Ciência e Agrotecnologia
SCIENCE AND AGROTECHNOLOGY
Ciência e Agrotecnologia, 45:e007921, 2021
http://dx.doi.org/10.1590/1413-7054202145007921
Agricultural Sciences
eISSN 1981-1829

Digital soil mapping: Predicting soil classes distribution in large areas based on existing soil maps from similar small areas

Mapeamento digital de solos: Predição da distribuição de classes de solo em grandes áreas baseando-se em mapas de solo preexistentes de áreas menores semelhantes

Thaís Gabriela Gonçalves[1], Nívea Adriana Dias Pons[1*], Eliane Guimarães Pereira Melloni[1], Marcelo Mancini[2], Nilton Curi[2]

[1]Universidade Federal de Itajubá/UNIFEI, Instituto de Recursos Naturais, Itajubá, MG, Brasil
[2]Universidade Federal de Lavras/UFLA, Departamento de Ciência dos Solos/DCS, Lavras, MG, Brasil
[*]Corresponding author: npons@unifei.edu.br
Received in April 13, 2021 and approved in July 8, 2021

ABSTRACT

There is an ever-growing need for soil maps, since detailed soil information is directly related to agricultural activities, urbanization and environmental protection. However, there is a lack of large-scale soil maps in developing tropical countries such as Brazil. Albeit there are soil maps for small areas, large regions usually have undetailed maps. Considering the importance of finding low-cost alternatives to overcome the lack of detailed soil information, the main objective of this work was to manually create a local soil map and extrapolate it to similar larger areas that lack detailed soil information. The Anhumas River Basin, in the municipality of Itajubá, southeast Brazil, was manually mapped and this map was used to predict soils distribution for the entire municipality. First, the prediction model was tested in the same basin and provided sufficient results, achieving 67% global accuracy and 0.62 Kappa coefficient. Second, the resulting map was used together with the soil map of the larger José Pereira Basin to map the entire municipality, achieving 54% global accuracy and 0.40 Kappa coefficient. Low resolution parent material information was found to confuse models; maps showed better results when this variable was removed. The Minas Gerais soil map presents general mapping units only for the Acrisol class and its associations with other soil classes in the area. The soil map predicted by this work identified more soil classes. Mapping representative areas and extrapolating these maps to larger similar areas constitute a promising alternative to overcome the lack of detailed soil maps.

Index terms: Soil classification; pedology; decision trees; spatial distribution.

RESUMO

Há uma crescente demanda por mapas de solo, já que informações detalhadas de solos estão diretamente relacionadas com agricultura, urbanização e conservação ambiental. Porém, há escassez de mapas de solo em grande escala em países tropicais em desenvolvimento como o Brasil. Apesar de existirem mapas para pequenas áreas, grandes regiões possuem mapas pouco detalhados. Considerando-se a importância de se buscarem alternativas para superar a falta de informações detalhadas de solo, o objetivo deste trabalho foi extrapolar mapas locais para áreas maiores semelhantes que necessitam de informações mais detalhadas. A bacia do rio Anhumas, no município de Itajubá, sudeste do Brasil, foi manualmente mapeada e este mapa foi utilizado para predizer a distribuição de solos para todo o município. Primeiro, o modelo de predição foi testado na mesma bacia e alcançou acurácia global de 67% e coeficiente Kappa de 0,62. Depois, os resultados foram usados conjuntamente com dados da bacia José Pereira para predizer as classes de solo de todo o município, alcançando acurácia global de 54% e coeficiente Kappa de 0,40. Dados de material de origem em baixa resolução confundiram os modelos; as predições obtiveram melhores resultados quando esta variável foi removida. O mapa de solos de Minas Gerais apresenta apenas a classe de solo Argissolo e suas associações com outros solos para a área. O mapa confeccionado por este trabalho, porém, identificou mais classes de solo. O mapeamento de áreas representativas e sua extrapolação para áreas maiores constitui alternativa promissora para superar a escassez de mapas detalhados de solo.

Termos para indexação: Classificação de solo; pedologia; árvores de decisão; distribuição espacial.

INTRODUCTION

In the last few decades, the disorganized urbanization process of Brazilian cities resulted in several problems regarding the quality of life of their residents, such as the progressive degradation of natural resources (Braga; Silva; Schaffrath, 2012). Problems related to erosion, sediments redeposition, soil degradation, biodiversity loss, water contamination, demand accurate and more detailed soil information to be properly managed (Zhang; Liu; Song,

2017). For adequate resource management, data about soil properties, their variability and spacialization through maps are crucial and can adequately be obtained by pedologic assessment. In Brazil, however, the generalized lack of more detailed soil data hinders environmental assessment and prevents the use of statistical techniques that can predict the distribution of soil classes and properties (Teske; Giasson; Bagatini, 2015).

Following the growing demand for information about the spatial distribution of soil classes and properties, modern computational techniques are being continuously developed and tested (Arruda et al., 2016; Camera et al., 2017; Minai; Libohova; Schulze, 2020; Silva et al., 2016; Peng et al., 2020). These newly developed digital methods improve soil mapping and are becoming powerful tools capable of providing detailed data in larger similar areas, which is key for the assessment of agronomical and environmental problems (Padarian; Minasny; McBratney, 2020; Piikki; Söderström; Stadig, 2017; Vincent et al., 2018).

One way of applying such techniques is to use soil information from smaller areas and extrapolate them to larger areas with similar physiographic characteristics (Afshar; Ayoubi; Jafari, 2018; Angelini et al., 2020; Dias et al., 2016). If the soil forming factors are similar, the models built based on reference areas can be applied elsewhere (Malone et al., 2016). Mallavan, Minasny and McBratney (2010) called this concept "homosoil", and argued that if the homology of soil-forming factors is assumed, soil information from different parts of the world can be used to infer data about soils in an area of interest (Scull; Franklin; Chadwick, 2005).

This concept can be useful for developing tropical countries like Brazil, where it is common to find soil maps for very specific locations, but more detailed soil classes distribution for larger areas is rare. As an example, the best available soil map for the area studied in this work (Itajubá, state of Minas Gerais) is at the scale of 1:650,000 (UFV-CETEC-UFLA-FEAM, 2010). This scale is considered too small to support adequate decisions regarding urbanization, agricultural activities and environmental management in the specific area of interest (Wolski et al., 2017). Hence, although most preexisting maps are useful for studies at smaller scales, municipalities that demand localized information and where detailed soil maps are not available could benefit from the application of these methods. This is the case of Itajubá, the municipality studied herein.

According to Höfig, Giasson and Vendrame (2014), the extrapolation of the soil-landscape relationships to similar adjacent areas based on smaller reference areas is scarce, even though this process can minimize the lack of large-scale soil maps for many countries, including Brazil. This process can use preexisting data from the literature, or data obtained from soil surveys in similar, but smaller areas, reducing cost. Hence, for countries that have insufficient resources to map large portions of their territory, the further exploration of extrapolation techniques would help to overcome the lack of financial resources and potentially provide low-cost alternatives to deliver the much needed detailed soil information for municipalities that require more adequate decision-making support, such as Itajubá.

Considering the necessity of more detailed soil data in order to support adequate land management, especially in Brazil and other development tropical countries, this work attempts to perform the extrapolation of existing soil maps from smaller reference areas to a larger area with physiographic similarities. The specific objectives of this work were to:

i) manually build a soil map for the Anhumas River Basin (a small basin within the municipality of Itajubá, Minas Gerais) based on 16 characterized soil profiles (4 catenas), soil analyses and experience of pedologists;

ii) use the manually built soil map in tandem with data from the literature to create a prediction model in order to extrapolate the soil classes distribution for the entire municipality of Itajubá.

We hypothesize that the use of existing soil maps from smaller reference areas to infer soil classes distribution across larger neighboring areas will provide reliable data to support more adequate urban development and environmental planning.

MATERIAL AND METHODS

Study area

The study was conducted in Itajubá, a municipality located in the south of the state of Minas Gerais, Brazil, between the coordinates 45º32'30" W 22º20'00" S and 45º14'30" W and 22º33'00" S (Figure 1). The region is situated at the meridional limits of the intertropical zone and influenced by high altitudes. The climate is Cwa – temperate humid with dry winter and hot summer – according to the Köppen classification system (Alvares et al., 2013). Mean annual precipitation is 1,409.5 mm, presenting more precipitation intensity from October to March, and the dry period ranges from April to September. The annual average temperature is 20.5 ºC.

The Basin of the Anhumas River was used in this study as the reference area (Figure 1). A soil survey was conducted in this area, and the resulting map was used to

predict soil classes distribution for the entire municipality. The Anhumas River is 1.23 km long with a drainage area of approximately 23.53 km², corresponding to 8% of Itajubá's total area.

The municipality of Itajubá is located in the Atlantic Forest biome. The Biological Reserve Serra dos Toledos (REBIO Serra dos Toledos) preserves its native rainforest. The rest of the municipality is occupied mainly by semiperennial tropical forest. The main land use in the region is grazing, occupying 42.22% of the territory.

