University of São Paulo
"Luiz de Queiroz" College of Agriculture

Geotechnologies applied in digital soil mapping

Wanderson de Sousa Mendes

Thesis presented to obtain the degree of Doctor in Science.
Area: Soil and Plant Nutrition

Piracicaba
2020

Wanderson de Sousa Mendes
Agronomist

Geotechnologies applied in digital soil mapping

Advisor:
Prof. Dr. **JOSÉ ALEXANDRE MELO DEMATTÊ**

Thesis presented to obtain the degree of Doctor in Science. Area: Soil and Plant Nutrition

Piracicaba
2020

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA – DIBD/ESALQ/USP

Mendes, Wanderson de Sousa

Geotechnologies applied in digital soil mapping / Wanderson de Sousa Mendes. - - Piracicaba, 2020.

163 p.

Tese (Doutorado) - - USP / Escola Superior de Agricultura "Luiz de Queiroz".

1. Pedometria 2. Sensoriamento remoto 3. Espectroscopia 4. Mineralogia 5. Mapeamento digital de solos I. Título

À minha família e amigos que acompanham toda a minha jornada de superações

DEDICO

AGRADECIMENTOS

À minha família, da qual fazem parte minha tia Josélia Silva Sousa, minha avó Olindina Silva Sousa e minha mãe Joselita Silva Sousa Mendes, por serem a base da minha formação social e por todo apoio moral e suporte.

Ao meu amigo Antônio Vicente da Silva pelo apoio moral e suporte no início do doutorado.

Ao meu orientador e amigo Professor Doutor José Alexandre Melo Demattê por ter me concedido a oportunidade de crescer profissionalmente e sempre motivar durante o doutoramento.

À Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP, processo nº. 2016/26124-6) pelos recursos investidos na minha formação.

À Escola Superior de Agricultura “Luiz de Queiroz” (ESALQ/USP) por fornecer os meios físicos para o desenvolvimento das atividades.

Ao grupo de pesquisa de Geotecnologias em Ciência do Solo (GeoCiS) pela paciência, companheirismo e conhecimento compartilhado ao longo desse período.

Ao Departamento de Ciência do Solo e ao Programa de Pós-Graduação de Solos e Nutrição de Plantas por manterem um ambiente de excelência e apoio administrativo ao longo desse período.

Ao Professor Doutor Antonio Carlos Saraiva da Costa do Departamento de Agronomia da Universidade Estadual de Maringá pela parceria nas análises de ferro total e susceptibilidade magnética, fundamentais para a complementação deste trabalho.

À empresa Raízen por dados fornecidos conforme projeto com a Fundação de Estudos Agrários Luiz de Queiroz (processo nº. 87017).

À Associação dos Fornecedores de Cana de Piracicaba (AFOCAPI), representados na figura do Sr. Gabriel de Souza Camarinha do Departamento Técnico, por dados fornecidos à complementação deste trabalho.

Ao Professor Doutor Marcos Rafael Nanni do Departamento de Agronomia da Universidade Estadual de Maringá pela parceria na disponibilização de sensores e dados, fundamentais para a complementação deste trabalho.

Ao Professor Doutor Peterson Ricardo Fiorio do Departamento de Engenharia de Biossistemas da Escola Superior de Agricultura "Luiz de Queiroz" pela parceria na disponibilização de dados, fundamentais para a complementação deste trabalho.

Aos Mestres Henrique Bellinaso, Diego Fernando Urbina Salazar, Karina Patrícia Prazeres Marques e ao Doutor Rodnei Rizzo do grupo GeoCiS pela parceria na disponibilização de dados, fundamentais para a complementação deste trabalho.

À todos que, com boa intenção, colaboraram para a realização e finalização deste trabalho.

EPÍGRAFE

"Ora, quem experimenta uma sensação de dúvida e de admiração reconhece que não sabe; e é por isso que também aquele que ama o mito é, de certo modo, filósofo: o mito, com efeito, é constituído por um conjunto de coisas admiráveis.

De modo que, se os homens filosofaram para libertar-se da ignorância; é evidente que buscavam o conhecimento unicamente em vista do saber e não por alguma utilidade prática.

E o modo como as coisas se desenvolveram o demonstra: quando já se possuía praticamente tudo de que se necessitava para a vida e também para o conforto e para o bem-estar, então se começou a buscar essa forma de conhecimento".

Aristóteles. Metafísica. 982b 10-25.

SUMÁRIO

RESUMO

Geotecnologias aplicadas no mapeamento digital de solos

A civilização vive em um mundo de mapas os quais são vitais em nível regional e local para o delineamento das melhores práticas agrícolas. O solo é o substrato para o crescimento das plantas e, portanto, fundamental para atender a demanda alimentar. No entanto, a escala cartográfica desses mapas de solos, que para a melhor prática agrícola de manejo (MPAM), tem que ser a mais detalhada possível os quais atualmente são escassos. O Mapeamento Digital de Solos (MDS) tornou-se a abordagem mais fácil e viável para atingir essa demanda. Apesar de estudos anteriores terem tentado caracterizar melhor as profundidades do solo, há espaço para aperfeiçoamento em sua dinâmica e mapeamento. Tendo como foco este objetivo, as tecnologias de Sensoriamento Remoto (SR) provam ser uma grande ferramenta nesta tarefa. No entanto, alguns aspectos dessa abordagem ainda precisam ser testados usando outros modelos híbridos, estocásticos e determinísticos para as previsões de susceptibilidade magnética (SM) e atributos do solo na superfície e subsuperfície. Portanto, o capítulo 1 apresenta a avaliação de nove algoritmos de aprendizado de máquinas (AAMs) para predizer o teor de ferro livre na superfície do solo (0 - 20 cm) usando a estrutura do MDS. Com base no melhor desempenho desses nove AAMs, selecionamos cinco. O capítulo 2 mostra o uso desses cinco AAMs com variáveis ambientais usuais e novas (por exemplo, DEM, rede de drenagem e espectroscopia do solo) para predizer SM e atributos do solo até 100 cm de profundidade. Quanto a mineralogia dos solos, a quantificação dos minerais do solo atualmente consistem na análise de laboratório tradicional de solos. No entanto, desenvolvimentos na interpretação e análise da refletância difusa do visível e do infravermelho próximo (VNIR) permitem quantificar alguns dos minerais do solo. No capítulo 3, implementa-se uma nova metodologia usando espectroscopia VNIR para quantificar os principais minerais do solo e avalia a aplicação da estrutura de mapeamento digital do solo para espacializar esses minerais. Por fim, mas não menos importante, o capítulo 4 apresenta a inovação de usar todos os componentes do solo como preditores das unidades de mapeamento de solo na região de Piracicaba-SP em escala de fazenda (1: 20.000), gerando o primeiro mapa digital detalhado do solo da região. Adicionalmente neste capítulo, foi criado o mapa digital de ambientes de produção para cana-de-açúcar. Assim, esta tese apresenta uma nova estrutura metodologica e integrativa na obtenção de mapas digitais de solo em escala detalhada para a MPAM e serve como guia para futuros levantamentos de solos em todo o mundo.

Palavras-chave: Pedometria, Sensoriamento remoto, Espectroscopia, Mineralogia, Mapeamento digital de solos

ABSTRACT

Geotechnologies applied in digital soil mapping

The civilisation lives in a world of maps and soil maps are vital at regional and farm levels to achieve best management agricultural practices. Soil is the substrate for plant growth and vital to the fulfilment of the food demand. However, the cartographic scale of those soil maps, which for the best management agricultural practice (BMAP) have to be the most detailed as possible and they are scarce. The Digital Soil Mapping (DSM) became the easiest and feasible approach to achieve such demand. Despite previous studies have tried to better characterise soil depths, there is space for improvements on its dynamics and mapping. Looking at this goal, Remote Sensing (RS) technologies have proven to be a great power on this task. Nevertheless, some aspects of that approach still need to be tested using another hybrid, stochastic, and deterministic models for the predictions of magnetic susceptibility (MS) and soil attributes at surface and subsurface. Therefore, chapter 1 presents the evaluation of nine machine learning algorithms (MLAs) to predict the free iron content at the soil surface (e.g. 0 – 20 cm) using the DSM framework. Based on the best performance of those nine MLAs, we selected five MLAs. Chapter 2 shows the use of those five MLAs with usual and new environmental variables (e.g. DEM, drainage network, and soil spectroscopy) to predict the MS and soil attributes up to 100 cm depth. Attempts on quantifying soil mineral consist of having an observation measured using traditional laboratory soil analysis. However, developments in interpreting and analysing the visible and near-infrared (VNIR) diffuse reflectance have allowed quantifying some soil minerals. In chapter 3, it implements a novel framework using VNIR spectroscopy to quantify the main soil minerals and evaluates the application of digital soil mapping framework to spatialise those soil minerals. Last but not least, the chapter 4 presents the novelty of using all predict soil components as predictors of the soil mapping units in the region of Piracicaba-SP at farm scale (1:20,000), generating the first detailed digital soil map of the region. Additionally in this chapter, it was created the digital yield environmental map for sugarcane production. Thus, this thesis presents a new integrative framework to achieve detail soil maps for the BMAP and serves as a guide for future soil surveys across the world.

Keywords: Pedometrics, Remote sensing, Soil spectroscopy, Mineralogy, Digital soil mapping

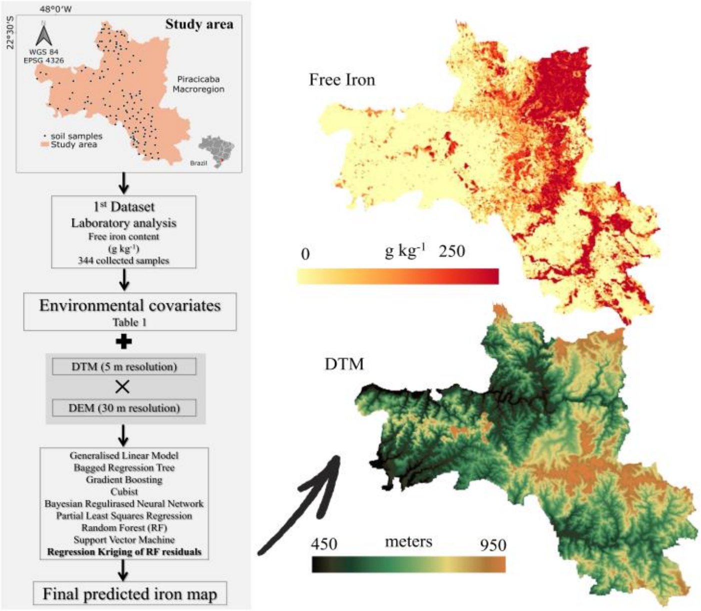
1. FREE IRON OXIDE CONTENT IN TROPICAL SOILS DETECTED BY INTEGRATIVE DIGITAL MAPPING

ABSTRACT

The free iron (FI) content in tropical soils is an important indicator of soil quality and can be used to infer soil genesis, classification, and distribution. Despite its importance in agriculture and pedology, laboratory analyses of soil iron content are costly and time-consuming. Remote sensing data combined with digital soil mapping procedures are useful tools for reducing the number of soil samples needed to characterise soil variability and, consequently, laboratory analysis costs. This study aimed to create a strategy for mapping FI content using a 35-year time series of Landsat images combined with topographic parameters at two spatial resolutions (5 and 30 m) as covariates and machine learning algorithms (MLAs) in a region with complex soil and geology in Brazil. The dataset comprised 344 FI observations at depths of 0–20 cm in a 2574 km² area. The dataset was split for calibration and external validation (85:15%), and the environmental covariates were chosen based on *scorpan* factors. We found that the temporal bare soil image improved model performance. Although 5 and 30 m resolution terrain data differed slightly, the best-fit model was obtained with the Random Forest 5 m resolution (root mean square error, 25.09 g kg^{-1}; adjusted R^2, 0.84). Among the evaluated MLAs, Random Forest was most suitable for predicting FI distribution in the study area. The FI map was crucial for identifying detailed soil types, which should be prioritised in future pedological studies.

Keywords: Machine Learning Algorithms; Regression Kriging; Pedometrics; Soil iron; Spectroscopy

Graphical Abstract



1.1. INTRODUCTION

Iron oxides are among the most abundant metallic oxides in soils, and iron is the element most frequently found in soil minerals (Hunt, 1980). Soil mineral forms, incidence, and concentrations vary spatially with environmental conditions (Schwertmann and Cornell, 2000; Viscarra Rossel et al., 2010). Iron species are most commonly classified as free, organic-bound, and amorphous iron. Goethite and haematite are the most common iron minerals in the tropics and are yellow-brown and red pigment agents in soils, respectively (Anda et al., 2008; Macedo and Bryant, 1989). Free iron (FI) is the main diagnostic group and comprises iron encrusted on the soil surface (Fe^{2+} and Fe^{3+}) but not embedded in the soil lattice structure (Fan et al., 2016). Thus, it is key to understanding soil genesis, classification, and distribution (Fan et al., 2016). Even though tropical soils contain trace amounts of iron oxides, these compounds can affect soil cation exchange capacity, particle aggregation, and colour (Viscarra Rossel et al., 2010).

Some pedogenetic processes are easily observed *in situ* because of the different chemical phases of iron oxides, which are directly influenced by the water table, temperature, pH, and soil redox conditions (Fan et al., 2014; Viscarra Rossel et al., 2010). For instance, in the Brazilian soil classification system, the pedogenetic process of gleisation produces Haplic Gleysols. This process is characterised by the presence of dissolved Fe^{2+} resulting from the stagnation of water in the soil profile and creating an anaerobic environment and grey soil layers. Ferralitisation, which involves the complete hydrolysis of iron to form haematite and goethite, characterises Haplic Ferralsols. Both processes occur under high soil iron concentrations and alteration of the water table.

Topographically, soils located in high, well-drained areas have a high haematite concentration and are typically reddish. Yellowish soils are generally found on foot slopes, where the anaerobic environment coupled with high organic matter content is conducive to high goethite concentrations (Macedo and Bryant, 1989); these are the typical patterns of tropical soils with high Fe_2O_3 (FI) concentrations. Furthermore, FI is an indicator of soil quality, fertility, and deposition age (Bartholomeus et al., 2007; Mulder et al., 2011). For example, most iron originates from mafic rocks, forming clayey and very clayey soils, which strongly affects other soil properties, such as water retention capacity and soil physical quality (Chagas et al., 2016). Therefore, FI is highly relevant to agricultural production, soil conservation, and pedology.

Despite the relevance of iron oxides to agriculture in the tropics, these minerals are generally poorly analysed in routine field assessments compared to other soil properties. This is likely because laboratory analyses are time-consuming and costly. Moreover, the chemical analysis procedure uses hazardous sulphuric acid as reagent. Nanni and Demattê (2006) pointed out that proximal and remote sensing can be alternatives to traditional soil analyses for quantifying soil attributes. Quantification of iron content using proximal (Demattê, 2002) and remote (Ben-Dor and Banin, 1995; Coyne et al., 1990) sensing has been found to be feasible. Coleman et al. (1993) were pioneers in quantifying soil iron based on Landsat Thematic Mapper data and found correlations of 0.10–0.29 between iron content and radiance data. Imaging spectrometry has also been used to locate and measure the specific absorption features of iron oxides and hydroxides in the visible and near-infrared (NIR) spectral region (Abrams and Hook, 1995; Mulder et al., 2011).

However, studies that solely used multispectral data from satellite sensors (e.g. Landsat data) did not accurately map iron content (Andrews Deller, 2006). Some spectral indices based on satellite data also did not provide accurate results. This was partially related to the loss of information when a single index instead of the entire soil spectrum was used (Levi and Rasmussen, 2014; Regmi and Rasmussen, 2018). Bartholomeus et al. (2007) used airborne hyperspectral data and found weak correlations between airborne spectral data and soil iron content. The reason for these weak relationships between spectral data and iron content was likely that external factors, such as partially covered or rough surfaces, physical and biological soil crusting, or atmospheric conditions, influenced the measurements, thus hampering the soil signals in the visible–NIR–short-wave infrared (SWIR) region (Xu et al., 2005).

Digital soil mapping (DSM) has improved with developing technology, becoming an excellent tool for decision makers, landowners, and landscape managers (McBratney et al., 2019). Modern technologies, such as proximal sensors (e.g. portable X-ray fluorescence), remote sensors (e.g. satellites), and big data analyses (e.g. new-generation algorithms and high-performance computers), can be used to create high-resolution soil maps. These tools can help to improve understanding of natural *in situ* resources and predict their variability and availability. FI is a well-studied soil attribute (Fan et al., 2016; Nandra, 1974; Shen et al., 2020) but has hardly been used in DSM. We therefore aimed to create a strategy for mapping FI by using multitemporal Landsat images to produce a bare soil image. This image, coupled with relief parameters, was used to train machine learning algorithms (MLAs) to predict FI content in a pedologically and geologically complex region in Brazil. The bare soil image could assist FI detection, since many laboratory spectroscopy studies have shown the strong relationship between FI content and soil spectral properties.

1.2. MATERIAL AND METHODS

1.2.1. Study area and dataset

The study area (2574 km^2) was located in São Paulo State, southeastern Brazil. The climate is dominated by two seasons, namely dry winters and rainy summers, with an annual average temperature of 20–22.5 °C and annual rainfall of 1200–1400 mm. Rolling uplands and undulating hills, with altitudes of 450–950 m, are common topographic characteristics. The main parent materials are Carboniferous siltstone, tillite, varvite, conglomerate, and sandstone (Tubarão Group), Permian shale, limestone, siltstone, and flint (Corumbataí Formation), Permian shale, dolomite, and siltite (Irati Formation), Jurassic sandstone, shale, and siltstone (Botucatu and Pirambóia Formations), and Cretaceous diabase and basalt (Serra Geral Formation) (Marconi, 1974). The diversity of parent materials and topography (plains to rolling hills) determines the varying FI content in the study area.

The dataset consisted of 344 field observations via samples collected with an auger at depths of 0–20 cm, based on the toposequence principle (Burrough, 2006). The sampling locations were distributed across different soil and landform types (Fig. 1). After sampling, the soil samples were oven-dried for 48 h at 50 °C, ground, and sieved through 2 mm mesh. The FI oxides were analysed by using sodium dithionite–citrate–bicarbonate extraction according to Mehra and Jackson (2013). Subsequently, 1 mL 0.1 N nitric acid solution was added to each sample to digest the organic matter. The samples were then placed in a block digester at 350 °C (slowly increasing) for 4 h, after which they were cooled and the solutions filtered. The FI content (Fe_2O_3) was analysed via atomic absorption spectrometry (Varian SpectrAA 10 Plus; λ, 324.80 nm; slit width, 0.2 nm).

The Brazilian Classification System (Santos et al., 2018) uses FI content to classify soils as follows: (i) hypoferric (< 80 g kg^{-1}), (ii) mesoferric (80–180 g kg^{-1}), (iii) ferric (180–360 g kg^{-1}), and (iv) perferric (> 360 g kg^{-1}). We classified our soil samples accordingly for exploratory analysis purposes. Furthermore, we used the predicted FI to relate our findings to the soil types of the Brazilian Classification System.

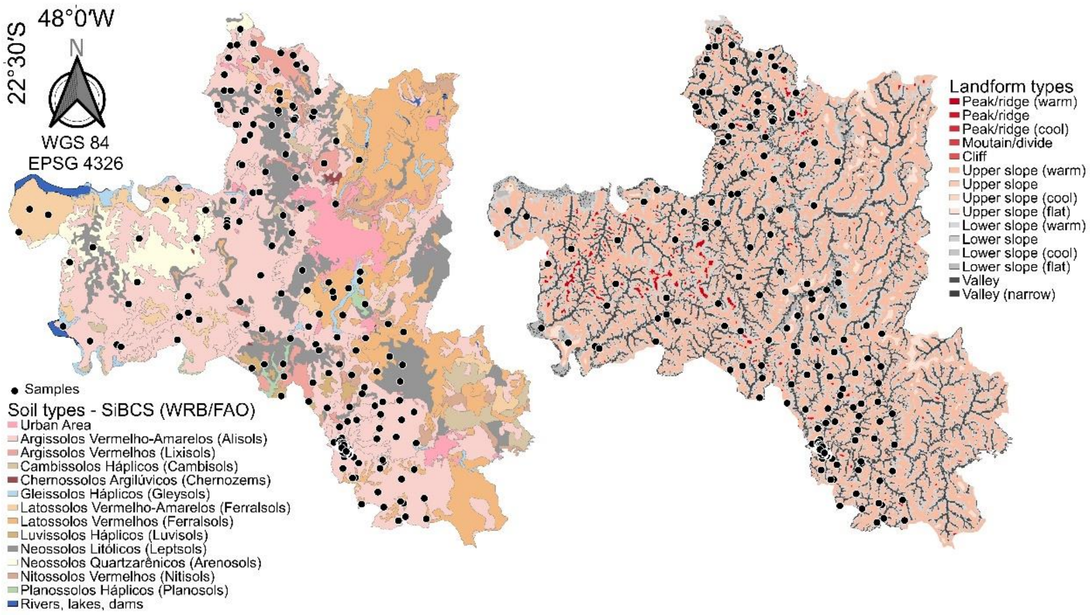

22°30'S
48°0'W
N
WGS 84
EPSG 4326
● Samples
Soil types - SiBCS (WRB/FAO)
Urban Area
Argissolos Vermelho-Amarelos (Alisols)
Argissolos Vermelhos (Lixisols)
Cambissolos Háplicos (Cambisols)
Chernossolos Argilúvicos (Chernozems)
Gleissolos Háplicos (Gleysols)
Latossolos Vermelho-Amarelos (Ferralsols)
Latossolos Vermelhos (Ferralsols)
Luvissolos Háplicos (Luvisols)
Neossolos Litólicos (Leptsols)
Neossolos Quartzarênicos (Arenosols)
Nitossolos Vermelhos (Nitisols)
Planossolos Háplicos (Planosols)
Rivers, lakes, dams
Landform types
Peak/ridge (warm)
Peak/ridge
Peak/ridge (cool)
Moutain/divide
Cliff
Upper slope (warm)
Upper slope
Upper slope (cool)
Upper slope (flat)
Lower slope (warm)
Lower slope
Lower slope (cool)
Lower slope (flat)
Valley
Valley (narrow)

Fig. 1. The distribution of the selected samples across soil types according to the Brazilian Classification System (SiBCS) and World Reference Base (WRB), and landform types from the SRTM landform in the study area.

1.2.2. Environmental covariates

The *scorpan* model is the basis of DSM (McBratney et al., 2003). To fulfil all aspects of the *scorpan* model, we generated a list of environmental covariates (Table 1) that were input into MLAs for soil prediction. The first was the *s* factor. Based on Demattê et al. (2020, 2018), we retrieved satellite images from Landsat archives from the United States Geological Survey (USGS) platform and created a unique temporal mosaic (dry season, July to September from 1984 to 2018) that represents pixels of bare soil areas. This mosaic was named Synthetic Soil Image (SYSI) and provided the spectral information of each pixel with bare soil on visible bands (1, blue; 2, green; 3, red), the NIR band (4), and SWIR bands (5, SWIR1; 7, SWIR2).

Table 1. Characteristics of the environmental covariates selected to be performed with the response variable, Total Free Iron content, in the digital soil mapping procedure.

***Scorpan*'s factors**	**Ancillary variables**	**Unit**	**Resolution (m)**	**Type of variable**	**Characteristics**
s	Synthetic Soil Image's bands	Spectral reflectance	30	Continuous	Bare soil areas
	RGB soil colour	adimensional	30	Continuous	Soil colour
c	Annual mean temperature	°C	30	Continuous	Temperature
	Annual mean precipitation	mm	30	Continuous	Rainfall
o	NDVI	adimensional	30	Continuous	Vegetation
	EVI	adimensional	30	Continuous	Vegetation
	Soil Relative Frequency	%	30	Continuous	Human activity
r	DEM from SRTM	m	30	Continuous	Relief
	Aspect from DEM	degree	30	Continuous	Downhill slope faces
	LS Factor from DEM	adimensional	30	Continuous	Component of the Revised Universal Soil Loss equation
	Plan Curvature from DEM	degree m^{-1}	30	Continuous	(-) concave/ (+) convex contours
	Profile Curvature from DEM	degree m^{-1}	30	Continuous	(-) convex/ (+) concave contours
	Slope from DEM	%	30	Continuous	Relief inclination
	Valley Depth from DEM	m	30	Continuous	Vertical distance to the base level of the channel network
	TWI from DEM	adimensional	30	Continuous	Soil water content
	DTM from IGC	m	5	Continuous	Relief
	Aspect from DTM	degree	5	Continuous	Downhill slope faces
	LS Factor from DTM	adimensional	5	Continuous	Component of the Revised Universal Soil Loss equation
	Plan Curvature from DTM	degree m^{-1}	5	Continuous	(-) concave/ (+) convex contours
	Profile Curvature from DTM	degree m^{-1}	5	Continuous	(-) convex/ (+) concave contours
	Slope from DTM	%	5	Continuous	Relief inclination
	Valley Depth from DTM	m	5	Continuous	Vertical distance to the base level of the channel network
	TWI from DTM	adimensional	5	Continuous	Soil water content
	Drainage Density	m^{-1}	30	Continuous	Drainage network
	Landforms	adimensional	30	Factor	Physiographic and landforms patterns
p	Gypsic index	adimensional	30	Continuous	Gypsiferous soil
	Natric index	adimensional	30	Continuous	Sodium rich soil
	Calcareous index	adimensional	30	Continuous	Discriminate calcareous sediments from igneous rocks or sediments
	Carbonate radicals index	adimensional	30	Continuous	Carbonate radicals
	Ferrous Fe index	adimensional	30	Continuous	Ferrous Fe
	Ferrous index	adimensional	30	Continuous	Ferrous
	Ferrous oxides index	adimensional	30	Continuous	Ferrous oxide
	Clay and hydroxides index	adimensional	30	Continuous	Clay and hydroxides
a	Geology	adimensional	30	Factor	Parent material
	Geomorphology	adimensional	30	Factor	Hillslope position
n	X	m	30	Continuous	Longitude
	Y	m	30	Continuous	Latitude

Regions without bare soil (e.g. forests, rivers, and perennial crops) were masked out and designated as 'not available' information, resulting in gaps without soil-surface spectral information. To fill these gaps, we applied the 'close gaps' function and Gaussian filter of the System for Automated Geoscientific Analyses version 2.3.2 (SAGA Development Team, 2016). This was performed to create a continuous SYSI map and obtain a complete representation of soil variability. In addition to the covariates that represent the *s* factor, we created a predicted soil RGB colour based on the soil spectral reflectance (Fig. S1).

For the *c* factor, we retrieved annual mean temperature and precipitation data from the WorldClim BIO Variables V1 (Hijmans et al., 2005) available from the Earth Engine Data Catalog. The normalised difference vegetation index (NDVI) (Rouse et al., 1973) and enhanced vegetation index (EVI) (Huete, 2004) were calculated by averaging 35-year Landsat bands (dry and moist seasons) and categorised as *o* factors. We used these spectral indices to summarise the Landsat bands into single representations determined by the index equations, which enhanced the differences between and signals of spatial patterns of different spectral regions. Additionally, we created an image that provided information on the number of times that each pixel was classified as a bare surface along with the historical collection, namely soil relative frequency (SRF, %). This product was used as an indicator of anthropogenic activities, because it shows the area of surface soils that have been disturbed over the historical period covered by SYSI.

For the *r* factor, we used and compared two terrain model sources. The first was retrieved from the Shuttle Radar Topography Mission from the USGS platform, which had a 30 m pixel resolution. The second digital elevation model (DEM) was created from the digitisation of a planialtimetric base map (with a scale of 1:10 000) with equidistant level curves of 5 m from the Geographic and Cartographic Institute of the State of São Paulo. Subsequently, we converted the vector to a raster with a 5 m pixel resolution using QGIS software (QGIS Development Team, 2020) and obtained a digital terrain model (DTM). The two DEMs were the basis for generating six relief features, namely aspect, LS factor, plan and profile curvature, slope, valley depth, and topographic wetness index. These were calculated in the System for Automated Geoscientific Analyses. Altogether, we obtained six relief covariates from the DEM with a spatial resolution of 30 m and another six from the DTM with a spatial resolution of 5 m. The DEM and DTM were used separately as covariates as well.

Another covariate generated as the *r* factor was drainage density and was created using 3D images of the study area. We used digital aerial photographs from the Geographic and Cartographic Institute and vectorised all channels (e.g. rivers, streams, and bases) in PHOTOMOD Lite 6.3 software. The following equation was used to calculate the drainage density (DD) in ArcGIS version 10.3 software: DD = total length channels (m)/basin area (m^2). The last covariate that represented the *r* factor was the Shuttle Radar Topography Mission landform (Theobald et al., 2015) retrieved from the Earth Engine Catalog. This covariate contained detailed multiscale data on physiographic and landform patterns with a spatial resolution of 90 m, which was downscaled and resampled to 30 m using the nearest neighbour method.

The SYSI spectral bands were used to create image indices that represented the *p* factor (Regmi and Rasmussen, 2018): gypsic index (SWIR1 – SWIR2/SWIR1 + SWIR2), natric index (SWIR1 – NIR/SWIR1 + NIR), calcareous sediment index (SWIR1 – Green/SWIR1 + Green), carbonate radicals (Red/Green), ferrous iron (SWIR2/Red), ferrous oxide (Red/Blue), ferrous (SWIR1/NIR), and clay/hydroxides (SWIR2/SWIR1). A geological map that represents the *p* factor was also used as a covariate. This map was created by Bonfatti et al. (2020), who used a DSM approach to characterise the soil parent material, which comprised alluvial deposits, sandstones, unconsolidated clay, basalt, shale, and siltstones.

