







Division - Soil in Space and Time | Commission - Pedometrics

Training pedologist for soil mapping: Contextualizing methods and its accuracy using the project pedagogy approach

Elias Mendes Costa^{(1)*} , Marcos Bacis Ceddia⁽¹⁾ , Felipe Nascimento dos Santos⁽¹⁾ , Laiz de Oliveira Silva⁽¹⁾ , Igor Prata Terra de Rezende⁽¹⁾  and Douglath Alves Correa Fernandes⁽¹⁾ 

⁽¹⁾ Universidade Federal Rural do Rio de Janeiro, Departamento de Solos, Seropédica, Rio de Janeiro, Brasil.

ABSTRACT: There is a growing demand for more detailed knowledge about soils, their functions, and connections with human activities and environmental services. In Brazil, where soil survey and mapping have been scarce since the 1990s, there is a remarkable sense of urgency. Recently, a national soil program was created (*PronaSolos*) to attend to the massive demand for soil information. *PronaSolos* is an effort to return to the systematic soil mapping of the national territory, which requires many pedologists who master the traditional knowledge of soil mapping, but above all, the modern and accurate digital soil mapping (DSM) techniques. Based on these aspects, this study aims to address the technical and educational aspects inherent in the training process of new pedologists by contextualizing different soil mapping methods using the pedagogy project approach (PPA). Specifically, the study sought to assess the following subjects: (i) evaluate the learning process of different apprentices in performing soil survey and mapping in a small training area; (ii) compare maps generated by conventional soil mapping (CSM) and DSM using two probabilistic design for validation (SRS - Simple Random Sampling and SSRS - Stratified Simple Random Sampling). The DSM techniques evaluated were: Multinomial Logistic Regression - MLR and Random Forest - RF. For the course, four apprentices were selected and trained in both CSM and DSM techniques. Finally, they were asked about the learning process in the PPA and improvement for future courses. This study showed that: a) the PPA is promising to train new pedologists since, by mixing theoretical activities and contextualized practices (a project in progress), it not only awakens great motivation and critical capacity but also develops the ability for apprentices to find solutions in a area in constant evolution; b) the quality of the maps changed significantly according to the validation sample design applied. The CSM present better quality than DSM, mainly when using SSRS. The RF presented equivalent accuracy to CSM using SRS. Irrespective to validation sample design, the MLR presented the lowest accuracy; c) The CSMs presented higher user's accuracy while the DSMs presented higher producer's accuracy; d) The quality of CSM generated by the apprentices was not clearly related to the previous experience and knowledge in soil science.



Keywords: digital soil mapping, conventional soil mapping, soil-landscape relationship, soil education, PronaSolos.

* **Corresponding author:**
E-mail: eliasmccosta@gmail.com

Received: July 30, 2020

Approved: December 14, 2020

How to cite: Costa EM, Ceddia MB, Santos FN, Silva LO, Rezende IPT, Fernandes DAC. Training pedologist for soil mapping: Contextualizing methods and its accuracy using the project pedagogy approach. Rev Bras Cienc Solo. 2021;45:e0200130. <https://doi.org/10.36783/18069657rbc20200130>

Editors: José Miguel Reichert  and Ricardo Simão Diniz Dalmolin 

Copyright: This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided that the original author and source are credited.



INTRODUCTION

Pedology focuses on understanding soil genesis, including classification, soil-landscape relationship, and mapping (Ma et al., 2019). Soil mapping is a complex task, once the pedologists have to present a model of the spatial distribution of soil types, which demands to put together the soil knowledge of a study site (soil-landscape relationship). Traditionally, the soil maps have been done by pedologists incorporating the soil knowledge into an implicit conceptual model that is used to infer soil variation (Scull et al., 2003). This method has been severely criticized in the scientific literature for two reasons: a) the conceptual model developed by the soil surveyor is primarily implicit, being constructed in a heuristic manner, and b) the influence of the soil classification over the modeling process and map representation. In the first criticism, researchers argue that the results of traditional soil maps are excessively dependent upon tacit knowledge and, as such, incomplete information exists relative to the derivation of the ultimate soil survey product. The second criticism, the influence of soil classification, in practice results that the spatial perception of soil seems that the soil classes are homogenous units with and sharply defined boundaries (Burrough and McDonnell, 1998).

More recently, considering these criticisms and the evolution of the research in soil mapping, researchers introduced the concept of Digital Soil Mapping (DSM) and Predictive Soil Mapping (PSM), McBratney et al. (2003), and Scull et al. (2003), respectively. Digital Soil Mapping involves traditional methods of classification and mapping and mathematical modeling for the creation and population of spatial information systems using field and laboratory data coupled with environmental covariates (McBratney et al., 2003; Ma et al., 2019). Although the DSM represents an evolution in the soil mapping process, the formation of new pedologists also represents a challenge, once besides the former knowledge, the training process must constantly be provided in new technologies (remote and proximal sensors, software, and hardware), and data analysis methods and algorithms (Dalmolin et al., 2020). Considering that the world requires more detailed soil maps and information about soil security, there is an increasing demand for pedologists to execute DSM.

In Brazil, few detailed soil surveys can support agricultural-environmental planning, but with the recent launch of the National Soil Program of Brazil (*PronaSolos*), there is an expectation of resuming pedological surveys in the country (Polidoro et al., 2016), which are scarce since the 1990s, with only a few specific studies (Nolasco-Carvalho et al., 2013). As a consequence of the interruption of a systematic soil survey project in Brazil, there is not only a high demand for soil maps but also pedologists. Another important aspect of this previous situation is that most part of the pedologists with the most elevated experience in soil survey and mapping in Brazil, was formed using the traditional soil mapping methods.

These experienced pedologists scattered along the country are very important to take advantage of their knowledge to plan and execute the newest soil maps and participate in the training process of new pedologists. Parallel to this stagnation of systematic soil survey projects in Brazil, in the last few years, a lot of researches and mapping studies have been done using DSM techniques. However, the number of new pedologists to meet the potential demand for mapping is still small. In short, the new soil maps are being done, most frequently, using digital techniques, but the experienced pedologists of Brazil, in general, are not up to date with the new techniques.

This aspect was one of the main issues behind the construction of the *PronaSolos* project. The *PronaSolos* Project planning encompassed many challenges, such as: (a) what is the best scale of the maps, considering the different Brazilian regions, new demands of information, and the availability of legacy data?; (b) Which techniques to apply (Traditional, Digital, or both)?; (c) How many new pedologists must be trained to attend the soil mapping process along the country?; (d) How must be a training course for new

pedologists to acquire not only the basic knowledge of soils and mapping techniques but also to prepare them to master new technologies and to do soil survey with efficiency and accuracy.

Since 2017, the Laboratory of Soil and Water in Agroecosystem (LASA-UFRRJ) began a soil survey and mapping project in areas of Amazon and Bahia States. The purpose of the project has some similarities with the *PronaSolos* project. Considering the huge territory to be mapped in both study sites, the project's coordination decides to train new pedologists to attend the demand of field activities of the soil survey and soil mapping process. In 2019, a group of four apprentices in Agronomy (Residency Program in Agronomy - REA-UFRRJ) was selected to be part of the project. Before going to the fieldwork in Bahia State, the apprentice took part in a soil survey and mapping course that was offered by professors of LASA. To develop the training of pedologists, the methodological proposal called Project Pedagogy was adopted. The method is based on experience, which proposes a connection between the student and a research project that arouses his interest. The role of the tutor is to favor the teaching based on the discoveries, fruit of the research carried out by the apprentices under his guidance.

The method is interesting to use in soil mapping training because it fits the purpose of learning during project execution. Still regarding the method, the following features are highlighted: (a) *intentionality* - the project is chosen based on the objectives that the tutor intends for the learners to reach and on the set of knowledge they need to build, soil mapping, soil assessment, and applicability (capability and agricultural-environmental planning); (b) *flexibility* - learners have completely different characteristics and prior knowledge. Therefore, it is likely that the reception, engagement, and even the outcome that each learner achieves will be completely different. The tutor needs to follow up on activities to notice these differences and modify the plan if necessary; and (c) *multidisciplinary* - problem-solving can rarely be achieved with knowledge provided by a single area (in this case, soil science, mathematical modeling, cartography, and statistics). Based on these aspects, this study aims to address the technical and educational aspects inherent in the training process of new pedologists by contextualizing different soil mapping methods (and their specificities) using the project pedagogy approach. Since young pedologists with different degrees of experience, even if they receive the same training, will present different reception, engagement, and produce different soil maps and consequently different map accuracy, this study aimed to: (i) evaluate the learning process of different technicians in performing soil survey and mapping using conventional techniques in a small training area; and (ii) compare maps generated by conventional and digital techniques (Logistic Regression and Random Forest) using for validation two probabilistic approaches, simple random and stratified simple random sampling.

MATERIALS AND METHODS

This study was conducted in the context of the project entitled "Digital Mapping of Soils in Oil and Gas Exploration and Production Areas - Case Studies of the North and Northeast Brazilian Fields". The project was funded by the National Petroleum Agency (ANP), under the agreement number 5850.0105881.17.9 (PETROBRAS/FAPUR/UFRRJ). The Laboratory of Soil and Water in Agroecosystem (LASA) was responsible for conducting the research between 2017 and 2021. Considering the areas' extent to be mapped on a detailed and semi-detailed scale (1:10,000 to 1:25,000), the coordination had to recruit and train four new pedologists. Specifically, the new pedologists were supposed to be ready to be at the fieldwork in 2019.

Considering the purpose of this study, this section was organized in two parts. The first part explains how the pedagogical approach was performed. The second describes the

soil mapping project in the training study site, encompassing the fieldwork and the technical procedures to execute the soil mapping.

Pedagogical project approaches

The pedagogical project approach was conducted following four steps: (i) selection of the apprentices and evaluation of their previous knowledge; (ii) preparation and planning; (iii) development, and (iv) project completion. The selection of four apprentices was made through a public context of the agronomy residency program of the Federal Rural University of Rio de Janeiro (UFRRJ, Portuguese). The apprentices were evaluated through theoretical tests on basic soil knowledge, curriculum analysis, and interview. Although the application of the test of expertise in soil, cartography and agricultural capability was important in the process, the program tutors already knew that agronomists who graduated at most two years would not have a great experience on these themes. Thus, the evaluation of the curriculum and the interview were decisive in selecting, where apprentices were sought motivated to work with multidisciplinary projects and willing to commit for two years in training activities and soil survey and mapping in different regions.

The selection process also considered positive the largest possible diversity of apprentice profiles (race, gender, and sexual orientation, for example). The four apprentices selected were alumni of Agronomic Engineer course at UFRRJ. All of them, during the undergraduate course, took part in the same formal soil science disciplines, such as: Pedology, Soil Physics, Soil Fertility, and Agricultural Capability of Brazilian Soils. The detailed profile of the four apprentices selected is presented in table 1.

Considering the information presented in table 1, the instructor ranked the apprentices according to the rank of experience (RE) in a scale varying from 0 to 1 that could help

Table 1. Profile of the four apprentices selected

Apprentice	Gender	Age	Academic Formation and Skills	Experience*
Apprentice A	Male	28	Agriculture and Livestock technician. Undergraduate in Agronomic Engineer. Master Degree in Soil Science (PPGA-CS/UFRRJ).	Basic experience with soil survey and pedology. RE = 0.7
Apprentice B	Male	27	Surveying technician. Undergraduate in Agronomic Engineer. Basic experience with geoscience and DSM.	Very basic level of experience with soil science. RE = 0.3
Apprentice C	Male	28	Undergraduate in Agronomic Engineer. Master Degree in with masters in Plant Science (PPGF/UFRRJ), area of concentration in plant production, research topic in management and production of crops with economic importance.	Very basic level of experience with soil science. RE = 0.5
Apprentice D	Female	25	Agriculture and Livestock technician. Undergraduate in Agronomic Engineer. Master Degree in with masters in Plant Science (PPGF/UFRRJ), area of concentration in agroecology, research topic in nutrient cycling on the soil.	Basic experience with soil fertility and soil classification during under graduation. RE=0.5

* Basic Experience: they took part in more disciplines in Pos Graduation Program in Agronomy-Soil Science (Soil Formation and Characterization, Soil Chemistry, Soil Physics, and Soil Fertility), however, they had not any experience in fieldwork related to soil survey and soil mapping; Very Basic level in Soil Science: they only took part in soil science disciplines of the under graduation course in Agronomic Engineer, and had not any experience in fieldwork related to soil survey and soil mapping. RE: Rank Experience - the rank was based on work experience and pos graduation course. The instructor gave the highest level to apprentice A (RE = 0.7 - 2 years master degree in soil science), intermediate level to apprentices C and D (RE = 0.5 - 2 years master science and plant production), and lowest level to apprentice B (RE=0.3 - newly formed in agronomy as he entered the residence program).

explain the results along the training course, namely: apprentice A (RE = 0.7) > C (RE = 0.5) = D (RE = 0.5) > B (RE = 0.3).

Based on the aims of the project and the profile of the four apprentices, the coordination organized the training project, which was scheduled in theoretical and practice modules. The theoretical course focused on the following subjects: (a) concepts of soil formation, soil characterization (laboratory and field methods); (b) methods of soil survey and mapping; (c) use of GIS, soil sensors, and soil database system; and (d) evaluation of validation methods to compare the accuracy of soil maps. Besides the theoretical content, the apprentices to new pedologists took part in all steps of a detailed soil survey and mapping project, which was conducted in a training site of about 10 hectares.

The development of the project last four months and the theoretical and practical modules took place simultaneously. At the beginning (first week), the apprentices receive theoretical classes; thereafter, they applied the knowledge in the study site. In some weeks, the actives were only practiced in field and laboratory (soil description, soil sampling, field tests, and soil analysis). In the final part, the apprentices applied the knowledge to analyze the field and laboratory data, literature review, writing reports, and mapping soils of the study site using traditional and digital soil mapping techniques.

The completion of the training project was made through the evaluation of the soil map presented by each apprentice (validation of each soil map, comparison of each other, and with the maps generated by DSM techniques). Besides, each apprentice presented her/his perception of the main challenges they faced and a personal evaluation of the training course. This last part of the evaluation was made using a structured questionnaire developed by the tutors of the program (Table 2).

