, and MLR. The results of MLR variety were significantly different from the results of the other models.

The evaluation of the model performance by uncertainty showed the same trend as the other indices (variety and OA), in which RF and GBM produced maps with greater areas with low uncertainty. On the other hand, MLR had more cells in the High and Very High uncertainty classes, except for Dataset05, which did not have a significant difference from the RF and GBM models. The RF model showed a decrease in the number of High and Very High uncertainty cells when the analysis included all datasets, while the other models did not present this trend. All the datasets presented RF with the best performance, followed by GBM, C50, and MLR. The tree learners are ensemble models that average across multiple trees and produced better results than the RLM.

The MLR presented the worst results in uncertainty for all datasets, except Dataset05, which used the method of splitting the data into training and validation by a single dataset. When a single dataset was used (Dataset05), all the model performances improved, mainly MLR, as a consequence of the inherited model type that improved when all samples were used together. The tree-based models also show the same trend, but on a smaller scale.

The map outputs to Dataset02 were considered a poor map result, even if the statistical results were not, and the inclusion of geographic UTM position Y did not improve the final result.

For Dataset03 and 04, which included distance rasters, the map outputs showed a large area with LVAd (red color) in the southwest, which is not consistent with pedologist expectations. The LAd unit (yellow color) was spread in the central area of the watershed, and this is also not consistent with pedologist expectations. These expert considerations for the four datasets conclude that the better map outputs are the ones produced by Dataset01, which can be compared with the ones produced by Dataset05 (Figure 9) because they have the same covariates.

The highest OA values were achieved by the GBM, C50, and RF models to Dataset05, following the same trend of the models with the other datasets.

Considering the qualitative evaluation of the MLR output map for Dataset05, the shape of the unit RLd+RckO (blue color) was spread over the entire area, which was not expected according to the team of pedologist experts who work in primary soil data collection. Thus, the MLR output map was put away. For Dataset05, the best output maps are represented by the models RF, GBM, and C50, and the same results were obtained by Camera et al. (2017), Heung et al. (2017), and Mosleh et al. (2017).

The greater values of coincidence are between the same model considering both Dataset01 and Dataset05. Despite this, the coincidence of cell prediction by the models is high, over 78 % for the RF and GBM models with Dataset01 and 05.

Based on the results, it is possible to conclude that the performance of RF and GBM models are equivalent to all datasets, and the final maps output may be decided on by pedologist experts, corroborating with Kempen et al. (2009), where pedological tacit knowledge remains a key factor in building a model that achieves both statistically and pedologically accurate mapped outputs.

## CONCLUSIONS

The geomorphologic covariates are the most important for use in this case because the soil distribution is relief-dependent and has a greater relation with these covariates. When the geographic position and distance raster were included in the covariate dataset, the

statistical parameters were improved, but the quality of the map outputs did not present the expected distribution according to the expert visual evaluation. The variability of the covariates within the ROICs is lower than the variability between the ROICs.

The ROIC methodology to split the dataset between training and validation subsets (Dataset01 to 04) showed worse performance, considering statistical parameters, compared to the single dataset (Dataset05). The use of a grouped sample selection showed lower OA and kappa than the single dataset; using the single dataset was the better procedure for developing digital soil unit maps.

Even so, the map outputs of RF and GBM for Dataset01 and Dataset05, with the same covariates, presented the same majority predictions for at least 78 % of the area. It seems that both methods produce consistent results in map outputs according to this methodology and pedologist expertise. The application of majority maps made possible to evaluate the uncertainty and support a consistent soil unit prediction.

In general, RF was the best model for classifying soil units, and GBM was similar but with slightly lower statistical values. The OA, kappa index, variety, majority, and uncertainty values can contribute to choosing the best model combined with pedologist expert evaluation about some artefact effects that can be produced by the models with particulars covariates addition.

Although the OA of the models for Dataset01 and Dataset05 is quite different, both models produced a good map output in a qualitative assessment by the pedologist. Additionally, the use of ROICs or not with this methodology produces very similar map outputs when using the RF or GBM model.

The increase in the number of covariates is not a guarantee in improvement in OA or kappa or in the quality of the map output, and the particular sampling design with conditioned Latin hypercube sampling may have an impact on the subsequent modeling and the linear model (MLR) might be disadvantaged compared to the machine learning models.

The geographic position and distance raster do not improve the quality of the map output for Dataset01 to Dataset04, representing no spatial dependency of soil units, according to a visual interpretation.

The results suggest that more studies need to be done to predict soil classes/units in tropical areas with a complex degree of soil distribution based on lithology, climate, topography, and vegetation cover.

## ACKNOWLEDGMENTS

We would like to acknowledge Project RHIMA, sponsored by the Science and Technology Ministry, for the funds for fieldwork and laboratory analysis.

## AUTHOR CONTRIBUTIONS

**Conceptualization:** Waldir de Carvalho Junior (lead), Elpidio Inacio Fernandes Filho (lead), and Nilson Rendeiro Pereira (supporting).

**Methodology:** Elpidio Inacio Fernandes Filho (lead), Waldir de Carvalho Junior (supporting), and Nilson Rendeiro Pereira (supporting).