An implicit age factor covariate (*a* factor) was created according to Marques et al. (2018). This covariate comprised a geomorphological map containing five groups: summit, shoulder, back slope, foot slope, and toe slope. For the *n* factor, we used geographical coordinates (X and Y) with the coordinate reference system EPSG code 32723.

1.2.3. Machine learning methods and geostatistical approach

Eight MLAs were tested: Cubist (Quinlan and Ross, 1993), Random Forest (RF) (Breiman, 2001), Generalised Linear Model (GLM) (Nelder, 1977), Bagged Regression Tree (BaRT) (Breiman, 1996), Stochastic Gradient Boosting (Friedman, 2002), Bayesian Regularised Neural Network (BRNN) (Ticknor, 2013), Partial Least Square Regression (PLSR) (Helland, 1988), and Support Vector Machine (SVM) (Vapnik, 2000). Cubist, RF, and BaRT are decision tree algorithms that differ based on their ways of dealing with variance reduction. RF and Cubist are widely used for DSM purposes and are well-described in the literature (Gray et al., 2016; Pouladi et al., 2019; Shahbazi et al., 2019a). BaRT is an algorithm that reduces the variation of a mathematical learning method as a general procedure (Keskin et al., 2019). GLMs attempt to adapt the model rather than changing the input data and involve a lengthening of linear regressions to accommodate non-normal response distributions (Lane, 2002).

Stochastic Gradient Boosting incorporates bagging aggregation and is a hybrid method that performs a small regression or grouping by using the residuals of the former trees building sequentially (Forkuor et al., 2017; Friedman, 2002). The BRNN trains and calculates non-trivial weights, converging them to a constant as network increases in size. The model complexity is reduced, and unnecessary linkages are changed to zero. The PLSR is similar to principal component analysis and integrates predictor variables, collinearly compressing them to construct a predictive model. Some PLSR factors may explain part of the variation between response and predictor variables. The SVM differs from decision tree methods because it uses kernel functions, converting their linear non-separable issues into separable ones (Bishop, 2006; Franklin, 2005). It is widely applied in DSM (Gomes et al., 2019; Liakos et al., 2018; Meier et al., 2018).

Additionally, we performed regression kriging (RK) using the best MLA result and its model residuals. RK, as highlighted by Keskin and Grunwald (2018), is extensively used in soil science because of its practicality and robustness as a hybrid spatial interpolator. Several studies have been conducted on RK, with the residuals of other machine learning methods in DSM (Angelini and Heuvelink, 2018; de Carvalho Junior et al., 2014; Knotters et al., 1995; Odgers et al., 2011; Pouladi et al., 2019; Sayão et al., 2018; Sindayihebura et al., 2017; Vasques et al., 2016).

1.2.4. Digital FI content mapping procedure

FI was calibrated using 44 environmental covariates (Table 1). MLAs and RK were performed using the 'caret' (Kuhn, 2008) and 'gstat' (Pebesma, 2004) packages in R software by randomly splitting the dataset into calibration (85%, 295 samples) and validation (15%, 49 samples) sets. The first set was used to calibrate the models. MLA parameters were optimised using a five-fold repeated cross-validation method, executed five times for each model to avoid the effects of environmental covariate autocorrelation (Bonfatti et al., 2020; Meyer et al., 2019). The best fit model was selected considering the lowest root mean square error (RMSE) and highest coefficient of determination (R^2) (Table 2). This was considered as internal validation. Subsequently, the final predicted maps were validated using the omitted samples. This procedure is known as external validation and provides a better indication of the model's generalisation.

Table 2. Model's parameters of the best fit for total iron content (g kg^{-1}) at 0 – 20 cm (layer A).

Models	**Terrain sources**	**Parameters**				
		$\mathrm{m_{try}}$		**$\mathrm{R^2_{train}}$**		
Random Forest	DEM	19		0.86		
	DTM	19		0.86		
		Committees		**Neighbours**		**$\mathrm{R^2_{train}}$**
Cubist	DEM	20		9		0.87
	DTM	20		9		0.86
		n. comp.		**$\mathrm{R^2_{train}}$**		
Partial Least Square Regression	DEM	3		0.66		
	DTM	3		0.66		
		Neurons		**$\mathrm{R^2_{train}}$**		
Bayesian Regularised Neural Network	DEM	2		0.85		
	DTM	2		0.85		
		$\mathrm{m_{trees}}$		**$\mathrm{R^2_{train}}$**		
Bagged Regression Tree	DEM	25		0.83		
	DTM	25		0.83		
		$\mathrm{n_{trees}}$		**Interaction depth**		**$\mathrm{R^2_{train}}$**
Stochastic Gradient Boosting	DEM	150		3		0.86
	DTM	150		3		0.86
		Cost		**$\mathrm{R^2_{train}}$**		
Support Vector Machine	DEM	1		0.81		
	DTM	0.25		0.80		
		AIC		**$\mathrm{R^2_{train}}$**		
Generalised Linear Model	DEM	1169.5		0.80		
	DTM	1170.9		0.81		
Regression Kriging Residuals		**model**	**psill**	**range**	**kappa**	**nugget**
Random Forest	DTM	Spherical	0.42	8	0.7	0

Shrinkage value (0.1) was constant in the Stochastic Gradient Boosting models. AIC means the Akaike Information Criterion.

1.2.5. Model evaluation

The dataset separated for external validation was used to assess the models' performance. The RMSE, adjusted R^2 value (R$^2_{\mathrm{adj}}$), ratio of performance to interquartile distance (RPIQ), and coefficient of efficiency (COE) were calculated to assess the amount of variation explained by each model. The RPIQ index was calculated via the difference between the third and first quartiles divided by the RMSE. We included the COE as proposed by Legates and McCabe (2013):

$$COE = 1.0 - \frac{\sum_{i=1}^{N}|O_i - P_i|}{\sum_{i=1}^{N}|O_i - \bar{O}'_i|}$$

where O_i is the observed value, $\bar{O}'_i$ the observed baseline, and P_i is the model predicted series with N pairs for evaluation.

The COE value ranged between 1 and -1, with a value close to 1 indicating that the model prediction had a good fit. A COE value close to zero or a negative value indicates that the model is not predictive. The bias was also calculated, which measured the distance between the average predicted and observed values. This validation metric compensated for a limitation of the R$^2_{\mathrm{adj}}$, revealing the model bias. Moreover, we qualitatively investigated the relationship between the predicted map of FI content and the Brazilian soil map.

1.3. RESULTS AND DISCUSSION

1.3.1. Exploratory analysis

The distribution of FI in the study area ranged from 0 to 250 g kg^{-1} and reflected the diversity of parent materials. The distribution of the covariate values among the Fe_2O_3 classes are shown in Table 3. The derivative covariates from both the DEM (30 m resolution) and DTM (5 m resolution) indicated variation within Fe_2O_3 classes. On the other hand, the geology was a combination of three parent materials (sandstone, diabase, and siltite) for the hypoferric class; the other two classes had only diabase as bedrock. This tendency corroborated the soil FI content found by Santos et al. (2018). Therefore, this exploratory analysis was the first to show how environmental covariates responded to variation in iron content before DSM procedures were executed.

Table 3. Summary of the ancillary variables selected and grouped by the total free iron classes according to the Brazilian Soil Classification System.

Parameters	Hypoferric (< 80 g kg^{-1})			Mesoferric (80-180 g kg^{-1})			Ferric (180-360 g kg^{-1})			Perferric (> 360 g kg^{-1})
	1st Quartile	Median	3rd Quartile	1st Quartile	Median	3rd Quartile	1st Quartile	Median	3rd Quartile	
SYSI B1[1]	646.40	716.10	824.50	456.30	498.40	575.50	463.60	506.00	552.00	No data available
SYSI B2[1]	1016.90	1111.50	1273.20	746.30	814.10	930.60	755.90	798.20	854.70	
SYSI B3[1]	1361.00	1486.00	1667.00	1096.70	1183.20	1296.20	1099.40	1163.70	1225.50	
SYSI B4[1]	2047.00	2228.00	2512.00	1583.00	1705.00	1829.00	1568.00	1655.00	1712.00	
SYSI B5[1]	2710.00	3045.00	3493.00	2038.00	2154.00	2325.00	2008.00	2113.00	2211.00	
SYSI B7[1]	2362.00	2659.00	3060.00	1812.00	1909.00	2035.00	1741.00	1832.00	1908.00	
SC Red[2]	127.08	137.67	144.28	113.71	119.61	125.31	113.45	118.89	124.49	
SC Green[2]	89.18	104.89	113.64	66.99	71.03	90.61	67.90	78.98	87.82	
SC Blue[2]	67.42	79.24	87.69	45.64	61.05	69.88	48.34	63.24	69.84	
AMT[3]	20.40	20.60	20.70	20.30	20.40	20.60	20.40	20.60	20.60	
AMP[4]	1166	1178	1245	1166	1166	1176	1166	1166	1166	
NDVI[5]	0.18	0.20	0.21	0.16	0.18	0.18	0.16	0.17	0.18	
EVI[6]	0.29	0.32	0.35	0.22	0.25	0.28	0.22	0.24	0.28	
SRF[7]	5.78	10.92	14.85	2.13	4.44	9.83	0.63	0.91	2.20	
DEM[8]	504.00	533.50	561.00	524.00	533.00	578.00	531.50	550.00	561.20	
ADEM[9]	61.39	152.41	271.68	53.49	89.53	276.68	52.24	79.65	278.54	
LSFDEM[10]	0.37	0.76	1.20	0.65	1.28	1.71	1.09	1.58	1.81	
PLCDEM[11]	-0.0093	0.0034	0.0141	-0.0087	0.0016	0.0083	-0.0044	0.0007	0.0077	
PRCDEM[12]	-0.0007	0.0000	0.0010	-0.0004	0.0004	0.0013	-0.0003	0.0006	0.0013	
SDEM[13]	5.04	7.48	10.02	6.44	9.92	12.93	9.76	12.72	15.64	
VDDEM[14]	5.24	12.65	22.67	5.01	8.22	14.94	0.64	4.95	11.13	
TWIDEM[15]	6.19	6.72	7.67	6.26	7.11	7.73	5.97	6.24	6.68	
DTM[16]	498.30	529.60	558.40	521.30	531.60	574.40	528.90	547.50	559.30	
ADTM[17]	62.98	152.03	269.80	51.03	95.93	286.78	62.05	100.43	284.88	
LSFDTM[18]	0.32	0.69	1.10	0.52	0.82	1.44	0.74	1.27	1.77	
PLCDTM[19]	-0.02	0.00	0.03	-0.0115	0.0094	0.0240	-0.0212	0.0025	0.0305	
PRCDTM[20]	0.00	0.00	0.00	-0.0027	-0.0004	0.0055	-0.0002	0.0024	0.0059	
SDTM[21]	4.23	7.00	10.13	5.41	7.62	11.69	7.35	10.64	15.03	
VDDTM[22]	0.13	1.21	4.22	0.17	0.52	3.72	0.02	0.50	3.56	
TWIDTM[23]	6.28	6.89	7.52	6.15	6.99	7.58	5.90	6.52	7.08	
DD[24]	5.73	7.36	8.48	6.82	7.44	8.76	8.48	8.92	9.02	
Landforms	Upper slope	Upper slope	Lower slope	Upper slope	Upper slope	Upper slope	Upper slope	Upper slope	Upper slope	
GYPI[25]	0.06	0.07	0.07	0.06	0.07	0.07	0.07	0.07	0.07	
NATI[26]	0.13	0.15	0.17	0.10	0.11	0.14	0.11	0.13	0.14	
CALI[27]	0.44	0.46	0.47	0.41	0.44	0.46	0.43	0.45	0.46	
CARI[28]	1.30	1.32	1.36	1.38	1.44	1.48	1.38	1.44	1.49	
FFI[29]	1.70	1.78	1.86	1.51	1.59	1.66	1.50	1.57	1.65	
FI[30]	1.31	1.35	1.41	1.21	1.26	1.32	1.25	1.30	1.33	
FOI[31]	1.99	2.06	2.17	2.20	2.37	2.51	2.12	2.23	2.47	
CHI[32]	0.87	0.87	0.88	0.87	0.88	0.89	0.86	0.87	0.87	
Geology	Sandstone	Diabase	Siltite	Diabase	Diabase	Diabase	Diabase	Diabase	Diabase	
Geomorph[33]	Shoulder	Shoulder	Footslope	Shoulder	Shoulder	Footslope	Backslope	Shoulder	Shoulder	
X[34]	218759	228008	228783	228309	229124	229633	229124	229280	229435	
Y[35]	7454661	7459606	7486013	7453767	7454165	7458770	7453646	7453764	7453990	

[1]Synthetic Soil Image bands (SYSI B1, B2, B3, B4, B5 and B7); [2]Predicted Soil RGB colours (SC Red, Green and Blue); [3]Annual Mean Temperature (AMT); [4]Annual Mean Precipitation (AMP); [5]Normalised Difference Vegetation Index (NDVI); [6]Enhanced Vegetation Index (EVI); [7]Soil Relative Frequency (SRF); [8]Digital Elevation Model (DEM); [9]Aspect from DEM (ADEM); [10]LS Factor from DEM (LSFDEM); [11]Plan Curvature from DEM (PLCDEM); [12]Profile Curvature from DEM (PRCDEM); [13]Slope from DEM (SDEM); [14]Valley Depth from DEM (VDDEM); [15]Topographic Wetness Index from DEM (TWIDEM); [16]Digital Terrain Model (DTM); [17]Aspect from DTM (ADTM); [18]LS Factor from DTM (LSFDTM); [19]Plan Curvature from DTM (PLCDTM); [20]Profile Curvature from DTM (PRCDTM); [21]Slope from DTM (SDTM); [22]Valley Depth from DTM (VDDTM); [23]Topographic Wetness Index from DTM (TWIDTM); [24]Drainage Density (DD); [25]Gypsic Index (GYPI); [26]Natric Index (NATI); [27]Calcareous Index (CALI); [28]Carbonate Radicals Index (CARI); [29]Ferrous Fe Index (FFI); [30]Ferrous Index (FI); [31]Ferrous Oxides Index (FOI); [32]Clay and Hydroxides Index (CHI); [33]Geomorphology (Geomorph.); [34]Longitude (X) and [35]Latitude (Y) coordinates.

The distribution of FI was assessed by calculating its mean, standard deviation, skewness, and kurtosis. FI content showed a left distribution tendency and high variability (coefficient of variability > 90%) in the training dataset (Fig. 2a). In regression models, the target variable would be better fitted if it were normally distributed. To achieve a normal distribution, we performed Box–Cox transformation (Fischer, 2016; Malone et al., 2013). The power transformation decreased variation by up to 45% (Fig. 2b). We subsequently modelled the transformed data using eight MLAs, and the best fit model was selected to map the residual via RK (e.g. simple kriging).

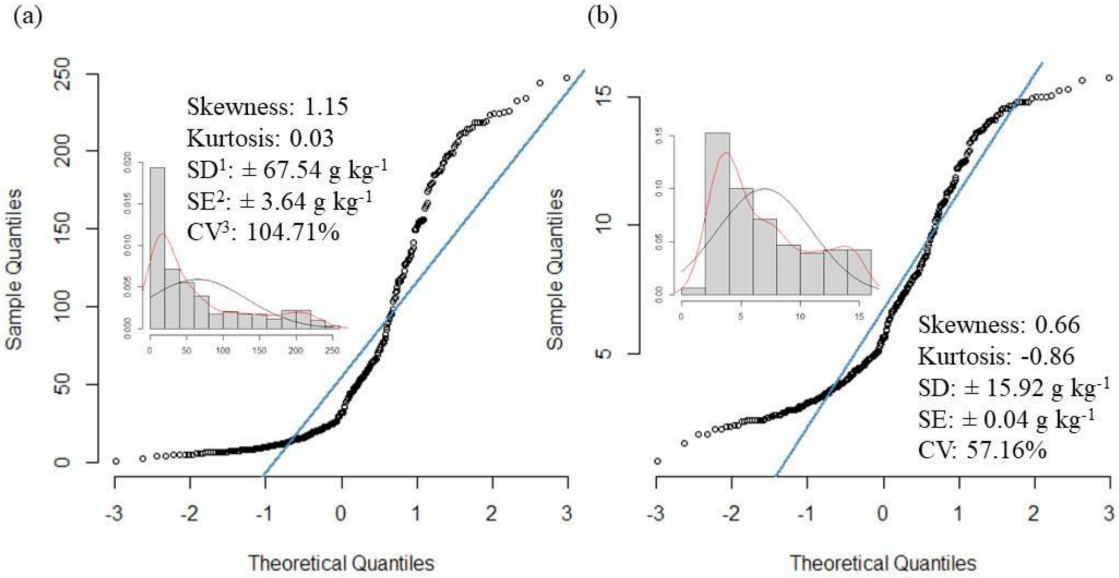


Fig. 2. Descriptive statistics of total iron content (a) and its power transformation $(\sqrt{Fe_2O_3})$ (b). Red line empirical density and black line normal density in the histogram. [1]SD, standard deviation. [2]SE, standard error. [3]CV, coefficient of variation.

FI content was positively correlated with carbonate radicals and ferrous oxides indices of over 0.4, i.e. an increase in these indices indicated increased Fe_2O_3 content (Fig. 3). Martínez-Graña et al. (2016) mapped the soils of Spain and found that the fluvial terraces had soil features such as carbonate accumulation and gleyic horizon development. The relationship between carbonates and iron is directly connected to the hydrologic cycle, where the water table plays an important role in two soil-forming processes, namely gleisation and ferralitisation, respectively. Gleisation involves the periodic wetting of reduced iron (Bockheim, 2018) or dissolved Fe^{2+}, which mostly remains reduced. Ferralitisation encompasses the total hydrolysis of Fe^{+3}, forming oxides (e.g. goethite and hematite) and hydroxides. These iron and aluminium compounds are formed by the removal of silica and bases. Both processes occur in the presence of high soil iron concentrations and alterations in the water table.

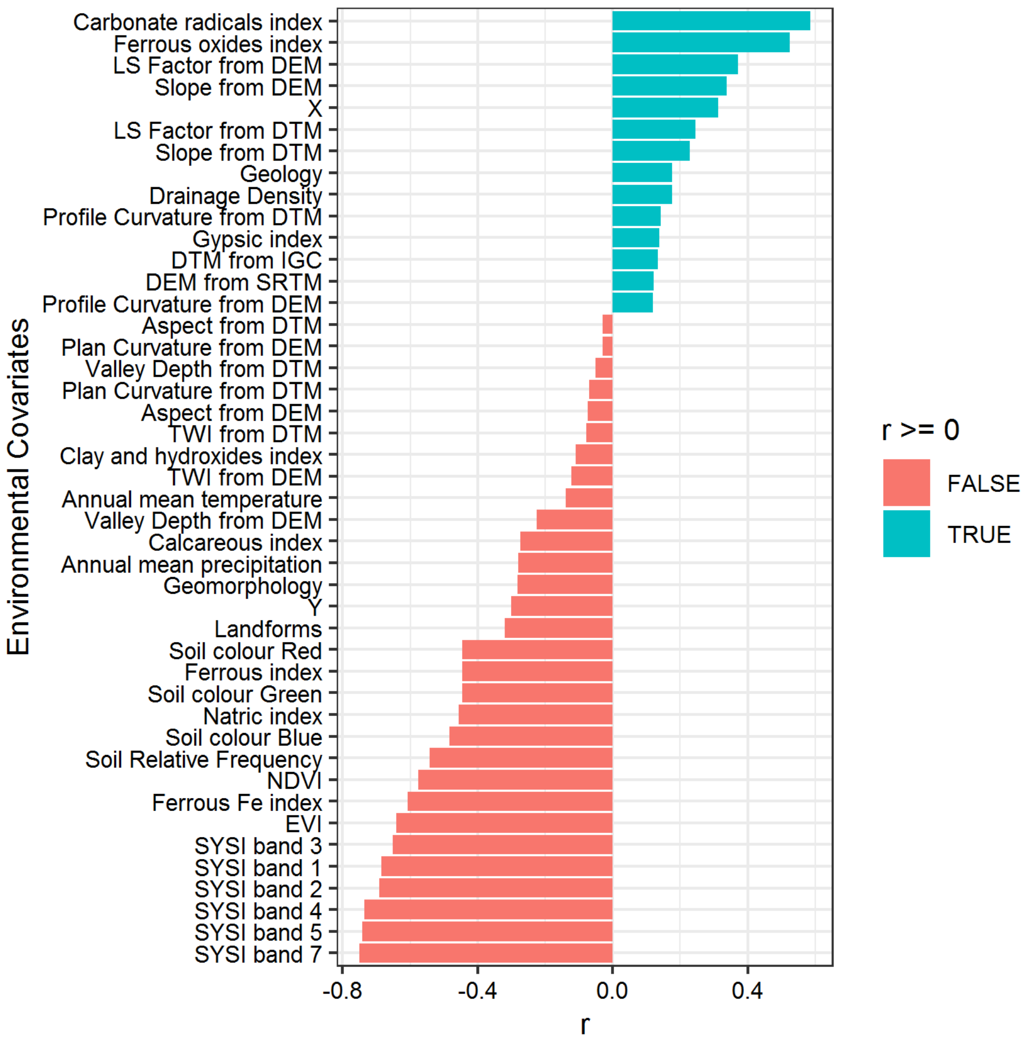


Fig. 3. Pearson's correlation index ($p < 0.01$) between the response variable (Fe_2O_3, g kg^{-1}) and the environmental covariates (*scorpan* factors). False means negative correlation; True means positive correlation.

Negative correlation coefficients between FI and soil RGB colour, ferrous and natric indices, NDVI, and SRF were found (Fig. 3). Similarly, ferrous iron index, EVI, and SYSI bands had negative correlation values of 0.6–0.8, i.e. the values of these covariates increased with decreasing iron content. The opposite was reported by Regmi and Rasmussen (2018), who predicted soil landscape units in the arid southwestern United States. They found that lower values of ferrous iron index derivatives from satellite bands indicated aeolian deposits. However, our results showed high FI contents, which could have led to a low ferrous iron index. The same reasoning applies to the EVI and NDVI, where areas with higher vegetation cover should represent high FI contents and vice versa. The SRF showed that areas with higher exposure frequency had lower FI contents, which makes sense, as bare soil areas lead to greater physical weathering (e.g. wind erosion) of bare soil areas (Dwivedi, 2001).

1.3.2. Model performance

To predict FI content, we used satellite spectral bands of SYSI and other environmental variables that matched the *scorpan* factors (Table 1). The predictive models were run using terrain derivatives from the DTM (Fig.

4a) and DEM (Fig. 4b). Table 4 presents the parameters used to evaluate the models' performance. Even though the RF model used only 19 of the 36 available environmental variables, it performed best, with RMSE, R^2_{adj}, RPIQ, and COE values of 25.09, 0.84, 0.72, and 0.69 respectively. The m_{try} (Table 3) explicitly shows the number of predictors needed to outperform the RF predictions. No meaningful differences existed between models that used pixel resolutions of 5 or 30 m as environmental covariates. This corroborates the results of Samuel-Rosa et al. (2015), who found that detailed environmental covariates did not significantly improve DSM performance. These authors concluded that it would be more useful to increase sampling density rather than devoting time and resources to generating detailed data, such as DEMs or DTMs.

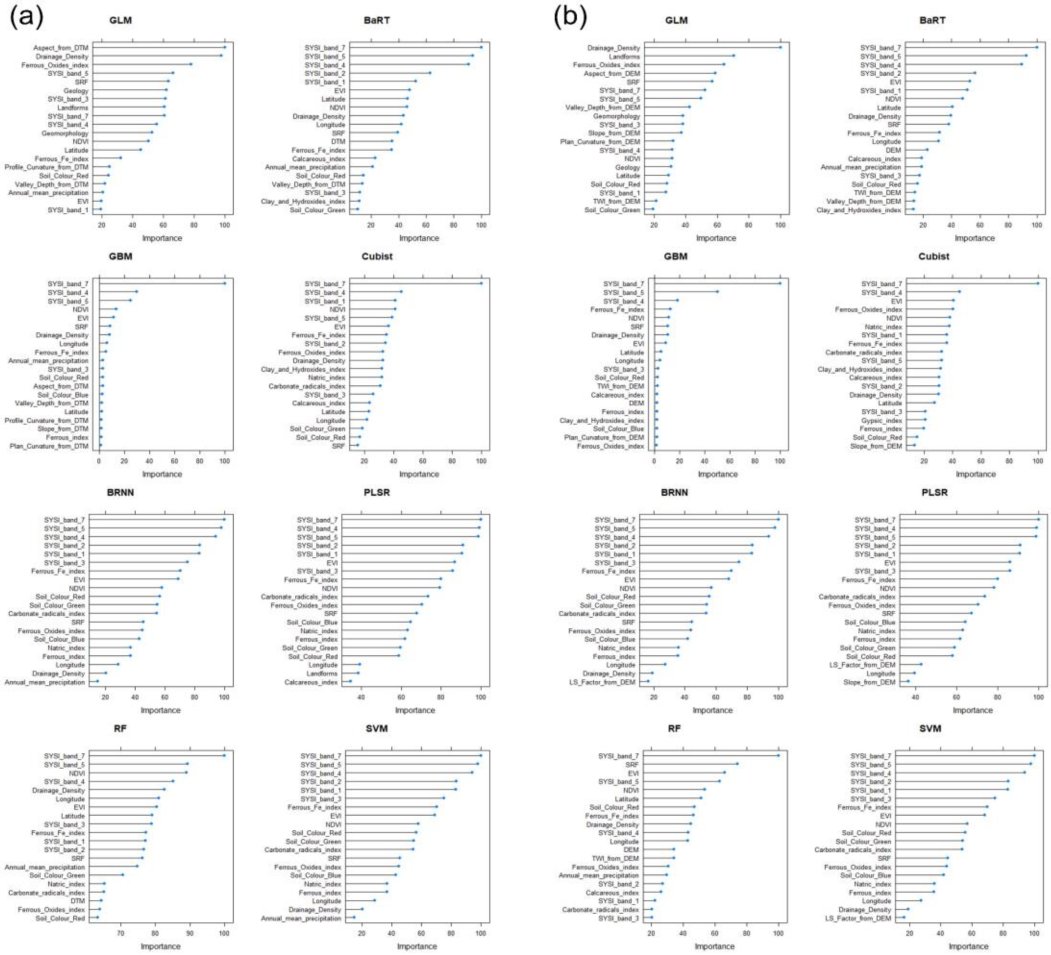


Fig. 4. The importance level, in percentage, of the top 20 environmental covariates after modelling the total free iron content within each machine learning algorithm based on 5 m (DTM, Digital Terrain Model) (a) and 30 m (DEM, Digital Elevation Model) (b) pixel resolution.

Table 4. Model evaluation at 0 – 20 depth (layer A) for predicted total free iron based on covariates from Digital Terrain Model (DTM) and Digital Elevation Model (DEM).

	Parameters	GLM[1]	BaRT[2]	GBM[3]	Cubist	BRNN[4]	PLSR[5]	RF[6]	SVM[7]	Regression Kriging best model	Parameters	RKRF DTM[8]
Fe_2O_3 (g/kg) DTM	RMSE[9]	34.81	25.90	31.15	31.24	37.47	39.70	25.09	35.73		RMSE	24.98
	R^2_{adj}[10]	0.71	0.83	0.76	0.75	0.66	0.61	0.84	0.69			
	RPD[11]	1.83	2.46	2.04	2.04	1.70	1.60	2.54	1.78		R^2_{adj}	0.84
	RPIQ[12]	0.52	0.69	0.58	0.58	0.48	0.45	0.72	0.50			
	COE[13]	0.57	0.68	0.60	0.61	0.54	0.46	0.69	0.54		RPD	2.55
Fe_2O_3 (g/kg) DEM	RMSE	36.86	26.19	31.00	33.58	39.13	39.69	25.97	37.60			
	R^2_{adj}	0.67	0.83	0.76	0.72	0.63	0.61	0.83	0.67		RPIQ	0.72
	RPD	1.73	2.43	2.05	1.90	1.63	1.60	2.45	1.69			
	RPIQ	0.49	0.69	0.58	0.54	0.46	0.45	0.69	0.48		COE	0.69
	COE	0.55	0.67	0.62	0.57	0.50	0.46	0.67	0.54			

Note. [1]GLM, Generalised Linear Model; [2]BaRT, Bagged Regression Tree; [3]GBM, Stochastic Gradient Boosting; [4]BRNN, Bayesian Regularised Neural Network; [5]PLSR, Partial Least Square Regression; [6]RF, Random Forest; [7]SVM, Support Vector Machine; [8]RKRF, Regression Kriging with RF. [9]RMSE, Root Mean Square Error; [10]R^2_{adj}, Adjusted Correlation Index; [11]RPD, Ratio of Performance to Deviation; [12]RPIQ, Ratio of Performance to Interquartile Distance; [13]COE, Coefficient of Efficiency.