Table 2. The questionnaire used to evaluate the program

Subject	Questions
1- Which stages of the course were most challenging?	Degree of challenge (1 for the highest challenge and 6 for lowest). a-() Theoretical classes of introduction to pedology, soil-landscape relationship, soil survey, and classification. b-() Practical classes to identify soil in the field (description and classification) and soil-landscape relationship in practice. c-() Field tests and laboratory analysis. d-() Theoretical classes involving digital soil mapping, GIS and environmental covariates processing. e-() Theoretical classes involving statistical modeling, sampling, prediction, and accuracy assessment. f-() Practical classes involving data organization, mental and mathematical model calibration, and prediction validation.
2- Concerning the time used for theoretical classes, field and lab practice.	-Did you think it was appropriate? -Could it have been another proportion of time for each part?
3- About the difference of previous knowledge of the participants in the area.	-Do you think it was limiting in the course?
4- About the multidisciplinary characteristic of the course.	-Do you think it is important? And/or was it limiting the course?
5- About automating processes and/or making parts of the survey more quantitative and less subjective.	-The use of pedometrics tools did help to define the mapping units? -Do you think it is valid?
6- About the Pedagogic Project Approach.	-Was it different? -Did you find it valid?
7- Your evaluation of the course.	-What can be improved?

Soil mapping project and the study site

The study was carried out in a small training area, 10 hectares, located at the Agroecological Farm (Agroecological Integrated System - SIPA, Portuguese) in Seropédica, Rio de Janeiro State. The SIPA is an experimental farming resulting from the association between the Brazilian Company of Agricultural Research (Embrapa Agrobiologia, Portuguese), Agricultural Research Company of the state of Rio de Janeiro (Pesagro-Rio, Portuguese) and UFRRJ, and it is located at the coordinates 22° 45' 0" S and 43° 40' 30" W (Figure 1). The training area is used for grazing, and its relief is predominantly smooth, ranging up to wavy, with altitudes varying between 18 and 52 m above sea level.

According to Alvares et al. (2013), the study site's climate is classified as AW (Tropical with dry winter - Köppen classification system). The SIPA has an average temperature of 23.5 °C and an average annual rainfall of 1354 mm. The highest precipitation values are verified in the months of November to January, and the lowest in the period between May and August (Oliveira-Júnior et al., 2014).

Environmental covariates

The environmental covariates available in the study site (detail level) were those available from the digital elevation model (DEM), which represents the relief on soil formation equation, and from an orbital image (RapidEye from March 2014) that presents the organisms factor. The DEM, with a spatial resolution of 2 m, was generated from the contour lines with 1 m equidistance and points of elevation distributed throughout the study site. The contour lines and points of elevation were elaborated from a planialtimetric survey with the aid of a DGPS - Differential Global Positioning System.

To reconcile the spatial resolution of the image (5 m orthorectified) with that of the DEM, the image was interpolated to a resolution of 2 m using the value of the neighboring.

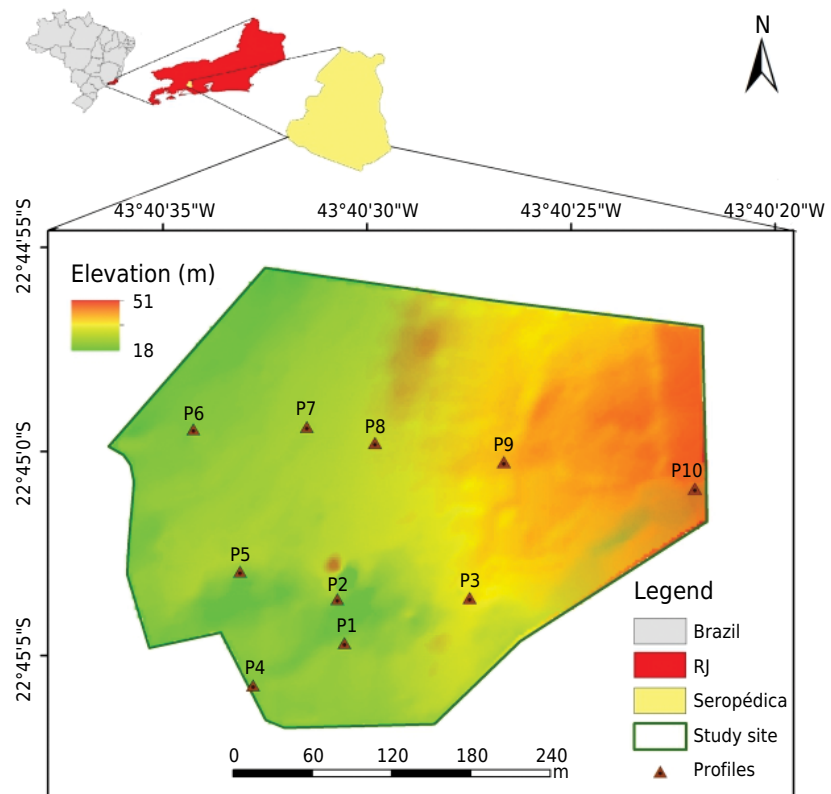


Figure 1. Study site and soil profiles distribution (Triangles).

The image was atmospherically corrected using the 6S (Second Simulation of Satellite Signal in the Solar Spectrum) model (Vermote et al., 1997) to convert radiance at the satellite level into a physical variable, surface reflectance, and remove the atmosphere effect (Antunes et al., 2014). After the atmospheric correction, besides the five spectral bands available, there were also obtained the normalized difference vegetation index (NDVI) and the soil adjusted vegetation index (SAVI) by using arithmetic operations (Equations 1 and 2, respectively).

$$NDVI = \frac{\rho_{nir} - \rho_{red}}{\rho_{nir} + \rho_{red}} \quad \text{Eq. 1}$$

$$SAVI = \frac{(1 + L)(\rho_{nir} - \rho_{red})}{\rho_{nir} + \rho_{red} + L} \quad \text{Eq. 2}$$

In which ρ_{nir} is the radiant flux reflected in the near-infrared, represented by the band 5 of the RapidEye sensor; ρ_{red} is the radiant flux reflected in the red, represented by the band 3. The constant L can present values from 0 to 1, varying according to its own biomass; the reference values of L are (Huete, 1988):

$L = 1$ (for low vegetation densities)

$L = 0.5$ (for medium vegetation densities)

$L = 0.25$ (for high vegetation densities)

Considering both the DEM and the RapidEye images, 20 covariates were used (13 from DEM and 7 from the orbital image - 5 spectral bands, NDVI, and SAVI). Details from de environmental covariates are in table 3.

Fieldwork, soil classification, and mapping units definition

Along the training study site, ten trenches were opened at different landscape points (Figure 1). The spatial distribution of the trenches was defined using the technique of free walk in toposequence. It was done to cover the main variability of the relief along the study site. The labels and positions of the trenches are presented, as follow: one pit at the top (P10), two in the middle third position (P3 and P9), one in the lower third (P8), three in the foothills (P2, P5, and P7) and three in the lowland area (P1, P4, and P6). Both the soil profiles and the soil samples were described and collected according to Santos et al. (2015). In each soil horizon, disturbed soil samples were collected to perform physical (particle size and particle density) and chemical analyses (sodium, potassium, calcium, magnesium, H + Al, aluminum, and total organic carbon). Soil properties were measured according to Teixeira et al. (2017). Besides, undisturbed samples were collected to determine soil bulk density, total porosity, and macro and micropores.

The profiles were then classified according to the Brazilian Soil Classification System (Santos et al., 2018). These soil profiles were used as modal profiles for the pedologists training. Once the soil was taxonomically classified, the group of tutors and apprentices begun to define the mapping units to be used in the soil map. The definition of soil mapping units is a key point in the definition of the soil map and its validation process. To reduce the subjectivity and turn the process of defining mapping units into a more interactive and didactic activity, it was performed using the experts' knowledge (tutors) and numerical methods. The Algorithm for Quantitative Pedology (APQ) was used to evaluate the dissimilarity between soil profiles, according to Beaudette et al. (2013). The 10 soil profiles were submitted to the dissimilarity analysis using both the available data of the landscape and the soil attributes (soil depth and sand, silt and clay content). This exercise allowed the tutors to show the relationship between soil and landscape and how this is associated with a given classification system, in this case, SiBCS (Brazilian Soil Classification System).

Table 3. Environmental covariates, soil formation factor that represents their sources, resolution, and definition

Formation factor	Covariate	Source	unit	Definition
Organism (O)	Bands (1,2,3,4 and 5)	RapidEye (2014)	ρ	Bands in the spectrum of 440 - 510 nm (Blue), 520 - 590 nm (Green), 630 - 685 nm (Red), 690 - 730 nm (Red Edge), 760 - 850 nm (Near IR)
	NDVI	RapidEye (2014)	dimensionless	$NDVI = (NIR - Red) / (NIR + Red)$
	SAVI	RapidEye (2014)	dimensionless	$SAVI = (1 + 0.5) (NIR - Red) / (NIR + Red + 0.5)$
Relief (R)	DEM	LASA	m	The digital elevation model of the area represents the terrain's surface made by interpolation of contour lines and elevation points.
	Slope	DEM-LASA	%	Gradient or rate of change of elevation between neighboring cells
	Aspect	DEM-LASA	degrees	Represents exposure faces, values in degrees (0 to 360°)
	Northernness	DEM-LASA	degrees	It indicates the direction of the slope relative to the northern. Northernness = $abs(180^\circ - Aspect)$
	Plan_curv	DEM-LASA	m^{-1}	The shape of the hillside on the horizontal plane (concave, rectilinear, or convex).
	Prof_curv	DEM-LASA	m^{-1}	The shape of the hillside on the vertical plane (concave, rectilinear, or convex).
	Convergence	DEM-LASA	%	The general shape of the hillside in all directions (concave, rectilinear, or convex)
	Cat_area	DEM-LASA	m^2	Related to the volume of flooding that reaches a certain cell
	TWI	DEM-LASA	dimensionless	Describes a tendency for a cell to accumulate water
	LS_factor	DEM-LASA	dimensionless	Attribute equivalent to the topographic factor of the Revised Universal Soil Loss Equation (RUSLE)
	RSP	DEM-LASA	dimensionless	Represents relative slope position based on the base channel network
CHND	DEM-LASA	m	Altitude above the channel network (CHNB- original elevation)	
CHNB	DEM-LASA	m	Interpolation of a channel network base level elevation	

NDVI: normalized difference vegetation index; SAVI: soil-adjusted vegetation index; DEM: digital elevation model; Plan_curv: plan curvature; Prof_curv: profile curvature; Convergence: convergence index; Cat_area: catchment area; TWI: topographic wetness index; LS_factor: LS factor; RSP: relative slope position; CHND: channel network distance; CHNB: channel network base level; ρ : surface reflectance.

Covariates selection and creation of environmental strata of the landscape

As presented along item 2.2, 20 environmental covariates were considered to be potentially usable in soil mapping, being 13 derivate from DEM and 7 from orbital image. However, the use of these covariates to develop soil prediction models depends on their effectiveness in explaining the soil variation along the landscape. Besides, some covariates can be redundant, that is, more than one of them represents similar relation with soil variation in the landscape and are highly correlated with each other.

Considering these aspects and facilitating the choice of the environmental covariates to train the predictive model, a principal component analysis (PCA) was performed to verify the best explanations for soil variation in the landscape. A correlation analysis was then performed between the previously selected covariates (those more explainable covariates) to eliminate redundant covariates and simplify the models.

The same analysis was used to create environmental strata of the landscape, which was useful not only for the apprentices to develop their mental models of soil mapping but also to establish the sampling designs to validate both the conventional and the digital soil maps (item 2.7).

Conventional soil mapping approach

The conventional soil mapping (CSM) was produced based on a soil survey and soil-landscape relationship. A lot of soil survey aspects were discussed during the training process, in theory and practice, always respecting the limitation (feeling and time) of each participant. After theoretical and practical classes in the field and laboratory, the apprentices went to the part of conducting a detailed soil survey and mapping project. The first step (office work) consisted of gathering information about the study site that could help the understanding of the soil-landscape relationship of the area. The apprentices could evaluate each map of the relief (terrain attributes) and the land-use history through satellite images and indexes and ten complete soil profiles along the landscape. After that, the soil profiles were collected, analyzed, and classified. To define the soil MUs a quantitative method using dissimilarity analysis was carried out and discussed during the training process. To delimited the boundaries of the MUs, each apprentice has used your level of knowledge/experience in soil science and the information obtained during the training. Afterward, some boundaries of MUs were verified and adjusted during the field survey depending on the apprentice feeling.

In summary, the criteria used for the apprentices to build the mental model were: soil taxon, at first until the second level, and soil attributes of each profile. After discussions, the third and fourth categorical levels of SiBCS and the specific attributes and characteristics of these levels, such as E horizon thickness, texture, soil color, waxiness, and others, were also considered for the definition of the MUs. In addition, characteristics such as position in the landscape, rock outcrop, presence of gravel pebbles, drainage, and others. The boundaries of the MUs were generated basically from the analysis of elevation, slope, contour lines, image bands, and overlapping theses maps.

The process of constructing the conceptual model of pedogenesis in the study site was guided by training process and prior knowledge of soil formation conditions of the study site; as all apprentices have studied at UFRRJ, they had some prior contact with landscapes presenting some similarities with the training study site.

Digital soil mapping approach

The use of DSM allowed the apprentices to get in contact with algorithms commonly used to predict soil MUs. Besides, it was a good exercise to evaluate the advantages, difficulties, and limitations of each technique and compare DSM with conventional maps. The tutors decided to use the two most commonly used techniques for predicting soil MUs: Multinomial Logistic Regression and Random Forest.

a) Multinomial Logistic Regression

Multinomial Logistic Regression (MLR) is a technique used exclusively for predicting categorical variables such as soil types. It is a parametric method that allows predicting the probability of occurrence of a response variable, considering the values of a series of independent variables (environmental covariates) that can be qualitative or quantitative. The logistic function is represented by:

$$\text{Logit}_j = \log \frac{\pi_j(s)}{\pi_k(s)} = \alpha_j \beta_j' x, j = 1, \dots, k - 1 \quad \text{Eq. 3}$$

In which logit_j is the natural logarithm of the ratio between the probability $\pi_j(s)$ of a given soil observation belonging to the j^{th} category, conditional on the values of the covariates

contained in the vector $x(s)$, and the probability $\pi_k(s)$ of that soil observation belongs to category k taken as reference (Agresti, 2002). The logit model intercepts adjusted for the j^{th} category is given by α_j , while β_j is a vector with the coefficients adjusted for each of the predictor variables whose values are contained in the x .