**Software:** Waldir de Carvalho Junior (lead), Elpidio Inacio Fernandes Filho (lead), and Nilson Rendeiro Pereira (supporting).

**Validation:** Waldir de Carvalho Junior (lead), Elpidio Inacio Fernandes Filho (supporting); Nilson Rendeiro Pereira (supporting), Braz Calderano Filho (supporting),

Helena Saraiva Koenow Pinheiro (supporting), Cesar da Silva Chagas (supporting), Silvio Barge Bhering (supporting), Vinicius Rendeiro Pereira (supporting), and Sara Lawall (supporting).

**Formal analysis:** Waldir de Carvalho Junior (lead), Elpidio Inacio Fernandes Filho (supporting), and Nilson Rendeiro Pereira (supporting).

**Investigation:** Waldir de Carvalho Junior (supporting), Elpidio Inacio Fernandes Filho (supporting), Nilson Rendeiro Pereira (lead), Braz Calderano Filho (supporting), Helena Saraiva Koenow Pinheiro (supporting), Cesar da Silva Chagas (supporting), Silvio Barge Bhering (supporting), Vinicius Rendeiro Pereira (supporting), and Sara Lawall (supporting).

**Resources:** Waldir de Carvalho Junior (lead), Elpidio Inacio Fernandes Filho (supporting), Nilson Rendeiro Pereira (lead), Braz Calderano Filho (supporting), and Helena Saraiva Koenow Pinheiro (supporting).

**Data curation:** Waldir de Carvalho Junior (supporting), Elpidio Inacio Fernandes Filho (supporting), Nilson Rendeiro Pereira (lead), Vinicius Rendeiro Pereira (supporting), and Sara Lawall (supporting).

**Writing – original draft:** Waldir de Carvalho Junior (lead), Elpidio Inacio Fernandes Filho (supporting), and Nilson Rendeiro Pereira (supporting).

**Writing – review and editing:** Waldir de Carvalho Junior (lead), Elpidio Inacio Fernandes Filho (supporting), and Nilson Rendeiro Pereira (supporting).

**Visualization:** Waldir de Carvalho Junior (lead), Elpidio Inacio Fernandes Filho (supporting), and Nilson Rendeiro Pereira (supporting).

**Supervision:** Waldir de Carvalho Junior (lead), Elpidio Inacio Fernandes Filho (supporting), and Nilson Rendeiro Pereira (supporting).

**Project administration**: Waldir de Carvalho Junior (lead), Elpidio Inacio Fernandes Filho (supporting), and Nilson Rendeiro Pereira (supporting).

**Funding acquisition:** Waldir de Carvalho Junior (lead), Elpidio Inacio Fernandes Filho (supporting), Nilson Rendeiro Pereira (supporting), and Sara Lawall (lead).

## REFERENCES

Adhikari K, Minasny B, Greve MB, Greve MH. Constructing a soil class map of denmark based on the FAO legend using digital techniques. Geoderma. 2014;214-215:101-13. https://doi.org/10.1016/j.geoderma.2013.09.023

Afshar FA, Ayoubi S, Jafari A. The extrapolation of soil great groups using multinomial logistic regression at regional scale in arid regions of Iran. Geoderma. 2018;315:36-48. https://doi.org/10.1016/j.geoderma.2017.11.030

Bhering SB, Chagas CD, Carvalho Junior W, Pereira NR, Calderano Filho B, Pinheiro HSK. Mapeamento digital de areia, argila e carbono orgânico por modelos Random Forest sob diferentes resoluções espaciais. Pesq Agropec Bras. 2016;51:1359-70. https://doi.org/10.1590/s0100-204x2016000900035

Burrough PA, Vangaans PFM, Hootsmans R. Continuous classification in soil survey: Spatial correlation, confusion and boundaries. Geoderma. 1997;77:115-35. https://doi.org/10.1016/S0016-7061(97)00018-9

Calderano Filho B, Polivanov H, Chagas CD, Carvalho Júnior W, Barroso EV, Guerra AJT, Calderano SB. Artificial neural networks applied for soil class prediction in mountainous landscape of the Serra do Mar. Rev Bras Cienc Solo. 2014;38:1681-93. https://doi.org/10.1590/S0100-06832014000600003

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

Carvalho Junior W, Lagacherie P, Chagas CS, Calderano Filho B, Bhering SB. A regional-scale assessment of digital mapping of soil attributes in a tropical hillslope environment. Geoderma. 2014;232:479-86. https://doi.org/10.1016/j.geoderma.2014.06.007

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

Conrad O, Bechtel B, Bock M, Dietrich H, Fischer E, Gerlitz L, Wehberg J, Wichmann V, Bohner J. System for automated geoscientific analyses (SAGA) v. 2.1.4. Geosci Model Dev Discuss. 2015;8:2271-312. https://doi.org/10.5194/gmdd-8-2271-2015

Emre İE, Erol N, Ayhan Yİ, Özkan Y, Erol Ç. The analysis of the effects of acute rheumatic fever in childhood on cardiac disease with data mining. Int J Med Inform. 2019;123:68-75. https://doi.org/10.1016/j.ijmedinf.2018.12.009

Friedman JH. Stochastic gradient boosting. Comput Stat Data An. 2002;38:367-78. https://doi.org/10.1016/S0167-9473(01)00065-2

Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat. 2001;29:1189-232.