The final maps of all predicted models for FI content were evaluated using an external dataset (Table 4 and Fig. S2). The PLSR showed the poorest performance. Likewise, the BRNN, GLM, and SVM did not show satisfactory performance. On the other hand, Cubist, Stochastic Gradient Boosting, BaRT, and RF adequately predicted FI content.

We performed RK on the best model, i.e. the RF model (Fig. S3). RF residuals were spatialised over the entire site. The RK approach showed higher accuracy than the RF model alone did (Table 4).

Shahbazi et al. (2019) monitored soil crystalline iron oxides in Iran via Cubist, multiple linear regression, and decision trees. Their estimated sampling density was one point per 0.02 km^2, whereas ours was one point per 7.48 km^2. Despite our low sampling density, almost all our models showed satisfactory performance. Shahbazi et al. (2019) indicated that decision tree prediction was superior to multiple linear regression and Cubist, corroborating our findings. To enhance DSM performance, sampling collection has to be well-distributed and represent most of the variability of the study area.

We attained the model importance level within MLAs in our study by using the top 20 most relevant environmental covariates (Fig. 4). The main covariates were SYSI bands 7 (2064–2345 nm) and 4 (772–898 nm), which contained the SWIR and NIR spectral responses of bare soil. Therefore, soil FI content was directly linked to these spectral bands. Demattê et al. (2017) analysed the spectral behaviour of wetland soil properties and genesis in two Brazilian biomes and found that hematite had absorption features at 750 and 1050 nm. This could explain the importance level of band 4 in predicting FI content, because it comprised hematite and goethite responses. Regmi and Rasmussen (2018) mapped the relationship between soil attributes and landscape in the southwestern United States and found high ferrous iron concentrations (Landsat ETM + bands 7/3 and 5/4 ratios) in alluvial deposits of metamorphic rocks and lower concentrations in aeolian deposits. This supports our finding of SYSI band 7 being the most important covariate for predicting FI.

Drainage density was also one of the foremost environmental covariates in the MLAs. This makes sense as the role of the water table in the natural chemical reactions of iron in pedogenetic processes is well-known (Klingebiel, 1958; Vogt et al., 2003). The temporal images, SYSI, and their band indices were fundamental to obtaining good results in our study. The relationship between the environmental covariates used to predict FI content was in line with established soil formation factors and processes (Ma et al., 2019; Schaetzl and Anderson, 2005).

1.3.3. Final FI content map and relationship with soils

Using the best-predicted map of FI content, we identified zones with high iron content from mafic rocks and low iron content from sedimentary or metamorphic rocks (Fig. 5). Goethite and haematite are the most common iron minerals in tropical soils and are respectively described as yellow-brown and red pigment agents in soils (Anda et al., 2008; Macedo and Bryant, 1989). High FI contents indicate higher haematite concentrations, and low FI contents indicate higher goethite concentrations. Demattê (2002) reported that soils with low iron content showed an increasing tendency from B1 to B7 in Landsat 5 Thematic Mapper. This agreed with the FI spectral curves in our study, which allowed us to differentiate the iron classes (Fig. 5c).

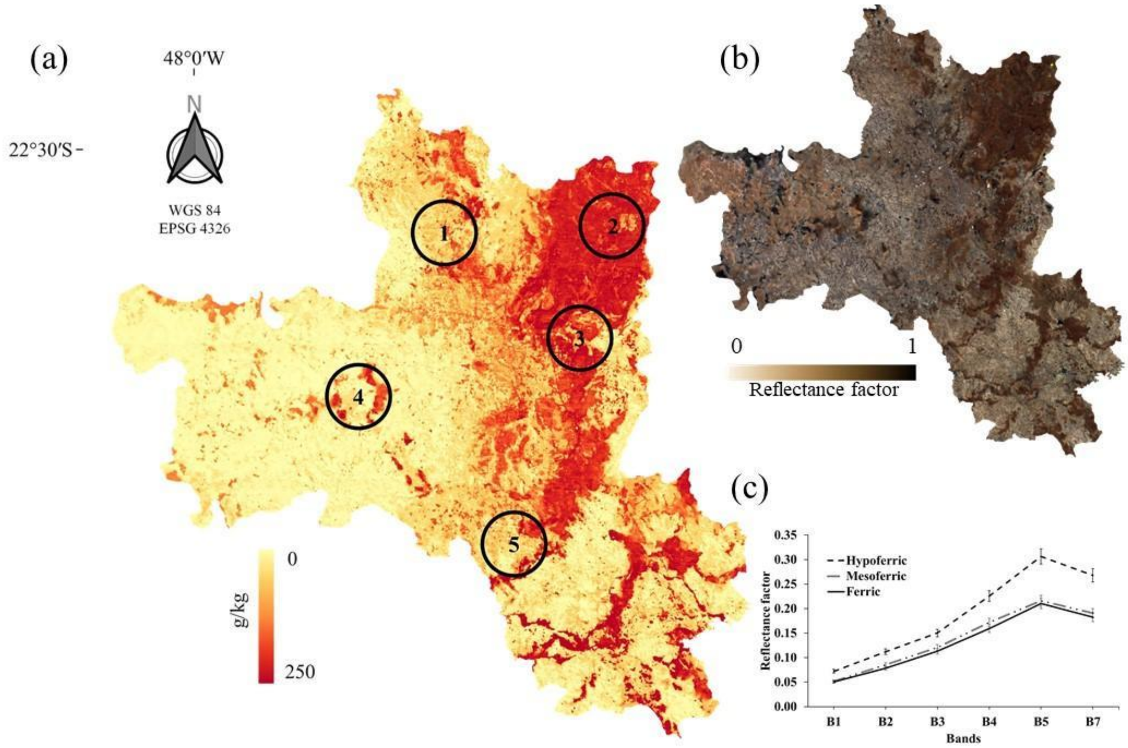


Fig. 5. Final predicted map of total free iron content using regression kriging of Random Forest algorithm with Digital Terrain Model (DTM). The highlighted areas are described in Fig. 7.

We selected five sites and overlaid the soil map to qualitatively analyse the patterns (Figs. 5a and 6). Sites 1–5 had high FI contents and were characterised by Typic Haplorthox (orange) and Typic Eutrorthox (red) soil types. The latter is derived from basalt rocks, with a very clayey texture and FI content of up to 180 g kg^{-1}. Typic Paleudalf and Arguidoll have lower FI contents than Rhodic Paleudalf, Typic Eutrorthox, Typic Haplorthox, and Typic Paleudalf do, because they are derived from argillite, siltite, and sandstone. Typic Paleudalf is characterised by a red-yellow colour because of its high goethite and low FI contents (Galvão et al., 1997). This pattern strongly agreed with the predicted iron content in this study. Moreover, the predicted FI content map indicated areas with high and low soil iron concentrations, which could represent another soil type not mapped in the traditional soil survey. This likely happened because the traditional soil map scale represented 100 ha, whereas the predicted FI content map represented 0.09 ha (30 × 30 m pixel resolution), displaying more detailed information.

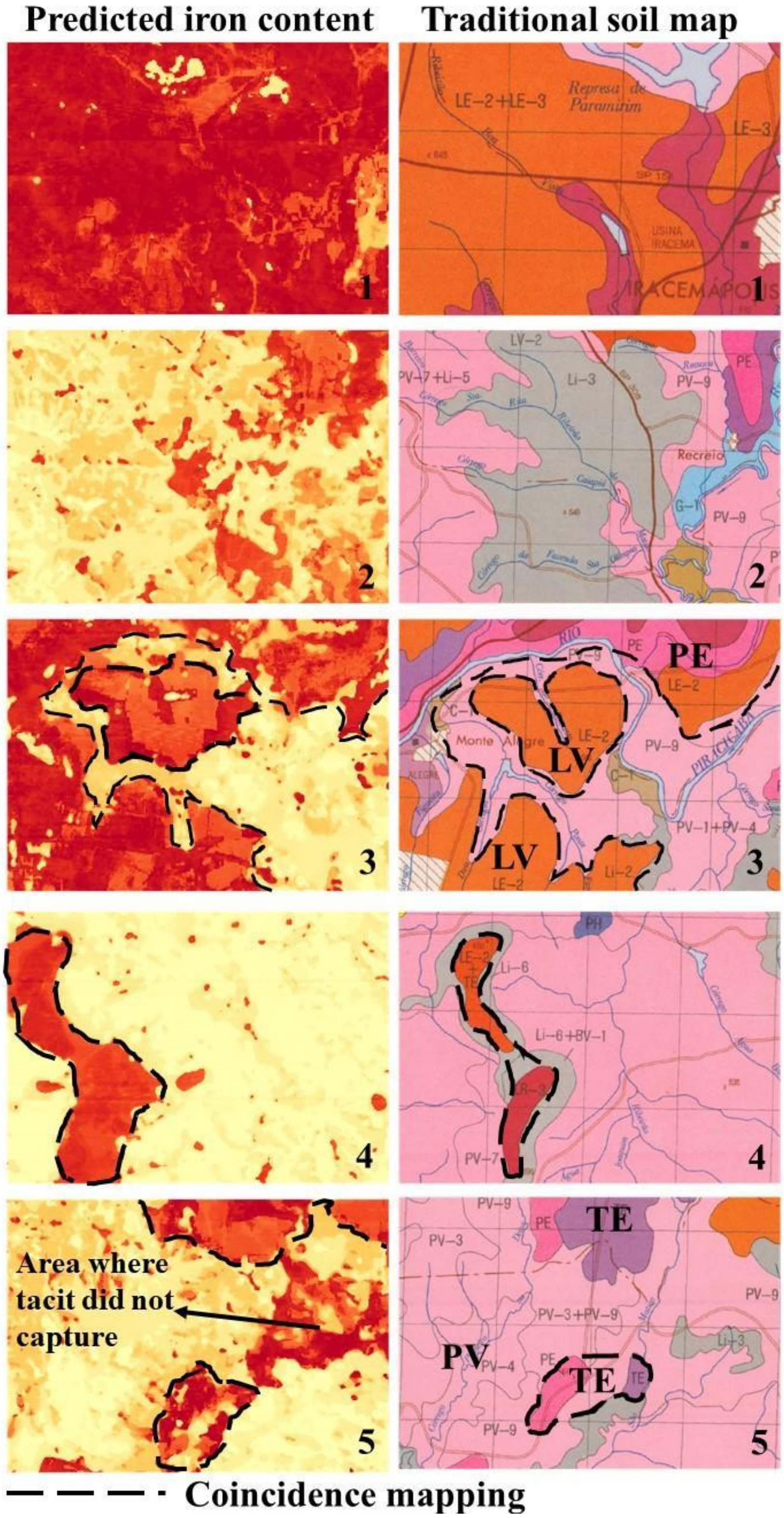


Fig. 6. Visual association of the predicted total free iron content with the Brazilian traditional soil map 1/100,000 scale from Agronomic Institute of Campinas (IAC, Portuguese acronym). LE: Typic Haplorthox, LR: Typic Eutrorthox, PV: Typic Paleudalf, Li: Lithic Distrochrept, PE: Typic Paludult, TE: Rhodic Paleudalf, and BV: Typic Arguidoll.

The results showed the potential of DSM and remote sensing–derived products to produce detailed FI maps for tropical regions. The bare soil composite was an important predictor that revealed spatial changes in topsoil colour, a morphological feature highly related to soil mineralogy and iron content (Gray et al., 2016; Shahbazi et al., 2019b). Furthermore, the soil iron cycle is strongly affected by hill slope water drainage, where crystalline forms occur on top of the relief with good drainage (oxidative environments), whereas reduced forms and less crystallised iron minerals are more abundant in moister areas (reductive environments) in the terrain (Bartholomeus et al., 2007). The parent material is another important factor affecting the soil iron cycle and was depicted by the bare soil composite. Therefore, the combination of spectral features derived from bare soil composites and terrain attributes could indicate the main patterns in FI content at the landscape level.

1.4. CONCLUSIONS

The mapping of FI content in tropical soils using a DSM approach proved to be feasible. It is vital to determine the best MLAs to better understand the relationship between the covariates and response variables for a specific site. However, no unique algorithm exists for mapping at local, regional, or national scales. In this study, we found that five models, namely Cubist, Stochastic Gradient Boosting, BaRT, RF, and RK, could accurately predict FI content. Detailed relief data at a 5 m resolution were not superior to 30 m resolution data. Therefore, for regional mapping, data available from the USGS is sufficient.

The use of the temporal images from Landsat archives merged to a unique bare soil image was vital to improving model performance. The final predicted FI content map was important for identifying detailed soil types, which should be considered in future pedological studies.

ACKNOWLEDGMENTS

This work was supported by the São Paulo Research Foundation (grant numbers 2016/26124-6 and 2014/22262-0) and forms part of the first author's PhD thesis. The authors wish to thank the Geotechnologies in Soil Science Group (https://esalqgeocis.wixsite.com/english).

APPENDIX A. SUPPLEMENTARY DATA

Methodology and Results of the Soil RGB colour environmental variable

S factor soil RGB colour prediction procedure

Additional to the covariates that represent the *s* factor of the *scorpan* model, we created a predicted soil RGB colour based on the soil spectral reflectance. In this case, we used a part of the Brazilian Soil Spectral Library (Demattê et al., 2019) with 1,431 soil samples. The spectral data were acquired using a Fieldspec 3 sensor, and the methodology is reported in Demattê et al. (2019). As described in Viscarra Rossel et al. (2006), the soil spectra from the laboratory in the wavelength range between 450 and 780 nm (visible spectral region) were converted to RGB values corresponding to the Landsat bands 1, 2, and 3 by taking the average reflectance between 450 – 520 nm (band 1, Blue), 520 – 600 nm (band 2, Green), and 630 – 690 nm (band 3, Red). Subsequently, the dataset was split into 75% for calibration (1,073 samples) and 25% for external validation (358 samples). The SYSI's bands were used as covariates in a Cubist model to predict RGB soil colour for the entire area using the following hyperparameters: Red, committees = 10 and neighbours = 0; Green, committees = 10 and neighbours = 0; Blue, committees = 20 and neighbours = 9.

Soil RGB colour as environmental covariates of "s" factor

Soil colour of the region was predicted and spatialized from soil spectra data for mapping FI (Fig. S3). We performed the cubist model to predict the soil RGB colour from the laboratory using the SYSI's bands as covariates. Hence, only spectral data were used to generate maps of red, green and blue colour components. The R^2_{adj} for the predicted red, green, and blue were 0.53, 0.69 and 0.64, respectively. The RMSE values were 14.07, 13.27 and 11.91 for red, green and blue, respectively. This soil RGB colour was used to boost our predictive models of FI.

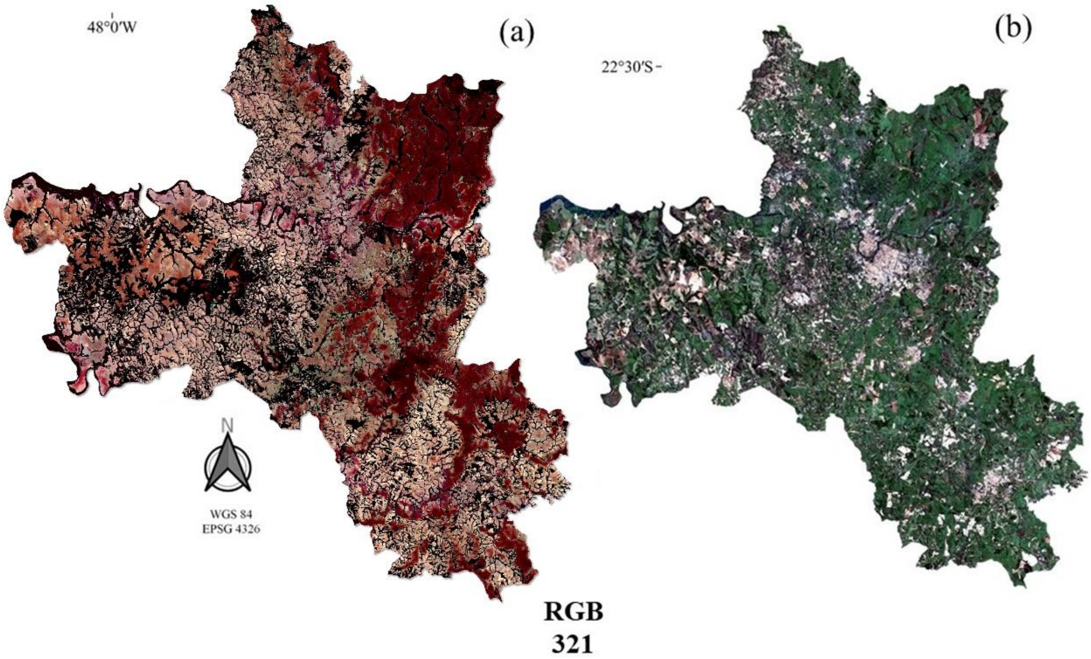


Fig. S1. Final predicted map of soil colour using Cubist algorithm with spectral reflectance of SYSI bands (a) and satellite image (b).

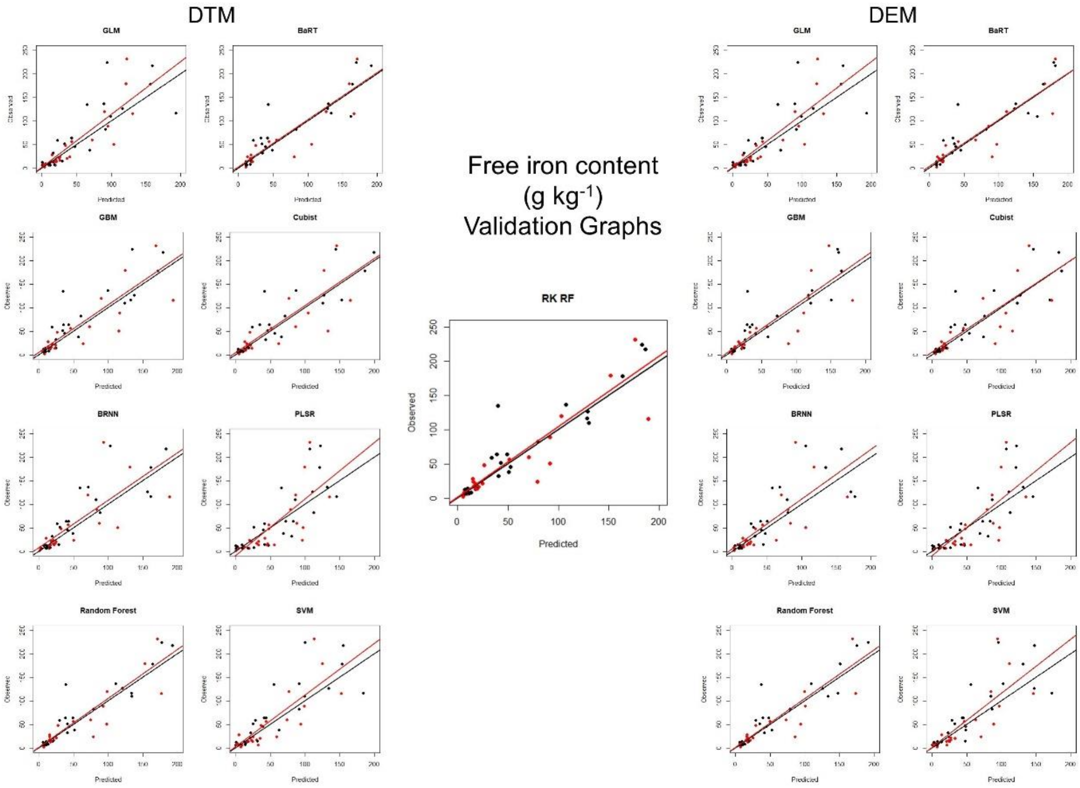


Fig. S2. Graphs of predicted versus observed values of free iron content for each model using Digital Terrain Model (DTM) and Digital Elevation Model (DEM)

.

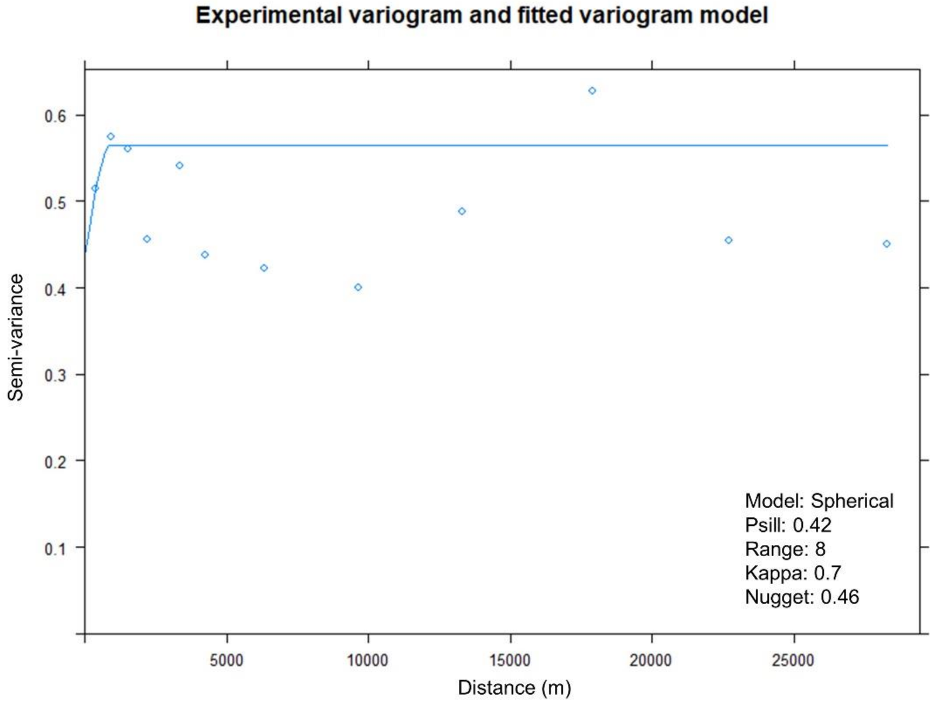


Fig. S3. Semivariogram of the Regression Kriging of the Random Forest model, which was the best predictive model of free iron content.

REFERENCES

Abrams, M., Hook, S.J., 1995. Simulated Aster data for geologic studies. IEEE Trans. Geosci. Remote Sens. 33, 692–699. https://doi.org/10.1109/36.387584

Anda, M., Shamshuddin, J., Fauziah, C.I., Omar, S.R.S., 2008. Mineralogy and factors controlling charge development of three Oxisols developed from different parent materials. Geoderma 143, 153–167. https://doi.org/10.1016/j.geoderma.2007.10.024

Andrews Deller, M.E., 2006. Facies discrimination in laterites using Landsat Thematic Mapper, ASTER and ALI data-examples from Eritrea and Arabia. Int. J. Remote Sens. 27, 2389–2409. https://doi.org/10.1080/01431160600586050

Angelini, M.E., Heuvelink, G.B.M., 2018. Including spatial correlation in structural equation modelling of soil properties. Spat. Stat. 25, 35–51. https://doi.org/10.1016/J.SPASTA.2018.04.003

Bartholomeus, H., Epema, G., Schaepman, M., 2007. Determining iron content in Mediterranean soils in partly vegetated areas, using spectral reflectance and imaging spectroscopy. Int. J. Appl. Earth Obs. Geoinf. 9, 194–203. https://doi.org/10.1016/j.jag.2006.09.001

Ben-Dor, E., Banin, A., 1995. Quantitative analysis of convolved Thematic Mapper spectra of soils in the visible near-infrared and shortwave-infrared spectral regions (0·4–2·5 μm). Int. J. Remote Sens. 16, 3509–3528. https://doi.org/10.1080/01431169508954643

Bishop, C.M.C., 2006. Pattern recognition and machine learning. Springer.

Bockheim, J.G., 2018. Diversity of diagnostic horizons in soils of the contiguous USA: A case study. CATENA 168, 5–13. https://doi.org/10.1016/J.CATENA.2017.10.016

Bonfatti, B.R., Demattê, J.A.M., Marques, K.P.P., Poppiel, R.R., Rizzo, R., Mendes, W. de S., Silvero, N.E.Q., Safanelli, J.L., 2020. Digital mapping of soil parent material in a heterogeneous tropical area. Geomorphology 107305. https://doi.org/10.1016/j.geomorph.2020.107305

Breiman, L., 2001. Random Forests.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24, 123–140. https://doi.org/10.1007/BF00058655

Burrough, P.A., 2006. Chapter 41 The Display of Digital Soil Data, 1976–2004, in: Digital Soil Mapping - An Introductory Perspective. pp. 555–633. https://doi.org/10.1016/S0166-2481(06)31041-0

Chagas, C. da S., de Carvalho Junior, W., Bhering, S.B., Calderano Filho, B., 2016. Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions. Catena 139, 232–240. https://doi.org/10.1016/j.catena.2016.01.001

Coleman, T.L., Agbu, P.A., Montgomery, O.L., 1993. Spectral differentiation of surface soils and soil properties. Soil Sci. 155, 283–293. https://doi.org/10.1097/00010694-199304000-00007

Coyne, L.M., Bishop, J.L., Scattergood, T., Banin, A., Carle, G., Orenberg, J., 1990. Quantifying Iron and Surface Water in a Series of Variably Cation-Exchanged Montmorillonite Clays, in: Near-Infrared Correlation Spectroscopy. American Chemical Society, pp. 407–429. https://doi.org/10.1021/bk-1990-0415.ch021

de Carvalho Junior, W., Lagacherie, P., da Silva Chagas, C., Calderano Filho, B., Bhering, S.B., 2014. A regional-scale assessment of digital mapping of soil attributes in a tropical hillslope environment. Geoderma 232–234, 479–486. https://doi.org/10.1016/j.geoderma.2014.06.007

Demattê, J.A.M., 2002. Characterization and discrimination of soils by their reflected electromagnetic energy. Pesqui. Agropecuária Bras. 37, 1445–1458. https://doi.org/10.1590/S0100-204X2002001000013

Demattê, J.A.M., Fongaro, C.T., Rizzo, R., Safanelli, J.L., 2018. Geospatial Soil Sensing System (GEOS3): A powerful data mining procedure to retrieve soil spectral reflectance from satellite images. Remote Sens. Environ. 212, 161–175. https://doi.org/10.1016/j.rse.2018.04.047

Demattê, J.A.M., Horák-Terra, I., Beirigo, R.M., Terra, F. da S., Marques, K.P.P., Fongaro, C.T., Silva, A.C., Vidal-Torrado, P., 2017. Genesis and properties of wetland soils by VIS-NIR-SWIR as a technique for environmental monitoring. J. Environ. Manage. 197, 50–62. https://doi.org/10.1016/J.JENVMAN.2017.03.014

Demattê, J.A.M., Safanelli, J.L., Poppiel, R.R., Rizzo, R., Silvero, N.E.Q., Mendes, W. de S., Bonfatti, B.R., Dotto, A.C., Salazar, D.F.U., Mello, F.A. de O., Paiva, A.F. da S., Souza, A.B., Santos, N.V. dos, Maria Nascimento, C., Mello, D.C. de, Bellinaso, H., Gonzaga Neto, L., Amorim, M.T.A., Resende, M.E.B. de, Vieira, J. da S., Queiroz, L.G. de, Gallo, B.C., Sayão, V.M., Lisboa, C.J. da S., 2020. Bare Earth's Surface Spectra as a Proxy for Soil Resource Monitoring. Sci. Rep. 10, 4461. https://doi.org/10.1038/s41598-020-61408-1

Dwivedi, R.S., 2001. Soil resources mapping: A remote sensing perspective. Remote Sens. Rev. 20, 89–122. https://doi.org/10.1080/02757250109532430

Fan, J.X., Wang, Y.J., Liu, C., Wang, L.H., Yang, K., Zhou, D.M., Li, W., Sparks, D.L., 2014. Effect of iron oxide reductive dissolution on the transformation and immobilization of arsenic in soils: New insights from X-ray photoelectron and X-ray absorption spectroscopy. J. Hazard. Mater. 279, 212–219. https://doi.org/10.1016/j.jhazmat.2014.06.079

Fan, S.-S., Chang, F.-H., Hsueh, H., Ko, T.-H., 2016. Measurement of Total Free Iron in Soils by H2S Chemisorption and Comparison with the Citrate Bicarbonate Dithionite Method. J. Anal. Methods Chem. 2016, 1–7. https://doi.org/10.1155/2016/7213542

Fischer, C., 2016. Comparing the Logarithmic Transformation and the Box-Cox Transformation for Individual Tree Basal Area Increment Models. For. Sci. 62, 297–306. https://doi.org/10.5849/forsci.15-135

Forkuor, G., Hounkpatin, O.K.L., Welp, G., Thiel, M., 2017. High Resolution Mapping of Soil Properties Using Remote Sensing Variables in South-Western Burkina Faso: A Comparison of Machine Learning and Multiple Linear Regression Models. PLoS One 12, e0170478. https://doi.org/10.1371/journal.pone.0170478

Franklin, J., 2005. The elements of statistical learning: data mining, inference and prediction. Math. Intell. 27, 83–85. https://doi.org/10.1007/BF02985802

Friedman, J.H., 2002. Stochastic gradient boosting 38, 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

Galvão, L.S., Vitorello, Í., Formaggio, A.R., 1997. Relationships of spectral reflectance and color among surface and subsurface horizons of tropical soil profiles. Remote Sens. Environ. 61, 24–33. https://doi.org/10.1016/S0034-4257(96)00219-2

Gomes, L.C., Faria, R.M., de Souza, E., Veloso, G.V., Schaefer, C.E.G.R., Filho, E.I.F., 2019. Modelling and mapping soil organic carbon stocks in Brazil. Geoderma 340, 337–350. https://doi.org/10.1016/J.GEODERMA.2019.01.007

Gray, J.M., Bishop, T.F.A., Wilford, J.R., 2016. Lithology and soil relationships for soil modelling and mapping. CATENA 147, 429–440. https://doi.org/10.1016/j.catena.2016.07.045

Helland, I.S., 1988. On the structure of partial least squares regression. Commun. Stat. - Simul. Comput. 17, 581–607. https://doi.org/10.1080/03610918808812681

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A., 2005. Very high resolution interpolated climate surfaces for global land areas. Int. J. Climatol. 25, 1965–1978. https://doi.org/10.1002/joc.1276

Huete, A.R., 2004. Remote Sensing for Environmental Monitoring, in: Environmental Monitoring and Characterization. Elsevier, pp. 183–206. https://doi.org/10.1016/B978-012064477-3/50013-8

Hunt, G.R., 1980. Electromagnetic radiation: the communication link in remote sensing, in: Siegal, B.S., Gillespie, A.R. (Eds.), Remote Sensing in Geology. Wiley & Sons, New York, pp. 5–45.