Soil mapping methodology

This work was conducted in two main steps (Figure 2): i) the soil survey and the identification of catenas, followed by sampling and laboratory analyses to classify the studied soil profiles; and ii) the mapping of the Anhumas River Basin, followed by its extrapolation to create the map for the municipality of Itajubá. Each step of the mapping process will be further detailed in the following sections.

Soil sampling and laboratory analyses

Across the Anhumas River Basin 16 soil profiles were morphologically described, sampled and characterized in the laboratory, resulting in a sampling density of 0.68 samples/km². Sampling was performed along four soil catenas (Figure 3) within the basin and, in each catena, four landscape positions were selected: interfluve, shoulder, midslope and footslope, in places with minimum anthropogenic interference. The criteria for selecting the soil catenas were based upon previous literature information and intensive field work, aiming to assess the maximum representativeness and variability of the environment as possible.

Samples were collected according to the Manual of Description and Sampling of Soils in the Field (Santos et al., 2015). 1.5 kg of disturbed soil was sampled for each identified horizon, along with undisturbed soil samples using an Uhland sampler for bulk density and porosity analyses. Soils were classified according to the Brazilian Soil Classification System, as described by (Santos et al., 2018), and according to the World Reference Base for Soil resources (WRB) (IUSS, 2015).

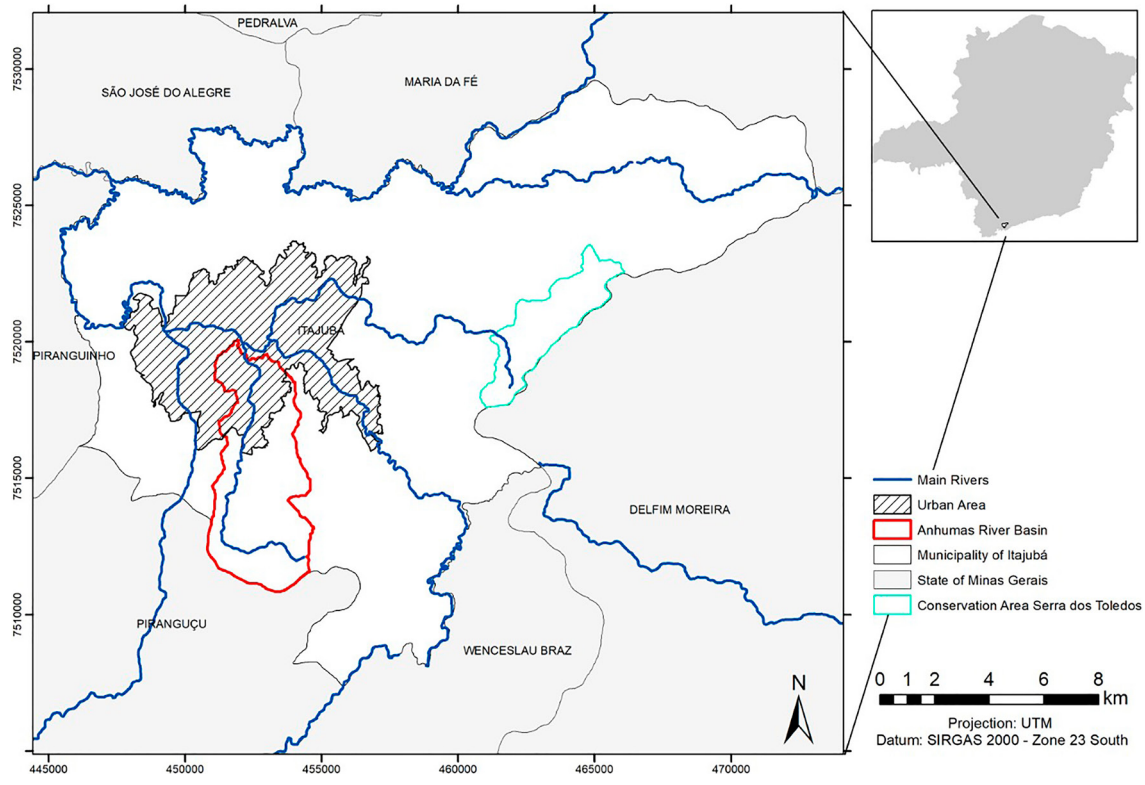

Figure 1: Municipality of Itajubá and the Anhumas River Basin where the soil survey was conducted, situated in the state of Minas Gerais, Brazil.

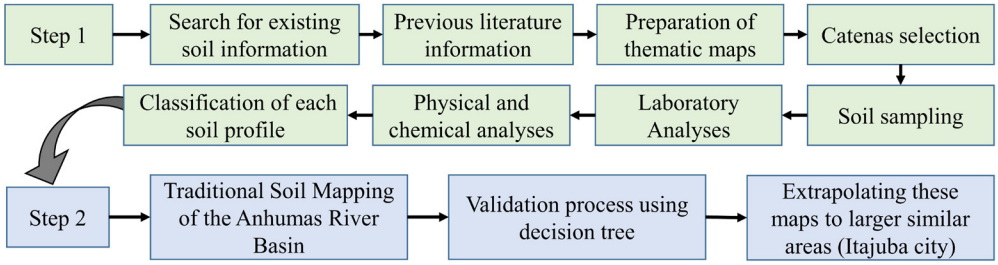

Figure 2: Illustration of the followed steps to create soil maps for the Riberão Anhumas Basin and the municipality of Itajubá, located in the state of Minas Gerais, Brazil.

Disturbed samples were air-dried and passed through a 2-mm sieve. Next, they were submitted to the following analyses: pH in water at 1:2.5 (soil:water) ratio according to McLean (1982); exchangeable Ca^{2+}, Mg^{2+} and Al^{3+} were extracted with 1 mol L^{-1} KCl (Teixeira et al., 2017); available P and K^{+} were extracted with Melich-1 extractant and determined by colorimetry and flame photometry, respectively (Teixeira et al., 2017). Soil organic carbon was determined according to Walkley and Black (1934). Base saturation, sum of bases and aluminum saturation were calculated and their values interpreted in accordance with the Soil Fertility Committee of Minas Gerais (Ribeiro; Guimarães; Alvarez, 1999). Values for effective CEC (t) and CEC at pH 7.0 (T) were obtained indirectly from potential acidity, exchangeable bases and aluminum (Vettori, 1969). Particle size distribution was determined by the hydrometer method (Gee; Bauder, 1986).

Soil mapping of the Anhumas River Basin

This work aimed to produce maps at the scale of 1:25,000, to match the map created by Lima (2021), who studied and mapped the larger basin of José Pereira, in the same municipality. Lima (2021) used a sampling density of 0.78 samples/km² to produce maps at the scale of 1:25,000; accordingly, this work defined a similar sampling density – 0.68 samples/km² – in order to achieve the same map scale.

Soil data obtained from physical and chemical analyses, and morphological characterization were associated to the variables: lithology (Serviço Geológico do Brasil - CPRM, 2014), elevation, slope, profile curvature, and plan curvature obtained from the Shuttle Radar Topographic Mission (SRTM) using a Geographic Information System. Soil classes data and the mentioned variables were associated and spatialized across the entire basin using the software ArcGIS 10.2 in order to create a soil classes map for the Anhumas River Basin based on the experience of pedologists in correlating terrain attributes and soil classes distribution. The delimitation was done manually via geoprocessing. The variables were sliced in ArcGIS according to information from each sampled profile. Units with the same characteristics and position in the landscape receive the same classification as the soil.

The soil map of the Minas Gerais State (UFV-CETEC-UFLA-FEAM, 2010) (Figure 4) was used to assist in the delimitation of soil mapping units. The Minas Gerais State soil map is available at the scale of 1:650,000 and justifies the need for more detailed soil maps to support sustainable agronomical and environmental policies.

Digital soil mapping

First, the digital mapping technique was tested within the basin before extrapolating to the whole municipality. Digital mapping was performed considering the fourth categorical level according to the Brazilian Soil Classification System (Santos et al., 2018). A dataset with eight geomorphometric variables obtained from the SRTM was created. Utilized variables were: parent material, elevation, slope, plan curvature, profile curvature, aspect, flow direction, and flow accumulation. Note that these variables are not the same as those used in the manual mapping process (section 2.4), as some variables are more useful for modeling algorithms than for human interpretation (i.e., flow direction and accumulation). A random sampling grid was created over the Anhumas River Basin comprising 4,706 points, representing 2 points per hectare, from which the mentioned variables were extracted.

Next, the software Weka (Witten; Frank, 2005) was utilized to train decision trees via the J48 algorithm. The training process was done using cross-validation, where points were divided into 10 subsets: 9 were used to generate the trees and 1 was used to validate the model. This process was repeated 10 times with each different subset, and therefore every subset participated in modeling and validation processes until de best model was selected.