Gruber FE, Baruck J, Geitner C. Algorithms vs. surveyors: a comparison of automated landform delineations and surveyed topographic positions from soil mapping in an Alpine environment. Geoderma. 2017;308:9-25. https://doi.org/10.1016/j.geoderma.2017.08.017

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

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

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

Kuhn M. Package “C50”. R package version 0.1.1; 2017 [cited 2017 December 1]. Available from: http://cran.r-project.org/web/packages/C50/C50.pdf.

Leite CAS, Perrotta MM, Silva LC, Heineck CA, Salvador AD, Vieira VS, Lopes RC, Silva MGM. Carta geológica do Brasil ao milionésimo: folha SF-23. Brasília, DF: Programa Geologia do Brasil; 2004.

Liaw A, Wiener M. randomForest: Breiman and Cutler’s random forests for classification and regression. R package version 4.6-10; 2018 [cited 2018 March 25]. Available from: http://cran.r-project.org/web/packages/randomForest/randomForest.pdf.

Liaw A, Wiener M. Classification and regression by randomForest. R News. 2002;2/3:18-22.

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. Latin hypercube sampling as a tool for digital soil mapping. Dev Soil Sci. 2007;31:153-65. https://doi.org/10.1016/S0166-2481(06)31012-4

Mosleh Z, Salehi MH, Jafari A, Borujeni IE, Mehnatkesh A. Identifying sources of soil classes variations with digital soil mapping approaches in the Shahrekord plain, Iran. Environ Earth Sci. 2017;76:748. https://doi.org/10.1007/s12665-017-7100-0

Pandya R, Pandya J. C5.0 algorithm to improved decision tree with feature selection and reduced error pruning. Int J Comput Appl T. 2015;117:18-21.

Pásztor L, Laborczi A, Bakacsi Z, Szabo J, Illes G. Compilation of a national soil-type map for Hungary by sequential classification methods. Geoderma. 2018;311:93-108. https://doi.org/10.1016/j.geoderma.2017.04.018

Ridgeway G. Package 'gbm'. R package version 2.1.3; 2017 [cited 2019 January 14]. Available from: http://cran.r-project.org/web/packages/gbm/gbm.pdf.

Ripley B, Venables W. R package 'nnet'. R package version 7.3.12; 2016 [cited 2016 February 2]. Available from: http://cran.r-project.org/web/packages/nnet/nnet.pdf.

Rossiter DG. Digital soil mapping as a component of data renewal for areas with sparse soil data infrastructures. In: Hartemink AE, McBratney A, Mendonça-Santos ML, editors. Digital Soil Mapping with Limited Data. New York: Springer; 2008. p. 69-80.

Rossiter DG, Zeng R, Zhang G-L. Accounting for taxonomic distance in accuracy assessment of soil class predictions. Geoderma. 2017;292:118-27. https://doi.org/10.1016/j.geoderma.2017.01.012

Santos HG, Jacomine PKT, Anjos LHC, Oliveira VA, Lumbreras JF, Coelho MR, Almeida JA, Araújo Filho JC, Oliveira JB, Cunha TJF. Sistema brasileiro de classificação de solos. 5. ed. rev. ampl. Brasília, DF: Embrapa; 2018.

Silva LC, Cunha HCS. Geologia do Estado do Rio de Janeiro: texto explicativo do mapa geológico do Estado do Rio de Janeiro. 2. ed. Brasília, DF: CPRM; 2001.

Sim S, Im J, Park S, Park H, Ahn MH, Chan P-w. Icing detection over east Asia from geostationary satellite data using machine learning approaches. Remote Sens. 2018;10:631. https://doi.org/10.3390/rs10040631

Taghizadeh-Mehrjardi R, Minasny B, Mcbratney AB, Triantafilis J, Sarmadian F, Toomanian N. Digital soil mapping of soil classes using decision trees in central Iran. In: Minasny B, Malone BP, MacBratney AB. Digital soil assessment and beyond. Boca Raton: CRC Press; 2012. p. 197-202.

Teske R, Giasson E, Bagatini T. Comparação de esquemas de amostragem para treinamento de modelos preditores no mapeamento digital de classes de solos. Rev Bras Cienc Solo. 2015;39:14-20. https://doi.org/10.1590/01000683rbcs20150344

Witten IH, Frank E. Data mining: practical machine learning tools and techniques. 2nd ed. San Francisco: Elsevier; 2005.

Wolski MS, Dalmolin RSD, Flores CA, Moura-Bueno JM, ten Caten A, Kaiser DR. Digital soil mapping and its implications in the extrapolation of soil-landscape relationships in detailed scale. Pesq Agropec Bras. 2017;52:633-42. https://doi.org/10.1590/S0100-204X2017000800009