Keskin, H., Grunwald, S., 2018. Regression kriging as a workhorse in the digital soil mapper's toolbox. Geoderma 326, 22–41. https://doi.org/10.1016/j.geoderma.2018.04.004

Keskin, H., Grunwald, S., Harris, W.G., 2019. Digital mapping of soil carbon fractions with machine learning. Geoderma 339, 40–58. https://doi.org/10.1016/j.geoderma.2018.12.037

Klingebiel, A.A., 1958. Soil Survey Interpretation-Capability Groupings. Soil Sci. Soc. Am. J. 22, 160–163. https://doi.org/10.2136/sssaj1958.03615995002200020019x

Knotters, M., Brus, D.J., Oude Voshaar, J.H., 1995. A comparison of kriging, co-kriging and kriging combined with regression for spatial interpolation of horizon depth with censored observations. Geoderma 67, 227–246. https://doi.org/10.1016/0016-7061(95)00011-C

Kuhn, M., 2008. Building Predictive Models in R Using the caret Package. J. Stat. Softw. 28, 1–26. https://doi.org/10.18637/jss.v028.i05

Lane, P.W., 2002. Generalized linear models in soil science. Eur. J. Soil Sci. 53, 241–251. https://doi.org/10.1046/j.1365-2389.2002.00440.x

Legates, D.R., McCabe, G.J., 2013. A refined index of model performance: a rejoinder. Int. J. Climatol. 33, 1053–1056. https://doi.org/10.1002/joc.3487

Levi, M.R., Rasmussen, C., 2014. Covariate selection with iterative principal component analysis for predicting physical soil properties. Geoderma 219–220, 46–57. https://doi.org/10.1016/j.geoderma.2013.12.013

Liakos, K., Busato, P., Moshou, D., Pearson, S., Bochtis, D., Liakos, K.G., Busato, P., Moshou, D., Pearson, S., Bochtis, D., 2018. Machine Learning in Agriculture: A Review. Sensors 18, 2674. https://doi.org/10.3390/s18082674

Ma, Y., Minasny, B., Malone, B.P., Mcbratney, A.B., 2019. Pedology and digital soil mapping (DSM). Eur. J. Soil Sci. 70, 216–235. https://doi.org/10.1111/ejss.12790

Macedo, J., Bryant, R.B., 1989. Preferential Microbial Reduction of Hematite Over Goethite in a Brazilian Oxisol. Soil Sci. Soc. Am. J. 53, 1114. https://doi.org/10.2136/sssaj1989.03615995005300040022x

Malone, B.P., McBratney, A.B., Minasny, B., 2013. Spatial Scaling for Digital Soil Mapping. Soil Sci. Soc. Am. J. 77, 890. https://doi.org/10.2136/sssaj2012.0419

Marconi, A., 1974. Mineralogia de solos das séries Paredão Vermelho, Ribeirão Claro e Saltinho, do município de Piracicaba, SP. An. da Esc. Super. Agric. Luiz Queiroz 31, 403–418. https://doi.org/10.1590/s0071-12761974000100031

Marques, K.P.P., Demattê, J.A.M., Miller, B.A., Lepsch, I.F., 2018. Geomorphometric segmentation of complex slope elements for detailed digital soil mapping in southeast Brazil. Geoderma Reg. 14, e00175. https://doi.org/10.1016/J.GEODRS.2018.E00175

Martínez-Graña, A.M., Goy, J.L., Zazo, C., Silva, P.G., 2016. Soil map and 3D virtual tour using a database of soil-forming factors. Environ. Earth Sci. 75, 1402. https://doi.org/10.1007/s12665-016-6225-x

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

McBratney, A., de Gruijter, J., Bryce, A., 2019. Pedometrics timeline. Geoderma 338, 568–575. https://doi.org/10.1016/j.geoderma.2018.11.048

Mehra, O.P., Jackson, M.L., 2013. Iron oxide removal from soils and clays by a dithionite-citrate system buffered with sodium bicarbonate, in: Clays and Clay Minerals. Elsevier, pp. 317–327. https://doi.org/10.1016/B978-0-08-009235-5.50026-7

Meier, M., Souza, E. de, Francelino, M.R., Fernandes Filho, E.I., Schaefer, C.E.G.R., Meier, M., Souza, E. de, Francelino, M.R., Fernandes Filho, E.I., Schaefer, C.E.G.R., 2018. Digital Soil Mapping Using Machine Learning Algorithms in a Tropical Mountainous Area. Rev. Bras. Ciência do Solo 42. https://doi.org/10.1590/18069657rbcs20170421

Meyer, H., Reudenbach, C., Wöllauer, S., Nauss, T., 2019. Importance of spatial predictor variable selection in machine learning applications – Moving from data reproduction to spatial prediction. Ecol. Modell. 411. https://doi.org/10.1016/j.ecolmodel.2019.108815

Mulder, V.L.L., De Bruin, S., Schaepman, M.E.E., Mayr, T.R.R., 2011. The use of remote sensing in soil and terrain mapping — A review. Geoderma 162, 1–19. https://doi.org/10.1016/j.geoderma.2010.12.018

Nandra, S.S., 1974. Free iron oxide content of a tropical soil. Plant Soil 40, 453–456. https://doi.org/10.1007/BF00011532

Nanni, M.R., Demattê, J.A.M., 2006. Spectral Reflectance Methodology in Comparison to Traditional Soil Analysis. Soil Sci. Soc. Am. J. 70, 393. https://doi.org/10.2136/sssaj2003.0285

Nelder, J.A., 1977. A Reformulation of Linear Models. J. R. Stat. Soc. Ser. A 140, 48. https://doi.org/10.2307/2344517

Odgers, N.P., McBratney, A.B., Minasny, B., 2011. Bottom-up digital soil mapping. II. Soil series classes. Geoderma. https://doi.org/10.1016/j.geoderma.2011.03.013

Pebesma, E.J., 2004. Multivariable geostatistics in S: The gstat package. Comput. Geosci. 30, 683–691. https://doi.org/10.1016/j.cageo.2004.03.012

Pouladi, N., Møller, A.B., Tabatabai, S., Greve, M.H., 2019. Mapping soil organic matter contents at field level with Cubist, Random Forest and kriging. Geoderma 342, 85–92. https://doi.org/10.1016/j.geoderma.2019.02.019

QGIS Development Team, 2020. QGIS geographic information system. Open source geospatial foundation project.

Quinlan, J.R. (John R., Ross, J., 1993. C4.5 : programs for machine learning. Morgan Kaufmann Publishers.

Regmi, N.R., Rasmussen, C., 2018. Predictive mapping of soil-landscape relationships in the arid Southwest United States. Catena 165, 473–486. https://doi.org/10.1016/j.catena.2018.02.031

Rouse, J.W., Hass, R.H., Schell, J.A., Deering, D.W., 1973. Monitoring vegetation systems in the great plains with ERTS. Third Earth Resour. Technol. Satell. Symp. 1, 309–317. https://doi.org/citeulike-article-id:12009708

SAGA Development Team, 2016. SAGA GIS.

Samuel-Rosa, A., Heuvelink, G.B.M.B.M., Vasques, G.M.M., Anjos, L.H.C.H.C., 2015. Do more detailed environmental covariates deliver more accurate soil maps? Geoderma 243–244, 214–227. https://doi.org/10.1016/j.geoderma.2014.12.017

Santos, H.G., Jacomine, P.K.T., Anjos, L.H.C. dos, Oliveira, V.A. de, Lumbreras, J.F., Coelho, M.R., Almeida, J.A., Araújo Filho, J.C. de, Oliveira, J.B., Cunha, T.J.F., 2018. Sistema Brasileiro de Classificação de Solos, 5 ed. ed. Embrapa, Brasília - DF.

Sayão, V.M., Demattê, J.A.M., Bedin, L.G., Nanni, M.R., Rizzo, R., 2018. Satellite land surface temperature and reflectance related with soil attributes. Geoderma 325, 125–140. https://doi.org/10.1016/j.geoderma.2018.03.026

Schaetzl, R.J., Anderson, S., 2005. Soils genesis and geomorphology. Cambridge University Press, New York, NY.

Schwertmann, U., Cornell, R.M., 2000. Iron Oxides in the Laboratary. Wiley-VCH Verlag GmbH, Weinheim, Germany, Germany. https://doi.org/10.1002/9783527613229

Shahbazi, F., Hughes, P., McBratney, A.B., Minasny, B., Malone, B.P., 2019a. Evaluating the spatial and vertical distribution of agriculturally important nutrients — nitrogen, phosphorous and boron — in North West Iran. CATENA 173, 71–82. https://doi.org/10.1016/j.catena.2018.10.005

Shahbazi, F., McBratney, A., Malone, B., Oustan, S., Minasny, B., 2019b. Retrospective monitoring of the spatial variability of crystalline iron in soils of the east shore of Urmia Lake, Iran using remotely sensed data and digital maps. Geoderma 337, 1196–1207. https://doi.org/10.1016/j.geoderma.2018.11.024

Shen, Q., Demisie, W., Zhang, S., Zhang, M., 2020. The Association of Heavy Metals with Iron Oxides in the Aggregates of Naturally Enriched Soil. Bull. Environ. Contam. Toxicol. 104, 144–148. https://doi.org/10.1007/s00128-019-02739-2

Sindayihebura, A., Ottoy, S., Dondeyne, S., Van Meirvenne, M., Van Orshoven, J., 2017. Comparing digital soil mapping techniques for organic carbon and clay content: Case study in Burundi’s central plateaus. CATENA 156, 161–175. https://doi.org/10.1016/j.catena.2017.04.003

Theobald, D.M., Harrison-Atlas, D., Monahan, W.B., Albano, C.M., 2015. Ecologically-Relevant Maps of Landforms and Physiographic Diversity for Climate Adaptation Planning. PLoS One 10, e0143619. https://doi.org/10.1371/journal.pone.0143619

Ticknor, J.L., 2013. A Bayesian regularized artificial neural network for stock market forecasting. Expert Syst. Appl. 40, 5501–5506. https://doi.org/10.1016/j.eswa.2013.04.013

Vapnik, V.N., 2000. The Nature of Statistical Learning Theory. Springer New York, New York, NY. https://doi.org/10.1007/978-1-4757-3264-1

Vasques, G.M., Coelho, M.R., Dart, R.O., Oliveira, R.P., Teixeira, W.G., 2016. Mapping soil carbon, particle-size fractions, and water retention in tropical dry forest in Brazil. Pesqui. Agropecuária Bras. 51, 1371–1385. https://doi.org/10.1590/s0100-204x2016000900036

Viscarra Rossel, R.A., Bui, E.N., De Caritat, P., McKenzie, N.J., 2010. Mapping iron oxides and the color of Australian soil using visible-near-infrared reflectance spectra. J. Geophys. Res. Earth Surf. 115. https://doi.org/10.1029/2009JF001645

Vogt, J. V, Colombo, R., Bertolo, F., 2003. Deriving drainage networks and catchment boundaries: a new methodology combining digital elevation data and environmental characteristics. Geomorphology 53, 281–298. https://doi.org/10.1016/S0169-555X(02)00319-7

Xu, M., Watanachaturaporn, P., Varshney, P., Arora, M., 2005. Decision tree regression for soft classification of remote sensing data. Remote Sens. Environ. 97, 322–336. https://doi.org/10.1016/j.rse.2005.05.008

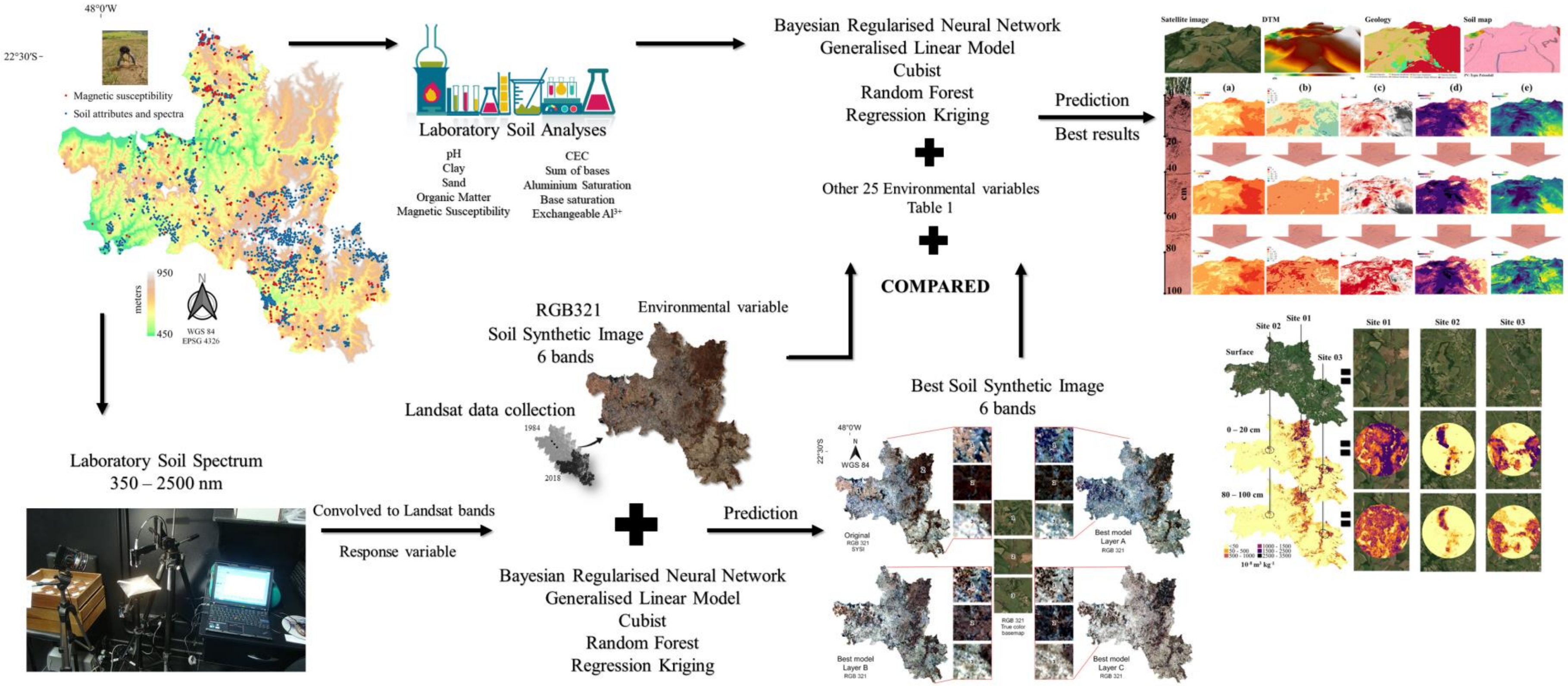
2. INTEGRATION OF MULTISPECTRAL AND HYPERSPECTRAL DATA TO MAP MAGNETIC SUSCEPTIBILITY AND SOIL ATTRIBUTES AT DEPTH: A NOVEL FRAMEWORK

ABSTRACT

The understanding of attributes and magnetic susceptibility (χ) at soil surface, mainly subsurface, is crucial due to their role to identify climate changes, soil degradation, soil classification systems, soil fertility, and pedogenesis. The integration of proximal sensing (PS) and remote sensing (RS) data sources could increase the efficiency of Digital Soil Mapping. Nevertheless, products of this integration need to be evaluated in hybrid, stochastic, and deterministic models to predict soil attributes and χ at surface and subsurface. This study investigates the PS and RS integration by applying four deterministic (e.g. Bayesian Regularised Neural Network, Generalised Linear Model, Random Forest and Cubist) and hybrid models (e.g. Regression Kriging of residuals of the best-fitted model) to create a new environmental variable, the Best Synthetic Soil Image (BSSI), at three soil depths (e.g. 0 – 20, 40 – 60 and 80 – 100 cm) that quantitatively represent the soil spectral signature. We also used the BSSI in a comparison with bare soil surface (e.g. SYSI - Synthetic Soil Image) to predict soil attributes and χ by performing the deterministic and hybrid models. We hypothesize that the BSSI, which integrates PS and RS data, enhances soil modelling predictions at subsurface by selecting the best model approach. The BSSI demonstrated original and valuable contribution to increase the predictive model power at deeper layers, while SYSI was effective at upper layers. The PS and RS integration helped to identify the main soil patterns horizontally and vertically, which traditional soil surveys have not been capable of representing.

Keywords: soil spectroscopy; soil mapping; remote sensing; pedometrics; pedology

Graphical Abstract



Laboratory Soil Analyses

pH
Clay
Sand
Organic Matter
Magnetic Susceptibility

CEC
Sum of bases
Aluminium Saturation
Base saturation
Exchangeable Al^{3+}

Bayesian Regularised Neural Network
Generalised Linear Model
Cubist
Random Forest
Regression Kriging

Other 25 Environmental variables
Table 1

COMPARED

Prediction

Best results

RGB321
Soil Synthetic Image
6 bands

Environmental variable

Landsat data collection

Laboratory Soil Spectrum
350 – 2500 nm

Convolved to Landsat bands

Response variable

Prediction

Bayesian Regularised Neural Network
Generalised Linear Model
Cubist
Random Forest
Regression Kriging

Best Soil Synthetic Image
6 bands

2.1. INTRODUCTION

The knowledge of soil attributes and properties at the surface, mainly subsurface, is crucial for pedogenesis. Soil attributes, such as clay and sand contents, soluble aluminium (Al^{3+}), aluminium (Al_{sat}) and base saturation (B_{sat}), cation exchange capacity (CEC), soil organic matter (SOM), sum of bases (SB), and the pH play a crucial role to identify climate changes (Gray and Bishop, 2019; Minasny et al., 2017), soil degradation (Chen and Rao, 2008; Lal, 2015; Nampak et al., 2018), soil classification systems (Demattê et al., 2019; Rizzo et al., 2020), soil fertility (Demattê et al., 2017; Li et al., 2018), and soil security (Bennett et al., 2019; McBratney et al., 2014). Likewise, soil property, such as the magnetic susceptibility (χ), which stems mainly from soil maghemite and magnetite contents, helps to understand pedogenesis (De Jong et al., 2000; de Souza Bahia et al., 2017; Jordanova, 2016; Lourenço et al., 2014; Maher, 1998; Torrent et al., 2010) and map soil classes (Ramos et al., 2017; Silvero et al., 2019; Teixeira et al., 2018).

The Digital Soil Mapping (DSM) is an easy and feasible approach to improve the understanding of soil attributes and properties. The DSM has been applied to predict soil classes (Silva et al., 2016; Triantafilis et al., 2009; Vincent et al., 2018), morphology (Demattê, 2016; Hartemink et al., 2020), parent materials (Bonfatti et al., 2020), and attributes, such as clay (Loiseau et al., 2019), SOM (Gray et al., 2016), pH (Dharumarajan et al., 2020), among others (Padarian et al., 2017; Poppiel et al., 2019b), contributing to the improvement of management practices (Minasny et al., 2017; Tajik et al., 2020). The DSM basis was formalised in the *scorpan* model by McBratney et al. (2003) and it takes into account the model of soil formation established by Jenny (1941). This framework deals with the spatial prediction of soil attributes and properties, which could be performed by the stochastic or deterministic models and by combining both approaches (hybrid). Therefore, the soil attribute predicted is the response variable and its predictors are environmental variables that explain its spatial behaviour in the landscape.

The remote sensing (RS) and proximal sensing (PS) products provide valuable information to enhance the environmental variables, which simulate characteristics of the environmental conditions (e.g. vegetation, climate, soil, etc.), compromising the DSM efficiency. The Digital Elevation Models (DEM) is an example of RS data retrieved from different satellite data sensors (e.g. Shuttle Radar Topography Mission, Light Detection and Ranging), the source of relief features for soil modelling. The DEM are the basis for the topographic wetness index, slope, curvature, and many other relief features, which are related to the soil attribute mapped. Another example of RS data is the spectral information retrieved by satellite sensors. Demattê et al. (2018) retrieved bare soil surface using temporal Landsat collection and showed its potential for predicting soil attributes. This potential was assessed by Fongaro et al. (2018) by presenting model improvements for clay and sand quantification. Gallo et al. (2018) also evaluated and proved the potential of using bare soil surface image as a predictor of sand, clay, CEC, and SOM.

Most studies on DSM applied environmental variables derived from multispectral data (e.g. satellite sensors) and digitalised maps (Behrens et al., 2014; Gray et al., 2016; Minasny and McBratney, 2016; Rutgers et al., 2019); however, these studies did not find models that were well-fitted at soil subsurface (Table 1), a limitation for soil mapping. The integration of PS and RS data sources could overcome this limitation. The successful integration of multi and hyperspectral data is key for a better prediction of soil properties (Crucil et al., 2019; Demattê et al., 2015; Poppiel et al., 2019a). Mendes et al. (2019) mapped subsurface using subsurface soil reflectance as an environmental variable and involved the laboratory spectra of soil samples at 80 – 100 cm depth into Landsat TM bands, the response variables. The bands of the Synthetic Soil Image were the environmental variables, as described at Demattê et al. (2018). This environmental variable represents surface bare soils retrieved from Landsat TM time-series. The soil spectra were predicted at subsurface by using the DSM approaches and by applying the multiple linear regression and geographically weighted regression. Afterwards, the predicted bands of soil spectra at subsurface were used as environmental variables to map soil attributes at the same depth. This procedure of predicting an environmental variable at soil subsurface was pioneer and was named spectral pedotransfer (SPEDO) (Mendes et al., 2019).

Table 1. Comparison of studies predicting soil properties using a variety of deterministic, stochastic and hybrid methods in Digital Soil Mapping.

Soil attribute	Depth (cm)	1 sample/x km^2	Methods	RMSE	R^2	Reference
pH	0-30	0.73	GLM	0.15	0.05	(Mosleh et al., 2016)
	0-30	227.00	RK	0.68-0.84	0.06-0.14	(Malone et al., 2014)
	60-100	316.28	BMA	1.09-1.13	0.08-0.11	(Malone et al., 2014)
	0-30	15.38	RF, RK	0.75-0.83	0.71-0.75	(Vaysse and Lagacherie, 2015)
	30-60	17.76	RF, RK	0.82-0.83	0.71	(Vaysse and Lagacherie, 2015)
	0-20	0.55	Lasso, georob, geoGAM, BRT, RF, MA	0.83-0.93	0.35-0.45	(Nussbaum et al., 2018)
	40-60	0.83	Lasso, georob, geoGAM, BRT, RF, MA	1.07-1.25	-0.22-0.30	(Nussbaum et al., 2018)
Clay ($g\ kg^{-1}$)	0-15	191.29	QRF	76.5	0.45	(Loiseau et al., 2019)
	30-60	195.22	QRF	~100	0.35	(Loiseau et al., 2019)
	60-100	619.18	QRF	~130	0.20	(Loiseau et al., 2019)
	0-60	0.04	MLR	3.68-12.37	0.37-0.52	(Godinho Silva et al., 2016)
	0-20	1.45	GWR, MLR	82.87-91.88	0.54-0.62	(Mendes et al., 2019; Rizzo et al., 2020)
	80-100	1.45	GWR, MLR	88.24-101.43	0.54-0.63	(Mendes et al., 2019; Rizzo et al., 2020)
	0-20	3.27	PLSR	89.84	0.75	(Gallo et al., 2018)
	0-20	15.90	Cubist, RF	65.01-97.73	0.61-0.83	(Fongaro et al., 2018)
	0-30	0.73	ANN	87	0.22	(Mosleh et al., 2016)
	0-100	66.86	MLR	41.70-69.00	0.19-0.63	(Angelini et al., 2017)
	0-10	0.27	Lasso, georob, geoGAM, BRT, RF, MA	57.76-66.98	0.23-0.42	(Nussbaum et al., 2018)
	50-100	0.34	Lasso, georob, geoGAM, BRT, RF, MA	88.71-97.06	-0.16-0.02	(Nussbaum et al., 2018)
Sand ($g\ kg^{-1}$)	0-30	60.91	QRF	158.76-178.30	0.43-0.53	(Dharumarajan et al., 2020)
	60-100	60.91	QRF	142.50-166.10	0.36-0.54	(Dharumarajan et al., 2020)
	0-20	15.90	Cubist, RF	79.99-128.52	0.63-0.86	(Fongaro et al., 2018)
	0-20	3.27	PLSR	151.70	0.56	(Gallo et al., 2018)
	0-10	0.19	Lasso, georob, geoGAM, BRT, RF, MA	52.18-58.96	0.06-0.26	(Nussbaum et al., 2018)
	50-100	0.23	Lasso, georob, geoGAM, BRT, RF, MA	104.07-117.70	0.02-0.23	(Nussbaum et al., 2018)
OM ($g\ kg^{-1}$)	0-100	66.86	MLR	0.06-4.10	0.04-0.28	(Angelini et al., 2017)
	0-20	3.27	PLSR	22.31	0.34	(Gallo et al., 2018)
	0-10	0.19	Lasso, georob, geoGAM, BRT, RF, MA	31.58-35.04	0.08-0.25	(Nussbaum et al., 2018)
	50-100	0.23	Lasso, georob, geoGAM, BRT, RF, MA	60.90-75.36	-0.22-0.20	(Nussbaum et al., 2018)
	0-20	0.01	MLR	6.72-10.62	0.15-0.55	(Sayão et al., 2018)
	0-25	*28.5/ha	RK(Cubist and RF), Cubist, RF, Kriging	2.27-4.20	0.88-0.91	(Pouladi et al., 2019)
CEC ($mmol_c\ kg^{-1}$)	0-20	0.55	Lasso, georob, geoGAM, BRT, RF, MA	72.29-83.04	0.26-0.44	(Nussbaum et al., 2018)
	40-60	0.83	Lasso, georob, geoGAM, BRT, RF, MA	51.36-83.24	-0.68-0.36	(Nussbaum et al., 2018)
	0-20	3.27	PLSR	58.60	0.40	(Gallo et al., 2018)
	80-100	1.45	GWR, MLR	69.82-84.37	0.02-0.35	(Mendes et al., 2019; Rizzo et al., 2020)
	5-15	50.72	DSMART	51.12	0.28	(Ellili Bargaoui et al., 2019)
	30-60	54.78	DSMART	21.50	0.34	(Ellili Bargaoui et al., 2019)
	0-20	0.02	MLR, OK, RK	15.40-22.80	0.13-0.60	(Sun et al., 2019)

	0-30	8/ha	Bayesian model	7.2	**0.69	(Li et al., 2018)
	60-90	8/ha	Bayesian model	26.2	**0.86	(Li et al., 2018)
SB (mmol$_c$ kg^{-1})	5-15	353.71	Cubist, RF	47.86-97.72	**0.46-0.78	(Gray et al., 2016)
	0-30	403.30	MLR	53.70	**0.72	(Gray and Bishop, 2019)
	30-100	403.30	MLR	61.65	**0.74	(Gray and Bishop, 2019)
Al^{3+} (mmol$_c$ kg^{-1})	80-100	1.45	GWR, MLR	11.73-14.56	0.07-0.32	(Mendes et al., 2019; Rizzo et al., 2020)
AS (%)	80-100	1.45	GWR, MLR	17.01-19.16	0.11-0.27	(Mendes et al., 2019; Rizzo et al., 2020)
	0-20	104.55	RF	20	0.26	(Poppiel et al., 2019)
	20-60	104.55	RF	21	0.45	(Poppiel et al., 2019)
	80-100	104.55	RF	20	0.56	(Poppiel et al., 2019)
BS (%)	0-20	1.45	GWR, MLR	14.83-16.04	0.24-0.31	(Mendes et al., 2019; Rizzo et al., 2020)
	80-100	1.45	GWR, MLR	20.19-21.09	0.05-0.12	(Mendes et al., 2019; Rizzo et al., 2020)
	0-20	104.55	RF	20	0.18	(Poppiel et al., 2019)
	20-60	104.55	RF	15	0.30	(Poppiel et al., 2019)
	80-100	104.55	RF	14	0.36	(Poppiel et al., 2019)

*Small areas represented in hectares instead of square kilometres. **Lin's concordance index.