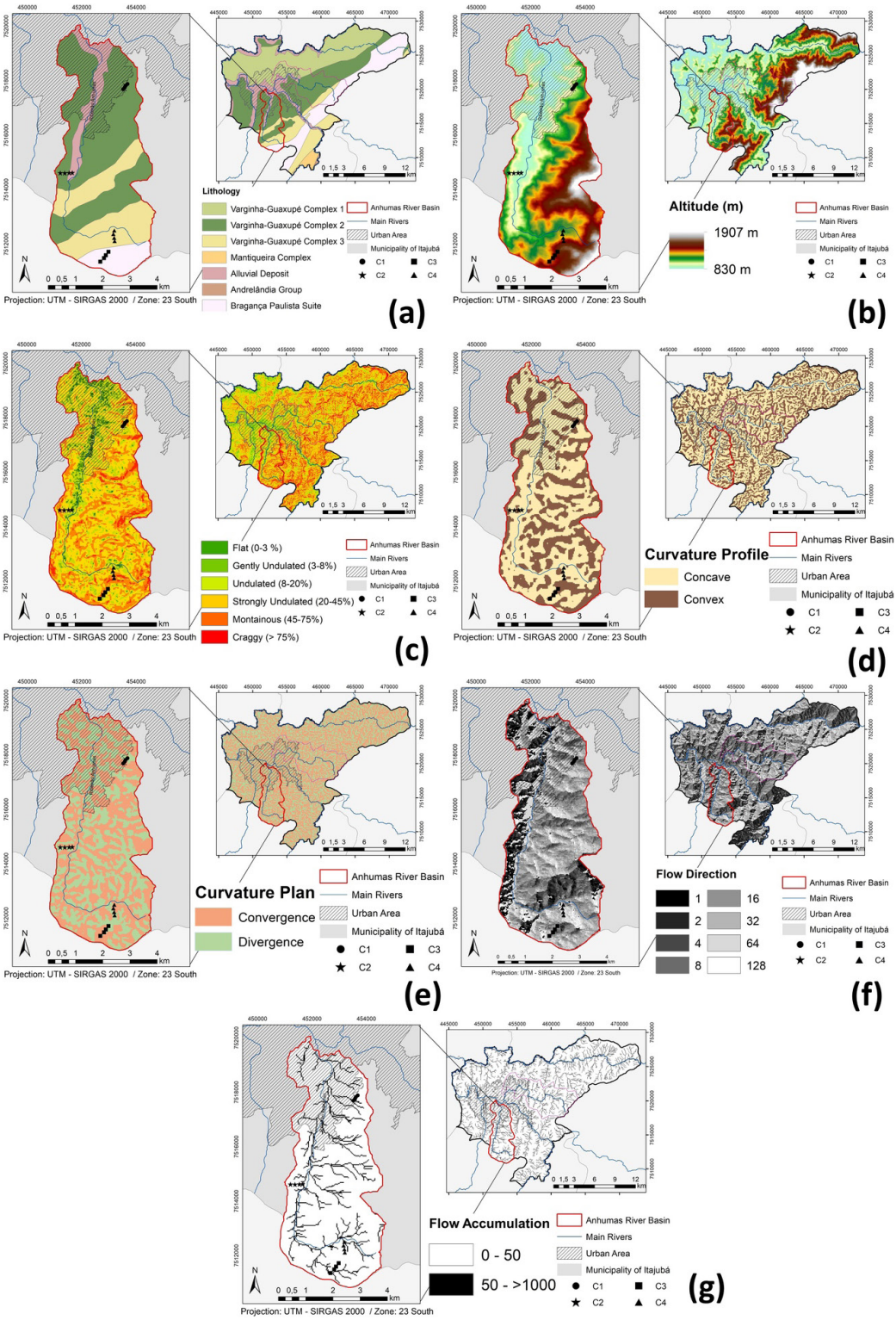

Figure 3: Lithology (a), Altitude (b), Relief (c), Curvature profile (d), Curvature plan (e), Floe direction (f) and Flow accumulation (g) of the 4 studied soil catenas at the Anhumas River Basin at the Itajubá municipality, Brazil.

After generating the decision tree, the classification model created by the J48 algorithm was converted into conditional tests (classification rules that can be used by other software), so it could be loaded into ArcGIS. The conditional tests were then applied to each pixel and their respective geomorphometric variables across the mapped region, generating the digital soil map.

This methodology was first applied to the Anhumas River Basin and, subsequently, to the whole municipality of Itajubá. To apply the method to the entire municipality, the map generated for the Anhumas River Basin was used together with the map created by Lima (2021) for the neighboring larger basin of the José Pereira River, that covers an area of 39.74 km², corresponding to 14% of the municipality's territory. Such use of literature data had the objective of improving prediction performance by increasing the amount of available data, considering that the neighboring basin studied by Lima (2021) has similar pedogenetic characteristics (Mallavan; Minasny; McBratney, 2010).

The mapping process was tested with and without parent material information, as the available parent material data are not detailed. Additionally, the municipality map was first created with mapping units containing soil classifications up to the fourth categorical level, but models were also tested with simplified mapping units up only to the second categorical level, to test if combining soils into larger groups and reducing the specificity of soil classes might improve prediction performance.

Both basins are within the municipality's limits and comprise together around 22% of its territory. A random sampling grid of 12,654 points was created, also representing 2 points per hectare, and the same aforementioned method was performed to create a digital soil map extrapolating the soil information from both basins to the entire municipality. The total absence of soil profiles described and classified in other areas of the municipality prevented the possibility of external validation methods.

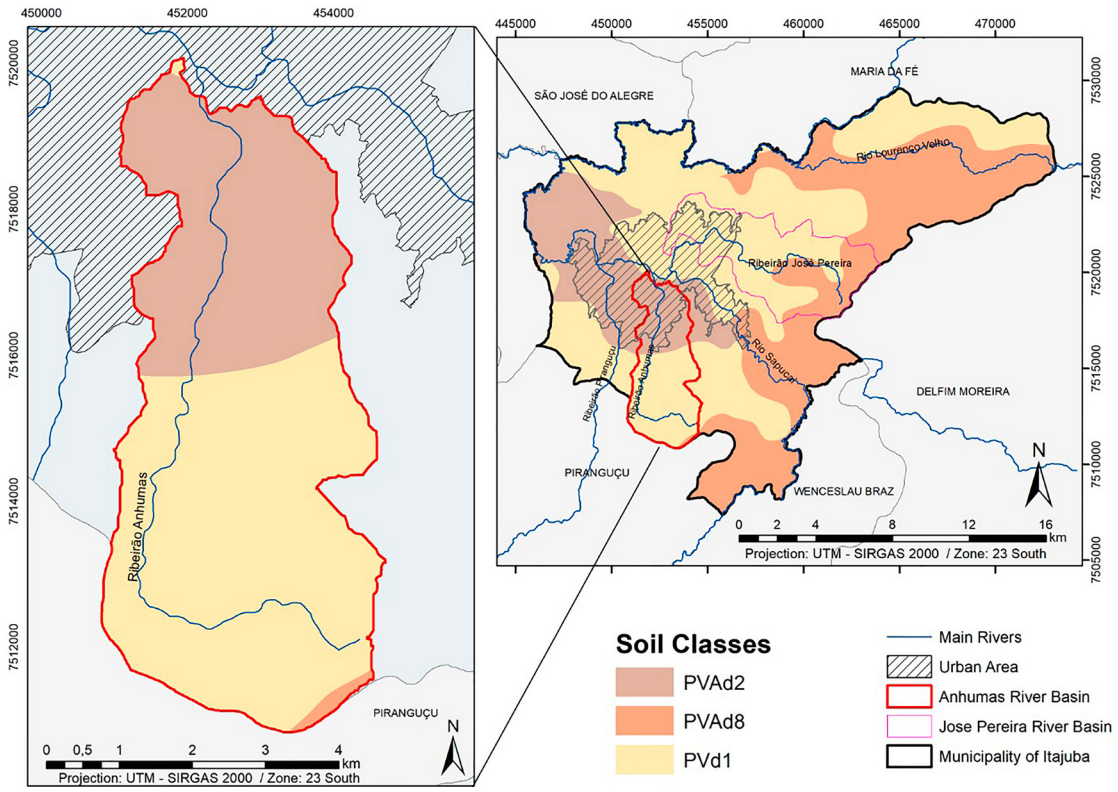

Figure 4: Soil classes map for the Itajubá municipality, Brazil, at the scale of 1:650,000 (UFV-CETEC-UFLA-FEAM, 2010). PVAd2 – Haplic Acrisol (gently undulated relief); PVAd8 – Haplic Acrisol (undulated relief and strongly undulated relief) associated with the classes Ferralsol and Dystric/Eutric Cambisol; PVd – Rhodic Acrisol.