The logistic model belongs to the family of generalized MLR and was used to model the relationships between the different soil types (map units) as categorical dependent variables and the environmental covariates as independent variables. A model with previous covariates selected was fitted to predict the spatial distribution of soil types.

b) Random Forest

Random Forest (RF) is an algorithm developed by (Breiman, 2001). It is based on regression and classification trees, where it built various regression or classification trees with bootstrap sampling on the input variables and internal validation (Grimm et al., 2008; Yang et al., 2016). RF depends only on three user-defined parameters: the number of trees in the forest, the minimum number of data points on each terminal node (nodesize), and the number of variables used to produce each tree (mtry). The values indicated by the literature are $n_{\text{tree}} = 500$, $n_{\text{nodesize}} = 5$, and $m_{\text{try}} = \text{one-third of the total number of predictors}$ (Grimm et al., 2008; Were et al., 2015; Yang et al., 2016). As done for MLR, a model with the same previous covariates selected was fitted in RF using the parameters indicated by the literature.

Sampling design for map validation

The evaluation of the soil maps' accuracy is an important issue, and the apprentices also took part in all steps of this task. In the literature, it is already established that the probabilistic sampling is recommended for the validation of spatial predictions (Brus et al., 2011; Brus, 2019). The design of probabilistic sampling can influence the results of the map validation. According to Brus et al. (2011), five basic types of sampling design can be considered to evaluate the accuracy of the maps (Simple Random Sampling - SRS, Stratified Simple Random Sampling - SSRS, Systematic Random Sampling - SY, Cluster Random Sampling - CL, and Two-Stage random Sampling - TS). Considering the logistic support and the work capacity of the group, it was decided to test the SRS and SSRS sample designs.

Another important aspect of the sampling for map validation is the number and the type of field observations used. Before deciding on the number of field observations, an assessment was made of what kind of observation would be used to validate the map (auger holes, mini trenches or regular trenches). Although auger holes are simpler sampling and allow greater yield in the field, these observations do not allow the ideal visualization of some important soil morphological aspects used to classify soils, such as the thickness of the horizons, shape, size, and grade of soil structure, quantity and size of mottles and clay films. On the other hand, although ideal for soil evaluation, conventional pit are very laborious for opening and closing, resulting in less field yield.

Thus, it was decided to use mini trenches with dimensions of 0.60 m deep, 0.50 m wide, and 0.60 m long; once this kind of observation could combine two important aspects during the soil evaluation, that is, greater yield in the field and better morphological evaluation of the soil. The mini pit dimensions can be adapted to a specific study site to attend the most important differentiation of soil types. Considering mini trenches as field observation, the group realized that it was possible to open, manually, 60 mini pits along the study site, 30 of them for each sampling design.

a) Simple random sampling

The SRS is the sampling technique where all the elements that compose the sample universe have the same probability of being selected for the sample, that is, the same likelihood of inclusion and are completely independent of each other. It would be like making a fair draw among the individuals of the universe: In the specific case of soil science would be, for example, to select any *pedon* of a given type of soil (class). In this type of sampling, only the number of samples (n) is defined. In this study, it was used $n = 30$ (Figure 2a).

Despite being a totally probabilistic method (not biased), which is an ideal condition for validation, in this selected method, the spatial distribution of the mini trenches can be irregularly scattered along the study site. It means that there may be strong grouping of some locations sampling, in addition to the presence of large voids between sample sites. Consequently, simple random samples are not always the best option, especially when searching for spatially representative samples of a study site and/or covering the variation of terrain and landscape characteristics.

b) Stratified simple random sampling

Just as SRS, the SSRS belongs to the same probabilistic family and consists of dividing the entire population or the “object of study” into different subgroups or strata so that an individual can only be part of a single stratum or layer. The SSRS approach was tested because if the subareas (strata) are internally homogeneous and heterogeneous between strata, the use of stratified random sampling reduces the sample error and may be somewhat more efficient for validation, since you certify that there are no empty spaces and that all soil types (or soil attributes) have at least one representative individual for validation. When using SSRS, it is assumed that the environmental strata is related to different soils types, which could not have been perfectly sampled using SRS. One important aspect using this sample design is the delineation of the strata. Commonly, the map unities are used as strata to distribute the validation points (Brus et al., 2011), however as the definition of map unities was being developed concurrently, and apprentices could generate different maps, it was decided to create environmental strata that could be associated to soil variation along the study site (item 2.4).

In the specific case, the study site was subdivided into ten strata using an unsupervised classification, k-means algorithm, and previous selected covariates, which are the environmentally homogeneous areas (Figure 2b). Once the ten strata were defined, the 30 soil samples were distributed following a proportional stratification, where the number of points is proportional to the stratum area (Figure 2b).

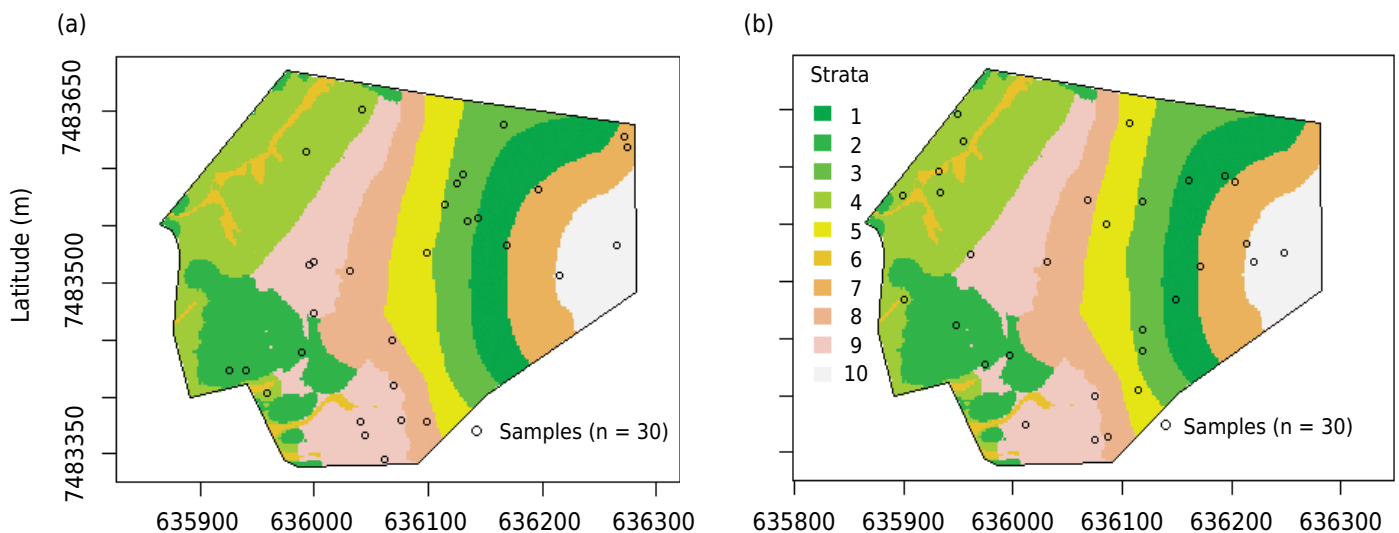


Figure 2. Sampling design applied to the study site. (a) Simple Random Sampling – SRS; (b) Stratified Simple Random Sampling – SSRS.

Quality measures for categorical soil maps

The performance of each soil map predicted by the three different approaches (MLR, RF, and 4-CSM) was evaluated using the same independent validation dataset, being a probabilistic sample selected by SRS and SSRS. For assessing the quality of the predicted soil maps, the following quality measures were used: Overall Accuracy, Kappa coefficient of agreement, User's Accuracy, and Producer's Accuracy. All of them were based on the confusion matrix (Brus et al., 2011) and are calculated as the proportion of the samples or soil types that were correctly predicted over the total number of validation locations (reference field data).

The overall accuracy was given by:

$$OA = \frac{\sum_{i=1}^c E_{ij}}{n} \quad \text{Eq. 4}$$

In which E is the confusion or error matrix of dimensions $c \times c$; and n is the number of samples (observations). In the literature, overall accuracy is also called overall purity, map purity, global accuracy, and general accuracy.

User's accuracy given by:

$$UA = \frac{E_u}{E_{iu}} \quad \text{Eq. 5}$$

In which E_{iu} denotes the number of points mapped as the mapping unit u that is, the sum of the rows in the confusion matrix; and E_u are the classes correctly classified in that unit u , the main diagonal of the confusion matrix. The complement of UA ($1 - UA$) is referred to as the error of commission (inclusion), that is, the error ruled by the inclusion of pixels from other classes in the class in question. In the literature, other synonyms are also used for User's Accuracy, such as map unit purity (Brus et al., 2011), which is about predicted classes (map).

Producer's accuracy given by:

$$PA = \frac{E_u}{E_{ju}} \quad \text{Eq. 6}$$

In which E_{ju} denotes the number of points mapped as the mapping unit u , that is, the sum of the columns in the confusion matrix, and E_u are the classes correctly classified in that unit u , main diagonal of the confusion matrix. The complement of PA ($1 - PA$) is referred to as the omission errors (exclusion), that is, when a pixel ceases to be classified correctly in that mapping unit and is incorrectly classified as another unit. In the literature, other synonyms are also used for producer's Accuracy, such as class representation on terrestrial truth (reference field data).

Kappa index was given by:

$$\hat{k} = \frac{n \sum_{i=1}^c E_{ij} - \sum_{i=1}^c E_i E_j}{n^2 - \sum_{i=1}^c E_i E_j} \quad \text{Eq. 7}$$

Where c is the number of classes on the matrix, E_{ij} values on the row i and column j , E_i total on the row i and E_j total on column j , and n the number of samples (observations).

Finally, as different soil maps of the same study site using CSM and DSM techniques were compared, the indexes WPAI (Weighted Producer Accuracy Index) and WUAI (Weighted

User Accuracy Index) were also computed (Equations 8 and 9, respectively). Both user and producer accuracy are commonly calculated for each mapping unit.

The WUAI and WPAI indices' generation aims to give a global view of the user and producer accuracy for each map. Thus, as in each map, the mapping units have different territorial expression, both indexes are weighted averages of user and producer accuracy. The weighting is done by multiplying this accuracy by the area of each mapping unit (MU), divided by the total area of the map (A). The WUAI and WPAI indices allow us to know how the types of errors are distributed (commission or omission, respectively) in each map (give an overview of these errors for a specific map). Thus, after comparing the global accuracy and the kappa index, before entering into the detailed evaluation of the types of errors per mapping unit (which is conventionally done), we used the WUAI and WPAI indices to compare the relevance of commission and omission errors on each map (4 CSM generated by apprentices and 2 DSM generated by RF and MLR). The index values range from 0 to 1, with 0 lack of accuracy and 1 maximum accuracy.

$$WPAI = \frac{\sum_j^n \frac{E_{ju}}{E_{ju}} \times a_u}{A} \quad \text{Eq. 8}$$

In which E_{ju} denotes the number of points mapped as the mapping unit u that is, the sum of the columns in the confusion matrix and E_u are the classes correctly classified in that unit u , main diagonal of the confusion matrix, a_u is the surface area of the mapped unit u and A is the total surface area of the map.

$$WUAI = \frac{\sum_i^n \frac{E_{iu}}{E_{iu}} \times a_u}{A} \quad \text{Eq. 9}$$

In which E_{iu} denotes the number of points mapped as the mapping unit u that is, the sum of the rows in the confusion matrix and E_u are the classes correctly classified in that unit u , main diagonal of the confusion matrix, a_u is the surface area of the mapped unit u , and A is the total surface area of the map. For simplicity, we will use the terms overall accuracy (for global accuracy) and user and producer accuracy (for commission and omission errors, respectively).

Software used

The Spring 5.2.5 (Câmara et al., 1996) and 6S software (Vermote et al., 1997; Antunes et al., 2014) were used for atmospheric correction of the RapidEye satellite image.

The R software (R Core Team, 2019) was used for the covariates preparation and statistical modeling. The following packages were used: *raster*, *rgdal*, *maptools*, and *RSAGA* for data management, preparation, and visualization (Brenning et al., 2018; Bivand and Lewin-Koh, 2019; Bivand et al., 2019; Hijmans, 2019); *randomForest* for RF (Liaw and Wiener, 2002); and *nnet* (Venables and Ripley, 2002) for MLR modeling; *factoextra* (Kassambara and Mundt, 2017) for principal component analysis (PCA); *cluster* (Maechler et al., 2019) for cluster analysis; *aqp* (Beaudette et al., 2013) for dissimilarity analysis and profile plots; and *sampling* (Tillé and Matei, 2016) for soil sampling selection.

RESULTS

Soil landscape relationship and map unity definition

Figure 3 presented the ten soil profiles and their respective position according to the toposequence. As shown, the soil color has a strict relationship with the topography, which, in turn, is directly related to soil drainage and soil moisture characteristics. The profiles of the highest part of the landscape (P10 - *Argissolo Vermelho Distrófico nitossólico*; and P9 and P3 - *Argissolo Vermelho Eutrófico nitossólico*) have a reddish color, passing

through yellowish (P8 - *Argissolo Amarelo Eutrófico típico*) and light yellow or grey with or without mottled (P1 - *Cambissolo Háplico Ta Distrófico típico*; P4 and P6 - *Planossolo Háplico Distrófico gleissólico*; P2 and P7 - *Planossolo Háplico Distrófico arênico*; and P5 - *Planossolo Háplico Distrófico espessarênico*).

The results of the dissimilarity analysis using AQP to create five MUs is presented in figure 4. Further information about each MU is presented in table 4. From the five MUs, only one is composed, that is, more than one soil class integrate this MU (MU1, Figure 4).