CEC, Cation Exchange Capacity; SB, Sum of Bases; AS, Aluminium Saturation; BS, base saturation; OM, Organic Matter; Al^{3+}, exchangeable Al^{3+}.

GLM, Generalised Linear Model; RK, Regression Kriging; RF, Random Forest; QRF, Quantile Regression Forest; MLR, Multiple Linear Regression; PLSR, Partial Least Square Regression; ANN, Artificial Neural Network; KNN, K-Nearest Neighbour; Lasso, grouped least absolute shrinkage and selection operator; georob, robust external-drift kriging; geoGAM, boosted geoadditive model; BRT, boosted regression tree; MA, model averaging; DSMART, Disaggregation and Harmonisation of Soil Map Units Through Resampled Classification Trees; OK, Ordinary Kriging.

Another major aspect of DSM is to select the best modelling algorithm. Improvements in modelling soil formation factors and processes have been performed via geospatial and spatial analyses, characterised by the use of some particular treatments (e.g. data transformation; Hengl et al. 2004) and field designs (e.g. Latin hypercube; Carré et al. 2007) to reduce uncertainty, an intrinsic part of the soil taken as a natural system (McBratney et al., 2000). The knowledge generated from the stochastic, deterministic, and hybrid models has uncovered quantitative patterns of soil attributes and formation factors. These models comprise the Bayesian Regularised Neural Network (Poggio et al., 2016; Tien Bui et al., 2012), Generalised Linear Model (McKenzie and Austin, 1993; Tajik et al., 2020), Random Forest (Castro-Franco et al., 2018; Hengl et al., 2018), Cubist (Bonfatti et al., 2016; Malone et al., 2018), Regression Kriging (Keskin and Grunwald, 2018), and among other untested algorithms in soil science. Machine learning algorithms classify instances of unknown identity using samples of known targets (Cracknell, 2007).

The Bayesian Regularised Neural Network (BRNN) relies on supervised learning, a technique in the artificial neural network approach. The BRNN is a mathematical technique that converts nonlinear systems into a unique solution, which changes behaviour continuously in relation to the initial conditions. According to Ticknor (2013), the BRNN trains and calculates on non-trivial weights, converging them to a constant, as the network increases. The model complexity is penalised and unnecessary linkages are driven to zero, separating the unnecessary linkages (leaving those apart). Therefore, we conducted a comprehensive review of this approach to DSM (Table 1). The Generalised Linear Model (GLM) adapts the model instead of changing the input data. Besides, the GLM is a lengthening of linear regressions, accommodating non-normal response distributions (Lane, 2002; McBratney et al., 2003). The Random Forest (RF) and Cubist models are decision tree algorithms that differ from each other in terms of the way to deal with variance reduction. The decision tree methods, RF and Cubist, are widely and well-posed in the literature for DSM (Gray et al., 2016; Pouladi et al., 2019; Shahbazi et al., 2019). The Regression Kriging (RK), as highlighted by Keskin and Grunwald (2018), is extensively known in soil science because of its practicality and robusticity as a hybrid spatial interpolator. Several studies have investigated RK with the residuals of other machine learning on DSM (Angelini and Heuvelink, 2018; de Carvalho Junior et al., 2014; Knotters et al., 1995; Odgers et al., 2011; Pouladi et al., 2019; Sayão et al., 2018; Sindayihebura et al., 2017; Vasques et al., 2016).

This study investigates the SPEDO method by applying four deterministic (e.g. Bayesian Regularised Neural Network, Generalised Linear Model, Random Forest, and Cubist) and hybrid models (e.g. Regression Kriging of residuals of the best-fitted model) to create a new environmental variable at three soil depths (e.g. 0 – 20, 40 – 60 and 80 – 100 cm) that represents quantitatively the soil spectral signature. We also used the new environmental variable and compared it with the bare soil surface variable to predict soil attributes, namely clay, sand, and SOM contents, pH in water, CEC, SB, Al^{3+}, Al_{sat}, and B_{sat}, as well as χ by performing the deterministic and hybrid models. Our hypothesis is that the new environmental variable, which integrates PS and RS data, enhances soil modelling predictions at subsurface by selecting the best model approach.

2.2. MATERIAL AND METHODS

2.2.1. Characterising the study area

The study site covers approximately 2,574 km^2 in the municipalities of Piracicaba, Charqueada, Iracemápolis, Saltinho, Rio das Pedras, Mombuca, Rafard and Capivari, São Paulo State, Brazil (Fig. 1). The climate in the region is characterized by dry winters and rainy summers, with an annual average temperature between 20.1 and 22.5 °C, and annual average rainfall between 1200 and 1400 mm (INMET, 2020). The relief in the region consists of undulating hills and rolling uplands with altitude ranging from 450 to 950 m. The region has great diversity of parent material, such as Carboniferous materials composed of siltstones, tillites, varvites, conglomerates, and sandstones (Tubarão Group); Permian composed of shales, limestones, siltstones and flint (Corumbataí Formation); Permian consisting of shales, dolomite, siltite and pyrombetuminosite (Irati Formation); Jurassic consisting of sandstone, shales and siltstones (Botucatu and Pirambóia Formation); as well as Cretaceous constituted of diabase and basalt (Serra Geral Formation) (IGC, 2018). This diversity of parent materials and topography confer a great variety of soils to the region. The economic activity of the region is predominantly agriculture with till and no-till farming, meaning that some sites have the soil revolved up to 60 cm of depth along the year (Rudorff et al., 2010).

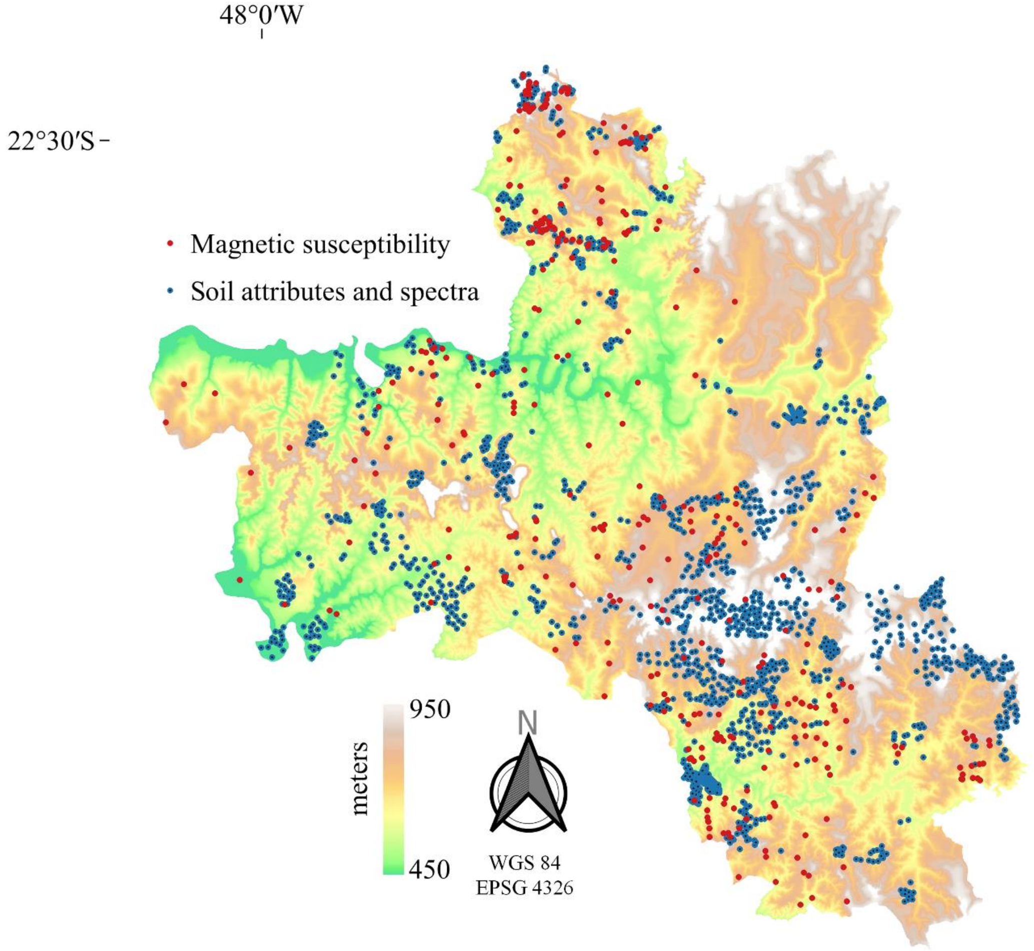


Fig. 1. Study area.

2.2.2. Soil and spectral data

The dataset consisted of the laboratory chemical and physical analyses of soil spectra (Table 2). The soil samples were collected using an auger at 0-20 cm (layer A), 40-60 cm (layer B), and 80-100 cm (layer C) depths based on the toposequence method, which considers relief and geological variation along the landscape. The toposequence method consists of selecting soil samples by a topographic profile along a transect crossing a map from summit to toeslope (Gobin et al., 2000). For the soil chemical and physical analyses, we collected 5,689 samples at layers A (2,229 samples), B (1,796 samples), and C (1,664 samples). The number of samples for each layer differed because some of the sampling points presented soil compaction or reached the parent material.

Table 2. Number of observations randomly split by 80% for calibration and 20% for validation at three different soil depths.

Soil data source	Calibration			Validation		
	A	B	C	A	B	C
Magnetic Susceptibility	279	-	153	71	-	37
Laboratory Spectrum	808	656	701	200	164	172
Laboratory Analyses	1,785	1,438	1,333	444	358	331

Note. Laboratory Analyses include pH in water; clay, sand, and OM content; Al^{3+}; sum of bases; cation exchange capacity; aluminium and base saturation. A: 0 – 20 cm depth. B: 40 – 60 cm depth. C: 80 – 100 cm depth.

The samples were oven-dried at 45°C for 48 h and then, ground, sieved at 2-mm mesh, and analysed. The chemical attributes determined were pH in water, exchangeable bases (Ca^{2+}, Mg^{2+} and K^{+}), soluble aluminium (Al^{3+}), potential acidity (H^{+} + Al^{3+}), and SOM, according to the methodology described by Camargo et al. (2009). We calculated the sum of bases (SB, $mmol_c$ kg^{-1}), CEC (CEC, $mmol_c$ kg^{-1}), percentage of base saturation (B_{sat}, %), and percentage of Al saturation (Al_{sat}, %). The clay and sand contents were determined using the densimeter method and sieving, respectively, as described in Camargo et al. (2009). For the soil spectral analysis, 2,701 out of 5,689 soil samples at layers A (1008 samples), B (820 samples), and C (873 samples) were placed on Petri dishes. The Fieldspec 3 sensor (Analytical Spectral Devices, Boulder, Colorado, USA) was used to obtain soil reflectance spectra in the spectral range of 350 – 2500 nm in the laboratory. The sensor, vertically positioned at 8 cm from the platform, spotted the energy reflected from two 50-W halogen lights with no-collimated beam to the target plane. The lights were positioned at 35 cm from the platform at a zenith angle of 30°. Three measurements were carried out for each sample turning the Petri dishes 90° between the sensor reading intervals. A white plate was used as a white reference (100% reflectance). Afterwards, the average of all three readings, along with the white reference reflectance, was used to calculate the final reflectance factor for each sample. The magnetic susceptibility (χ, 10^{-8} m^3 kg^{-1}) was analysed using 540 out of 5,689 soil samples at layers A (350 samples) and C (190 samples). This soil property was determined in 10 g of air-dried fine soil and the clay fraction using the Bartington MS2 equipment coupled to a Bartington MS2B sensor at low and high frequencies of 0.47 kHz and 4.7 kHz, respectively, according to the methodology described by Dearing (1999).

2.2.3. Modelling approach

2.2.3.1. Environmental variables

The DEM was retrieved from the Shuttle Radar Topography Mission (USGS, 2018) and used to derive relief variables, such as aspect, LS factor, plan and profile curvatures, slope, valley depth, and the topographic wetness index in QGIS (QGIS Development Team, 2020). The Normalised Difference Vegetation Index (NDVI) and the Enhanced Vegetation Index (EVI) were calculated by averaging 35-year Landsat images (moist and dry seasons). The Drainage Density was calculated (total length channels, m / basin area, m^2) in the ArcGIS version 10.3 by vectorising all channels of digital aerial photographs from the Geographic and Cartographic Institute of Sao Paulo (IGC, 2018) in PHOTOMOD Lite 6.3 software. The gypsic, natric, calcareous, carbonate, ferrous iron, ferrous, ferrous oxide, and clay (hydroxides) indices, which typify the soil chemical composition, were calculated using the bands of the Synthetic Soil Image according to the formulas described by Regmi and Rasmussen (2018). The munsell colour was acquired from the study of Silvero et al. (2021). The soil relative frequency was created based on how many times each pixel of 35-year Landsat collection was classified as a bare soil surface. Geology and geomorphology were acquired from Bonfatti et al. (2020). We summarised the environmental variables selected to predict the soil spectra and attributes at surface and subsurface in Table 3.

Table 3. Characteristics of the environmental variables selected as predictors of the soil attributes in the digital soil mapping procedure.

Ancillary variables	Unit	Resolution (m)	Type of variable	Characteristics	Reference
Synthetic Soil Image's bands	Spectral reflectance	30	Continuous	Bare soil areas	(Demattê et al., 2018)
Munsell colour	Dimensionless	30	Factor	Soil colour	(Silvero et al., 2021)
NDVI	Dimensionless	30	Continuous	Vegetation	(Rouse et al., 1973)
EVI	Dimensionless	30	Continuous	Vegetation	(Huete, 2004)
Soil Relative Frequency	%	30	Continuous	Human activity	(Demattê et al., 2018)
DEM from SRTM	m	30	Continuous	Relief	(USGS, 2018)
Aspect from DEM	degree	30	Continuous	Downhill slope faces	(Conrad et al., 2015)
LS Factor from DEM	Dimensionless	30	Continuous	Component of the Revised Universal Soil Loss equation	
Plan Curvature from DEM	degree m^{-1}	30	Continuous	(-) concave/ (+) convex contours	
Profile Curvature from DEM	degree m^{-1}	30	Continuous	(-) convex/ (+) concave contours	
Slope from DEM	%	30	Continuous	Relief inclination	
Valley Depth from DEM	m	30	Continuous	Vertical distance to the base level of the channel network	
TWI from DEM	Dimensionless	30	Continuous	Soil water content	
Drainage Density	m^{-1}	30	Continuous	Drainage network	
Gypsic index	Dimensionless	30	Continuous	Gypsiferous soil	(Regmi and Rasmussen, 2018)
Natric index	Dimensionless	30	Continuous	Sodium rich soil	
Calcareous index	Dimensionless	30	Continuous	Discriminate calcareous sediments from igneous rocks or sediments	
Carbonate radicals index	Dimensionless	30	Continuous	Carbonate radicals	
Ferrous Fe index	Dimensionless	30	Continuous	Ferrous Fe	
Ferrous index	Dimensionless	30	Continuous	Ferrous	
Ferrous oxides index	Dimensionless	30	Continuous	Ferrous oxide	
Clay and hydroxides index	Dimensionless	30	Continuous	Clay and hydroxides	
Geology	Dimensionless	30	Factor	Parent material	(Bonfatti et al., 2020)
Geomorphology	Dimensionless	30	Factor	Hillslope position	

2.2.3.2. Algorithms

The Bayesian Regularised Neural Network, Generalised Linear Model, Random Forest and Cubist models were performed using the "caret" R package (Kuhn, 2008), which required the "brnn" (Gianola et al., 2011), "glm" (Venables and Ripley, 2002), "ranger" (Breiman, 2001), and "Cubist" (Quinlan and Ross, 1993) R packages, respectively. The Regression Kriging of residuals of the best-fitted model was performed using the "automap" R package (Gianola et al., 2011). The model parameters are presented in Table S1. The algorithms were performed using the R programming (R Development Core Team, 2020).

2.2.3.3. Synthetic Soil Image and Bare Soil Image at three depths

The soil spectral pattern is represented by the Synthetic Soil Image (SYSI) as part of the environmental variables to predict soil attributes; however, we predicted a bare soil spectral image at layers A, B, and C using the laboratory soil spectrum (LSS) in the visible, near, and shortwave infrared (Vis-NIR-SWIR; 350 – 2500 nm) region. The SYSI is a mosaic of bare soil surface retrieved from Landsat images during the dry season (July to September) between 1984 and 2018. The dry-season reduced cloud coverage, providing a higher absolute frequency of bare soil areas also reducing moisture influence on the spectra. The entire method to generate the SYSI is described at Demattê et al. (2018) and represents bare soil areas at the soil surface (layer A). To create a bare soil image at the three different depths, we used the SYSI bands as predictors and the LSS as a response variable. The soil spectra data measured were convolved to Landsat spectra bands (excluding the thermal band), as described in Ben-Dor and Banin (1995), and Demattê et al. (2018). Thus, laboratory soil spectra for band 1 (B1, blue, 0.45 – 0.52 μm), band 2 (B2, green, 0.52 – 0.60 μm), band 3 (B3, red, 0.63 – 0.69 μm), band 4 (B4, near-infrared, 0.76 – 0.90 μm), band 5 (B5, shortwave infrared 1, 1.55 – 1.75 μm), and band 7 (B7, shortwave infrared 2, 2.08 – 2.35 μm) were the spectral average of the wavelength allocated to each Landsat spectral band, respectively. Based on the DSM principles, the SYSI bands were the environmental variables applied to the response variable for each spectral band, using the machine learning algorithms (MLAs) and the Regression Kriging of residuals of the best model resulted from four MLAs. Mendes et al. (2019) named this process of predicting soil spectra at depth as the Spectral Pedotransfer Function (SPEDO).

2.2.3.4. Model calibration and evaluation

The soil chemical, spectral, and physical data were split into internal calibration, which is used to calibrate the models, and external validation sets, used to assess the models performance (Table 2). The calibration data for model fitting were set by using a fivefold repeated cross-validation method, executed five times, to avoid the effects of environmental variables autocorrelation, as described by Meyer et al. (2019). The model tuning parameters were selected based on the lowest root mean squared error (RMSE) and highest adjusted coefficient of determination (R^2_{adj}) in the "caret" R package. The metrics of model assessment used in this study were the RMSE, R^2_{adj}, concordance correlation coefficient (CCC), and bias. Each parameter explains the relationship between the predicted and observed values in distinct ways. The RMSE (Eq. (1)) explains the proximity of the predicted values to the real values by using the square root of squares of the residuals, which sum up the degree of residuals. The R^2_{adj} verifies the variance proportion of covariates that affect the response variable by the approximated line of regression that the model explains (Eq. (2)). This metric allows digital soil mappers to compare model performances between distinct target variables. The Bias is calculated as the distance from the average prediction and observed values (Eq. (3)). This other validation metric fulfils one of the limitations of R^2_{adj}, showing the model bias. The last validation metric of model performance is Lin's CCC (Eq. (4)), which assesses the agreement between the predicted and observed values and could be a more appropriate metric than R^2_{adj}.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{1}$$

$$R_{adj}^2 = 1 - \frac{(SS_{res}/df_e)}{(SS_{tot}/df_t)} \tag{2}$$

$$Bias = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i) \tag{3}$$

$$CCC = \frac{2\rho\sigma_{pred}\sigma_{obs}}{\sigma_{pred}^2 + \sigma_{obs}^2 + \left(\mu_{pred} - \mu_{obs}\right)^2} \tag{4}$$

Where *n*, y_i, and $\hat{y}_i$ are sample size, observed values, and predicted values of the response variable, respectively. SS_{res}, SS_{tot}, df_e, and df_t are respectively the sum of squares of the regression residual, the sum of the square of the total residual, the degrees of freedom of the estimated population error variance, and the degree of freedom of the estimated population variance of the dependent variable. σ_{pred}^2 and σ_{obs}^2 are the prediction and observation variances, respectively, μ_{pred} and μ_{obs} are the means of the predicted and observed values. ρ is the correlation coefficient between the predicted and observed values.

2.3. RESULTS AND DISCUSSION

2.3.1. Integrating Proximal and Remote Sensing data

The performance results of the models for layers A, B, and C are presented in Fig. 2. For layer A (Fig. 2a and Table S2), the BRNN and Cubist algorithms had the best overall performance compared with GLM, RF, and RK of BRNN residuals for all spectral bands. The RF showed the worst performance with higher RMSE and Bias, lower R^2_{adj} and CCC. For layer B (Fig. 2b and Table S3), the GLM presented the best metrics of model assessment than any other models. The RF did not fit well the model for layer A. For layer C (Fig. 2c and Table S4), the model efficiency in all algorithms was rather similar, except for the RK of the GLM residuals. This shows that any of the models could be applied to predict soil spectra in layer C, where sampling density was 0.27 samples/km^2. Mendes et al. (2019) found sampling density of 0.55 samples/km^2 mapping soil spectra using the Multiple Linear Regression and Geographically Weighted Regression (GWR). These authors obtained the best result using GWR with R^2_{adj} and RMSE of 0.62 and 0.02, 0.72 and 0.03, 0.69 and 0.04, 0.69 and 0.05, 0.69 and 0.11, 0.67 and 0.05 for bands 1, 2, 3, 4, 5, and 7, respectively. Sampling densities were twofold the densities found in our study. We achieved similar performance of the models using MLAs (i.e. Cubist and RF) for bands 1, 2, 3, 4, 5, and 7 with R^2_{adj} and RMSE of 0.22 and 0.04, 0.19 and 0.07, 0.16 and 0.08, 0.20 and 0.10, 0.19 and 0.14, 0.22 and 0.10, respectively.

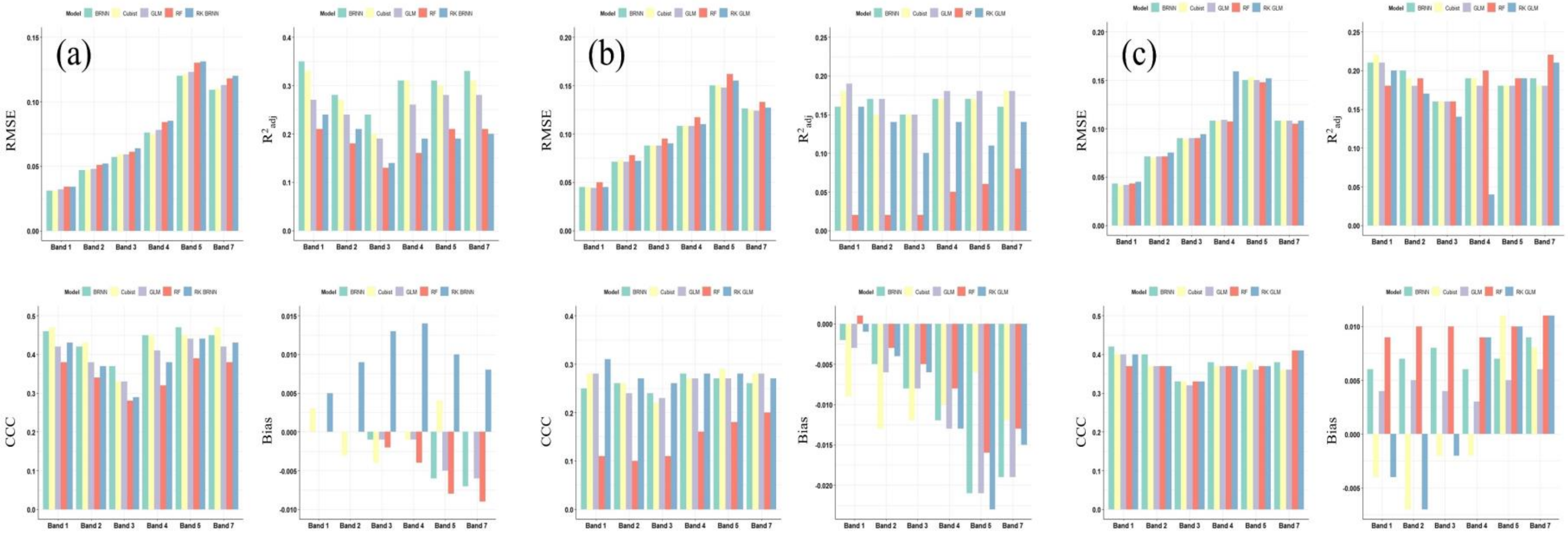


Fig. 2. Model evaluation at 0 – 20 (a), 40 – 60 (b) and 80 – 100 (c) cm depth for the predicted reflectance spectra in the Vis-NIR-SWIR based on observed laboratory convolved spectra. GLM, Generalised Linear Model; BRNN, Bayesian Regularised Neural Network; RF, random forest; RK, regression kriging of the residuals.

The mean spectral curve feature with 5% of a confidence interval for the three layers and six bands (Fig. 3) showed that BSSI had a similar reflectance factor to LSS. The spectral features from LSS, BSSI, and SYSI were qualitatively analogous; however, their spectral reflectance differed. The spectral curve of BSSI in layer A (Fig. 3a) followed the same pattern of the convolved spectra obtained in the laboratory. This tendency continued in layers B (Fig. 3b) and C (Fig. 3c). We highlight that there is limited evidence on subsurface reflectance prediction from surface spectra (Mendes et al., 2019).

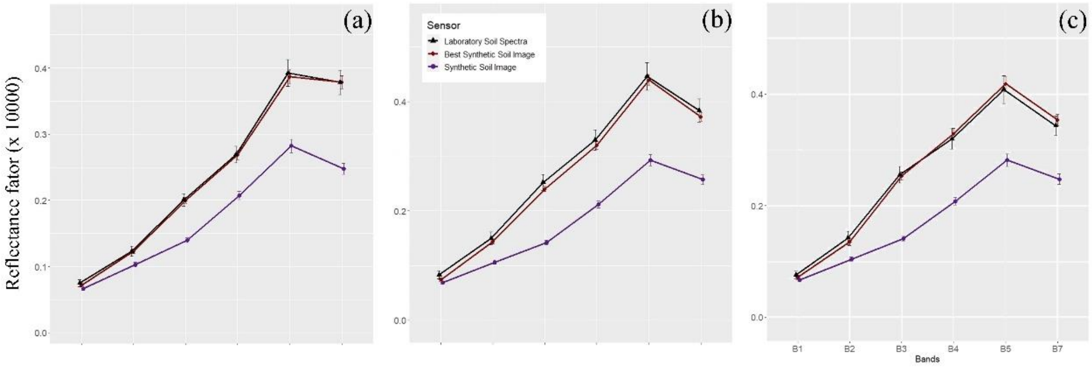


Fig. 3. Interval of confidence at 95% for the best model, laboratory convolved spectra and original SYSI in the layer A at 0 – 20 cm depth (a), layer B at 40 – 60 cm depth (b), and layer C at 80 – 100 cm depth (c) based on the mean values for each spectral band.

2.3.2. Soil dataset

Nine soil attributes and one soil property were analysed and mapped. The descriptive statistics of these soil components are shown in Fig. 4 (Tables S5 and S6). Clay, sand, SOM, pH in water, and B_{sat} had skewness close to zero, meaning that data is almost symmetrically or normally distributed in layer A. Al^{3+} was strongly right skewed or unimodal right, while SB, CEC, Al_{sat} were right skewed. In layer B, SOM, Al^{3+}, SB, CEC and B_{sat} were not normally distributed. In layer C, SOM, Al^{3+}, SB, and CEC were right skewed, meaning that they were not normally distributed. Normal distribution is an important characteristic that needs evaluation before modelling because the response variable could be better fitted, if normally distributed (Malone et al., 2013). Another significant statistical measure is the coefficient of variation (CV), which provides information on the dispersion frequency of the target variable. The relationship between accuracy of regression-kriging models and CV were well investigated by Keskin and Grunwald (2018). The authors highlighted that a higher CV means lower model accuracy. Soil attributes in the three layers presented CV up to 6%, meaning low variability. The study of Silvero et al. (2021) corroborates this finding.