Model validation

The validation of the prediction maps used as reference the manually-made soil map for the Anhumas Basin, as well as the map created by Lima (2021), as those are the most accurate references available for the studied area. The validation process was done using the Cohen's Kappa coefficient (Equation 1) and global accuracy (number of correct predictions/total number of tested samples), both calculated from the confusion matrix generated by Weka through the modeling process. The Kappa coefficient evaluates the agreement or reproducibility between two datasets and can be described by the following equation:

$$Kappa = \frac{Po - Pe}{1 - Pe} \quad (1)$$

Where Pe is the probability of random agreement and Po is the observed agreement (Landis; Koch, 1977). Results vary between -1 and 1, indicating more prediction reliability as the value approaches 1.

Additionally, the Producer's and User's accuracies were calculated (Story; Congalton, 1986) (Equations 2 and 3). Both scores indicate the correlation between soil classes mapped via traditional soil map and digital soil map. Producer's accuracy calculates how often each reference class (map of soil classes) is being correctly predicted, and the user's accuracy indicates how often one predicted class (class from digital soil map) corresponds to the correct reference.Where x_{ii} and x_{jj} represent the correctly predicted samples, x_{ij} indicates the sum of samples on a row or column, and L is the length of a row or column in the confusion matrix.

$$Producer's\ Accuracy = \frac{x_{ii}}{\sum_{i=1}^{L} x_{ij}} \quad (2)$$

$$User's\ Accuracy = \frac{x_{jj}}{\sum_{j=1}^{L} x_{ij}} \quad (3)$$

RESULTS AND DISCUSSION

Soil catenas

Table 1 shows the physical aspects of each sampled point. After classifying all 16 soils in this study, many of the found soil classes were not represented in the available map for the state of Minas Gerais (Figure 5), which was expected due to its small scale (1:650,000). Soil classifications are presented at Table 2.

The soils of C1 were located near the urbanized region of the basin, between altitudes of 921 and 1,000 m. The main land use in this catena is grazing. According to the available soil map for the region (Figure 5), all soils in this catena would be included in the Acrisol class, but soil classification results here also showed the classes Lixisol and Cambisol (Table 2). The parent materials of these soils were migmatitic orthogneisses associated with ultramafic and granitoid rocks.

Soils found in C2 are situated in gentle relief, between altitudes of 887 and 996 m. Parent materials are mostly the same as C1, but includes influence of alluvial deposits for C2S3 and C2S4 (Figure 2). The available soil map indicated this region would be occupied by the class Acrisol (Figure 5). Results here also showed the presence of soil classes Rhodic Ferralsol and Haplic Ferralsol, as well as Rhodic Abruptic Lixisols (Table 2).

C3 was located south of the basin. Hilltops and springs are well-preserved despite the intense grazing activities in this area. Soils were situated at altitudes between 1,157 and 1,263 m. Parent material is represented mainly by tonalite and granodiorite. According to the state's soil map, this catena was occupied by the class Rhodic Acrisol under strongly undulated relief. Still, our results showed the presence of classes Haplic Acrisol and Gleysol, which were not present in the available soil map (Figure 5).

Soils of C4 were located at altitudes between 1,049 and 1,185 m. This catena presents mountainous relief with slopes surpassing 45% (Table 1). Parent materials of these soils are represented by paragneiss, biotite, schists, and quartzite. The available soil map showed that this catena would be comprised of Rhodic Acrisols, but our classifications indicated Ferralsols and Abruptic Acrisols as well.

Albeit the Minas Gerais State soil map indicated major presence of Acrisols, our study showed that there was much more soil variability in different parts of the landscape. This emphasizes the importance and the need of finding ways to extrapolate this kind of more detailed soil information to larger similar areas.

Traditional soil mapping

The map created manually using the data obtained from the soil classification showed a soil classes distribution in agreement with Ruhe (1956) (Figure 5), with shallower and less developed soils occurring at higher and steeper positions in the landscape, whilst thicker and more developed soils being found at lower and smoother portions of the landscape, with exception of the fluvial valley.

Table 1: Physical aspects of sampling sites of 4 soil catenas at the Anhumas River Basin, located at Itajubá, state of Minas Gerais, Brazil.

Site	Coordinates (UTM)		Position in the landscape	Slope	Altitude	Land use	Vegetation
				Catena 1 (C1)			
C1S1	453682.2	7517893	Interfluve	30.31%	1,000 m	Grazing	Grassland
C1S2	453616.6	7517831	Shoulder	55.40%	966 m	Grazing	Grassland
C1S3	453585.3	7517773	Midslope	39.37%	939 m	Grazing	Grassland
C1S4	453522.7	7517699	Footslope	35.98%	921 m	Preservation area	Grassland/Bushes
				Catena 2 (C2)			
C2S1	451217.2	7514620	Interfluve	58.03%	996 m	Grazing	Grassland/Bushes
C2S2	451380.1	7514614	Shoulder	27.19%	931 m	Grazing	Grassland
C2S3	451540.1	7514621	Midslope	12.83%	908 m	Grazing	Grassland
C2S4	451654.4	7514628	Footslope	25.28%	887 m	Grazing	Grassland
				Catena 3 (C3)			
C3S1	452983.5	7511726	Interfluve	28.18%	1,263 m	Grazing	Grassland
C3S2	452889.6	7511599	Shoulder	38.29%	1,212 m	Grazing	Grassland
C3S3	452818.5	7511495	Midslope	54.57%	1,181 m	Grazing	Bushes
C3S4	452684.3	7511364	Footslope	1.29%	1,157 m	Preservation area	Trees
				Catena 4 (C4)			
C4S1	453210.9	7512117	Interfluve	54.50%	1,185 m	Grazing	Grassland
C4S2	453190.6	7512234	Shoulder	69.25%	1,127 m	Grazing	Grassland
C4S3	453161.6	7512375	Midslope	57.74%	1,078 m	Grazing	Grassland
C4S4	453195.5	7512504	Footslope	9.07%	1,049 m	Grazing	Grassland

The resulting map shows that Acrisol is the predominating class in the studied basin, which was expected, considering that it was present in all studied catenas. These findings are similar to those obtained by Lima (2021) when studying the neighboring larger basin of José Pereira. The author reported that in 8 studied catenas, 52% of soil profiles were classified as Acrisols, followed by the classes Cambisol (26%) and Gleysol (7%) – a soil classes distribution similar to that found in this work. The predominance of the Acrisol class in this region is also supported by maps at smaller scales such as the available map for the state of Minas Gerais (Figure 5), which similarly indicates that most mapping units include the referred soil class.

Rhodic Lixisols and Acrisols were mainly present at the highest parts of hillsides across the right margin of the Anhumas River, followed by the soil class Cambisol under strongly undulated and mountainous relief. At the left side of Anhumas River, where hilly relief predominates, the class Rhodic Lixisols and Acrisols were also found at hilltops. However, at lower portions of hillsides, Ferralsols predominated, especially Haplic Ferralsols under gently undulated relief. The valley of Anhumas River was characterized by the presence of Gleysols, classes that remain waterlogged during most part of the year. Reduction mottles could be observed in the studied Gleysol profiles, reflecting the redoximorphic conditions due to the perched water table.

Digital soil mapping

Part I - Basin soils map

The predicted map for the Anhumas River Basin derived from the manually created map (section 3.2) achieved a global accuracy value of 66.96% and a Kappa coefficient of 0.62 (Table 3 and Figure 6). Obtained results were similar to those observed by Coelho and Giasson (2010) and Ten Caten et al. (2011). Coelho and Giasson (2010) obtained a global accuracy of 67.31% and a Kappa coefficient of 0.39 when using the J48 algorithm to evaluate the agreement between the map created via digital soil mapping and the preexisting soil map for their study area. Ten Caten et al. (2011) observed a global accuracy of 61.79% and a Kappa coefficient of 0.46 using the same algorithm.