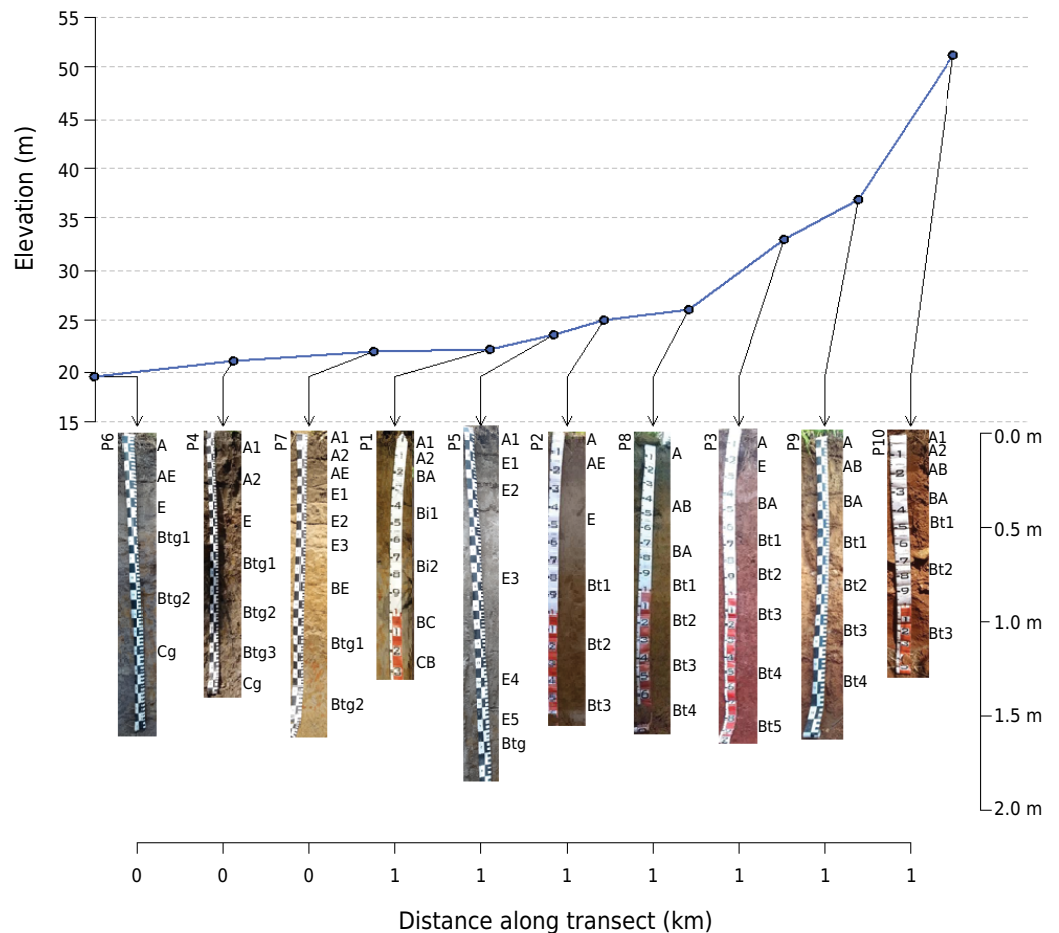


Figure 3. Soil profiles and position on the landscape.

Table 4. Soil map units and landscape properties

MU	Taxonomic unit		n	Elev	Slope
	IUSS-WRB	SiBCS			
MU1 - SXd + GXvd	Gleyic Planosols + Gleyic Cambisols	<i>Planossolo Háplico Distrófico gleissólico</i> (P4 and P6) + <i>Cambissolo Háplico Ta Distrófico típico</i> (P1)	3	20.9	2.7
MU2 - SXd	Stagnic Planosols	<i>Planossolo Háplico Distrófico arênico</i> (P2 and P7) + <i>Planossolo Háplico Distrófico espessarênico</i> (P5)	3	23.5	5.1
MU3 - PAd	Nitic Lixisols	<i>Argissolo Amarelo Eutrófico típico</i> (P8)	1	26.1	10.1
MU4 - PVe	Nitic Lixisols	<i>Argissolo Vermelho Eutrófico nitossólico</i> (P3 and P9)	2	35.0	15.2
MU5 - PVd	Nitic Acrisols	<i>Argissolo Vermelho Distrófico nitossólico</i> (P10)	1	51.3	7.8

MU: Map units; IUSS-WRB: World Reference Base updated in 2015 (IUSS Working Group WRB, 2015); SiBCS: Brazilian Soil Classification System (Santos et al., 2018); n: number of complete soil profile data; Elev: elevation (m); Slope (%).

The soils with expressive sandy layer were grouped in MU2, with all profiles classified as *Planossolo Háplico* (P1, P4, and P6). These soils are associated with the lower third of the landscape. The MU1, however, despite also having soil classified as *Planossolos Háplicos*, they are closest to the order of *Gleissolos* and/or *Cambissolo Háplico*, where the hidromorphism is prominently defining the soil properties. These soils (P1, P4, and P6) are in the lower part of the landscape. The MU3 (P8) represents the soil pattern between the middle third and lower third of the landscape. The soil is well-drained, with a yellowish color and clayey subsurface horizon (*Argissolo Amarelo*). The color is strongly related to the soil moisture and geology; in this case, there is a predominance of goethite as mineral. The soils of MU4 (P3 and P9), despite belonging to the same taxonomic class (*Argissolo Vermelho*) as the soil of MU5 (P10 - top of the landscape), differ in depth, presence of gravel, bases saturation. Both on MU4 and MU3, the presence of hematite is remarkable, resulting in redder colors to the soil which are directly related to the better drainage condition.

Selecting covariates to simplify soil mapping

Irrespective of the method used to generate soil maps (DSM or CSM), one of the great challenges of soil mapping is not only how to select good environmental covariates among many available, but also how to make the soil mapping a simple way creating a parsimonious model, especially for learners.

Analyzing the PCA, we see that component 1 (dimension 1) explains 34.8 % of the data variation, and two explains 27 %, together 61.8 % of the data (Figure 5). Component 1 is strongly related to the bands of the sensor 1, 2, 3, and 5 and some derivate indices, NDVI and SAVI (on the horizontal). Component 2 (dimension 2) features terrain attributes. Basically DEM, rsp, chnd twi, and cat_area.

Although these components explain most of the area's environmental variation, some of them are redundant since they are highly correlated (with a correlation greater than 90 % EX: SAVI and NDVI). To simplify the model that will be used in the stratification of the area it was used only environmental variables with low correlation (less than 0.9; Figure 6). Since DEM has high correlation with chnd (greater than 0.9) (Figure 6), it was

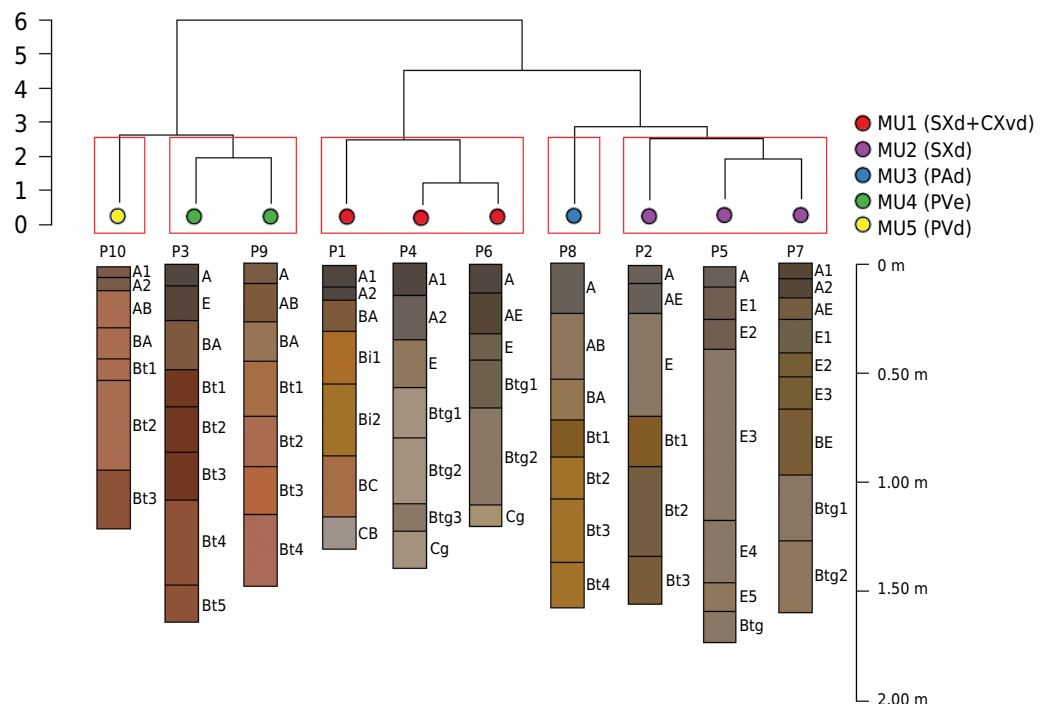


Figure 4. The dissimilarity of soil profiles based on the soil key properties and landscape relationship.

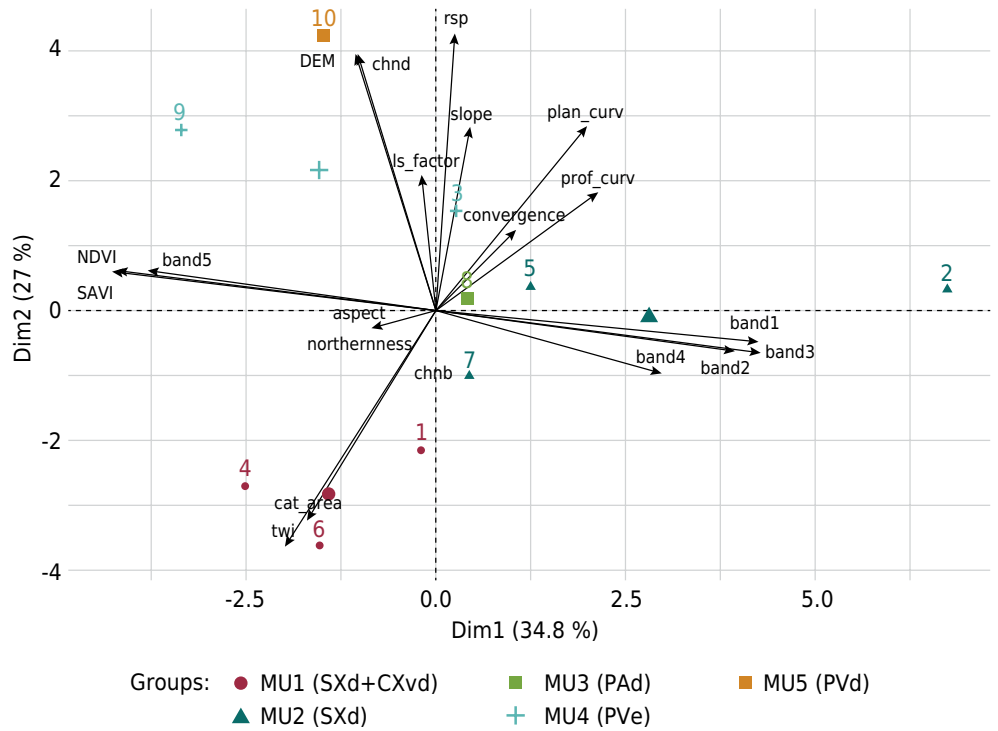


Figure 5. Plot of loading corresponding to the first two principal components, the environmental covariates and soil mapping units related. The number corresponds to the profile number, as defined on the soil survey and the symbol corresponds the MU.

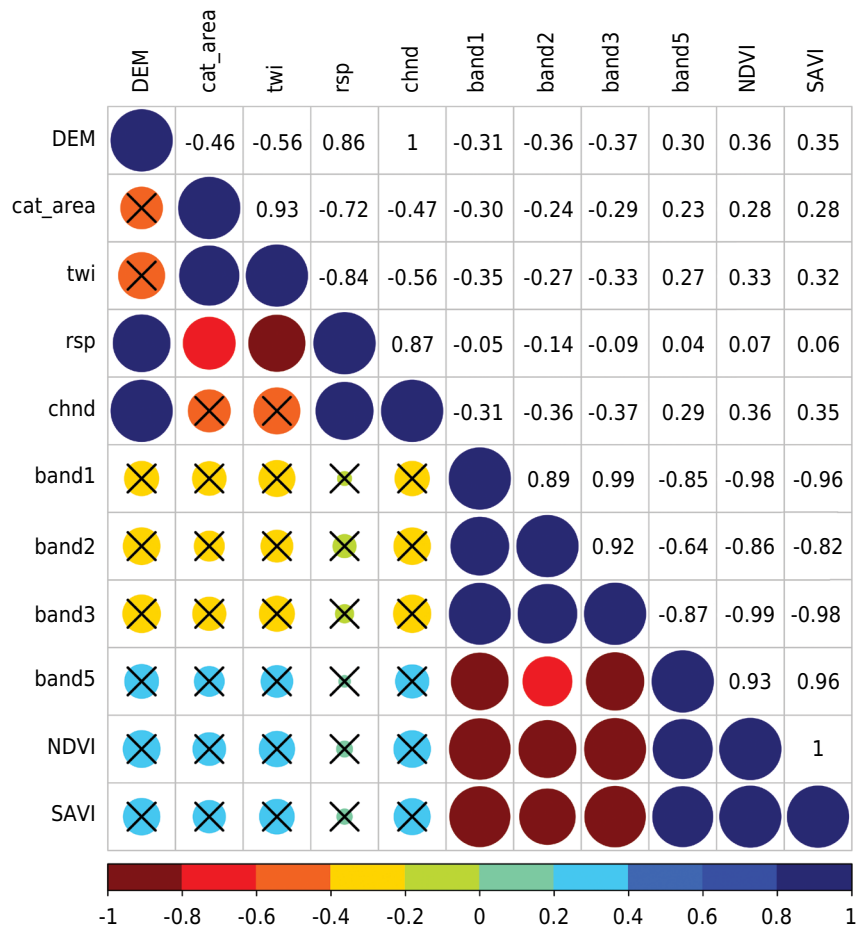


Figure 6. Correlation analysis of the environmental covariates with a higher score on the PCA analysis.

decided to maintain DEM because its use in DSM is more common. As *cat_area* has high correlation with *twi*, it was chosen to maintain the *twi* that has more common use in the DSM literature.

Since bands 1, 3, and 5 have a high relation between themselves and/or NDVI and SAVI (greater than 0.9) (Figure 6), it was decided to use only SAVI, which already uses bands 3 and 5 of the RapidEye sensor, and has a correction of the soil compared to NDVI, an advantage over NDVI use. Band 2, despite having a high correlation with SAVI, it was less than 0.9. In summary: the covariates of the relief selected were: DEM, *rsp*, and *twi*. From the satellite image (organism soil factor) band 2 and SAVI.

Soil maps and their performances

The maps generated by apprentices (conventional maps) and digital (RF and MLR) are shown in figure 7. In general, there is a clear difference in the patterns of the maps when comparing the conventional soil maps (Figures 7a and 7d) with the digital ones (Figures 7e and 7f). In the first case (conventional soil maps), the mapping units' distribution has a pattern more similar to the contour lines of the study area.