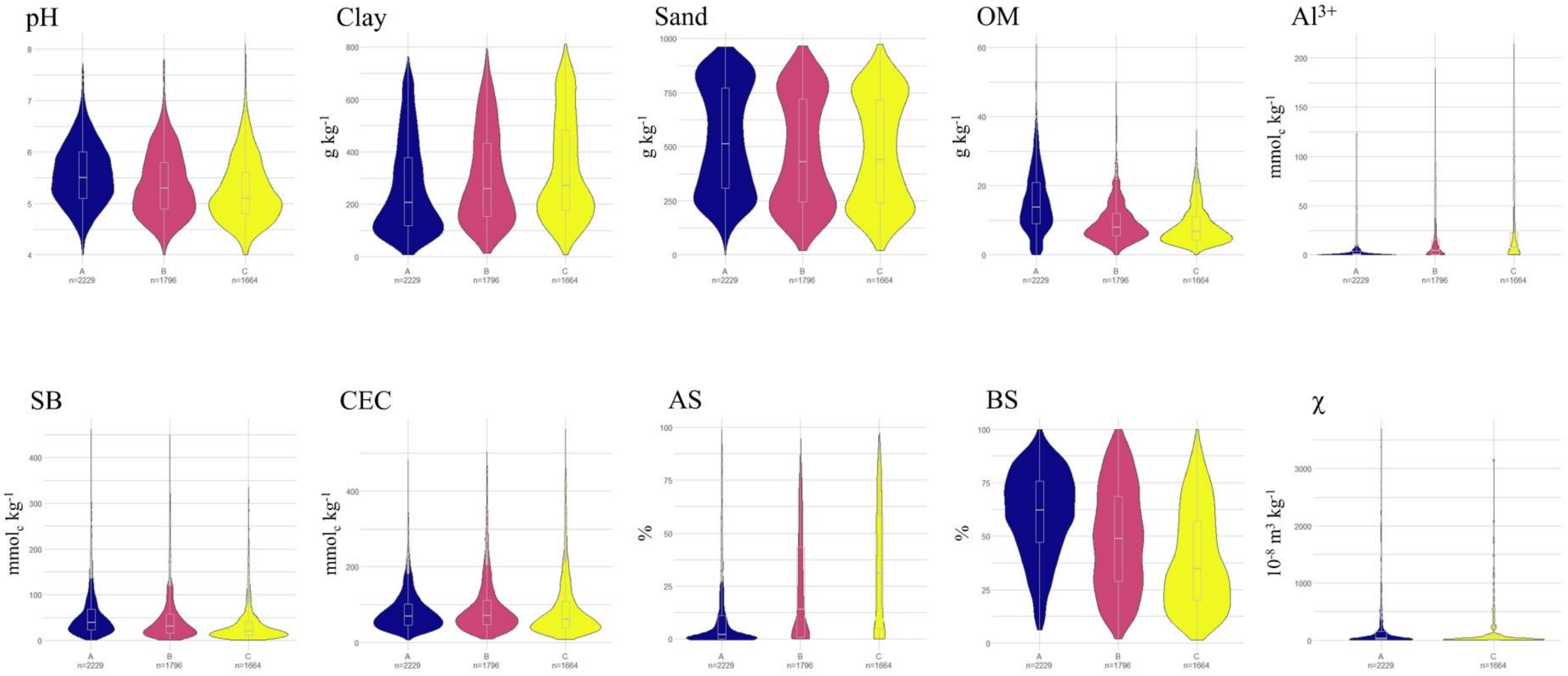


Fig. 4. Descriptive statistics of the soil attributes and property analysed in the study area at 0 – 20, 40 – 60 and 80 – 100 cm depth. OM, Organic Matter; Al^{3+}, soluble Al; SB, sum of bases; CEC, Cation Exchange Capacity; AS, Aluminium Saturation; BS, Base Saturation; χ, Magnetic Susceptibility.

Moreover, soil attributes reasonably represented the characteristics of the study site, which has very clayey to sandy soils (Silvero et al., 2021). Magnetic susceptibility (χ) at soil surface showed readings from 2.1 to 3,689 with a CV, mean value, and standard deviation of 2.32%, 239, and 555.9, respectively. Studies carried out using samples from soil surface in Brazil presented χ values varying from 48 to 9,670 and higher CV values compared to our data (Godinho Silva et al., 2016; Silvero et al., 2019; Teixeira et al., 2018). In layer C, χ had a CV, mean value, and standard deviation of 2.53%, 164.3, and 416.4, respectively. The diversity of the parent material probably explains the high χ variability. For instance, sandstones are rich in quartz and present low χ value, while basalts are rich in ferrimagnetic minerals and show high χ value. Basic (e.g. amphibolite, andesite, basalt, olivine-feldspar-basalt, charnockite, diabase, dolerite, gabbro, gneiss phomolite), and ultrabasic (e.g. gabbro, phonolite and serpentinite) igneous rocks exhibit median χ values around 1,000 or higher (Preetz et al., 2008). Nevertheless, the median χ values < 50 indicate soils derived from shales, clay-stones, phyllites, and mainly sandstones. Detection occurs because of the presence of magnetite in soils, as magnetite is weathering-resistant and thus its interference on susceptibility is constant or even increases caused by residual fortification (Friedrich et al., 1992). Furthermore, the χ data presented low variability and were slightly unimodal right in both soil layers.

The initial relationship between the χ values (Figs. S2 and S3) and soil attributes (Figs. S4, S5, and S6) and environmental variables was measured by the Pearson's correlation index. The χ values and soil physical attributes presented a satisfactory correlation with the environmental covariates, positively and negatively, in all three layers. Conversely, soil chemical attributes presented a low correlation with environmental variables.

2.3.3. Soil attributes prediction

The model assessment (e.g. R^2_{adj} and CCC) for soil attributes and χ in all three layers are shown in Figs. 5 and 6. In layer A, the best fitted-model used SYSI as the environmental variable and the Cubist algorithm for CEC, pH, and χ, and the RF algorithm for Al_{sat}, B_{sat}, SOM, sand, and SB (Figs. 5a and 6a). The BSSI bands only contributed for a better prediction of Al^{3+} and clay by applying the Cubist and RF algorithms, respectively (Figs. 5b and 6b). In layer B, the Cubist and RF algorithms were the most efficient models. The SYSI bands improved predictions of Al_{sat}, B_{sat}, CEC, clay, pH, sand, and SB, whereas the BSSI bands increased the model performance for Al^{3+} and SOM. As explained in the methodology, there was no available data for χ in this soil layer. In layer C, the RF and Cubist algorithms presented the best-fit model for soil attributes and χ, similar to observations for layers A and B. The BSSI bands improved the model performance to predict all soil attributes, except for CEC and χ. We do not recommend the GLM algorithm for predicting soil attributes and χ, because this algorithm displayed the worst performance in our study. Residuals of regression kriging (RK) of the best models presented similar metrics of the best model in predicting soil attributes and χ in all three layers (Fig. 7).

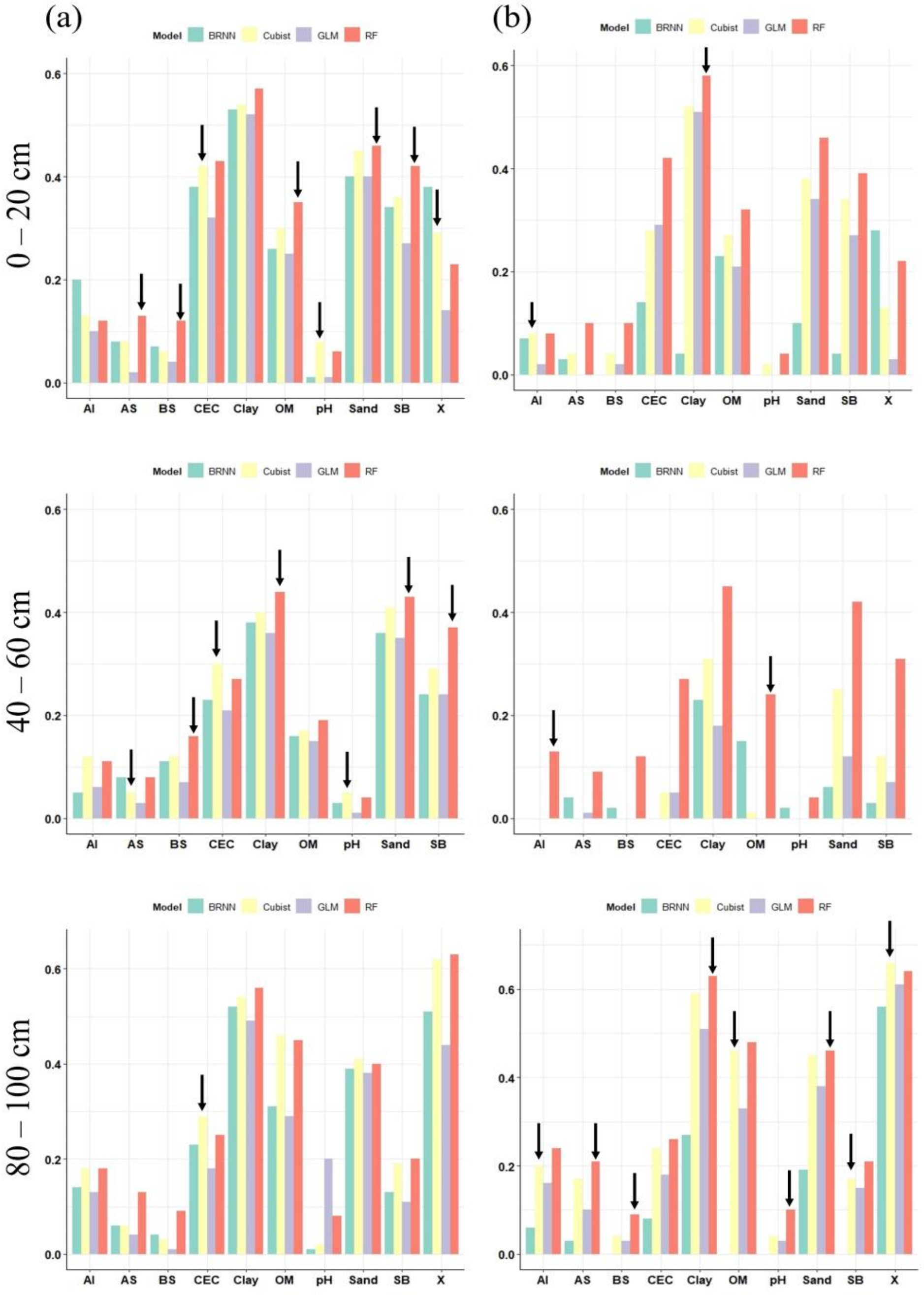


Fig. 5. Model evaluation of soil attributes and property predicted using the Synthetic Soil Image (a) and the Best Synthetic Soil Image (b) as an environmental variable accessed by the adjusted correlation index. Black arrow indicates the model that the residuals were used in the Regression Kriging. GLM, Generalised Linear Model; BRNN, Bayesian Regularised Neural Network; RF, Random Forest.

Further information on the metrics of model performance for soil attributes and χ are in the supplementary materials (Tables S7 – S16). Most studies on the mapping of soil physical attributes in different soil layers have reported a decreasing performance at increasing depths (Table 1). Despite the use of a variety of models or improvements of the quality of environmental variables (e.g. Digital Terrain Model) or sampling density, most studies have not reached reasonable improvements in the predictive power of soil properties at subsurface, which could be attributed to the lack of environmental variables that represent soil distribution and patterns at lower depths. Since the creation of SYSI (Demattê et al., 2018), some studies have proven its feasibility and great contribution to predicting clay, sand, and SOM contents at 0 – 20 cm depth (Fongaro et al., 2018; Gallo et al., 2018). Recently, Mendes et al. (2019) convolved the laboratory spectra to satellite bands at 80 – 100 cm depth and used this information as a target variable by DSM framework with the SYSI bands as environmental variables predicting the subsurface soil spectra of an area covering 478.82 km^2. Afterwards, the bands obtained were used as predictors to map clay and other soil attributes, improving subsurface prediction. Our findings corroborate these results, as the models at lower depths had a better performance than the ones in the upper layers.

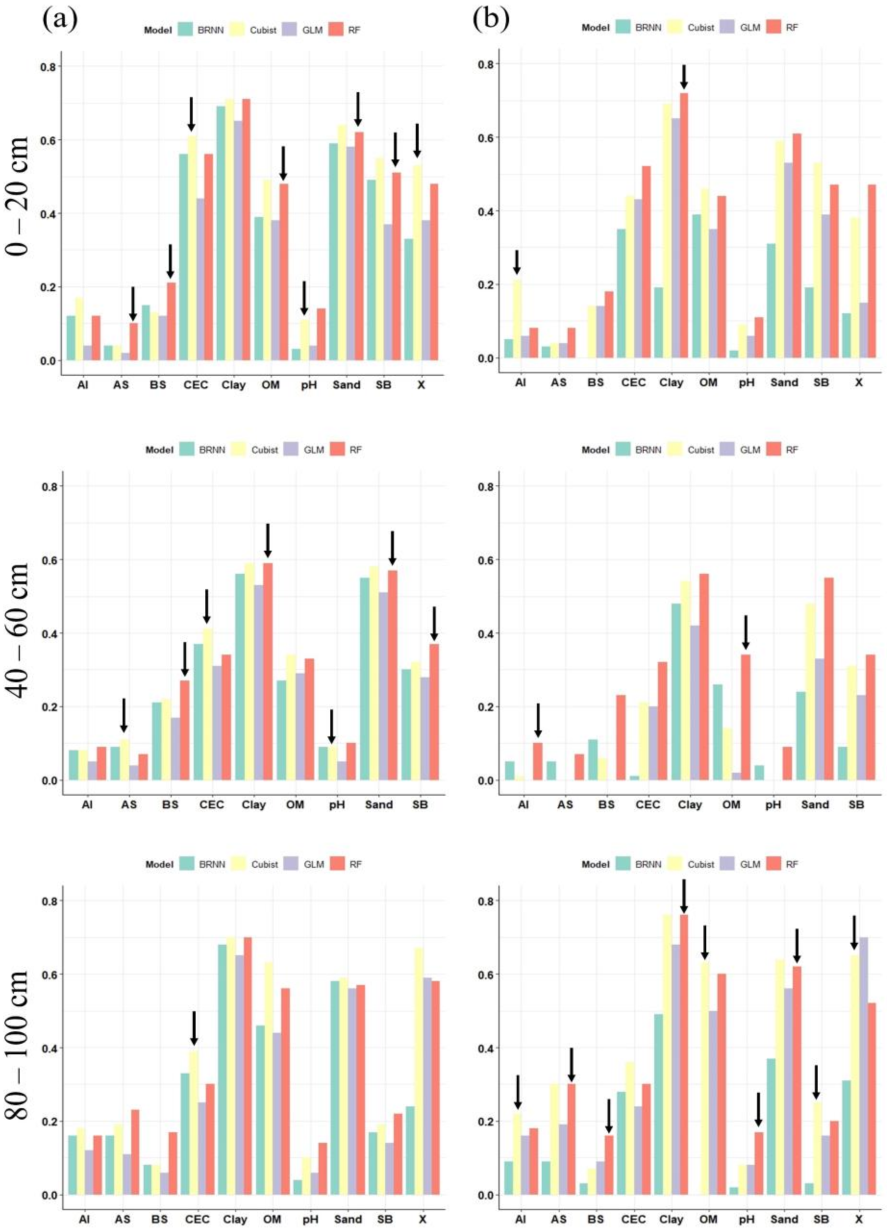


Fig.6. Model evaluation of soil attributes and property predicted using the Synthetic Soil Image (a) and the Best Synthetic Soil Image (b) as an environmental variable accessed by the Lin's concordance correlation coefficient. Black arrow indicates the model that the residuals were used in the Regression Kriging. GLM, Generalised Linear Model; BRNN, Bayesian Regularised Neural Network; RF, Random Forest.

The integration between PS and RS allowed to overcome issues, such as sampling density and detailed environmental variables, reducing costs and time to process data. Even though our sampling density was considered rather lower to predict soil chemical attributes, the models fitted well the predictions (Figs. 5 and 6). It is a consensus that soil chemical attributes are better mapped by increasing sampling density because of their dynamics in the soil matrix, as reported in the literature (Table 1). This trend remained for all chemical attributes mapped in our study; thus, as depth increased, the predictive power decreased. The Al^{3+} content is largely explored in plant nutrition because some plants are responsive to circa 20% of the effective CEC filled by Al (Buol et al., 2011). Thus, similar to other chemical attributes have for agricultural management, Al^{3+} should receive the same relevance to spatialize and locate sites for planting some crops.

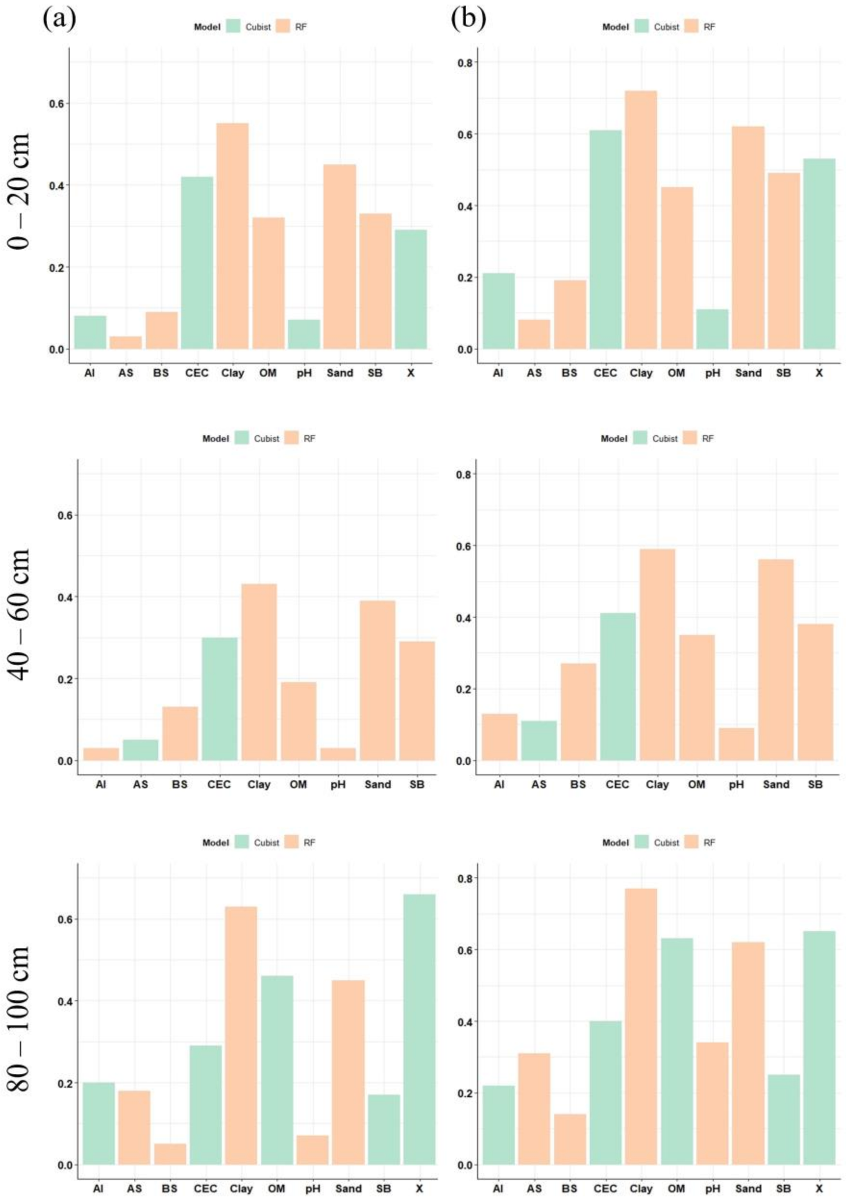


Fig. 7. Model evaluation of the Regression Kriging of the residuals of the best models for soil attributes and property predicted accessed by the adjusted correlation index (a) and the Lin's concordance correlation coefficient (b).

Studies mapping χ have used only the geostatistical framework (de Souza Bahia et al., 2017; Ramos et al., 2017; Siqueira et al., 2014; Teixeira et al., 2018). Siqueira et al. (2014) studied the minimal sampling density necessary to characterise Oxisols attributes (e.g. magnetic susceptibility) in São Paulo State and highlighted that density should be less than one sample per 7 ha to avoid the loss of χ spatial variability. Thus, this technique works well in small areas or when sampling has a volume large enough to meet the minimal density. Our findings proved that the use of residuals of MLAs and RK allows to overcome this limitation, even at lower sample density.

2.3.4. Interpreting the predicted maps of soil attributes and magnetic susceptibility

We selected three sites in the study area to discuss the spatial variability, vertically and horizontally, of soil attributes (Fig. 8) and χ (Fig. 9). Site 2 (Fig. 9) was selected to explain both soil attributes and χ, as it shows the diversity of the study area and highlights that traditional soil surveys (Oliveira and Prado, 1989) were not capable of capturing some important soil patterns along depth. The soil attributes selected were clay, SOM, pH, CEC, and B_{sat} because of their significant role in soil genesis and fertility. Site 2 has different parent materials, such as basalt, sandstone, sandy siltstone, and siltstone. Basaltic areas generate clayey soils (Fig. 8a) and sandy-to-medium-texture soils in sandstone and siltstones areas. Ferromagnesian minerals tend to form clayey, reddish colour, and base-rich soils with high χ (Fig. 9). Sandstones are normally dominated by quartz and coarse-to-sandy residues (Schaetzl and Anderson, 2005). In layer B, there is a clear increase of clay contents, probably forming a textural gradient. This gradient results from lessivage and clay eluviation in kaolinitic soils (Buol et al., 2011). As the depth increased, the clay content decreased in some parts in layer C. The SOM showed a strong and positive correlation with soil texture (Figs. 8a and 8b), while clayey soils have high SOM contents and vice-versa. This pattern is common in tropical environments and well-described in Lepsch et al. (1994). Furthermore, the SOM content of most soils decreases as the depth increases (Detwiler, 1986), corroborating our findings. Conversely, in the clay content patterns, the soil pH in tropical environments is generally acidic in deeper (layer C) than in upper layers (Fig. 8c). The physical weathering of mineral rocks into primary minerals increases the amount of acidic pH from the lower to upper layers, which is in agreement with our findings.

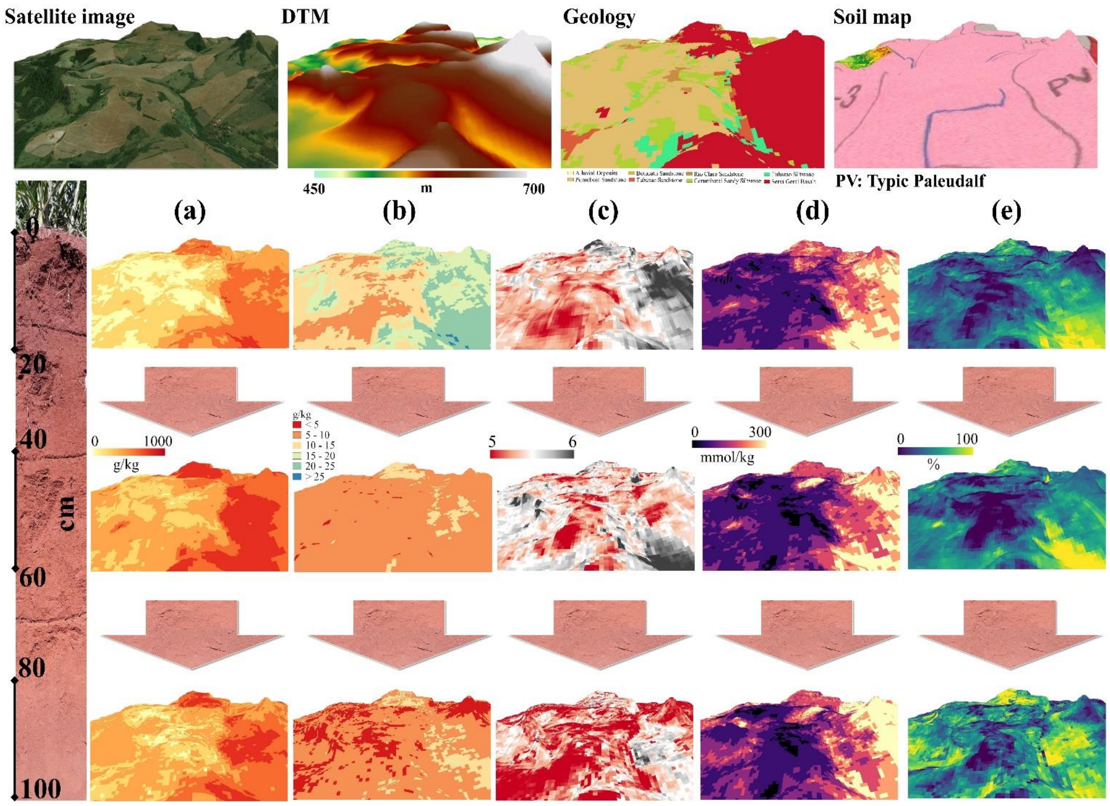


Fig. 8. Relationship of the predicted (a) clay, (b) organic matter, (c) pH in water, (d) cation exchange capacity, and (e) base saturation contents with geology, traditional soil survey (scale of 1: 100,000), and digital terrain model (DTM) at three different soil layers.

Moreover, the number of primary minerals, hydric regime, and relief in layer B could also explain the decreasing pH values, which increased soil acidity. The CEC is an indicator of low- and high-activity clay in soils. If CEC values are very low, it means low-activity clay in soils and vice-versa. As shown in Fig. 8d, there is a slight change in CEC values at increasing depths. Clay minerals, humic substances, Fe and Al oxides have a specific exchangeable surface and are thus responsible for CEC in tropical environments. These findings highlight soil-to-plant interaction where high values of CEC and B_{sat} indicate ideal soils for plant nutrition. Lepsch et al. (1994) analysed the relationship between carbon storage and other soil attributes in natural and cultivated areas in São Paulo State and underscored that B_{sat} values tend to increase at lower depths due to the downward movement of Ca with anions from fertilizers, such as sulphates (Fig. 8e). Although the authors reported no changes in pH in their study, B_{sat} values tended to increase, corroborating our results. Clayey soils retrieve more anions than sandy soils by the patterns observed between the Fig. 8a and 8e. The B_{sat} is used to differentiate some soil types in the Brazilian Classification System (Santos et al., 2018), because Brazilian soils have few morphological features and low fertility, and B_{sat} indicates soil fertility and weathering levels (Buol et al., 2011).

Analysing the magnetic susceptibility horizontally (Fig. 9), high values of χ are directly related to soils derived from ultrabasic and basic rocks. In site 1, it is possible to distinguish basaltic soils and other parent materials, such as sandstones (low χ values). In site 2, there is a huge transecting spot of soils derived from igneous rocks captured by the mapping of χ. In site 3, the mapping of χ reveals soils originated from mafic rocks in the west and east, and sedimentary rocks in the north and south of the selected area. The values of magnetic susceptibility decrease vertically with increasing soil depth. This trend has been reported in the literature (De Jong et al., 2000; Lu et al., 2019; Torrent et al., 2010) due to magnetic susceptibility increase due to the burning of topsoil vegetation and high SOM concentration in upper soil layers. The SOM fermentation process provides electron donors to bacterial Fe reduction, transforming ferrihydrite into maghemite, which later turns into haematite (Barrn and Torrent, 2002; Maher, 1998; Torrent et al., 2010).

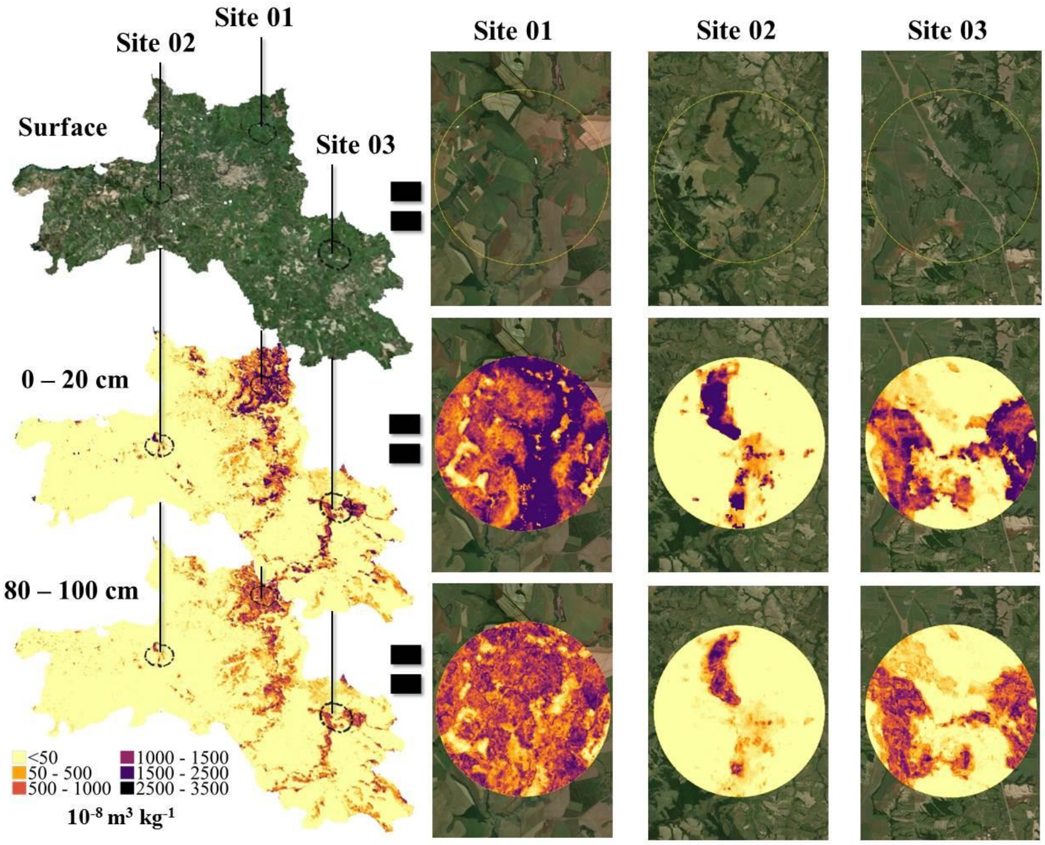


Fig. 9. Satellite soil surface and the final predicted maps of the best machine learning algorithms performed using the Synthetic Soil Image's and the Best Synthetic Soil Image's bands for magnetic susceptibility at 0 – 20 and 80 – 100 cm depths in the study area. Sites 01, 02 and 03 display the spatial variability of the magnetic susceptibility at the two depths.