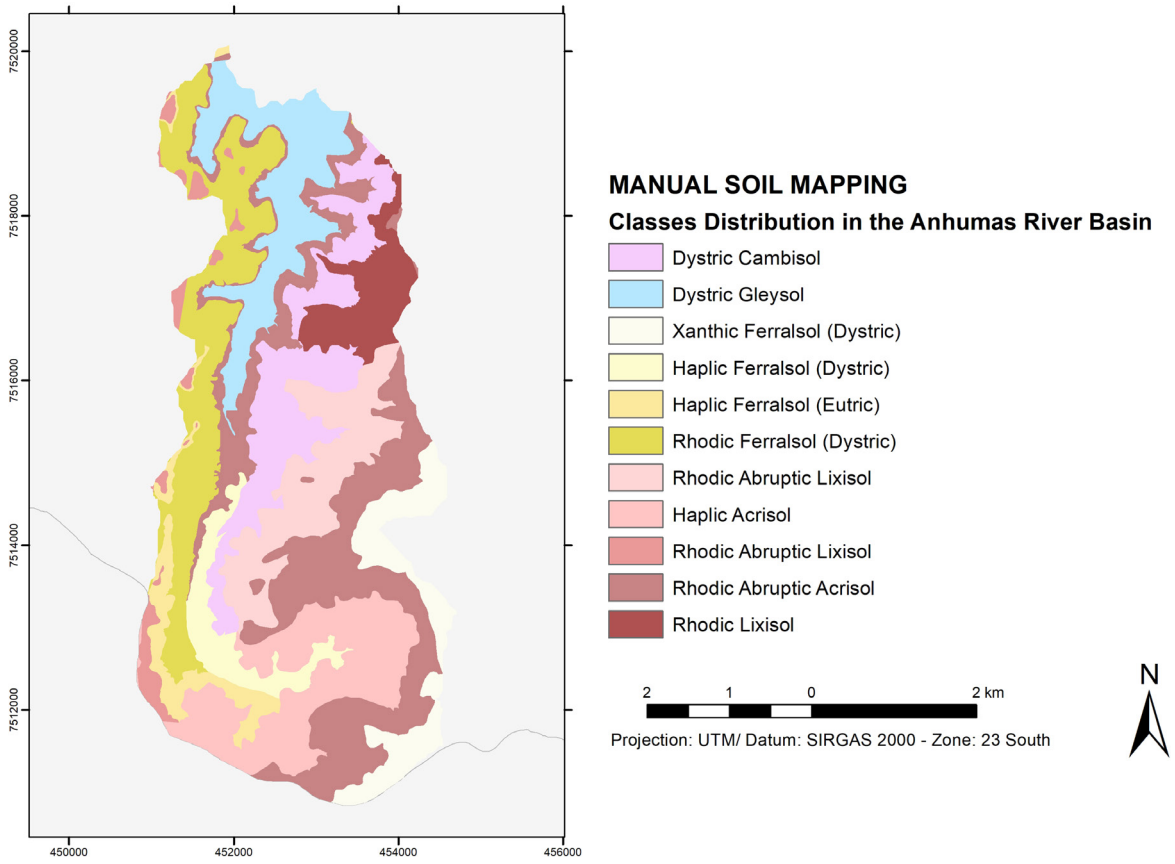

Figure 5: Soil classes distribution at the Anhumas River Basin, state of Minas Gerais, Brazil, created manually.

Table 2: Soil classification according to the Brazilian Soil Classification System for soils of the 4 catenas in the municipality of Itajubá, state of Minas Gerais, Brazil.

	Catena 1	Catena 2	Catena 3	Catena 4
	Soil Classification			
Site 1	Rhodic Lixisol	Rhodic Abruptic Lixisol	Rhodic Abruptic Acrisol	Xanthic Ferralsol (Dystric)
Site 2	Abruptic Acrisol	Haplic Ferralsol (Eutric)	Haplic Acrisol	Rhodic Abruptic Acrisol
Site 3	Eutric Cambisol	Rhodic Ferralsol (Dystric)	Haplic Acrisol	Abruptic Acrisol
Site 4	Dystric Cambisol	Rhodic Abruptic Acrisol	Dystric Gleysol	Haplic Ferralsol (Dystric)

Errors in prediction maps tend to occur between neighboring soil classes in the landscape (Ten Caten et al., 2011). Digital soil mapping groups available information during the machine learning phase and creates a classification model; however, classes that occupy only a small portion of the area, or classes that occupy areas with similar landscape attributes are more difficult to be correctly grouped, causing confusion for the algorithm. That can be observed, for instance, in classes PVAd1 and PVAd2, which presented low accuracy compared to other soil classes. These two classes are situated in very similar portions of the landscape (Table 3). This difficulty was noted by Láng et al. (2016), who mentioned that some soil classes may have been formed under similar environments, making them difficult to separate and consequently reducing the prediction efficiency of models.

Table 3: Confusion matrix scores of the soil classes prediction map created for the Anhumas River Basin, Itajubá, in the state of Minas Gerais, Brazil.

Mapping Unit	Soil Classes	User's Accuracy (%)	Producer's Accuracy (%)
CXd	Dystric Cambisol	78.75	87.26
GXd	Dystric Gleysol	75.36	70.8
LAd	Xanthic Ferralsol (Dystric)	51.56	50.11
LVAd	Haplic Ferralsol (Dystric)	20.21	48.72
LVAe	Haplic Ferralsol (Eutric)	32.7	57.5
LVd	Rhodic Ferralsol (Dystric)	57.22	51.2
PVAd1	Abruptic Acrisol	21.1	43.86
PVAd2	Haplic Acrisol	71.01	56.07
PVAe	Rhodic Abruptic Lixisol	73.48	73.48
PVd	Rhodic Abruptic Acrisol	86.97	80.43
PVe	Rhodic Lixisol	74.33	68.54
Kappa Coefficient		0.6205	
Global Accuracy		0.6696	

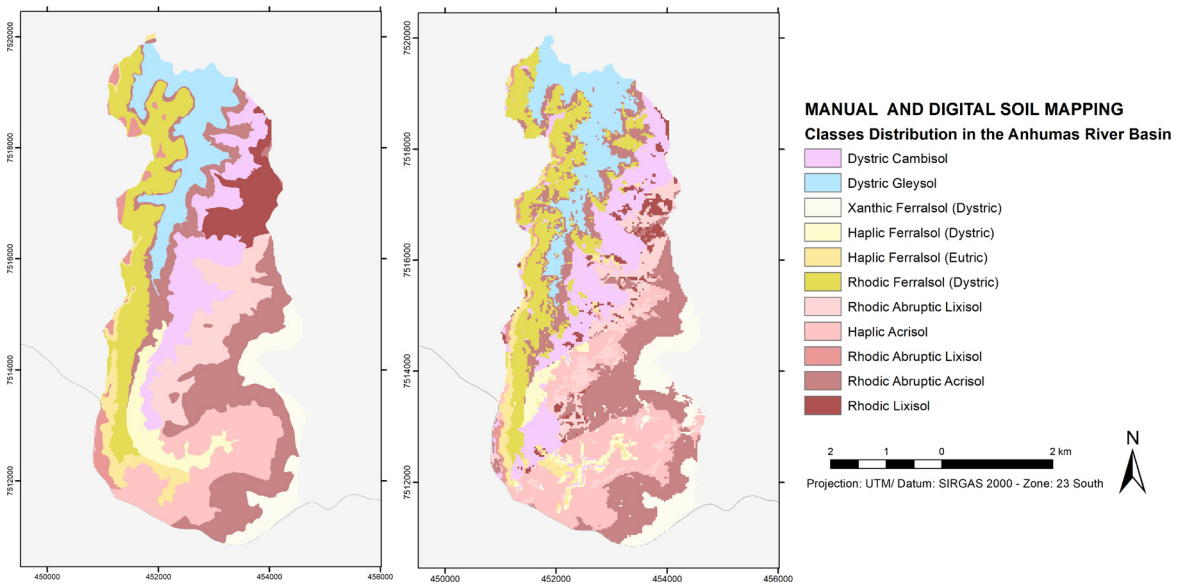

Figure 6: Comparison between (a) a soil map created manually via geoprocessing and (b) by digital soil mapping created by decision trees algorithm for the Anhumas River Basin, Itajubá, state of Minas Gerais, Brazil.

Conversely, the opposite behavior can be observed with the soil class GXd. This soil class occurs in the fluvial valleys, which have morphologic characteristics that are very specific. Therefore, GXd is easily identified by the algorithm, not creating the confusions seen between classes such as PVAd1 and PVAd2. The same occurs with the soil class CXd, which is usually formed under steeper slopes, and hence is less prone to be mistaken with other soil classes. As a result, both of these classes how high producer's accuracy values compared to the remaining ones: 87.26% (CXd) and 70.80% (GXd) (Table 3). It must be noted, however, that greater geographical expression also influences accuracy, raising the producer's accuracy values of classes such as CXd and PVd (80.40%). Láng et al. (2016) have also observed better

validation scores for Acrisols predictions, as this class had the biggest number of samples in the authors' dataset.

The soil classes PVAd1 and LVAd presented the lowest producer's accuracy values, with results below 50%, indicating high omission error. This can be explained by the presence of both soil classes in very similar landscape positions, located under undulated relief with predominance of convex pedoforms. Both classes also present the lowest user's accuracy values – 21.10% and 20.21%, respectively – indicating the difficulty of the algorithm in predicting them.

Part II - Municipality soils map

The map resulting from the extrapolation of the Anhumas River Basin soil map in tandem with data from Lima (2021) (Figure 7) using all variables, including parent material, showed that confusion occurred when predicting lithologic transitions. Soil classes abruptly changed, as could be observed especially in the northern part of the map. This is not expected to happen because soil classes tend to occur as a continuum in the landscape. Also, the scale of the parent materials map is too small.

A similar difficulty was reported by Chagas, Vieira and Fernandes Filho (2013) when trying to map soils digitally using neural networks. The authors stated that the lithologic heterogeneity and poor geologic map resolution were the main causes of the unsatisfactory soil maps. Parent materials dictate the substrate will be worked out by the other four soil forming factors and by the processes of soil formation, originating the current soil classes, and are therefore considered key variables for soil classification algorithms. Adhikari et al. (2014) reported that lithology was among the most important variables for their decision tree model when mapping Denmark, and Gray, Bishop and Wilford (2016) showed that lithology is often the most important variable to predict several soil attributes.