This pattern can be explained by the strong relationship between soil distribution and toposequence, which the apprentices observed during the fieldwork. Another experience that probably influenced the apprentices' soil map patterns was the development of the stratification of the area through PCA analyses. This is commonly referred to among pedologists by the term "catching the pattern". That is, when the pedologist, in the learning process, connects theoretical knowledge with the distribution of soils in the landscape and manages to build a mental model of distribution and prediction of soil types in the landscape (interpretation of soils in the landscape). This cognitive process is subjective and quite different among pedologists since cognition includes functions, such as learning, attention, memory, language, reasoning, decision making, etc., which

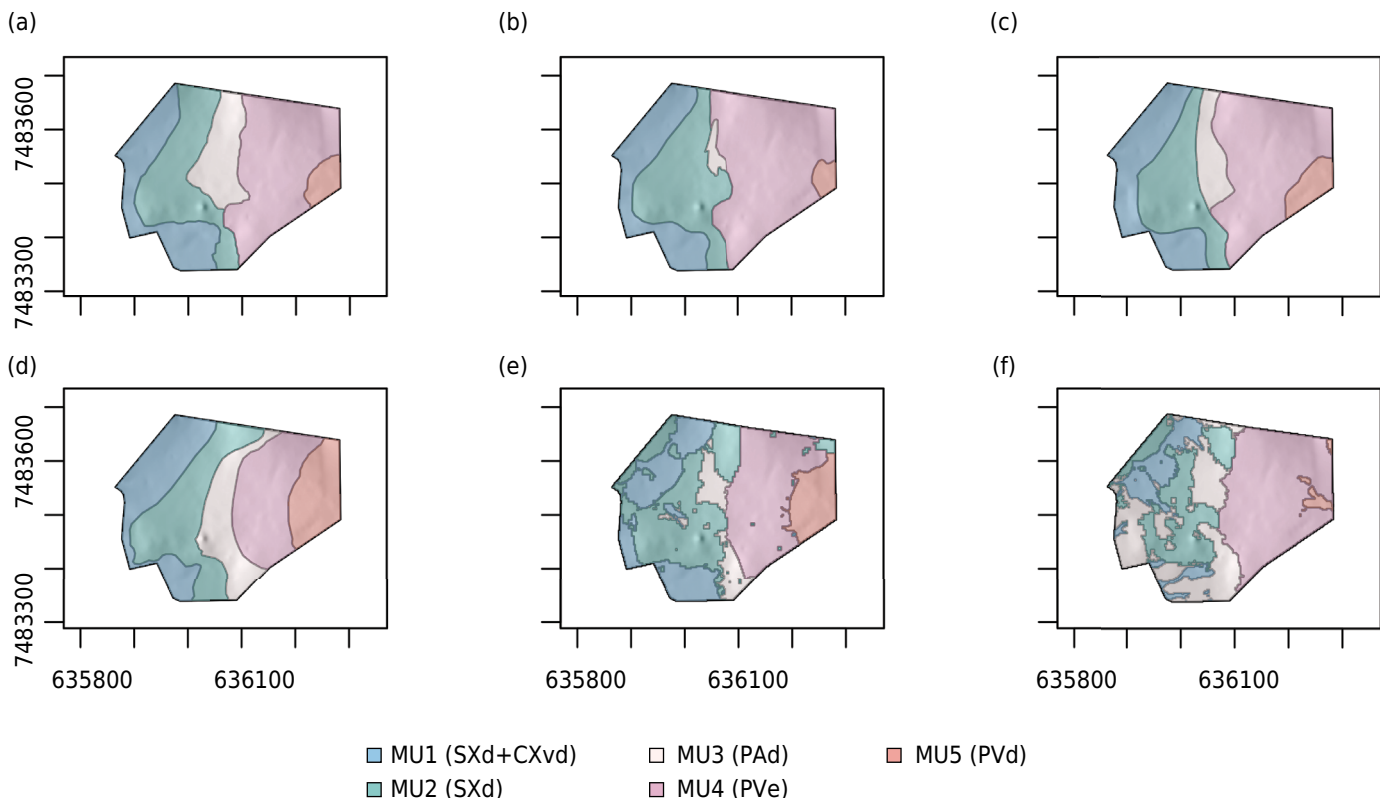


Figure 7. Conventional soil maps: A: apprentice A; B: apprentice B; C: apprentice C; D: apprentice D; E: random Forest; F: Multiple Logistic Regression.

are part of our intellectual development and experiences. Thus, observing the maps of the apprentices, it is noted that the MUs are more strongly inferred in the form of “zones” of soil types. These zones, due to the diversity in the cognitive process between the apprentices, present different territorial expressions and formats (limits), especially for MU2 and MU3 from apprentice B comparing to the others and MU4 and MU5 from apprentice D comparing to the others (Figures 7 and 8).

On the other hand, DSM algorithms (RF and MLR) predict soil types through scanning pixel by pixel. This process ends up making the map look “salt and pepper”, that is, the classification by pixel treats each individualized pixel of its neighbors and will have a prediction probably different from the surroundings. This effect is more explicit, the greater the spatial resolution of the map (greater detail). Commonly, this “salt and pepper” happens also in land use mapping, and the effect is mitigated through the application of majority filters, which are usually implemented in GIS tools.

Figure 8a presents the surface area of each map unity for CSM and DSM and the number of validation samples per MU for SRS (Figure 8b) and SSRS design (Figure 8c).

As expected, there are significant differences in the areas of each MU for the different maps. However, the smaller territorial expression of MU3 and 5 is common in different maps (except MU3 for MLR). In both cases, only one modal profile was used to develop the prediction model. It is also noteworthy that the area of MU1 is much smaller on the map generated by the MLR algorithm than on the other maps (around half of that shown on the other maps). These differences in territorial expression are reflected in the results of the validation, especially when evaluating different sample designs (Figures 8b and 8c). In some cases, as an area is so small, no field validation point was selected (reference data) to evaluate the results, for example, MU5 in apprentice B for SRS and SSRS (Figures 8b and 8c) and MLR for SRS (Figure 8b).

The main difference between the two sets of samples is that the SSRS set tended to allocate a similar number of samples between MU1 and MU2 for the CSMs and allocate one sample in MU5 for the MLR map that was left with no sample in the SRS. As expected, the areas with the lowest number of samples were the smaller areas MU3 and MU5 in both the SRS and SSRS.

In addition to the differences in the visualization pattern of the maps, the accuracy of the referred maps are presented through the conventional metrics (Kappa and overall accuracy) and the weighted producer accuracy index (WPAI) and weighted user accuracy index (WUAI) (Figures 9a, 9b, 9c, and 9d). Also, to facilitate the analysis, in figures 9a, 9c, and 9d are shown the RE. Figures 9a and 9c refer to the quality measures of the maps (WPAI, WUAI, overall accuracy, and Kappa index) and the RE, using validation based on SRS design, while figures 9b and 9d refers to the same indexes using validation based on SSRS design.

Comparing the measurements of agreement indices, it is possible to notice the great difference that the same map presents depending on the sample design adopted for validation. If the accuracy of the maps is evaluated based on SRS, considering the Kappa index classification, the quality of the maps has the following order: CSM-D > DSMRF > CSM-C > CSM-A > CSM-B > DSM-MLR (Figure 9a) while RE follows the sequence A > D = C > B. In this case, the conventional map generated by the apprentice D (RE = 0.5) and the digital algorithm RF are equivalent, and a substantial agreement between what was predicted and what was observed in the field is observed (Kappa index between 0.61 and 0.8). The maps generated by apprentices A (RE = 0.7) and C (RE = 0.5) are classified as of moderate agreement (Kappa index between 0.41-0.6), while the maps generated by apprentices B (RE = 0.3) and the MLR technique are of fair agreement (Kappa index 0.21-0.40). Analyzing the results based on SRS design, it is possible to affirm that the map generated by apprentice B (lowest RE) present the lowest values of both the OA and the Kappa index

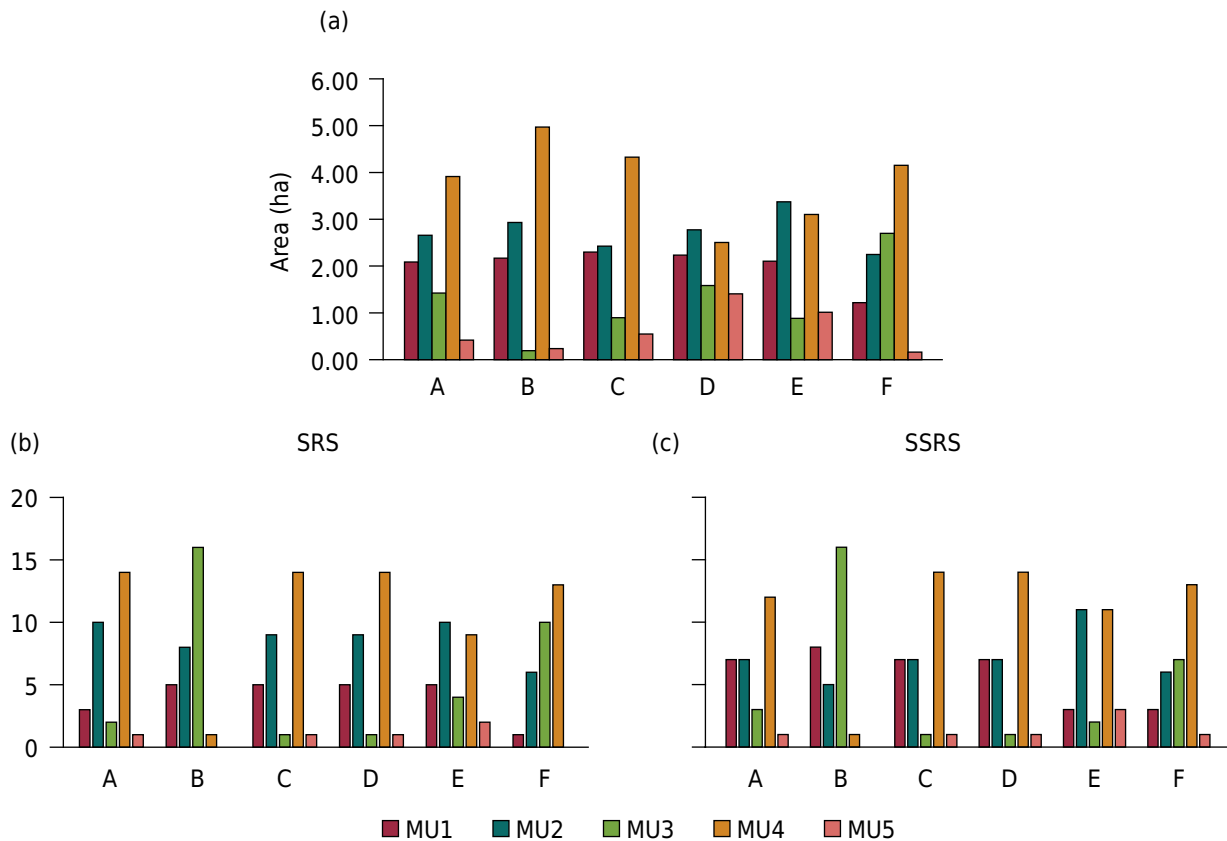


Figure 8. Map unity area of each soil map and number of validation samples per MU. A: apprentice A; B: apprentice B; C: apprentice C; D: apprentice D; E: Random Forest; F: Multiple Logistic Regression.

of the CSM (Figure 9a), which was not repeated in the SSRS sampling (Figure 9b). Using the same sample design (SRS), the map generated by apprentice D (with RE = 0.5) was the best, and the map generated by apprentice A (RE = 0.7) did not significantly surpass the quality of the map generated by apprentice C (RE = 0.5). In this case, the master degree in soil science (apprentice A) did not result in a better soil quality map.

On the other hand, the interpretation changes if the same maps' quality is analyzed based on a SSRS design (Figure 9b). In this case, the conventional maps generated by the apprentices A, B, C, and D show, irrespective of the difference of experience (RE), a strong agreement (Kappa index ≥ 0.70). These CSMs are substantially better than the maps generated by the RF and MLR techniques (moderate Kappa index). Although the digital map generated through the MLR algorithm presents the worst and second worst performance in SRS and SSRS, respectively, indicating a worse performance compared to the others, the same cannot be said about the maps generated through the RF algorithm and the conventional by apprentices A, B, and D. The maps generated by apprentices A and B showed significant improvement when comparing the OA and Kappa indexes through validation using SSRS sampling. For example, the OA and the kappa index of the apprentice A increased from 0.63 and 0.51 to 0.80 and 0.62 (best accuracy), respectively. Similarly, the map generated by apprentice B increased from 0.57 and 0.40 (penultimate in the validation with SRS design) to 0.73 and 0.61 (second-best accuracy in the SSRS design), respectively.