Most Brazilian soils are Latosols and Argosols, virtually similar to Oxisols, as well as Ultisols and low-activity clay Alfisols of Soil Taxonomy. The latter have the textural B horizon (argillic and kandic horizons), according to Soil Survey Staff (2014), which is a diagnostic horizon and classifies the soils as Argosols (Santos et al., 2018). Fig. 10 shows how clay mapping at surface and subsurface could help identify the textural B horizon by calculating the ratio of clay in layers C and A. Values higher than 1.5, in general, characterise pixels with the argillic horizon. As this diagnostic horizon is a classifier for Typic Paleudalf and Ultisol, traditional soil survey was not capable of capturing a high detailed information on these soils. Therefore, we highlight that the DSM improves qualitatively and quantitatively soil surveys because it is capable of capturing unseen soil features in a detailed framework without further cost and environmental impacts.

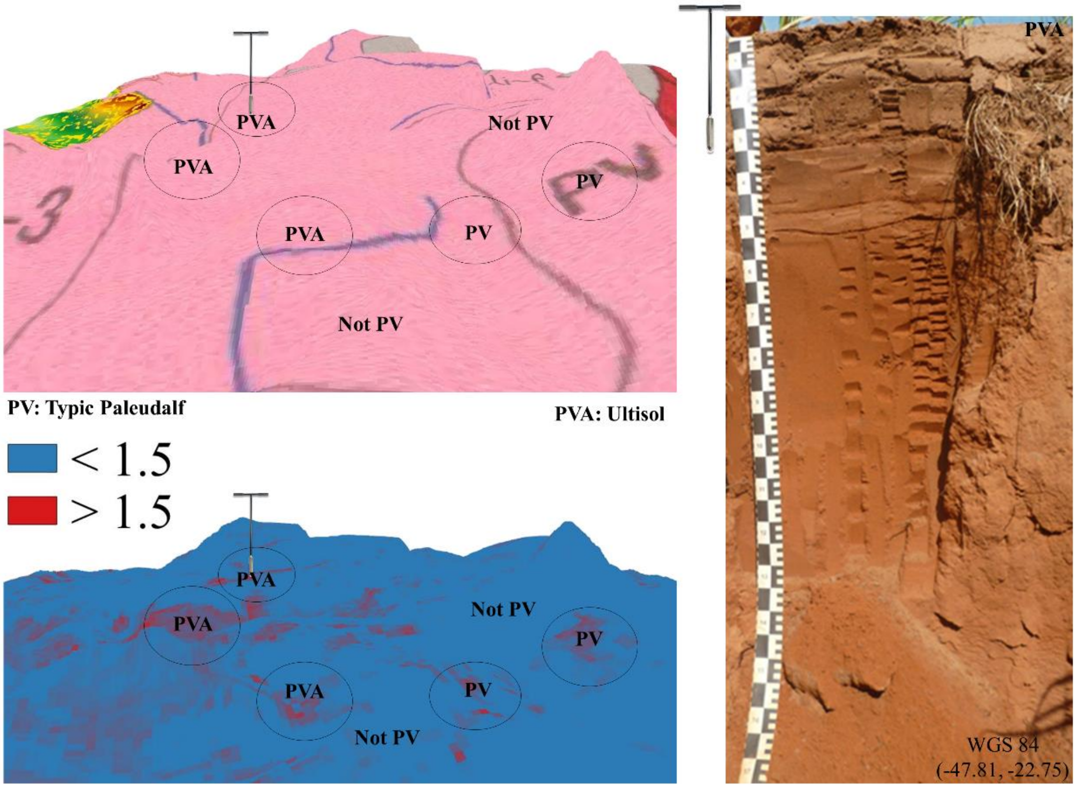


Fig. 10. Traditional soil survey (scale of 1: 100,000) from Agronomic Institute of Campinas and calculated textural B horizon from the predicted clay maps from surface (0 – 20 cm) and subsurface (80 – 100 cm) with a soil profile collected in the site 02 identified in Fig. 9.

2.4. CONCLUSION

Integrating proximal and remote sensing allowed to overcome issues, such as sampling density and detailed environmental variables. These frameworks combined could reduce costs and time to process data. Therefore, predicting soil spectra at 0 – 20, 40 – 60, and 80 – 100 cm depths using the DSM tools and the Synthetic Soil Image as environmental variables allowed creating variables that actually represent soil behaviour in deeper layers. For most soil attributes mapped in this study, the Best Synthetic Soil Image demonstrated original and highly valuable contribution to increase predictive model power at soil depth and the Synthetic Soil Image at upper layers. The models presented better predictive performance in deeper than in upper layers, possibly due to the scarcity of predictors that adequately describe soil subsurface variations of soil components, such as clay, sand, magnetic susceptibility, and soil organic matter. Nevertheless, despite improvements in the subsurface environmental variables, increasing sampling density of soil chemical attributes is necessary to improve their predictions at deeper layers.

Our findings also helped to identify the best algorithms that could be easily accessed and performed to map the main soil attributes and χ. Random Forest and Cubist models are well-stated and were confirmed here as preferable algorithms for mapping soil chemical and physical attributes. The Bayesian Regularised Neural Network and Generalised Linear Model presented better and/or similar predictive power that Random Forest and Cubist for mapping soil spectra. We also recommend using a Generalised Linear Model to predict magnetic susceptibility, besides the Random Forest, and Cubist and hybrid models rather than the sole use of the stochastic framework. Integrating PS and RS helped to identify horizontally and vertically the main soil patterns, which traditional soil surveys were not capable of capturing. Therefore, our findings evidenced the viability of using this integration with tacit knowledge from soil scientists.

ACKNOWLEDGMENTS

This research is part of PhD thesis of the first author, presented to the Graduate Program of Soil and Plant Nutrition at the Luiz de Queiroz College of Agriculture, University of São Paulo. We acknowledge the financial support of the São Paulo Research Foundation (FAPESP grant numbers 2016/26124-6 and 2014/22262-0). The authors also thank the members of the Geotechnologies in Soil Science Group (GeoCIS/GeoSS) (https:// esalqgeocis.wixsite.com/english) and the joint project of Raízen Company with the Luiz de Queiroz Agricultural Studies Foundation grant number 87017.

APPENDIX A. SUPPLEMENTARY DATA

Table S1. Tunning parameters of the machine learning methods applied for best synthetic soil image's bands, magnetic susceptibility (χ), pH, clay, sand, organic matter, cation exchange capacity, sum of bases, exchangeable Al^{3+}, aluminium saturation and base saturation.

Target	Main Source	Depth	Generalised Linear Model		Cubist				Bayesian Regularised Neural Network			Random Forest		
			RMSE	R^2	Committee	Neighbours	RMSE	R^2	Neurons	RMSE	R^2	M_{try}	RMSE	R^2
B1	SYSI	0-20	0.03	0.24	10	0	0.03	0.26	3	0.03	0.27	2	0.03	0.20
		40-60	0.04	0.14	10	0	0.04	0.15	3	0.04	0.15	2	0.04	0.05
		80-100	0.05	0.23	1	0	0.05	0.24	3	0.05	0.25	2	0.05	0.12
B2	SYSI	0-20	0.05	0.22	10	0	0.05	0.22	3	0.05	0.24	2	0.05	0.18
		40-60	0.07	0.12	10	0	0.07	0.13	3	0.07	0.13	2	0.07	0.04
		80-100	0.07	0.23	20	0	0.07	0.25	3	0.07	0.27	2	0.08	0.12
B3	SYSI	0-20	0.06	0.19	20	0	0.06	0.18	2	0.06	0.20	2	0.06	0.13
		40-60	0.08	0.11	10	0	0.08	0.10	2	0.08	0.12	2	0.09	0.03
		80-100	0.09	0.19	20	0	0.09	0.21	3	0.09	0.21	2	0.10	0.11
B4	SYSI	0-20	0.08	0.22	20	0	0.08	0.22	3	0.08	0.23	2	0.08	0.17
		40-60	0.11	0.14	20	0	0.11	0.14	2	0.10	0.15	2	0.11	0.06
		80-100	0.11	0.24	10	0	0.11	0.26	2	0.11	0.25	2	0.12	0.16
B5	SYSI	0-20	0.13	0.24	20	0	0.12	0.25	3	0.12	0.25	2	0.13	0.20
		40-60	0.15	0.13	20	0	0.15	0.13	2	0.15	0.14	2	0.16	0.06
		80-100	0.15	0.25	20	0	0.15	0.27	3	0.15	0.27	2	0.16	0.19
B7	SYSI	0-20	0.11	0.22	20	0	0.11	0.23	3	0.11	0.23	2	0.12	0.18
		40-60	0.12	0.15	10	0	0.12	0.153	1	0.12	0.15	2	0.12	0.08
		80-100	0.11	0.25	20	0	0.11	0.26	3	0.11	0.25	2	0.11	0.21
χ_{log}	SYSI	0-20	0.98	0.68	20	0	0.88	0.74	1	0.92	0.71	16	0.89	0.73
		80-100	1.19	0.55	20	9	1.05	0.64	2	1.04	0.63	16	1.03	0.64
	BSSI	0-20	0.98	0.68	10	5	0.89	0.73	2	0.90	0.73	16	0.87	0.74
		80-100	1.26	0.49	20	9	1.02	0.65	2	1.02	0.65	31	0.96	0.68
pH	SYSI	0-20	1.23	0.01	10	0	1.23	0.04	3	1.23	0.01	2	1.20	0.05
		40-60	1.23	0.03	10	0	1.23	0.04	3	1.23	0.03	31	1.23	0.09
		80-100	1.23	0.00	10	9	1.23	0.03	2	1.23	0.00	16	1.23	0.06
	BSSI	0-20	1.23	0.02	10	0	1.23	0.04	3	1.23	0.02	16	1.23	0.04
		40-60	1.23	0.03	10	0	1.23	0.03	3	1.23	0.03	31	1.23	0.07
		80-100	1.23	0.01	10	0	1.23	0.03	2	1.23	0.02	31	1.23	0.07
Clay	SYSI	0-20	3.38	0.48	20	0	3.31	0.50	2	3.31	0.49	16	3.16	0.53
		40-60	141.77	0.35	10	9	136.55	0.40	3	140.03	0.36	16	134.55	0.41
		80-100	3.09	0.43	20	0	3.02	0.46	3	3.02	0.46	16	2.95	0.48
	BSSI	0-20	3.46	0.45	20	0	3.31	0.49	3	3.31	0.48	31	3.16	0.52
		40-60	148.20	0.30	20	9	138.12	0.38	3	141.62	0.35	31	136.22	0.40
		80-100	2.95	0.48	10	9	2.88	0.51	3	2.95	0.48	16	2.81	0.52
Sand	SYSI	0-20	192.70	0.38	20	0	184.11	0.44	3	189.01	0.41	16	183.91	0.44
		40-60	203.46	0.36	20	9	193.27	0.42	2	197.90	0.40	16	192.26	0.43
		80-100	193.11	0.41	20	0	187.24	0.44	3	188.66	0.43	31	184.53	0.46
	BSSI	0-20	196.25	0.37	20	0	185.20	0.43	3	190.95	0.39	16	183.71	0.44
		40-60	214.00	0.30	20	9	197.67	0.40	3	203.05	0.36	16	193.81	0.42

		80-100	191.46	0.42	20	9	182.80	0.47	3	189.40	0.43	31	179.45	0.49
OM	SYSI	0-20	8.01	0.21	20	9	7.84	0.26	3	7.86	0.24	16	7.59	0.29
		40-60	3.55	0.14	20	9	3.23	0.25	3	3.38	0.18	16	3.23	0.26
		80-100	3.80	0.18	20	9	3.31	0.35	3	3.71	0.21	31	3.23	0.37
	BSSI	0-20	8.29	0.18	20	9	7.75	0.27	3	7.97	0.22	16	7.57	0.30
		40-60	3.55	0.13	10	9	3.31	0.22	2	3.55	0.14	16	3.23	0.25
		80-100	3.71	0.21	20	9	3.23	0.36	3	3.63	0.25	31	3.23	0.35
CEC	SYSI	0-20	3.02	0.30	20	0	2.88	0.35	3	2.95	0.34	16	2.88	0.37
		40-60	3.63	0.30	20	9	3.55	0.34	3	3.63	0.31	31	3.46	0.35
		80-100	4.26	0.26	20	9	3.89	0.36	3	4.17	0.30	16	3.80	0.37
	BSSI	0-20	3.02	0.29	20	9	2.95	0.33	2	3.02	0.31	16	2.88	0.36
		40-60	3.71	0.27	20	9	3.55	0.33	2	3.63	0.31	16	3.55	0.33
		80-100	4.46	0.22	20	9	4.07	0.33	3	4.26	0.27	31	3.89	0.35
SB	SYSI	0-20	4.89	0.26	10	9	4.78	0.29	3	4.78	0.28	31	4.67	0.31
		40-60	6.16	0.30	20	0	5.88	0.33	2	6.02	0.31	16	5.62	0.36
		80-100	7.08	0.16	20	0	6.60	0.23	3	6.76	0.20	31	6.16	0.29
	BSSI	0-20	5.12	0.23	20	0	4.89	0.26	2	4.89	0.26	31	4.67	0.31
		40-60	6.30	0.28	20	9	5.88	0.33	2	6.02	0.30	31	5.62	0.36
		80-100	7.08	0.16	20	9	6.45	0.25	3	7.08	0.18	31	6.30	0.27
Al	SYSI	0-20	9.77	0.05	20	9	9.54	0.09	2	9.54	0.06	16	8.91	0.12
		40-60	22.38	0.08	20	0	21.37	0.10	3	21.37	0.11	16	19.95	0.14
		80-100	20.89	0.12	20	0	20.89	0.14	3	20.89	0.14	16	19.05	0.18
	BSSI	0-20	10.23	0.04	20	9	9.77	0.08	1	9.54	0.06	31	8.91	0.11
		40-60	22.38	0.08	20	0	21.87	0.10	3	21.87	0.10	2	20.41	0.12
		80-100	20.89	0.14	20	0	19.50	0.17	3	20.41	0.15	31	19.05	0.18
AS	SYSI	0-20	21.37	0.02	20	0	20.89	0.06	2	20.89	0.03	31	18.62	0.09
		40-60	37.15	0.04	20	9	38.02	0.06	3	36.30	0.06	16	33.88	0.09
		80-100	28.22	0.05	20	9	27.68	0.10	3	28.09	0.06	31	26.95	0.12
	BSSI	0-20	24.54	0.02	20	0	21.37	0.05	2	20.89	0.03	31	19.49	0.08
		40-60	36.30	0.05	20	9	38.02	0.06	1	36.30	0.04	16	34.67	0.07
		80-100	27.65	0.08	20	0	27.19	0.11	2	27.69	0.07	31	26.94	0.12
BS	SYSI	0-20	19.39	0.07	10	0	19.46	0.07	2	19.29	0.08	16	19.04	0.10
		40-60	22.70	0.11	20	0	22.20	0.15	2	22.50	0.13	16	21.81	0.18
		80-100	22.88	0.02	10	0	23.12	0.02	1	22.77	0.02	16	22.16	0.07
	BSSI	0-20	20.24	0.06	20	0	19.51	0.07	2	19.38	0.07	31	19.15	0.09
		40-60	22.80	0.11	20	0	22.33	0.14	2	22.47	0.13	16	22.08	0.15
		80-100	22.84	0.02	20	9	23.08	0.04	1	22.84	0.01	31	22.35	0.05

Note. B1, B2, B3, B4, B5 and B7, band 1, 2, 3, 4, 5, and 7 of convolved laboratory soil spectrum (reflectance factor); χ_{log}, magnetic susceptibility logarithm; pH, pH in water; clay, (g kg^{-1}); sand, (g kg^{-1}); OM, organic matter (g kg^{-1}); CEC, cation exchange capacity (mmol$_c$ kg^{-1}); SB, sum of bases (mmol$_c$ kg^{-1}); Al, exchangeable Al^{3+} (mmol$_c$ kg^{-1}); AS, aluminium saturation (%); BS, base saturation (%); SYSI, bands of Soil Synthetic Image; BSSI, bands of Best Synthetic Soil Image (B1, B2, B3, B4, B5 and B7); RMSE, Root Mean Square Error; R^2_{adj}, Adjusted Correlation Index; CCC, Lin's Concordance Correlation Coefficient.

Table S2. Model evaluation at 0 – 20 cm depth (layer A) for predicted reflectance spectra in the Vis-NIR-SWIR based on observed laboratory convolved spectra. Highlighted the selected models within each band to generate the best final predicted synthetic soil image. Original values of SYSI were also extracted with the validation points for comparing the model performance.

	Parameters	GLM	Cubist	BRNN	Random Forest		Parameters	BRNN	SYSI
Band 1	RMSE	0.032	0.031	**0.031**	0.034	Regression kriging of the best fitted model	RMSE	0.034	0.035
	R^2_{adj}	0.27	0.33	**0.35**	0.21		R^2_{adj}	0.24	0.20
	CCC	0.42	0.47	**0.46**	0.38		CCC	0.43	0.33
	Bias	0.000	0.003	**0.000**	0.000		Bias	0.005	-0.008
Band 2	RMSE	0.048	0.047	**0.047**	0.051		RMSE	0.052	0.054
	R^2_{adj}	0.24	0.27	**0.28**	0.18		R^2_{adj}	0.21	0.19
	CCC	0.38	0.43	**0.42**	0.34		CCC	0.37	0.28
	Bias	0.000	-0.003	**0.000**	0.000		Bias	0.009	-0.020
Band 3	RMSE	0.059	0.059	**0.057**	0.061		RMSE	0.064	0.086
	R^2_{adj}	0.19	0.20	**0.24**	0.13		R^2_{adj}	0.14	0.14
	CCC	0.33	0.33	**0.37**	0.28		CCC	0.29	0.14
	Bias	-0.001	-0.004	**-0.001**	-0.002		Bias	0.013	-0.061
Band 4	RMSE	0.078	0.075	**0.076**	0.084		RMSE	0.085	0.102
	R^2_{adj}	0.26	0.31	**0.31**	0.16		R^2_{adj}	0.19	0.21
	CCC	0.41	0.45	**0.45**	0.32		CCC	0.38	0.26
	Bias	-0.001	-0.001	**0.000**	-0.004		Bias	0.014	-0.062
Band 5	RMSE	0.123	0.121	**0.120**	0.130		RMSE	0.131	0.169
	R^2_{adj}	0.28	0.30	**0.31**	0.21		R^2_{adj}	0.19	0.22
	CCC	0.44	0.45	**0.47**	0.39		CCC	0.44	0.25
	Bias	-0.005	0.004	**-0.006**	-0.008		Bias	0.010	-0.110
Band 7	RMSE	0.113	0.110	**0.109**	0.118		RMSE	0.120	0.176
	R^2_{adj}	0.28	0.31	**0.33**	0.21		R^2_{adj}	0.20	0.21
	CCC	0.42	0.47	**0.45**	0.38		CCC	0.43	0.19
	Bias	-0.006	0.000	**-0.007**	-0.009		Bias	0.008	-0.136

Note. BRNN, Bayesian Regularised Neural Network; GLM, Generalised Linear Model; SYSI, Synthetic Soil Image. RMSE, Root Mean Square Error (Reflectance factor x 10000); R^2_{adj}, Adjusted Correlation Index; CCC, Lin's Concordance Correlation Coefficient.

Table S3. Model evaluation at 40 – 60 cm depth (layer B) for predicted reflectance spectra in the Vis-NIR-SWIR based on observed laboratory convolved spectra. Highlighted the selected models within each band to generate the best final predicted synthetic soil image. Original values of SYSI were also extracted with the validation points for comparing the model performance.

	Parameters	GLM	Cubist	BRNN	Random Forest		Parameters	GLM	SYSI
Band 1	RMSE	**0.044**	0.045	0.045	0.050	Regression kriging of the best fitted model	RMSE	0.045	0.049
	R^2_{adj}	**0.19**	0.18	0.16	0.02		R^2_{adj}	0.16	0.10
	CCC	**0.28**	0.28	0.25	0.11		CCC	0.31	0.18
	Bias	**-0.003**	-0.009	-0.002	0.001		Bias	-0.001	-0.014
Band 2	RMSE	**0.071**	0.073	0.071	0.078		RMSE	0.072	0.086
	R^2_{adj}	**0.17**	0.15	0.17	0.02		R^2_{adj}	0.14	0.08
	CCC	**0.24**	0.26	0.26	0.10		CCC	0.27	0.12
	Bias	**-0.006**	-0.013	-0.005	-0.003		Bias	-0.004	-0.043
Band 3	RMSE	**0.088**	0.088	0.088	0.095		RMSE	0.090	0.142
	R^2_{adj}	**0.15**	0.15	0.15	0.02		R^2_{adj}	0.10	0.08
	CCC	**0.23**	0.22	0.24	0.11		CCC	0.26	0.06
	Bias	**-0.008**	-0.012	-0.008	-0.005		Bias	-0.006	-0.110
Band 4	RMSE	**0.108**	0.108	0.108	0.117		RMSE	0.110	0.162
	R^2_{adj}	**0.18**	0.17	0.17	0.05		R^2_{adj}	0.14	0.12
	CCC	**0.27**	0.27	0.28	0.16		CCC	0.28	0.11
	Bias	**-0.013**	-0.010	-0.012	-0.008		Bias	-0.013	-0.118
Band 5	RMSE	**0.148**	0.150	0.150	0.162		RMSE	0.155	0.217
	R^2_{adj}	**0.18**	0.17	0.17	0.06		R^2_{adj}	0.11	0.11
	CCC	**0.27**	0.29	0.27	0.18		CCC	0.28	0.13
	Bias	**-0.021**	-0.006	-0.021	-0.016		Bias	-0.023	-0.154
Band 7	RMSE	**0.124**	0.125	0.126	0.133		RMSE	0.127	0.180
	R^2_{adj}	**0.18**	0.18	0.16	0.08		R^2_{adj}	0.14	0.13
	CCC	**0.28**	0.28	0.26	0.20		CCC	0.27	0.14
	Bias	**-0.019**	-0.012	-0.019	-0.013		Bias	-0.015	-0.127

Note. BRNN, Bayesian Regularised Neural Network; GLM, Generalised Linear Model; SYSI, Synthetic Soil Image. RMSE, Root Mean Square Error (Reflectance factor x 10000); R^2_{adj}, Adjusted Correlation Index; CCC, Lin's Concordance Correlation Coefficient.

Table S4. Model evaluation at 80 – 100 cm depth (layer C) for predicted reflectance spectra in the Vis-NIR-SWIR based on observed laboratory convolved spectra. Highlighted the selected models within each band to generate the best final predicted synthetic soil image. Original values of SYSI were also extracted with the validation points for comparing the model performance.

	Parameters	GLM	Cubist	BRNN	Random Forest		Parameters	GLM	SYSI
Band 1	RMSE	0.042	**0.041**	0.043	0.043	Regression kriging of the best fitted model	RMSE	0.045	0.043
	R^2_{adj}	0.21	**0.22**	0.21	0.18		R^2_{adj}	0.20	0.16
	CCC	0.40	**0.40**	0.42	0.37		CCC	0.40	0.27
	Bias	0.004	**-0.004**	0.006	0.009		Bias	-0.004	-0.009
Band 2	RMSE	0.071	**0.070**	0.071	0.071		RMSE	0.075	0.082
	R^2_{adj}	0.18	**0.19**	0.20	0.19		R^2_{adj}	0.17	0.13
	CCC	0.37	**0.37**	0.40	0.37		CCC	0.37	0.18
	Bias	0.005	**-0.007**	0.007	0.010		Bias	-0.007	-0.038
Band 3	RMSE	0.090	**0.089**	0.090	0.090		RMSE	0.094	0.148
	R^2_{adj}	0.16	**0.16**	0.16	0.16		R^2_{adj}	0.14	0.07
	CCC	0.32	**0.33**	0.33	0.33		CCC	0.33	0.06
	Bias	0.004	**-0.002**	0.008	0.010		Bias	-0.002	-0.115
Band 4	RMSE	0.109	0.108	0.108	**0.107**		RMSE	0.159	0.159
	R^2_{adj}	0.18	0.19	0.19	**0.20**		R^2_{adj}	0.04	0.11
	CCC	0.37	0.37	0.38	**0.37**		CCC	0.37	0.12
	Bias	0.003	-0.002	0.006	**0.009**		Bias	0.009	-0.112
Band 5	RMSE	0.150	0.153	0.150	**0.148**		RMSE	0.152	0.198
	R^2_{adj}	0.18	0.18	0.18	**0.19**		R^2_{adj}	0.19	0.14
	CCC	0.36	0.38	0.36	**0.37**		CCC	0.37	0.19
	Bias	0.005	0.011	0.007	**0.010**		Bias	0.010	-0.126
Band 7	RMSE	0.108	0.108	0.108	**0.105**		RMSE	0.108	0.146
	R^2_{adj}	0.18	0.18	0.19	**0.22**		R^2_{adj}	0.21	0.15
	CCC	0.36	0.36	0.38	**0.41**		CCC	0.41	0.21
	Bias	0.006	0.008	0.009	**0.011**		Bias	0.011	-0.096

Note. BRNN, Bayesian Regularised Neural Network; GLM, Generalised Linear Model; SYSI, Synthetic Soil Image. RMSE, Root Mean Square Error (Reflectance factor x 10000); R^2_{adj}, Adjusted Correlation Index; CCC, Lin's Concordance Correlation Coefficient.

Table S5. Exploratory analysis of the soil attributes.

0 – 20 cm	pH	Clay	Sand	OM	Al^{3+}	SB	CEC	AS	BS
	-		g kg^{-1}			$mmol_c$ kg^{-1}		%	
Minimum	4.00	7.00	0.00	0.00	0.00	1.30	6.00	0.00	0.00
1st Quartile	5.14	118.00	308.10	8.10	0.10	23.00	46.19	0.10	47.00
Median	5.50	207.00	514.00	13.00	1.00	39.60	68.80	2.00	62.05
Mean	5.58	260.40	533.40	14.48	4.64	53.50	82.14	9.72	60.16
3rd Quartile	6.00	374.00	772.00	20.50	3.90	68.00	102.00	11.00	75.72
Maximum	7.72	765.00	963.00	61.00	913.00	461.90	481.90	99.00	100.00
SD	0.60	174.44	245.35	9.03	23.75	46.66	52.55	16.52	20.06
CV	0.11	0.67	0.46	0.62	5.11	0.87	0.64	1.69	0.33
Skewness	0.33	0.81	0.04	0.68	31.85	2.22	1.92	2.42	-0.47
Kurtosis	0.00	-0.32	-1.37	0.65	1202.91	7.73	6.08	6.00	-0.38
Skewness log	0.88	-0.31	-1.29	-1.34	1.19	-0.17	0.04	0.55	-2.20
Kurtosis log	-0.16	-0.38	7.36	1.72	1.19	0.03	0.00	-0.96	9.54
40 – 60 cm	pH	Clay	Sand	OM	Al^{3+}	SB	CEC	AS	BS
	-		g kg^{-1}			$mmol_c$ kg^{-1}		%	
Minimum	4.00	14.00	20.00	0.00	0.00	0.60	9.00	0.00	3.23
1st Quartile	4.90	153.00	245.20	5.40	0.40	16.12	47.00	0.67	29.00
Median	5.30	263.60	431.00	8.00	4.40	31.00	72.00	13.09	48.48
Mean	5.38	301.20	472.70	9.28	13.56	48.35	92.58	23.43	49.10
3rd Quartile	5.80	430.00	720.00	12.00	16.00	59.00	112.39	42.71	69.00
Maximum	7.70	794.00	967.00	50.00	189.00	448.80	504.00	94.00	100.00
SD	0.63	175.26	254.13	5.84	23.03	52.20	70.91	25.45	24.07
CV	0.11	0.58	0.54	0.63	1.69	1.08	0.76	1.08	0.49
Skewness	0.57	0.57	0.20	1.65	3.00	2.87	2.12	0.85	0.11
Kurtosis	0.02	-0.69	-1.33	5.16	10.99	11.37	5.64	-0.48	-1.00
Skewness log	0.34	-0.59	-0.66	-0.70	0.32	-0.05	0.15	-0.26	-0.99
Kurtosis log	-0.37	0.11	-0.03	1.94	-1.02	-0.06	-0.05	-1.48	0.75
80 – 100 cm	pH	Clay	Sand	OM	Al^{3+}	SB	CEC	AS	BS
	-		g kg^{-1}			$mmol_c$ kg^{-1}		%	
Minimum	4.00	7.00	24.00	0.00	0.00	0.60	7.57	0.00	1.32
1st Quartile	4.80	176.00	240.00	4.10	1.10	11.00	38.00	5.00	20.00
Median	5.10	268.00	440.00	6.40	7.70	21.00	62.00	31.00	34.67
Mean	5.24	331.70	471.50	8.12	18.80	33.39	88.63	34.06	39.02
3rd Quartile	5.60	483.00	714.00	10.40	22.20	40.00	107.00	58.00	57.00
Maximum	8.10	811.00	975.00	36.00	214.10	319.50	564.00	97.12	100.00
SD	0.61	193.27	250.67	5.60	28.18	38.32	79.77	28.73	22.94
CV	0.11	0.58	0.53	0.69	1.49	1.14	0.90	0.84	0.58
Skewness	0.79	0.58	0.13	1.31	2.55	2.96	2.28	0.36	0.47
Kurtosis	0.68	-0.85	-1.36	1.61	7.90	11.86	6.23	-1.15	-0.79
Skewness log	0.52	-0.53	-0.75	-0.65	0.06	0.02	0.36	-0.81	-0.63
Kurtosis log	0.07	0.48	0.12	1.48	-1.04	-0.05	-0.27	-0.80	-0.14

Note. OM, Organic Matter; SB, the Sum of Bases; CEC, Cation Exchange Capacity; AS, Aluminium Saturation; BS, Base Saturation. SD, Standard Deviation; CV, Coefficient of Variation.