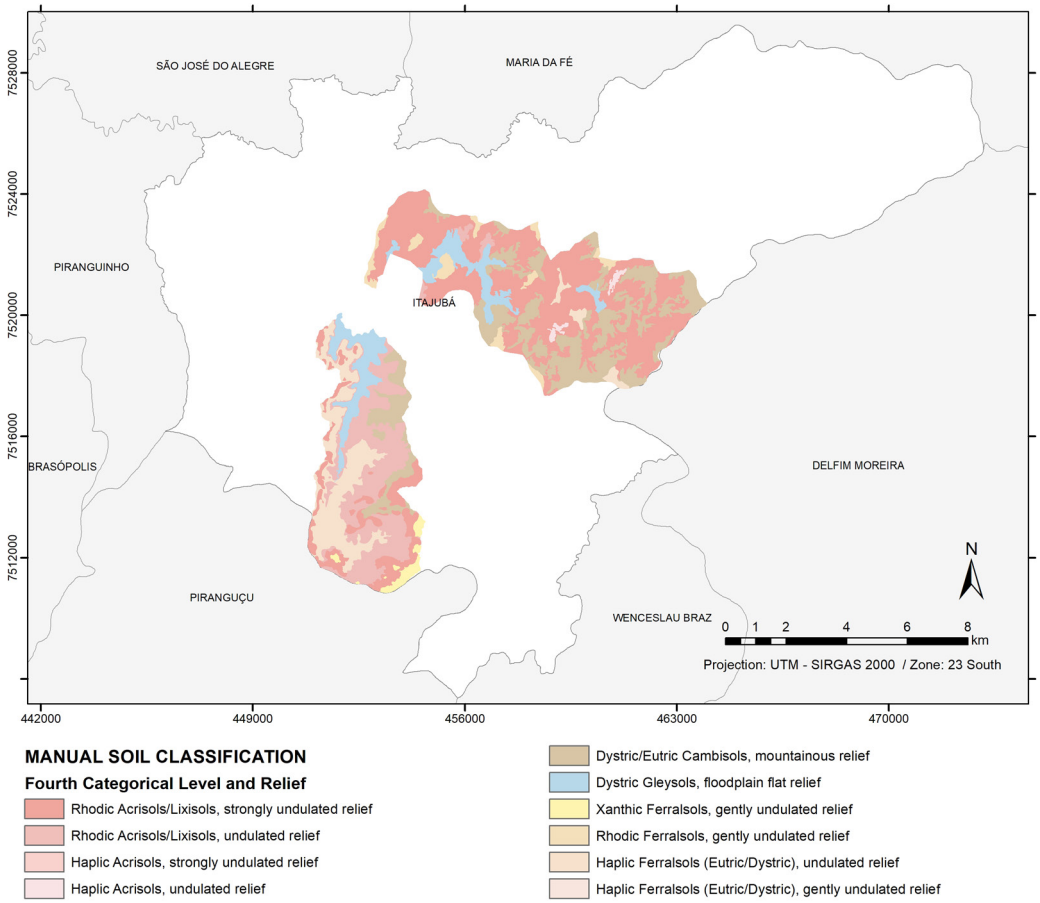

Figure 7: Soil maps used to train soil classes prediction models for the entire municipality of Itajubá, Brazil. The Anhumas River Basin map was created by this work and the José Pereira Basin map was adapted from Lima (2021).

Yet, the lack of detailed parent material maps is a recurring problem in Brazil (Mancini et al., 2019). The available parent material information for this area, as well for many regions of Brazil, is often not detailed enough for municipality soil maps. Höfig, Giasson and Vendrame (2014) mention this difficulty, reporting that the small scale of available lithological maps was not adequate to their work.

Hence, another map was created without parent material information. The defined map units are specified in Table 4. The map without parent material data showed more coherent results (Figure 8). As expected, the soil class Acrisol predominated in areas of strongly undulated relief, and Gleysols were present in the fluvial valleys. Transitions between soil classes were smoother and considered more realistic. This map was considered representative and was therefore validated. The obtained global accuracy was 50.80%, and the Kappa coefficient was 0.38 (Table 4).

The trained model had difficulty separating PVA1, LA, and LVA1 (Table 4). These soil classes occur at the same landscape positions, and hence are more difficult to be accurately predicted. The lowest values for User's accuracy were from PVA2 and CXd, 4.37% and 6.96%, respectively. These two soil classes also had the lowest results for producer's accuracy: 23.88% and 34.10%, respectively. They were mainly mistaken by PV2. This might be explained by the high representativeness of PV2 in this area: the classes Rhodic Acrisol and Lixisol occupies great part of the area in the same landscape positions as PVA2 and CXd.

To minimize this confusion between soil classes, another attempt was done. Considering that the fourth categorical level (Subgroup) in the Brazilian Classification System is too specific and demands too much information to accurately extrapolate to a much larger similar area, a map was created using mapping units that combine classes into broader groups, with classifications only up to the second categorical level (Suborder) (Tables 4 and 5). By reducing the specificity of soil classes, the algorithms might be less confused by soils occurring at similar landscape positions and therefore may provide more reliable predictions. The new soil classes prediction map is presented in Figure 9.

It must be noted that reducing the taxonomic levels does not incur losing the information gain when comparing the preexisting map for the area and the one produced by this work. The referred soil map for the region comprises three categorical levels, but is available at the scale of 1:650,000, a scale not sufficient to grasp the plurality of classes in the area and support localized decision-making. Despite including two categorical levels, the map generated by this work (Figure 9), at the scale of 1:25,000, illustrates a more accurate and realistic representation of the distribution of soil classes in the studied area.

Table 4: Confusion matrix scores of the soil classes prediction map created for the municipality of Itajubá, in the state of Minas Gerais, Brazil.

Mapping Unit	Soil Classes	User's Accuracy (%)	Producer's Accuracy (%)
CX	Dystric/Eutric Cambisols, mountainous relief	6.96	30.16
GX	Dystric Gleysols, floodplain flat relief	22.32	34.1
LA	Xanthic Ferralsols, gently undulated relief	52.89	45.74
LV	Rhodic Ferralsols, gently undulated relief	28.78	37.9
LVA1	Haplic Ferralsols (Eutric/Dystric), undulated relief	76.66	56.97
LVA2	Haplic Ferralsols (Eutric/Dystric), gently undulated relief	18.12	42.28
PV1	Rhodic Acrisols/Lixisols, undulated relief	73.27	65.4
PV2	Rhodic Acrisols/Lixisols, strongly undulated relief	28.95	70.97
PVA1	Haplic Acrisols, undulated relief	36.8	46.76
PVA2	Haplic Acrisols, strongly undulated relief	4.37	23.88
Kappa Coefficient	0.38		
Global Accuracy	0.508		

[1] With addition of the relief phase.

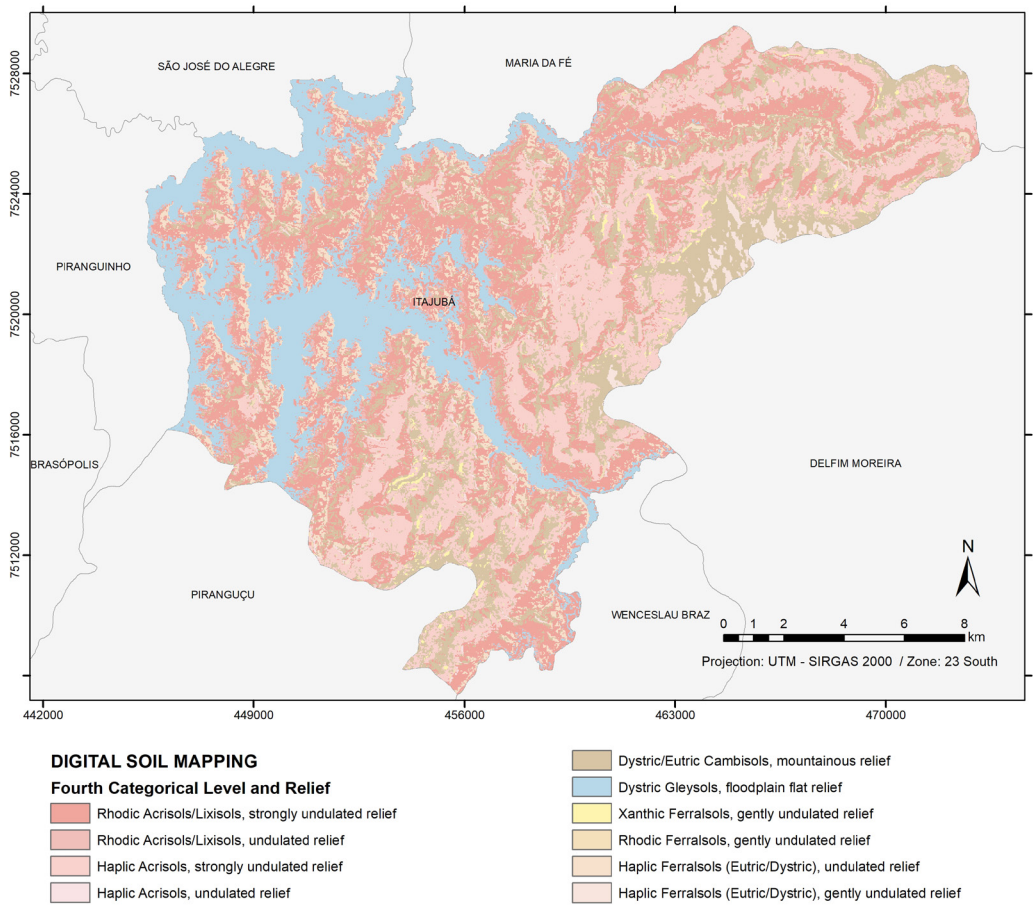

Figure 8: Digital soil mapping considering the fourth categorical level and relief.