In the opposite direction, the maps generated by apprentice D and the RF algorithm showed a decrease in accuracy, especially the RF that reduced OA and the Kappa index from 0.70 and 0.61 (second-best accuracy in SRS design) to 0.57 and 0.42 (worst accuracy with SSRS design), respectively. These results demonstrate that the evaluation of the quality of a map can vary considerably depending on the sample design used during the

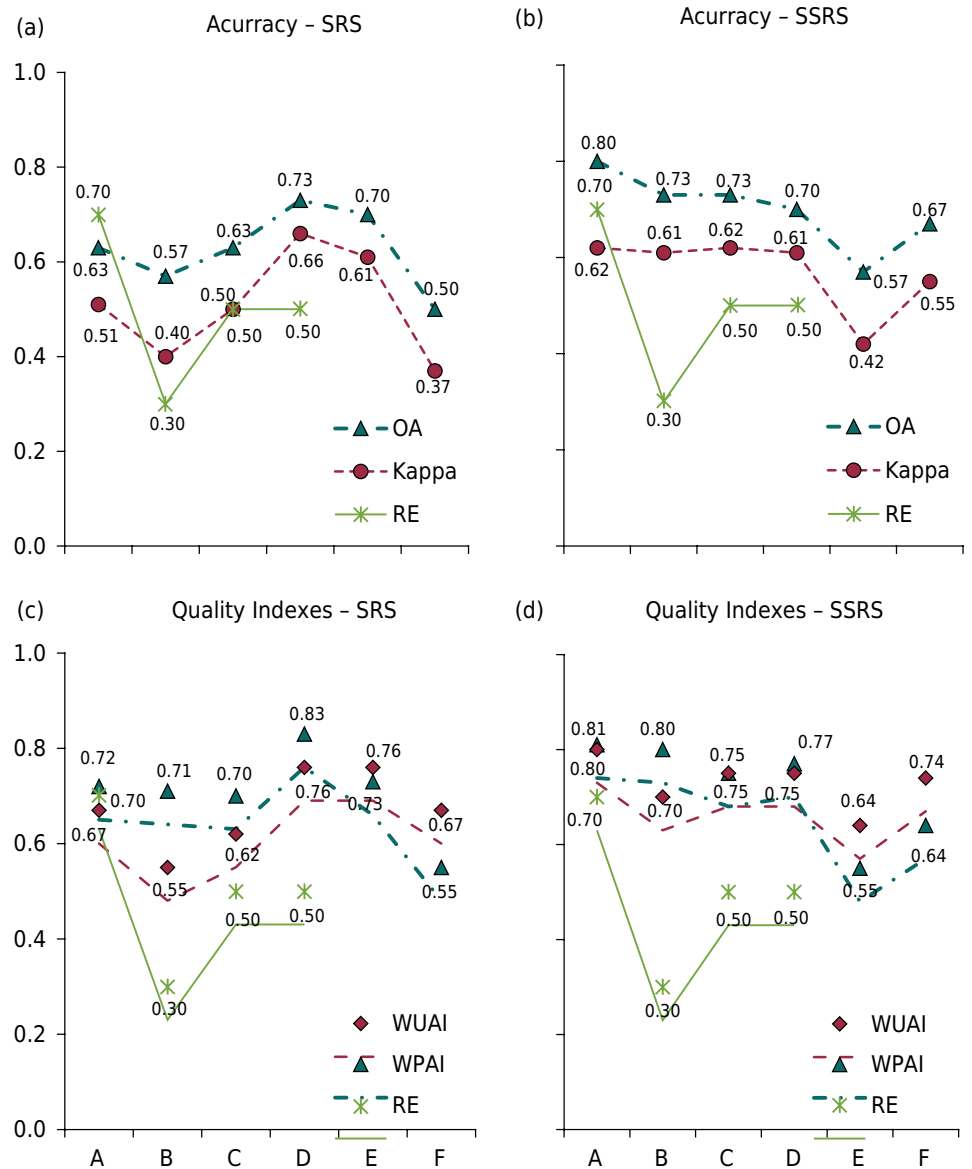


Figure 9. Quality indexes measures of the maps. A: apprentice A; B: apprentice B; C: apprentice C; D: apprentice D; E: Random Forest; F: Multiple Logistic Regression; WPAI: weighted producer accuracy index; WUAI: weighted user accuracy index; RE: Rank of experience. The producer and user accuracy values for each map (conventional and digital) in figures 9c and 9d were calculated using the weighted average of these indices, based on each mapping unit's area in relation to the total area of the maps. The values PA and UA were scaled in the same range from 0 to 1 for comparative visualization.

validation and that other aspects must be observed, such as the types of errors observed and the consequences of these errors in the final use of the map.

The producer accuracy (WPAI) is systematically higher than the user accuracy (WUAI) in conventional soil maps (greater accuracy of omission than commission), based on SRS design. On the other hand, digital soil maps presented the inverse relation ($WUAI > WPAI$), especially using the SSRS design (Figure 9d). Except the map generated by RF algorithm and the apprentice D ($RE = 0.5$), all maps presented better quality (higher WUAI and WPAI values) when the SSRS design for validation was adopted. Besides, proportionally, the sample design change from SRS to SSRS resulted in a greater relative improvement of user accuracy (WUAI) of conventional soil maps. On the other hand, the MLR algorithm's map presented a higher relative improvement of both producer's and user's accuracy (WPAI and WUAI).

By evaluating the confusion matrix in detail, it is possible to observe that the MU with less producer accuracy for the SRS validation dataset was MU2, where apprentices A, B, and C classified some profiles of this class as being MU1. Apprentice D classified most of the profiles of MU3 as being MU1 (Table 5). For DSM techniques, the lowest producer accuracy was for MU5, which was confused with MU2 in the RF as seen in the upper right corner of the map (Figure 7e) and MU1, which had some profiles classified as MU3 in the MLR (Table 5). Still, for the same validation dataset (SRS), the biggest commission errors, that is, lower user accuracy values were observed in MU1 for apprentices A, B, and D, which mainly MU2 profiles were classified as being MU1 (Table 5). As for apprentice C, the worst performance was for MU5, as some profiles of MU4 were classified as being MU5 and for the two DSM methods, the lowest accuracy of the user was seen in MU3 because several points of MU1 were classified in this MU (Table 5).

Table 5. Confusion matrix of soil mapping units using SRS validation dataset

A		Reference data (Field)					Total	User
		MU1	MU2	MU3	MU4	MU5		
Map	MU1	3	6	0	1	0	10	30
	MU2	0	4	1	0	0	5	80
	MU3	0	0	1	1	0	2	50
	MU4	0	0	0	10	0	10	100
	MU5	0	0	0	2	1	3	33.3
	Total	3	10	2	14	1	30	---
Producer		100	40	50	71.4	100	---	---
Overall accuracy								63.33
Kappa								0.51
B		Reference data (Field)					Total	User
		MU1	MU2	MU3	MU4	MU5		
Map	MU1	3	4	0	3	0	10	30
	MU2	2	3	0	0	0	5	60
	MU3	0	1	*	1	0	2	---
	MU4	0	0	0	10	0	10	100
	MU5	0	0	0	2	1	3	3.33
	Total	5	8	0	16	1	30	---
Producer		60	37.5	---	62.5	100	---	---
Overall accuracy								56.66
Kappa								0.4
C		Reference data (Field)					Total	User
		MU1	MU2	MU3	MU4	MU5		
Map	MU1	4	4	1	1	0	10	40
	MU2	1	4	0	0	0	5	80
	MU3	0	1	*	1	0	2	---
	MU4	0	0	0	10	0	10	100
	MU5	0	0	0	2	1	3	33.3
	Total	5	9	1	14	1	30	---
Producer		80	44.4	---	71.4	100	---	---
Overall accuracy								63.33
Kappa								0.5

Continue

Continuation

D		Reference data (Field)						
		MU1	MU2	MU3	MU4	MU5	Total	User
Map	MU1	3	4	3	0	0	10	30
	MU2	0	5	0	0	0	5	100
	MU3	0	0	2	0	0	2	100
	MU4	0	0	0	9	1	10	90
	MU5	0	0	0	0	3	3	100
	Total	3	9	5	9	4	30	---
Producer		100	55.6	40	100	75	---	---

Overall accuracy

73.33

Kappa

0.66

E		Reference data (Field)						
		MU1	MU2	MU3	MU4	MU5	Total	User
Map	MU1	5	0	0	0	0	5	100
	MU2	2	5	1	0	2	10	50
	MU3	3	0	1	0	0	4	25
	MU4	0	0	0	9	0	9	100
	MU5	0	0	0	1	1	2	50
	Total	10	5	2	10	3	30	---
Producer		50	100	50	90	33.3	---	---

Overall accuracy

70

Kappa

0.61

F		Reference data (Field)						
		MU1	MU2	MU3	MU4	MU5	Total	User
Map	MU1	1	0	0	0	0	1	100
	MU2	2	3	1	0	0	6	60
	MU3	7	2	1	0	0	10	10
	MU4	0	0	0	10	3	13	76.9
	MU5	0	0	0	0	*	0	---
	Total	10	2	2	10	3	30	---
Producer		10	50	50	100	---	---	---

Overall accuracy

50

Kappa

0.37

Producers: Producers accuracy (%); User: Users accuracy (%); Overall accuracy (%); A: apprentice A; B: apprentice B; C: apprentice C; D: apprentice D; E: Random Forest; F: Multiple Logistic Regression. * There is no reference class in the validation dataset.

When evaluating the SSRS sample dataset, the same standard was observed by the producer accuracy for apprentices A, B, and C in which the lowest producer accuracy was MU2, as some profiles of this class were classified mainly as being MU1 (Table 6). For apprentice D, the lowest producer accuracy was from MU5, which were predominantly classified as MU4. In this sample dataset, both the RF and the RLM showed a greater omission error in MU1 in which a good part of the profiles was classified as being from units 2 and 3 (Table 6).

When it comes to user accuracy for the same sample set, SSRS, there was a greater disagreement between apprentices and DSM algorithms. For apprentice A, the lowest accuracy was for MU1, which had part of the profiles of MU2 classified in this unit, while

apprentice B the worst result was for MU2, since it classified mainly profiles of MU1 in MU2. Apprentice C classified MU2 and MU4 profiles in MU3, which led to the worst performance of this class for this apprentice in terms of commission errors. Apprentice D, although like A, presented less user accuracy for MU1, different from it, classified mainly profiles of MU4 in MU1 (Table 6). In that case, the DSM methods differed among themselves. The RF showed less accuracy of MU5, since most of the profiles of MU4 were classified in this class. The RLM, on the other hand, classified MU1 profiles in MU3, which raised the commission error of this class. In general, most apprentices were confused by errors of commission and omission in MU1 and MU2, as these occupy a similar position in the landscape and the limits of each unit are difficult to define precisely.

Apprentices challenges

Based on the answer to the questionnaire, among the six items raised in the first question, the more challenging theme was: "Theoretical class involving statistical modeling,

Table 6. Confusion matrix of soil mapping units SSRS validation dataset

A		Reference data (Field)					Total	User	
		MU1	MU2	MU3	MU4	MU5			
Map	MU1	5	4	0	0	0	9	55.6	
	MU2	2	3	0	0	0	5	60	
	MU3	0	0	3	0	0	3	100	
	MU4	0	0	0	12	0	12	100	
	MU5	0	0	0	0	1	1	100	
	Total	7	7	3	12	1	30	---	
Producer		71.74	42.9	100	100	100	---	---	
Overall accuracy									80
Kappa									0.72

B		Reference data (Field)					Total	User	
		MU1	MU2	MU3	MU4	MU5			
Map	MU1	6	1	0	2	0	9	66.7	
	MU2	2	3	0	0	0	5	60	
	MU3	0	1	*	2	0	3	---	
	MU4	0	0	0	12	0	12	100	
	MU5	0	0	0	0	1	1	100	
	Total	8	4	0	16	1	30	---	
Producer		75	60	---	75	100	---	---	
Overall accuracy									73.33
Kappa									0.61

C		Reference data (Field)					Total	User	
		MU1	MU2	MU3	MU4	MU5			
Map	MU1	5	3	0	1	0	9	55.6	
	MU2	2	3	0	0	0	5	60	
	MU3	0	1	1	1	0	3	33.3	
	MU4	0	0	0	12	0	12	100	
	MU5	0	0	0	0	1	1	100	
	Total	7	7	1	14	1	30	---	
Producer		71.4	42.9	100	85.7	100	---	---	
Overall accuracy									73.33
Kappa									0.62

Continue

Continuation

D		Reference data (Field)					Total	User
		MU1	MU2	MU3	MU4	MU5		
Map	MU1	5	1	3	0	0	9	55.6
	MU2	1	4	0	0	0	5	80
	MU3	0	0	3	0	0	3	100
	MU4	0	1	1	8	2	12	66.7
	MU5	0	0	0	0	1	1	100
	Total	6	6	7	8	3	30	---
Producer		83.3	66.7	72.9	100	33.3	---	---

Overall accuracy

70

Kappa

0.61

E		Reference data (Field)					Total	User
		MU1	MU2	MU3	MU4	MU5		
Map	MU1	2	1	0	0	0	3	66.7
	MU2	5	4	2	0	0	11	36.4
	MU3	2	0	*	0	0	2	---
	MU4	0	0	1	10	0	11	90.9
	MU5	0	0	0	2	1	3	33.3
	Total	9	5	3	12	1	30	---
Producer		22.2	80	---	83.3	100	---	---

Overall accuracy

56.67

Kappa

0.42

F		Reference data (Field)					Total	User
		MU1	MU2	MU3	MU4	MU5		
Map	MU1	2	1	0	0	0	3	66.7
	MU2	3	3	0	0	0	6	50
	MU3	4	1	2	0	0	7	28.6
	MU4	0	0	1	12	0	13	92.3
	MU5	0	0	0	0	1	1	100
	Total	9	5	3	12	1	30	---
Producer		22.2	60	66.7	100	100	---	---

Overall accuracy

66.66

Kappa

0.55

Producers: Producers accuracy (%); User: Users accuracy (%); Overall accuracy (%); A: apprentice A; B: apprentice B; C: apprentice C; D: apprentice D; E: Random, Forest; F: Multiple Logistic Regression. * There is no reference class in the validation dataset.

sampling, prediction, and accuracy assessment". This item is among the top 3 challenging (number 1, 2, and 3 of rank) for all apprentices. Followed by the item "Practical class involving data organization, mental, and mathematical model calibration and prediction validation", which is also in the top three for most of them, only in one case where it was in fourth place (number rank 4) (Figure 10). The item "Theoretical class involving digital soil mapping, GIS and environmental covariates processing" was also assessed as challenging for most apprentices (on the top 3 for three of them A, C, and D), just for one of the apprentices (apprentice B), who already had a little experience in the topic placed the item as the least challenging (Figure 10).

Apprentices who are Agriculture and Livestock technician (apprentices A with RE = 0.7 and D with RE = 0.5) or have previous experience with soil science, classified the item "Theoretical classes of introduction to pedology, soil-landscape relationship, soil survey

and classification” and “Practical classes to identify soil in the field (description and classification) and soil-landscape relationship in practice” as being the least challenging and/or the easiest to understand, whereas apprentices (B with RE = 0.3 and C with RE = 0.5), who do not have this previous experience classified the practical classes as one of the most challenging. Despite the little experience in soil science of apprentice B, RE = 0.3, this is the apprentice who had the most exposure to related computer programs from GIS processing that reflected his rank (Figure 10).

DISCUSSION

Pedology and soil mapping

Irrespective of the technique applied to generate soil maps (conventional or digital), MU’s definition is a common process. For example, if only soil classes of a given taxonomic system are used for differentiation, some authors argue that the class is a theoretical concept and therefore cannot be mapped. Also, with an emphasis on profile characteristics, many soil scientists tend to have a tunnel vision, looking only inside the soil pits without integrating them in the landscape (Lepsch, 2013). On the other hand, if aspects of the landscape are considered (perhaps too much), some authors argue that the mapping is of landscape and not soils. So, one of the very important steps in the development of soil survey is a definition and characterization of the MU, which among other things, depend on the objective and the level of detail of the soil survey. The number of trenches and the training sampling design seems that it did not limit too much the performance of the CSM, but could interfere in the ability of the DSM algorithm to capture and differentiate the soil pattern along the study site. The number of soil observation during the soil survey is always a matter of discussion, which involves financial and logistic availability, but, for the purpose of this study, the trenches and the sampling design met the proposed goal of contextualizing the application of CSM and DSM techniques.

There is a strong relationship between soil types (its attributes such as color, horizon types, depth, and presence of gravel and pebbles) and the elevation along the study

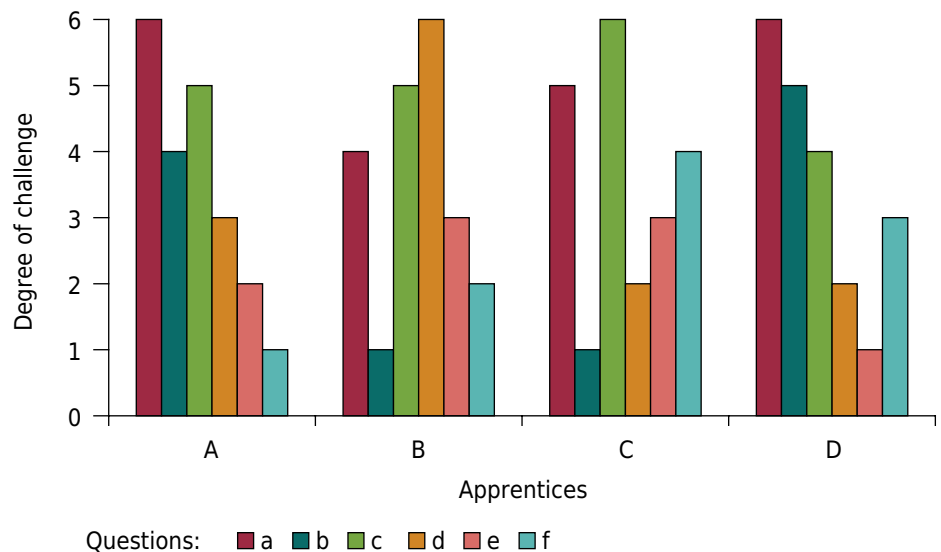


Figure 10. Degree of challenge in ascending order (1 for the highest challenge and 6 for lowest). a-Theoretical classes of introduction to pedology, soil-landscape relationship, soil survey and classification, b-Practical classes to identify soil in the field (description and classification) and soil-landscape relationship in practice, c- Field tests and laboratory analysis, d- Theoretical classes involving digital soil mapping, GIS and environmental covariates processing, e- Theoretical classes involving statistical modeling, sampling, prediction, and accuracy assessment, f- Practical classes involving data organization, mental and mathematical model calibration and prediction validation.

site. During the fieldwork, the apprentices were able to experience this variation of soil types throughout the toposequence, which was fundamental for each one to be able to develop a mental model of soil MU. This task is difficult since the model will transpose a specific knowledge (attributes of the landscape and soil type) to a spatial prediction of an entity called a MU, which can be simple (a taxonomic unit - Taxon) or compound (more than one taxonomic unit - Taxa).

In this sense, the dissimilarity analysis was useful to separate different groups of soil profiles (MU), based on key soil properties and landscape characteristics as proposed by Lepsch (2013) numerically and less subjectively as was done by (Pineiro et al., 2016, 2018). This approach is interesting and can help surpass an important gap in soil mapping, defining the mapping unit from a study site for detailed soil mapping. In addition, the pedometrics tools are very useful in the teaching and learning process in the formation of theoretical and practical knowledge of new pedologists, demonstrating relevance and synergy of traditional pedology and pedometrics tools in the soil mapping learning-teaching process (Ma et al., 2019).

Another important aspect of mapping is the selection of environmental covariates out of the huge amount available since the advance of technology and computer process capacity have boosted the production of data. In addition to simplifying the mental or mathematical model and creating the strata for SSRS design, the idea is to select those that are directly related to soil formation factors and pedological explanation. There are several selection methods in the DSM (Costa et al., 2020), but few techniques common to both, digital and conventional. Thus, the PCA technique combined to the correlation analysis was used, which can be common to two methods and strata creation, since it assumes a close relationship between soil types and the landscape. In summary, the covariates that most explain the distribution of soil in the landscape were selected without a pre-defined model as in DSM, that is, not an automatic process of a specific model (algorithm).

Different from studies that use the components instead of the covariate, if we use the covariates, we want to preserve the pedological explanation (ten Caten et al., 2011) and the possibility of the apprentices using it in the CSM. Both, PCA (ten Caten et al., 2011; Levi and Rasmussen, 2014) and correlation analysis (Jeune et al., 2018) are techniques commonly used in DSM to select covariates seeking to optimize models and improve prediction performance, and it can also be used to simplify the mental models used by pedologists in CSM. The analysis clearly shows that PCA can perform covariates selection and select a number of important environmental variables from all the feature components. For Levi and Rasmussen (2014) data reduction using a PCA combined with a cLHS design produced a sampling design that effectively captured the variability of soil types as a function of the relative area of the published soil map. Bringing this to validation data of this study, the PCA was equally useful to define the environmental strata used in the SSRS sampling and to optimize mental and digital soil model.

Along with analyzing the CSM and DSM results and the influence of the sampling method on these results, some important aspects should be highlighted. In a general assessment, when using the SRS data set, the DSM using the RF technique was as good as the CSM, and as observed by Jeune et al. (2018), the RF algorithm performed better than MLR. On the other hand, when the SSRS data set is used, the CSM remains relatively more accurate than the DSM, and among the algorithms, the MLR showed superior performance corroborating with Zeraatpisheh et al. (2017). This indicates that the sampling strategy for validation can result in different selected DSM models and different results between CSM and DSM. The different algorithms developed can present different performances between study areas and between mapping units within the same study area. That is, only by testing and comparing the best technique or combination of techniques, the best accuracy will be found.

Concerning the apprentices' performance, it is natural to believe that previous knowledge, especially when it comes to the basic concepts of pedology and the soil-landscape relationship, may be favorable. However, the results of this study are not so clear to sustain these beliefs. It was observed that the results of more experienced apprentices were similar to those with little experience in soil science, and again, the evaluation depends on the sample design used to validate the maps. Analyzing the results based on SRS design, it is possible to affirm that the map generated by apprentice B (lowest RE) present the lowest values of both the Overall Accuracy and the Kappa Index. In this case, the previous knowledge and experience, especially when it comes to the basic concepts of pedology and the soil-landscape relationship, may produce a better soil map. But at the same sample design (SRS), the map generated by apprentice D (with RE = 0.5) was the best, and the map generated by apprentice A (RE = 0.7) did not significantly surpass the quality of the map generated by apprentice C (RE = 0.5). In this case, the master degree in soil science (apprentice A) did not result in a better soil map. However, considering the SSRS design, this difference practically disappears. There is only a higher difference between apprentice A and the others in relation to the overall accuracy. The Kappa index is almost the same for all the apprentices.

Comparing the indexes WUAI and WPAI between the CSM, the map generated by the apprentice B presented the lowest WUAI both in SRS and SSRS sample design. It means that, in general, considering the SRS sample design, only 55 % of the surface area depicted on the map generated by apprentice B is that soil in reality. Considering the SSRS design, this value increased to 70 %. In terms of PA, the apprentice B presented equivalent results to the others.

In the specific case of DSM techniques, the greatest confusion was observed in MU with a small coverage area and/or with a small amount of modal profile in the unit. The small number of modal profiles in a MU can hamper the development of a prediction model because during the training step, the algorithm does not have many samples to establish relationships between the types of soils in that MU with the variations of the covariates along the landscape. A possible explanation for the better performance of conventional maps in predicting MUs with less data available may be because pedologists have several mental mechanisms to outline MUs that the machine algorithms do not have. For example, an experienced pedologist can (based on past work in regions with a similar pattern) use this knowledge to solve a local information limitation. According to Zare et al. (2018), the experienced surveyor can use few morphological field observations to classify soil profiles into predefined classification systems and extrapolates the soil types to make a map based on aerial photographs.

Furthermore, units with a smaller area occupy a position in the landscape that is difficult to discern, both by the mathematical model and by the apprentice, even using landscape characteristics that most separate these classes, which are elevation, slope and position in the landscape that influences the drainage condition, which are the covariates of the terrain selected by the PCA (DEM, *rsp*, and *twi*). Despite being selected in the PCA (band2 and SAVI), the covariates that represent the organism factor have a little contribution in the models since the variation in vegetation cover is small (predominating pasture).

Another important aspect is the level of detail of the soil survey. The map generated in the training site can be classified as detailed or ultra-detailed level. In this case, some soil properties used to differentiate soil mapping units have no direct relationship with the landscape. As an example, compare the classification of soil belonged to MU4 and MU5. In the first case (MU4), the modal profile was classified as Nitic Lixisol (*Argissolo Vermelho Eutrófico nitossólico - PVe*), while in the second case (MU5), the modal profile was classified as Nitic Acrisols (*Argissolo Vermelho Distrófico nitossólico - PVd*). The MU4 and MU5, in addition to differences in some morphological and physical

properties, differ between themselves in base saturation level of the soils, the presence of pebble and cobble, and the soil depths where these fragments occur. Analyzing the covariates used for DSM algorithm (DEM, RSP, TWI, Band 2, and SAVI), there is no clear relation of them with base saturation and pebbles and cobble presence. These aspects are an important reflection, as detailed and ultra-detailed soil maps are recommended to utilitarian decisions (like irrigation, drainage, and soil management). Maybe the DSM techniques should be more efficient in predicting soil properties (such as texture, water retention capacity, and nutrient content). Besides, considering soil maps with lower detailment level (1:25.000-1:250.000, for example), DSM could have similar or better results than those obtained by the conventional technique while providing additional information about each soil's landscape. In addition, DSM has the advantages of producing available information useful for applications in future surveys of similar areas (Bazaglia Filho et al., 2013) with associated uncertainty (Poggio and Gimona, 2017), automate and make mapping processes reproducible and easy to update with the entry of new soil data set or covariates. Finally, the DSM products can be used as input for further modeling in a number of areas related to agriculture and environmental science (Poggio and Gimona, 2017). The combination of DSM techniques and pedometrics tools and knowledge of bases of pedology can promote greater interactions between DSM and pedology and can help in forming new hypotheses and gaining new insights about soil and soil processes and mapping (Pinheiro et al., 2016, 2018; Maechler et al., 2019)

Pedagogic project and teaching-learning process

Based on the questionnaire's answer, among the six items raised in the first question the more challenging theme was: "Theoretical class involving statistical modeling, sampling, prediction, and accuracy assessment". This item is among the top three challenging for all apprentices. Followed by the item "Practical class involving data organization, mental and mathematical model calibration, prediction, and validation", which is also in the top three for most of them, only in one case where it was in fourth place.

The item "Theoretical class involving digital soil mapping, GIS, and environmental covariates processing" was also assessed as challenging for most apprentices (on the top 3 for three of them), just for the apprentice B, who already had experience in the topic, placed the item as the least challenging.

Apprentices A and D who have previous experience with soil science, classified the item "Theoretical classes of introduction to pedology, soil-landscape relationship, soil survey and classification" and "Practical classes to identify soil in the field (description and classification) and soil-landscape relationship in practice" as being the lowest challenging and/or the easiest to understand, whereas apprentices (B and C), which do not have this previous experience classified the practical classes as one of the most challenging.

In summary, for apprentices who had no experience with a theoretical and practical part involving the topics of digital soil mapping, GIS, modeling, statistics, mental and mathematical model calibration, and validation of predictions, these were the most difficult topics. For apprentices who had no experience with soil science other than the topics mentioned, practical classes in soil-landscape relations and description of soil profiles in the field were the most challenging and should pay particular attention.

When asked about the course time and time distribution between theoretical and practical classes, all participants found the course duration appropriate; but the ratio between theoretical and practical classes for each theme must be balanced according to the learners' needs. As in this course there was a lot of practical classes to describe the field and study the relationship between the soil and the landscape (one of the group's needs for use later on in the project), perhaps it would be interesting to have more theoretical and practical time (use of software) focused on statistical modeling and

digital tools. About having prior knowledge of the subjects covered, apprentices do not think that the lack of prior experience has been a limiting factor. However, it facilitates the teaching and learning process. In addition, it is very difficult to carry out a course in which all participants are aligned. For this, the course must start from a more basic theory to level the class and balance the activities according to the deficiency presented by each apprentice, then to reduce the effects of the lack of prior knowledge and speed up the learning process.

About the course be multidisciplinary, none of the participants found this a limiting factor, one factor that may have helped is that they all have a degree in Agronomy, a course that is already multidisciplinary and has disciplines in almost all departments of the university, so integrating information from different areas is not new to them. They also report that multidisciplinary is extremely important, since today's professional must have basic notions of different areas to solve problems in the modern world. Besides the fact that being multidisciplinary makes the course more interesting, the learning process is less tiring. For them, it is challenging to integrate information but extremely necessary.

When asked about automating processes using pedometric tools, which were previously done manually and subjectively in the past, they are motivated and see this as a real need, especially with regard to the optimization of financial and human resources; they report that the use of technologies available brings several benefits and facilities to the mapping process, reducing subjectivity facilitates soil mapping and makes it more dynamic and fluid. And this can be decisive for the maintenance of survey projects across the country, especially within the scope of *PronaSolos*. And this can be decisive for the maintenance of survey projects across the country, especially within the scope of *PronaSolos* (Dalmolin et al., 2020). We believe that the approach showed in this paper can be implemented with the *PronaSolos* program using the infrastructure that already exists in Brazilian Universities and the Residency Program in Agronomy, which is already implemented at UFRRJ.

About learning a subject while being involved in a research project on the theme, the apprentices believe that whenever it is possible to align a learning with its practical application in a project, the course utilization becomes greater, which is different from just reading and hearing about the theme, the opportunity to execute the newly acquired knowledge facilitates its fixation. This result corroborates that found by Hupy et al. (2005), who describe that students learn and understand relationships among physical landscape variables better by mapping them than they would in a classroom-based experience. Training using the project pedagogy approach can promote greater interactions between DSM and pedology (CSM) and demonstrate the importance of linking traditional pedology and pedometrics (Ma et al., 2019). Also, the use of DSM tools in GIS environment adds abstract analyses and quantitative assessment, which is a complementary learning style to fieldwork that mostly focuses on practical skills (Marra et al., 2017). In terms of previous experience, the project pedagogy approach can be a highly useful pedagogical tool and can be employed even among groups where the ability/skill levels are highly variable, as seen by our results and corroborated by Hupy et al. (2005).

The existence of the project increases the involvement with the activity learned and facilitates learning and makes a big difference in relation to the needs to qualify. As the theoretical and practical classes were totally focused on what they were going to do on the project, so they describe that although it is not possible to see everything during the course, for example, unforeseen problems that compromise the planning, the basis that they have today will help the team on to know how to identify problems and find solutions. Being involved in a project, the apprentices can learn in practice; in this case, the task does the teaching, not the professor.