Table 5: Confusion matrix of the soil classes prediction map created for the municipality of Itajubá, in the state of Minas Gerais, Brazil.

Mapping Unit	Soil Classes	User's Accuracy (%)	Producer's Accuracy (%)
CX	Cambisol	37.52	48.55
GX	Gleysol	60.52	51.19
LA	Xanthic Ferralsol	77.37	59.97
LV	Rhodic Ferralsol	4.89	24
LVA	Haplic Ferralsol	67.96	71.84
PV	Rhodic Acrisol/Lixisol	20.74	38.68
PVA	Haplic Acrisol	16.72	40.34
Kappa Coefficient		0.4038	
Global Accuracy		0.5469	

The new map (Figure 9) again shows remarkable presence of the soils pertaining to the class Acrisol, especially Haplic Acrisols, distributed across the whole Itajubá territory, where predominant relief varies from undulated to strongly undulated, along with the presence of Rhodic Acrisols/Lixisols close to the water bodies, neighboring the GXd class.

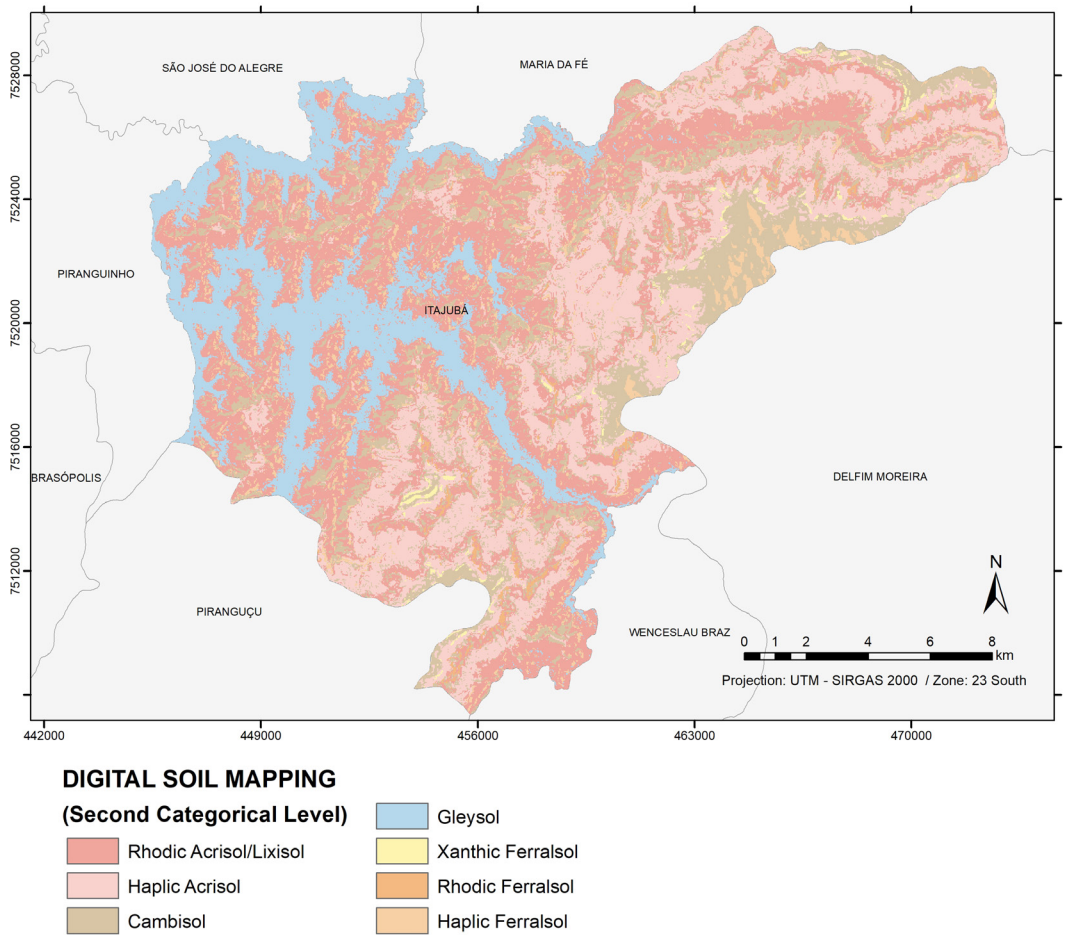

Figure 9: Soil map with mapping units up to the second categorical level (Suborder) for the municipality of Itajubá, Brazil, created by extrapolating soil data from two basins within the territory of the municipality.

This new soil map (Figure 9) achieved a global accuracy of 54.69% and a Kappa coefficient of 0.40 (Table 5). Ten Caten et al. (2011), attempting a similar technique, built a soil map at the scale of 1:50,000 via logistic regression using soil classification up to the second categorical level. Validation scores observed by the authors were: 39.29% global accuracy and 0.21 Kappa coefficient. Wolski et al. (2017) increased accuracy by classifying soils using the second categorical level via the J48 algorithm, achieving global accuracy of 66.1% and Kappa coefficient of 0.36 when validating with points obtained from field observations in a work at the scale of 1:50,000.

Compared to the previously available soil map (1:650,000), which featured only Haplic Acrisols, Rhodic Acrisols and their association with classes Cambisol and Ferralsol (4 classes in 3 mapping units) (Figure 4), the map produced by this work identified 7 different mapping units. The area occupied by each soil class is shown in Table 6.

The extrapolation of existing soil maps from small reference areas to larger similar target regions is an ongoing challenge (Krasilnikov; Targulian, 2019; Mello et al., 2021). The adoption of digital soil mapping techniques as a mean to improve existing soils mapping has been consolidating itself during the last few years, increasing its accuracy and reliability with time. Albeit validation scores may not be high (Table 5), the resulting maps are a significant improvement for areas where soil information is generalized. Maps produced by these methods make good use of available legacy data and are faster and inexpensive (Pásztor et al., 2018). Further studies on how to extrapolate soil data in developing tropical regions are needed and highly advised, as detailed soil data is a key factor for supporting decision-making in agricultural, environmental and urban endeavors.

Table 6: Area occupied by each soil class in the municipality of Itajubá, in the state of Minas Gerais, Brazil.

Map Unit	Soil Classes	Area (km²)	% of occurrence
CX	Cambisol	52.675	17.88
GX	Gleysol	47.142	16
LA	Xanthic Ferralsol	1.739	0.59
LV	Rhodic Ferralsol	2.891	0.98
LVA	Haplic Ferralsol	13.232	4.49
PV	Rhodic Acrisol/Lixisol	101.988	34.62
PVA	Haplic Acrisol	74.907	25.43
Total		294.575	100

CONCLUSIONS

The best map obtained from the extrapolation of the existing soil maps to the entire municipality of Itajubá using two reference areas (Anhumas River Basin and José Pereira River Basin) achieved a 54% global accuracy and a 0.40 Kappa coefficient. These validation scores are comparable to those found in the literature for similar methodologies. The resulting soil map was significantly more detailed than the state soil map for the area of interest. Parent material information and the use of very specific soil classes (up to the fourth categorical level - Subgroup -) confused models. Parent material information was not detailed enough and therefore could not help to infer soil classes distribution. The best prediction results were obtained by not using parent material data and reducing classification specificity to the second categorical level (Suborder). The main findings were: The state soils map (1:650,000) had only 3 pedological mapping units. Our soil map featured 7 pedological mapping units, resulting in more detailed soils information. Soil maps were built based on existing data from a neighboring larger basin and data obtained from a soil survey in a smaller area. This means reduced cost and time. This is important for every country, but is an important key especially for developing tropical countries like Brazil. Undetailed parent material information compromised results, yielding inaccurate maps. More detailed lithological information is needed in Brazil to assess its importance in soil prediction for tropical conditions. Reducing soil classes specificity (using second instead of fourth categorical level) avoids prediction errors and may result in more reliable maps.

REFERENCES

ADHIKARI, K. et al. Constructing a soil class map of denmark based on the FAO legend using digital techniques. **Geoderma**, 214-215(1):101-113, 2014.

AFSHAR, A. F.; AYOUBI, S.; JAFARI, A. The extrapolation of soil great groups using multinomial logistic regression at regional scale in arid regions of Iran. **Geoderma**, 315(1):36-48, 2018.

ALVARES, C. A. et al. Köppen's climate classification map for Brazil. **Meteorologische Zeitschrift**, 22(1):711-728, 2013.

ANGELINI, M. E. et al. Extrapolation of a structural equation model for digital soil mapping. **Geoderma**, 367(15):114226, 2020.

ARRUDA, G. P. de. et al. Digital soil mapping using reference area and artificial neural networks. **Scientia Agricola**, 73(3):266-273, 2016.

BRAGA, K. A. A. F.; SILVA, F. F.; SCHAFFRATH, V. R. Microbacia do Igarapé do Gigante: Unidade de planejamento para a gestão da bacia do Tarumã. **Revista em Agronegócios e Meio Ambiente**, 5(1):103-129, 2012.

CAMERA, C. et al. A high resolution map of soil types and physical properties for Cyprus: A digital soil mapping optimization. **Geoderma**, 285(1):35-49, 2017.

CHAGAS, C. S.; VIEIRA, C. A. O.; FERNANDES FILHO, E. I. Comparison between artificial neural networks and maximum likelihood classification in digital soil mapping. **Revista Brasileira de Ciência do Solo**, 37(2):339-351, 2013.

COELHO, F. F.; GIASSON, E. Métodos para mapeamento digital de solos com a utilização de sistemas de informação geográfica. **Ciência Rural**, 40(10):2099-2106, 2010.

DIAS, L. M. S. et al. Predição de classes de solo por mineração de dados em área da bacia sedimentar do São Francisco. **Revista de Pesquisa Agropecuária Brasileira**, 51(9):1396-1404, 2016.

GEE, G. W.; BAUDER, J. W. Particle-size Analysis. In: KLUTE, A. (Ed.). **Methods of soil analysis**: Physical and mineralogical methods. Madison: American Society of Agronomy, p.383-411, 1986.

GRAY, J. M.; BISHOP, T. F. A.; WILFORD, J. R. Lithology and soil relationships for soil modelling and mapping. **Catena**, 147:429-440, 2016.

HÖFIG, P.; GIASSON, E.; VENDRAME, P. R. S. Mapeamento digital de solos com base na extrapolação de mapas entre áreas fisiograficamente semelhantes. **Revista Agropecuária Brasileira**, 49(12):958-966, 2014.

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. World Soil Resources Reports No. 106. Rome: FAO, 2015. 203p.

KRASILNIKOV, P. V.; TARGULIAN, V. O. Towards "New Soil Geography": Challenges and solutions. A review. **Eurasian Soil Science**, 52:131-139, 2019.

LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. **Biometrics**, 33(1):159-174, 1977.

LÁNG, V. et al. Deriving world reference base reference soil groups from the prospective global soil map product - A case study on major soil types of Africa. **Geoderma**, 263(1):226-233, 2016.

LIMA, O. et al. Soil catenas in a pilot sub-basin in the region of Itajubá, Minas Gerais state, Brazil, for environmental planning. **Semina. Ciências Agrárias**, 42(3):1511-1528, 2021.

MALLAVAN, B. P.; MINASNY, B.; MCBRATNEY, A. B. Homosoil: A methodology for quantitative extrapolation of soil information across the globe. In: BOETTINGER, J. L. et al. (Eds.). **Digital soil mapping**: Bridging research, environmental application, and operation. London: Springer, p.137-149, 2010.

MALONE, B. P. et al. Comparing regression-based digital soil mapping and multiple-point geostatistics for the spatial extrapolation of soil data. **Geoderma**, 262(15):243-253, 2016.

MANCINI, M. et al. Parent material distribution mapping from tropical soils data via machine learning and portable X-ray fluorescence (pXRF) spectrometry in Brazil. **Geoderma**, 354:113885, 2019.

MCLEAN, E. O. Soil pH and lime requirement. In: PAGE, A. L. **Methods of soil analysis**. Madison: Soil Science Society of America, p.199-224, 1982.

MELLO, F. A. O. et al. Expert-based maps and highly detailed surface drainage models to support digital soil mapping. **Geoderma**, 384:114779, 2021.

MINAI, J.; LIBOHOVA, Z.; SCHULZE, D. G. Disaggregation of the 1:100,000 Reconnaissance soil map of the Busia Area, Kenya using a soil landscape rule-based approach. **Catena**, 195(1):104806, 2020.

PADARIAN, J.; MINASNY, B.; MCBRATNEY, A. B. Machine learning and soil sciences: A review aided by machine learning tools. **Soil**, 6(1):35-52, 2020.

PÁSZTOR, L. et al. Compilation of a national soil-type map for Hungary by sequential classification methods. **Geoderma**, 311:93-108, 2018.

PENG, Y. et al. Identifying and mapping terrons in Denmark. **Geoderma**, 363(1):114174, 2020.

PIIKKI, K.; SÖDERSTRÖM, M.; STADIG, H. Local adaptation of a national digital soil map for use in precision agriculture. **Advances in Animal Biosciences**, 8(2):430-432, 2017.

RIBEIRO, A. C.; GUIMARÃES, P. T. G.; ALVAREZ, V. H. Comissão de Fertilidade do Solo do Estado de Minas Gerais - CFSEMG. **Recomendação para o uso de corretivos e fertilizantes em Minas Gerais**: 5ª Aproximação. Viçosa, MG, Embrapa/UFV/SBCS, Cap. 5, p.25-32, 1999.

RUHE, R. V. Geomorphic surfaces and the nature of soils. **Soil Science**, 82(6):441-445, 1956.

SANTOS, H. G. dos. et al. **Sistema Brasileiro de Classificação de Solos**. 5ª ed. Brasília, DF: Embrapa, 2018. 353p.

SANTOS, R. D. et al. **Manual de descrição e coleta de solos no campo**. 6.ed. Viçosa: Sociedade Brasileira de Ciência do Solo, 2015. 102p.

SCULL, P.; FRANKLIN, J.; CHADWICK, O. A. The application of classification tree analysis to soil type prediction in a desert landscape. **Ecological Modeling**, 181(1):1-15, 2005.

SERVIÇO GEOLÓGICO DO BRASIL - CPRM. **Mapa de unidades litológicas**. Escala 1:100,000. Brasília: CPRM, 2014. Available in: <http://www.portalgeologia.com.br/index.php/mapa/>. Access in: Apr, 04, 2021.

SILVA, S. H. G. et al. Proximal sensing and digital terrain models applied to digital soil mapping and modeling of Brazilian latosols (Oxisols). **Remote Sensing**, 8(8):614, 2016.

STORY, M.; CONGALTON, R. G. Accuracy assessment: A user's perspective. **Photogrammetric Engineering and Remote Sensing**, 52(3):397-399, 1986.

TEIXEIRA, P. C. et al. **Manual de métodos de análise de solo**. 3ª ed. rev. e ampl. Brasília, DF: Embrapa, 2017. 574p.

TEN CATEN, A. et al. Extrapolação das relações solo-paisagem a partir de uma área de referência. **Ciência Rural**, 41(5):812-816, 2011.

TESKE, R.; GIASSON, E.; BAGATINI, T. Comparação de esquemas de amostragem para treinamento de modelos preditores no mapeamento digital de classes de solos. **Revista Brasileira de Ciências do Solo**, 39(1):14-20, 2015.

UFV-CETEC-UFLA-FEAM. **Mapa de solos do Estado de Minas Gerais**. Belo Horizonte, MG, 2010. 49p. Available in: http://www.feam.br/noticias/1/949-mapas-de-solo-do-estado-de-minas-gerais. Access in: April, 04, 2021.

VETTORI, L. **Métodos de análise de solo**. Rio de Janeiro: Ministério da Agricultura, 1969. 24p.

VINCENT, S. et al. Spatial disaggregation of complex Soil Map Units at the regional scale based on soil-landscape relationships. **Geoderma**, 311(1):130-142, 2018.

WALKLEY, A.; BLACK, I. A. An examination of the degtjareff method for determining soil organic matter, and a proposed modification of the chromic acid titration method. **Soil Science**, 37(1):29-38, 1934.

WITTEN, I. H.; FRANK, E. **Data mining**: Practical machine learning tools and techniques. 2.ed. Burlington: Morgan Kaufmann Publishers, 2005. 558p.

WOLSKI, M. S. et al. Digital soil mapping and its implications in the extrapolation of soil-landscape relationships in detailed scale. **Revista de Pesquisa Agropecuária Brasileira**, 52(8):633-642, 2017.

ZHANG, G.; LIU, F.; SONG, X. Recent progress and future prospect of digital soil mapping: A review. **Journal of Integrative Agriculture**, 16(12):2871-2885, 2017.