As for evaluating the course and what can be improved, everyone found the course good and suitable for the training of new pedologists, especially those who will work at *PronaSolos*. The main suggestions are: (a) increase the participation of other professionals, to add with regard to the multidisciplinary of the basic knowledge of the professional of Quantitative Pedology (Pedometrics); (b) the practical classes of soil description could approach more soil types so that a greater range of characteristics could be seen in the field, as this knowledge is of great help in the day-to-day soil survey; (c) increase the workload and deepen the practical part of modeling and DSM techniques, especially to meet those with little or no previous experience of the subject.

CONCLUSIONS

The project pedagogy approach is promising to train new pedologists since, by mixing theoretical activities and contextualized practices (a project in progress), it not only awakens great motivation and critical capacity, but also develops the ability for apprentices to find solutions in a area in constant evolution.

The quality of the maps (both CSM and DSM) changed significantly according to the validation sample design applied. The CSM present better quality than DSM, mainly when using SSRS validation design. Random Forest algorithm showed equivalent accuracy to CSM using SRS sample design. Irrespective to sample validation design, the MLR algorithm presented the lowest accuracy.



The CSMs presented higher producer's accuracy (WPAI), while the DSM presented higher user's accuracy (WUAI)



The quality of CSM generated by the apprentices was not clearly related to the previous experience and knowledge in soil science.







ACKNOWLEDGEMENTS







The authors acknowledge the National Petroleum Agency (ANP) for funding the project "Digital Mapping of Soils in Oil and Gas Exploration and Production Areas - Case Studies of the North and Northeast Brazilian Fields" under the agreement number 5850.0105881.17.9 (PETROBRAS/FAPUR/UFRRJ) that enabled the training of pedologists.







AUTHOR CONTRIBUTIONS



Conceptualization:  Elias Mendes Costa (equal) and  Marcos Bacis Ceddia (equal).

Methodology:  Elias Mendes Costa (equal) and  Marcos Bacis Ceddia (equal).







Software:  Elias Mendes Costa (lead),  Marcos Bacis Ceddia (supporting),  Felipe Nascimento dos Santos(supporting),  Laiz de Oliveira Silva(supporting),  Igor Prata Terra de Rezende(supporting), and  Douglath Alves Correa Fernandes (supporting).







Validation:  Elias Mendes Costa (equal),  Marcos Bacis Ceddia (equal),  Felipe Nascimento dos Santos(supporting),  Laiz de Oliveira Silva(supporting),  Igor Prata Terra de Rezende(supporting), and  Douglath Alves Correa Fernandes (supporting).







Formal analysis:  Elias Mendes Costa (equal),  Marcos Bacis Ceddia (equal),  Felipe Nascimento dos Santos(supporting),  Laiz de Oliveira Silva(supporting),  Igor Prata Terra de Rezende(supporting), and  Douglath Alves Correa Fernandes (supporting).



Investigation:  Elias Mendes Costa (equal) and  Marcos Bacis Ceddia (equal).



Resources:  Marcos Bacis Ceddia (lead).

Data curation:  Elias Mendes Costa (equal),  Marcos Bacis Ceddia (equal),  Felipe Nascimento dos Santos(supporting),  Laiz de Oliveira Silva(supporting),  Igor Prata Terra de Rezende (supporting), and  Douglath Alves Correa Fernandes (supporting).

Writing - original draft:  Elias Mendes Costa (equal),  Marcos Bacis Ceddia (equal),  Felipe Nascimento dos Santos (supporting),  Laiz de Oliveira Silva (supporting),  Igor Prata Terra de Rezende(supporting), and  Douglath Alves Correa Fernandes (supporting).

Writing - review and editing:  Elias Mendes Costa (equal),  Marcos Bacis Ceddia (equal),  Felipe Nascimento dos Santos (supporting),  Laiz de Oliveira Silva (supporting),  Igor Prata Terra de Rezende (supporting), and  Douglath Alves Correa Fernandes (supporting).

Visualization:  Elias Mendes Costa (lead) and  Marcos Bacis Ceddia (supporting).

Supervision:  Marcos Bacis Ceddia (lead) and  Elias Mendes Costa (supporting).

Project administration:  Marcos Bacis Ceddia (lead).

Funding acquisition:  Marcos Bacis Ceddia (lead).

REFERENCES

Agresti A. Categorical data analysis. 2nd ed. Florida: Gainesville; 2002.

Alvares CA, Stape JL, Sentelhas PC, Gonçalves JLM, Sparovek G. Koppen' s climate classification map for Brazil. Meteorol Z. 2013;22:711-28. <https://doi.org/10.1127/0941-2948/2013/0507>

Antunes MAH, Debiasi P, Siqueira JCS. Avaliação espectral e geométrica das imagens Rapideye e seu potencial para o mapeamento e monitoramento agrícola e ambiental. Rev Bras Cartogr. 2014;66:105-13.

Bazaglia Filho O, Rizzo R, Lepsch IF, Prado H, Gomes FH, Mazza JA, Demattê AJM. Comparison between detailed digital and conventional soil maps of an area with complex geology. Rev Bras Cienc Solo. 2013;37:1136-48. <https://doi.org/10.1590/S0100-06832013000500003>

Beaudette DE, Roudier P, O'Geen AT. Algorithms for quantitative pedology: A toolkit for soil scientists. Comput Geosci. 2013;52:258-68. <https://doi.org/10.1016/j.cageo.2012.10.020>

Bivand R, Keitt T, Rowlingson B. rgdal (R package): Bindings for the "Geospatial" Data Abstraction Library. 2019. [cited 19 Jan 2020]. Available from: <https://CRAN.R-project.org/package=rgdal>.

Bivand R, Lewin-Koh N. maptools (R package): Tools for Handling Spatial Objects. 2019. Breiman L. Random forests. Mach Learn. 2001;45:5-32. [cited 11 Feb 2020]. Available from: <https://doi.org/10.1023/A:1010933404324>

Brenning A, Bangs D, Becker M. RSAGA (R package): SAGA Geoprocessing and Terrain Analysis. 2018. [cited 19 Jan 2020]. Available from: <https://CRAN.R-project.org/package=RSAGA>

Brus DJ. Sampling for digital soil mapping: A tutorial supported by R scripts. Geoderma. 2019;338:464-80. <https://doi.org/10.1016/j.geoderma.2018.07.036>

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

Burrough PA, McDonnell RA. Principles of geographical information systems. Oxford: Oxford University Press; 1998.

- Câmara G, Souza R, Freitas U, Garrido J. Spring: integrating remote sensing and gis by object-oriented data modelling. *Comput Graph.* 1996;2:395-403. [https://doi.org/10.1016/0097-8493\(96\)00008-8](https://doi.org/10.1016/0097-8493(96)00008-8)
- Costa EM, Pinheiro HSK, Anjos LHC, Marcondes RAT, Gelsleichter YA. Mapping soil properties in a poorly-accessible area. *Rev Bras Cienc Solo.* 2020;44:e0190107. <https://doi.org/10.36783/18069657rbc20190107>
- Dalmolin RSD, Moura-Bueno JM, Samuel-Rosa A, Flores CA. How is the learning process of digital soil mapping in a diverse group of land use planners? *Rev Bras Cienc Solo.* 2020;44:e0190037. <https://doi.org/10.36783/18069657rbc20190037>
- Grimm R, Behrens T, Märker M, Elsenbeer H. Soil organic carbon concentrations and stocks on Barro Colorado Island - Digital soil mapping using Random Forests analysis. *Geoderma.* 2008;146:102-13. <https://doi.org/10.1016/j.geoderma.2008.05.008>
- Hijmans RJ. raster (R package): Geographic data analysis and modeling. 2019. <https://CRAN.R-project.org/package=raster>
- Huete AR. A soil-adjusted vegetation index (SAVI). *Remote Sens Environ.* 1988;25:295-308. [https://doi.org/10.1016/0034-4257\(88\)90106-X](https://doi.org/10.1016/0034-4257(88)90106-X)
- Hupy JP, Aldrich SP, Schaetzl RJ, Varnakovidia P, Arima EY, Bookout JR, Wiangwang N, Campos AL, McKnight KP. Mapping soils, vegetation, and landforms: An integrative physical geography field experience. *Prof Geogr.* 2005;57:438-51 <https://doi.org/10.1111/j.0033-0124.2005.00489.x>
- 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).
- Jeune W, Francelino MR, De 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/10.1590/18069657rbc20170133>
- Kassambara A, Mundt F. factoextra (R package): extract and visualize the results of multivariate data analyses. 2017. <https://CRAN.R-project.org/package=factoextra>
- Lepsch IF. Status of soil surveys and demand for soil series descriptions in Brazil. *Soil Horizons.* 2013;54:1-5. <https://doi.org/10.2136/sh2013-54-2-gc>
- Levi MR, Rasmussen C. Covariate selection with iterative principal component analysis for predicting physical soil properties. *Geoderma.* 2014;219-220:46-57. <https://doi.org/10.1016/j.geoderma.2013.12.013>
- Liaw A, Wiener M. Classification and Regression by randomForest (R package). *R News.* 2002;2:18-22. <https://CRAN.R-project.org/package=randomForest>
- Ma YX, Minasny B, Malone BP, McBratney AB. Pedology and digital soil mapping (DSM). *Eur J Soil Sci.* 2019;70:216-35. <https://doi.org/10.1111/ejss.12790>
- Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K. cluster (R package): Cluster analysis basics and extensions. 2019. [cited 27 Jan 2020]. Available from: <https://CRAN.R-project.org/package=cluster>
- Marra WA, Grint L, Alberti K, Karssenberg D. Using GIS in an Earth Sciences field course for quantitative exploration, data management and digital mapping. *J Geogr Higher Educ.* 2017;41:213-29. <https://doi.org/10.1080/03098265.2017.1291587>
- McBratney AB, Mendonça Santos ML, Minasny B. On digital soil mapping. *Geoderma.* 2003;117:3-52. [https://doi.org/10.1016/S0016-7061\(03\)00223-4](https://doi.org/10.1016/S0016-7061(03)00223-4)
- Olasco-Carvalho CC, Nunes FC, Antunes AMH. Histórico do levantamento de solos no Brasil: Da industrialização brasileira à era da informação. *Rev Bras Cartogr.* 2013;65:997-1013.
- Oliveira-Júnior JF, Delgado RC, Gois G, Lannes A, Dias FO, Souza JC, Souza M. Análise da precipitação e sua relação com sistemas meteorológicos em seropédica, Rio de Janeiro. *Floresta Ambient.* 2014;21:140-9. <https://doi.org/10.4322/floram.2014.030>

- Pinheiro HSK, Chagas CS, Carvalho Junior W, Anjos LHC. Ferramentas de pedometria para caracterização da composição granulométrica de perfis de solos hidromórficos. *Pesq Agropec Bras*. 2016;51:1326-38. <https://doi.org/10.1590/S0100-204X2016000900032>
- Pinheiro HSK, Helena L, Anjos C, Xavier PAM, Cesar S. Quantitative pedology to evaluate a soil profile collection from the Brazilian semi-arid region. *S AFR J Geomat*. 2018;1862:269-79. <https://doi.org/10.1080/02571862.2017.1419385>
- Poggio L, Gimona A. 3D mapping of soil texture in Scotland. *Geoderma Reg*. 2017;9:5-16. <https://doi.org/10.1016/j.geodrs.2016.11.003>
- Polidoro JC, Mendonça-Santos ML, Lumbrreras JF, Coelho MR, Carvalho Filho A, Motta PEF, Carvalho Junior W, Araújo Filho JC, Curcio GR, Correia JR, Martins ES, Spera ST, Oliveira SRM, Bolfe EL, Manzatto CV, Tosto SG, Venturieri A, Sá IB, Oliveira VA, Shinzato E, Anjos LHC, Valladares GS, Ribeiro JL, Medeiros PSC, Moreira FMS, Silva LSL, Sequinatto L, Aglio MLD, Dart RO. PronaSolos - Programa nacional de solos do Brasil (PronaSolos) - Dados eletrônicos. Rio de Janeiro: Embrapa Solos; 2016.
- R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2019.
- RapidEye. RapidEye Mosaic™ Product Specifications; 2012.
- Santos RD, Santos HG, Ker JC, Anjos LHC, Shimizu SH. Manual de descrição e coleta de solo no campo. 7. ed. Viçosa, MG: Sociedade Brasileira de Ciência do Solo; 2015.
- Santos HG, Jacomine PKT, Anjos LHC, Oliveira VA, Lumbrreras 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.
- 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>
- Teixeira PC, Donagemma GK, Fontana A, Teixeira WG. Manual de métodos de análise de solo. 3. ed. rev e ampl. Brasília, DF: Embrapa; 2017.
- ten Caten A, Dalmolin RSD, Pedron FA, Mendonça-Santos ML. Componentes principais como preditores no mapeamento digital de classes de solos. *Cienc Rural*. 2011;41:1170-6. <https://doi.org/10.1590/S0103-84782011000700011>
- Tillé Y, Matei A. sampling (R package): Survey sampling. 2016. <https://CRAN.R-project.org/package=sampling>
- Venables WN, Ripley BD. Modern applied statistics with S. 4th ed. New York: Springer; 2002.
- Vermote EF, Herman M, Morcrette J. Second simulation of the satellite signal in the solar spectrum, 6S: an overview. *IEEE T Geosci Remote*. 1997;35:675-86. <https://doi.org/10.1109/36.581987>
- Were K, Bui DT, Dick ØB, Singh BR. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afrotropical landscape. *Ecol Indic*. 2015;52:394-403. <https://doi.org/10.1016/j.ecolind.2014.12.028>
- Yang RM, Zhang GL, Liu F, Lu YY, Yang Fan, Yang Fei, Yang M, Zhao YG, Li DC. Comparison of boosted regression tree and random forest models for mapping topsoil organic carbon concentration in an alpine ecosystem. *Ecol Indic*. 2016;60:870-8. <https://doi.org/10.1016/j.ecolind.2015.08.036>
- Zare E, Ahmed MF, Malik RS, Subasinghe R, Huang J, Triantafyllis J. Comparing traditional and digital soil mapping at a district scale using residual maximum likelihood analysis. *Soil Res*. 2018;56:535-47. <https://doi.org/10.1071/sr17220>
- Zeraatpisheh M, Ayoubi S, Jafari A, Finke P. Comparing the efficiency of digital and conventional soil mapping to predict soil types in a semi-arid region in Iran. *Geomorphology*. 2017;285:186-204. <https://doi.org/10.1016/j.geomorph.2017.02.015>