Table S6. Descriptive statistics of Magnetic Susceptibility (χ) at 0 – 20 cm and 80 – 100 cm depths.

Parameters	Unit	0 – 20 cm	80 – 100 cm
Minimum	10^{-8} m^{3} kg^{-1}	2.13	0.64
1st Quartile		12.30	7.64
Median		36.21	19.59
Mean		239.05	164.36
3rd Quartile		157.44	86.71
Maximum		3689.40	3163.22
Skewness	Dimensionless	3.75	4.36
Kurtosis		15.11	22.64
Skewness log		0.53	0.70
Kurtosis log		-0.53	-0.22
SD	10^{-8} m^{3} kg^{-1}	555.89	416.40
CV	%	2.32	2.53

Note. SD, Standard Deviation; CV, Coefficient of Variation.

Table S7. Model evaluation for predicted pH in water using SYSI's and BSSI's bands as covariates based on laboratory analyses. In bold the best fitted model.

		0 – 20 cm			
	Parameters	GLM	Cubist	BRNN	Random Forest
SYSI	RMSE	0.57	**0.55**	0.56	0.55
	R^2_{adj}	0.01	**0.08**	0.01	0.06
	CCC	0.04	**0.11**	0.03	0.14
	Bias	-0.02	**-0.02**	-0.02	-0.01
	Parameters	GLM	Cubist	BRNN	Random Forest
BSSI	RMSE	0.64	0.58	0.57	0.56
	R^2_{adj}	0.00	0.02	0.00	0.04
	CCC	0.06	0.09	0.02	0.11
	Bias	-0.17	-0.09	-0.01	-0.01
		40 – 60 cm			
	Parameters	GLM	Cubist	BRNN	Random Forest
SYSI	RMSE	0.63	0.62	0.63	**0.62**
	R^2_{adj}	0.01	0.05	0.03	**0.04**
	CCC	0.05	0.09	0.09	**0.10**
	Bias	0.01	-0.04	0.02	**0.02**
	Parameters	GLM	Cubist	BRNN	Random Forest
BSSI	RMSE	0.75	0.68	0.63	0.62
	R^2_{adj}	0.00	0.00	0.02	0.04
	CCC	-0.08	-0.04	0.04	0.09
	Bias	0.01	-0.08	0.02	0.03
		80 – 100 cm			
	Parameters	GLM	Cubist	BRNN	Random Forest
SYSI	RMSE	0.60	0.61	0.60	0.58
	R^2_{adj}	0.20	0.02	0.01	0.08
	CCC	0.06	0.10	0.04	0.14
	Bias	-0.02	-0.02	-0.02	-0.01
	Parameters	GLM	Cubist	BRNN	Random Forest
BSSI	RMSE	0.59	0.60	0.66	**0.57**
	R^2_{adj}	0.03	0.04	0.00	**0.10**
	CCC	0.08	0.08	0.02	**0.17**
	Bias	-0.01	-0.07	-0.01	**0.00**

Regression kriging of the best fitted model

0 – 20 cm	
Parameters	Cubist
RMSE	0.55
R^2_{adj}	0.07
CC	0.11
Bias	-0.02
40 – 60 cm	
Parameters	Random Forest
RMSE	0.63
R^2_{adj}	0.03
CC	0.09
Bias	0.03
80 – 100 cm	
Parameters	Random Forest
RMSE	0.59
R^2_{adj}	0.07
CC	0.34
Bias	0.02

Note. GLM, Generalised Linear Model; BRNN, Bayesian Regularised Neural Network; SYSI, Synthetic Soil Image; BSSI, Best Synthetic Soil Image; RMSE, Root Mean Square Error; R^2_{adj}, Adjusted Correlation Index; CCC, Lin's Concordance Correlation Coefficient.

Table S8. Model evaluation for predicted clay content (g kg^{-1}) using SYSI's and BSSI's bands as covariates based on laboratory analyses. In bold the best fitted model.

		0 – 20 cm			
	Parameters	GLM	Cubist	BRNN	Random Forest
SYSI	RMSE	129.03	122.99	125.95	120.75
	R^2_{adj}	0.52	0.54	0.53	0.57
	CCC	0.65	0.71	0.69	0.71
	Bias	-38.69	-25.67	-31.72	-32.49
	Parameters	GLM	Cubist	BRNN	Random Forest
BSSI	RMSE	127.76	123.53	296.05	**119.33**
	R^2_{adj}	0.51	0.52	0.04	**0.58**
	CCC	0.65	0.69	0.19	**0.72**
	Bias	-25.45	-8.99	23.11	**-30.12**
		40 – 60 cm			
	Parameters	GLM	Cubist	BRNN	Random Forest
SYSI	RMSE	138.25	135.05	136.38	**129.74**
	R^2_{adj}	0.36	0.40	0.38	**0.44**
	CCC	0.53	0.59	0.56	**0.59**
	Bias	9.70	7.61	11.71	**12.86**
	Parameters	GLM	Cubist	BRNN	Random Forest
BSSI	RMSE	180.84	161.94	177.36	131.30
	R^2_{adj}	0.18	0.31	0.23	0.45
	CCC	0.42	0.54	0.48	0.56
	Bias	23.02	40.36	10.53	16.26
		80 – 100 cm			
	Parameters	GLM	Cubist	BRNN	Random Forest
SYSI	RMSE	138.98	130.99	134.81	130.72
	R^2_{adj}	0.49	0.54	0.52	0.56
	CCC	0.65	0.70	0.68	0.70
	Bias	-29.26	-24.37	-30.79	-33.07
	Parameters	GLM	Cubist	BRNN	Random Forest
BSSI	RMSE	136.54	123.86	169.76	**118.12**
	R^2_{adj}	0.51	0.59	0.27	**0.63**
	CCC	0.68	0.76	0.49	**0.76**
	Bias	-23.79	-16.52	-26.91	**-27.64**

Regression kriging of the best fitted model

0 – 20 cm	
Parameters	Random Forest
RMSE	119.36
R^2_{adj}	0.55
CC	0.72
Bias	-6.30
40 – 60 cm	
Parameters	Random Forest
RMSE	130.57
R^2_{adj}	0.43
CC	0.59
Bias	5.40
80 – 100 cm	
Parameters	Random Forest
RMSE	116.37
R^2_{adj}	0.63
CC	0.77
Bias	-5.72

Note. GLM, Generalised Linear Model; BRNN, Bayesian Regularised Neural Network; SYSI, Synthetic Soil Image; BSSI, Best Synthetic Soil Image; RMSE, Root Mean Square Error; R^2_{adj}, Adjusted Correlation Index; CCC, Lin's Concordance Correlation Coefficient.

Table S9. Model evaluation for predicted sand content (g kg^{-1}) using SYSI's and BSSI's bands as covariates based on laboratory analyses. In bold the best fitted model.

			0 – 20 cm		
	Parameters	GLM	Cubist	BRNN	Random Forest
SYSI	RMSE	188.32	180.51	188.84	**177.58**
	R^2_{adj}	0.40	0.45	0.40	**0.46**
	CCC	0.58	0.64	0.59	**0.62**
	Bias	10.45	17.01	7.92	**3.22**
	Parameters	GLM	Cubist	BRNN	Random Forest
BSSI	RMSE	198.55	193.21	306.24	178.30
	R^2_{adj}	0.34	0.38	0.10	0.46
	CCC	0.53	0.59	0.31	0.61
	Bias	-18.51	-15.52	-8.91	-1.06
			40 – 60 cm		
	Parameters	GLM	Cubist	BRNN	Random Forest
SYSI	RMSE	204.69	195.07	203.44	**192.80**
	R^2_{adj}	0.35	0.41	0.36	**0.43**
	CCC	0.51	0.58	0.55	**0.57**
	Bias	-12.18	-13.21	-12.02	**-9.31**
	Parameters	GLM	Cubist	BRNN	Random Forest
BSSI	RMSE	268.20	236.75	325.49	194.89
	R^2_{adj}	0.12	0.25	0.06	0.42
	CCC	0.33	0.48	0.24	0.55
	Bias	-54.16	-38.45	-22.80	-9.81
			80 – 100 cm		
	Parameters	GLM	Cubist	BRNN	Random Forest
SYSI	RMSE	192.67	186.42	191.49	188.31
	R^2_{adj}	0.38	0.41	0.39	0.40
	CCC	0.56	0.59	0.58	0.57
	Bias	-5.02	-2.40	-1.24	-1.66
	Parameters	GLM	Cubist	BRNN	Random Forest
BSSI	RMSE	192.89	181.87	221.74	**178.55**
	R^2_{adj}	0.38	0.45	0.19	**0.46**
	CCC	0.56	0.64	0.37	**0.62**
	Bias	-7.40	-7.49	-6.02	**-5.48**

Regression kriging of the best fitted model	0 – 20 cm	
	Parameters	Random Forest
	RMSE	179.79
	R^2_{adj}	0.45
	CC	0.62
	Bias	-1.04
	40 – 60 cm	
	Parameters	Random Forest
	RMSE	198.63
	R^2_{adj}	0.39
	CC	0.56
	Bias	-5.38
	80 – 100 cm	
	Parameters	Random Forest
	RMSE	180.82
	R^2_{adj}	0.45
	CC	0.62
	Bias	-3.45

Note. GLM, Generalised Linear Model; BRNN, Bayesian Regularised Neural Network; SYSI, Synthetic Soil Image; BSSI, Best Synthetic Soil Image; RMSE, Root Mean Square Error; R^2_{adj}, Adjusted Correlation Index; CCC, Lin's Concordance Correlation Coefficient.

Table S10. Model evaluation for predicted organic matter content (g kg^{-1}) using SYSI's and BSSI's bands as covariates based on laboratory analyses. In bold the best fitted model.

	0 – 20 cm				
	Parameters	GLM	Cubist	BRNN	Random Forest
SYSI	RMSE	8.32	8.05	8.30	**7.77**
	R$^2_{adj}$	0.25	0.30	0.26	**0.35**
	CCC	0.38	0.49	0.39	**0.48**
	Bias	-0.73	8.05	-0.81	**-0.39**
	Parameters	GLM	Cubist	BRNN	Random Forest
BSSI	RMSE	8.53	8.20	8.47	7.96
	R$^2_{adj}$	0.21	0.27	0.23	0.32
	CCC	0.35	0.46	0.39	0.44
	Bias	-0.63	-0.21	-0.98	-0.15
	40 – 60 cm				
	Parameters	GLM	Cubist	BRNN	Random Forest
SYSI	RMSE	4.73	4.68	4.71	4.62
	R$^2_{adj}$	0.15	0.17	0.16	0.19
	CCC	0.29	0.34	0.27	0.33
	Bias	-0.77	-0.62	-0.71	-0.75
	Parameters	GLM	Cubist	BRNN	Random Forest
BSSI	RMSE	13.20	6.76	4.76	**4.51**
	R$^2_{adj}$	0.00	0.01	0.15	**0.24**
	CCC	0.02	0.14	0.26	**0.34**
	Bias	0.08	-0.75	-0.92	**-0.84**
	80 – 100 cm				
	Parameters	GLM	Cubist	BRNN	Random Forest
SYSI	RMSE	4.63	3.98	4.57	4.13
	R$^2_{adj}$	0.29	0.46	0.31	0.45
	CCC	0.44	0.63	0.46	0.56
	Bias	-1.17	-0.70	-1.19	-1.09
	Parameters	GLM	Cubist	BRNN	Random Forest
BSSI	RMSE	4.49	**3.97**	12.90	4.00
	R$^2_{adj}$	0.33	**0.46**	0.00	0.48
	CCC	0.50	**0.63**	-0.07	0.60
	Bias	-1.03	**-0.72**	0.73	-1.05

Regression kriging of the best fitted model	0 – 20 cm	
	Parameters	Random Forest
	RMSE	8.06
	R$^2_{adj}$	0.32
	CC	0.45
	Bias	-1.48
	40 – 60 cm	
	Parameters	Random Forest
	RMSE	4.56
	R$^2_{adj}$	0.19
	CC	0.35
	Bias	0.15
	80 – 100 cm	
	Parameters	Cubist
	RMSE	3.97
	R$^2_{adj}$	0.46
	CC	0.63
	Bias	-0.73

Note. GLM, Generalised Linear Model; BRNN, Bayesian Regularised Neural Network; SYSI, Synthetic Soil Image; BSSI, Best Synthetic Soil Image; RMSE, Root Mean Square Error; R$^2_{adj}$, Adjusted Correlation Index; CCC, Lin's Concordance Correlation Coefficient.

Table S11. Model evaluation for predicted cation exchange capacity (mmol$_c$ kg^{-1}) using SYSI's and BSSI's bands as covariates based on laboratory analyses. In bold the best fitted model.

	0 – 20 cm				
	Parameters	GLM	Cubist	BRNN	Random Forest
SYSI	RMSE	39.53	**36.12**	37.17	36.06
	R^2_{adj}	0.32	**0.42**	0.38	0.43
	CCC	0.44	**0.61**	0.56	0.56
	Bias	-8.04	**-4.91**	-6.03	-6.93
	Parameters	GLM	Cubist	BRNN	Random Forest
BSSI	RMSE	39.69	39.72	45.95	36.77
	R^2_{adj}	0.29	0.28	0.14	0.42
	CCC	0.43	0.44	0.35	0.52
	Bias	-4.82	-2.98	-6.01	-7.19
	40 – 60 cm				
	Parameters	GLM	Cubist	BRNN	Random Forest
SYSI	RMSE	62.28	**58.74**	61.14	60.64
	R^2_{adj}	0.21	**0.30**	0.23	0.27
	CCC	0.31	**0.41**	0.37	0.34
	Bias	-11.99	**-10.85**	-10.24	-12.55
	Parameters	GLM	Cubist	BRNN	Random Forest
BSSI	RMSE	85.33	73.51	76.48	61.01
	R^2_{adj}	0.05	0.05	0.00	0.27
	CCC	0.20	0.21	0.01	0.32
	Bias	31.98	13.77	-17.61	-12.94
	80 – 100 cm				
	Parameters	GLM	Cubist	BRNN	Random Forest
SYSI	RMSE	75.60	**70.09**	72.46	72.65
	R^2_{adj}	0.18	**0.29**	0.23	0.25
	CCC	0.25	**0.39**	0.33	0.30
	Bias	-19.29	**-16.55**	-16.16	-17.35
	Parameters	GLM	Cubist	BRNN	Random Forest
BSSI	RMSE	75.78	72.22	105.53	72.98
	R^2_{adj}	0.18	0.24	0.08	0.26
	CCC	0.24	0.36	0.28	0.30
	Bias	-19.77	-16.32	12.41	-19.10

Regression kriging of the best fitted model

0 – 20 cm	
Parameters	Cubist
RMSE	36.12
R^2_{adj}	0.42
CC	0.61
Bias	-4.91
40 – 60 cm	
Parameters	Cubist
RMSE	58.74
R^2_{adj}	0.30
CC	0.41
Bias	-10.85
80 – 100 cm	
Parameters	Cubist
RMSE	70.10
R^2_{adj}	0.29
CC	0.40
Bias	-16.57

Note. GLM, Generalised Linear Model; BRNN, Bayesian Regularised Neural Network; SYSI, Synthetic Soil Image; BSSI, Best Synthetic Soil Image; RMSE, Root Mean Square Error; R^2_{adj}, Adjusted Correlation Index; CCC, Lin's Concordance Correlation Coefficient.

Table S12. Model evaluation for predicted sum of bases (mmol$_c$ kg^{-1}) using SYSI's and BSSI's bands as covariates based on laboratory analyses. In bold the best fitted model.

	0 – 20 cm				
	Parameters	GLM	Cubist	BRNN	Random Forest
SYSI	RMSE	36.52	33.75	34.36	**32.98**
	R^2_{adj}	0.27	0.36	0.34	**0.42**
	CCC	0.37	0.55	0.49	**0.51**
	Bias	-9.32	-6.45	-7.66	**-8.62**
	Parameters	GLM	Cubist	BRNN	Random Forest
BSSI	RMSE	36.32	34.16	44.23	33.94
	R^2_{adj}	0.27	0.34	0.04	0.39
	CCC	0.39	0.53	0.19	0.47
	Bias	-8.67	-6.41	-7.28	-9.13
	40 – 60 cm				
	Parameters	GLM	Cubist	BRNN	Random Forest
SYSI	RMSE	48.62	47.63	48.29	**46.02**
	R^2_{adj}	0.24	0.29	0.24	**0.37**
	CCC	0.28	0.32	0.30	**0.37**
	Bias	-11.71	-12.27	-11.65	**-12.37**
	Parameters	GLM	Cubist	BRNN	Random Forest
BSSI	RMSE	57.93	63.75	54.93	46.99
	R^2_{adj}	0.07	0.12	0.03	0.31
	CCC	0.23	0.31	0.09	0.34
	Bias	19.86	22.96	-15.15	-11.43
	80 – 100 cm				
	Parameters	GLM	Cubist	BRNN	Random Forest
SYSI	RMSE	42.43	41.27	41.79	40.55
	R^2_{adj}	0.11	0.19	0.13	0.20
	CCC	0.14	0.19	0.17	0.22
	Bias	-12.55	-12.33	-11.78	-11.56
	Parameters	GLM	Cubist	BRNN	Random Forest
BSSI	RMSE	41.95	**40.33**	107.54	40.93
	R^2_{adj}	0.15	**0.17**	0.00	0.21
	CCC	0.16	**0.25**	0.03	0.20
	Bias	-12.51	**-10.21**	17.74	-12.24

Regression kriging of the best fitted model		
	0 – 20 cm	
	Parameters	Random Forest
	RMSE	33.77
	R^2_{adj}	0.33
	CC	0.49
	Bias	1.72
	40 – 60 cm	
	Parameters	Random Forest
	RMSE	45.52
	R^2_{adj}	0.29
	CC	0.38
	Bias	-3.91
	80 – 100 cm	
	Parameters	Cubist
	RMSE	40.33
	R^2_{adj}	0.17
	CC	0.25
	Bias	-10.20

Note. GLM, Generalised Linear Model; BRNN, Bayesian Regularised Neural Network; SYSI, Synthetic Soil Image; BSSI, Best Synthetic Soil Image; RMSE, Root Mean Square Error; R^2_{adj}, Adjusted Correlation Index; CCC, Lin's Concordance Correlation Coefficient.

Table S13. Model evaluation for predicted Al (mmol$_c$ kg^{-1}) using SYSI's and BSSI's bands as covariates based on laboratory analyses. In bold the best fitted model.

	0 – 20 cm				
	Parameters	GLM	Cubist	BRNN	Random Forest
SYSI	RMSE	10.54	10.01	10.17	10.16
	R^2_{adj}	0.10	0.13	0.20	0.12
	CCC	0.04	0.17	0.12	0.12
	Bias	-2.72	-2.25	-2.61	-2.32
	Parameters	GLM	Cubist	BRNN	Random Forest
BSSI	RMSE	10.45	**10.17**	10.56	10.36
	R^2_{adj}	0.02	**0.08**	0.07	0.08
	CCC	0.06	**0.21**	0.05	0.08
	Bias	-1.76	**-1.72**	-2.80	-2.44
	40 – 60 cm				
	Parameters	GLM	Cubist	BRNN	Random Forest
SYSI	RMSE	21.52	21.16	21.30	21.12
	R^2_{adj}	0.06	0.12	0.05	0.11
	CCC	0.05	0.08	0.08	0.09
	Bias	-7.90	-7.72	-7.45	-7.76
	Parameters	GLM	Cubist	BRNN	Random Forest
BSSI	RMSE	51.49	23.70	21.93	**21.03**
	R^2_{adj}	0.00	0.00	0.00	**0.13**
	CCC	-0.04	0.01	0.05	**0.10**
	Bias	2.81	-4.36	-4.95	**-7.68**
	80 – 100 cm				
	Parameters	GLM	Cubist	BRNN	Random Forest
SYSI	RMSE	29.88	28.64	29.36	29.23
	R^2_{adj}	0.13	0.18	0.14	0.18
	CCC	0.12	0.18	0.16	0.16
	Bias	-10.71	-9.23	-10.19	-10.57
	Parameters	GLM	Cubist	BRNN	Random Forest
BSSI	RMSE	29.34	**28.16**	30.64	28.92
	R^2_{adj}	0.16	**0.20**	0.06	0.24
	CCC	0.16	**0.22**	0.09	0.18
	Bias	-10.51	**-8.97**	-11.13	-10.69

Regression kriging of the best fitted model

0 – 20 cm	
Parameters	Cubist
RMSE	10.19
R^2_{adj}	0.08
CC	0.21
Bias	-1.74
40 – 60 cm	
Parameters	Random Forest
RMSE	21.03
R^2_{adj}	0.03
CC	0.13
Bias	-0.46
80 – 100 cm	
Parameters	Cubist
RMSE	28.18
R^2_{adj}	0.20
CC	0.22
Bias	-9.03

Note. GLM, Generalised Linear Model; BRNN, Bayesian Regularised Neural Network; SYSI, Synthetic Soil Image; BSSI, Best Synthetic Soil Image; RMSE, Root Mean Square Error; R^2_{adj}, Adjusted Correlation Index; CCC, Lin's Concordance Correlation Coefficient.

Table S14. Model evaluation for predicted aluminium saturation (%) using SYSI's and BSSI's bands as covariates based on laboratory analyses. In bold the best fitted model.

			0 – 20 cm		
	Parameters	GLM	Cubist	BRNN	Random Forest
SYSI	RMSE	19.49	19.66	19.29	**18.64**
	R^2_{adj}	0.02	0.08	0.08	**0.13**
	CCC	0.02	0.04	0.04	**0.10**
	Bias	-7.16	-8.06	-7.17	**-6.54**
	Parameters	GLM	Cubist	BRNN	Random Forest
BSSI	RMSE	19.41	19.51	19.38	18.81
	R^2_{adj}	0.00	0.04	0.03	0.10
	CCC	0.04	0.04	0.03	0.08
	Bias	-5.15	-7.56	-7.04	-6.60
			40 – 60 cm		
	Parameters	GLM	Cubist	BRNN	Random Forest
SYSI	RMSE	29.64	**28.63**	28.77	29.30
	R^2_{adj}	0.03	**0.05**	0.08	0.08
	CCC	0.04	**0.11**	0.09	0.07
	Bias	-15.74	**-13.88**	-15.07	-15.86
	Parameters	GLM	Cubist	BRNN	Random Forest
BSSI	RMSE	48.23	33.19	29.66	29.47
	R^2_{adj}	0.01	0.00	0.04	0.09
	CCC	-0.12	-0.03	0.05	0.07
	Bias	-0.24	12.55	-15.92	-16.20
			80 – 100 cm		
	Parameters	GLM	Cubist	BRNN	Random Forest
SYSI	RMSE	27.05	27.22	26.94	25.73
	R^2_{adj}	0.04	0.06	0.06	0.13
	CCC	0.11	0.19	0.16	0.23
	Bias	1.24	1.33	1.12	1.71
	Parameters	GLM	Cubist	BRNN	Random Forest
BSSI	RMSE	26.21	25.14	27.27	**24.72**
	R^2_{adj}	0.10	0.17	0.03	**0.21**
	CCC	0.19	0.30	0.09	**0.30**
	Bias	1.13	-1.08	-0.91	**1.16**

Regression kriging of the best fitted model

0 – 20 cm	
Parameters	Random Forest
RMSE	18.19
R^2_{adj}	0.03
CC	0.08
Bias	-2.21
40 – 60 cm	
Parameters	Cubist
RMSE	28.65
R^2_{adj}	0.05
CC	0.11
Bias	-13.91
80 – 100 cm	
Parameters	Random Forest
RMSE	24.92
R^2_{adj}	0.18
CC	0.31
Bias	0.70

Note. GLM, Generalised Linear Model; BRNN, Bayesian Regularised Neural Network; SYSI, Synthetic Soil Image; BSSI, Best Synthetic Soil Image; RMSE, Root Mean Square Error; R^2_{adj}, Adjusted Correlation Index; CCC, Lin's Concordance Correlation Coefficient.

Table S15. Model evaluation for predicted base saturation (%) using SYSI's and BSSI's bands as covariates based on laboratory analyses. In bold the best fitted model.

		0 – 20 cm			
	Parameters	GLM	Cubist	BRNN	Random Forest
SYSI	RMSE	20.05	19.99	19.71	**19.19**
	R^2_{adj}	0.04	0.06	0.07	**0.12**
	CCC	0.12	0.13	0.15	**0.21**
	Bias	0.60	3.04	0.71	**0.45**
	Parameters	GLM	Cubist	BRNN	Random Forest
BSSI	RMSE	21.62	20.51	22.38	19.40
	R^2_{adj}	0.02	0.04	0.00	0.10
	CCC	0.14	0.14	-0.01	0.18
	Bias	-1.33	2.27	0.75	0.60
		40 – 60 cm			
	Parameters	GLM	Cubist	BRNN	Random Forest
SYSI	RMSE	23.30	22.68	22.85	**22.11**
	R^2_{adj}	0.07	0.12	0.11	**0.16**
	CCC	0.17	0.22	0.21	**0.27**
	Bias	0.96	0.00	1.12	**0.93**
	Parameters	GLM	Cubist	BRNN	Random Forest
BSSI	RMSE	30.97	30.13	25.11	22.61
	R^2_{adj}	0.00	0.00	0.02	0.12
	CCC	0.00	0.06	0.11	0.23
	Bias	10.77	9.74	4.31	1.17
		80 – 100 cm			
	Parameters	GLM	Cubist	BRNN	Random Forest
SYSI	RMSE	22.77	22.84	22.41	21.81
	R^2_{adj}	0.01	0.03	0.04	0.09
	CCC	0.06	0.08	0.08	0.17
	Bias	-0.75	-4.38	-0.74	-0.45
	Parameters	GLM	Cubist	BRNN	Random Forest
BSSI	RMSE	22.47	22.85	22.86	**21.79**
	R^2_{adj}	0.03	0.04	0.00	**0.09**
	CCC	0.09	0.07	0.03	**0.16**
	Bias	-0.63	-4.79	-0.56	**-0.06**

Regression kriging of the best fitted model	
0 – 20 cm	
Parameters	Random Forest
RMSE	19.57
R^2_{adj}	0.09
CC	0.19
Bias	1.14
40 – 60 cm	
Parameters	Random Forest
RMSE	22.66
R^2_{adj}	0.13
CC	0.27
Bias	0.23
80 – 100 cm	
Parameters	Random Forest
RMSE	22.38
R^2_{adj}	0.05
CC	0.14
Bias	0.91

Note. GLM, Generalised Linear Model; BRNN, Bayesian Regularised Neural Network; SYSI, Synthetic Soil Image; BSSI, Best Synthetic Soil Image; RMSE, Root Mean Square Error; R^2_{adj}, Adjusted Correlation Index; CCC, Lin's Concordance Correlation Coefficient.

Table S16. Model evaluation for predicted magnetic susceptibility (χ, 10^{-8} m^3 kg^{-1}) using SYSI's and BSSI's bands as covariates based on laboratory analyses. In bold the best fitted model.

	0 – 20 cm					Regression kriging of the best fitted model	0 – 20 cm	
	Parameters	GLM	Cubist	BRNN	Random Forest		Parameters	Cubist
SYSI	RMSE	370.21	**313.30**	309.47	333.70			
	R^2_{adj}	0.14	**0.29**	0.38	0.23		RMSE	313.28
	CCC	0.38	**0.53**	0.33	0.48			
	Bias	-49.86	**-34.83**	-107.16	-33.60		R^2_{adj}	0.29
	Parameters	GLM	Cubist	BRNN	Random Forest			
BSSI	RMSE	835.06	405.40	348.63	329.22		CC	0.53
	R^2_{adj}	0.03	0.13	0.28	0.22			
	CCC	0.15	0.38	0.12	0.47		Bias	-34.77
	Bias	116.21	-11.52	-127.63	-36.91			
	80 – 100 cm						80 – 100 cm	
	Parameters	GLM	Cubist	BRNN	Random Forest		Parameters	Cubist
SYSI	RMSE	244.84	213.18	301.83	231.40			
	R^2_{adj}	0.44	0.62	0.51	0.63		RMSE	212.84
	CCC	0.59	0.67	0.24	0.58			
	Bias	-56.74	-61.48	-113.35	-73.47		R^2_{adj}	0.66
	Parameters	GLM	Cubist	BRNN	Random Forest			
BSSI	RMSE	207.67	**212.80**	288.14	245.35		CC	0.65
	R^2_{adj}	0.61	**0.66**	0.56	0.64			
	CCC	0.70	**0.65**	0.31	0.52		Bias	-60.18
	Bias	-54.24	**-60.11**	-107.81	-84.09			

Note. Note. MS, Magnetic Susceptibility; GLM, Generalised Linear Model; BRNN, Bayesian Regularised Neural Network; SYSI, Synthetic Soil Image; BSSI, Best Synthetic Soil Image; RMSE, Root Mean Square Error; R^2_{adj}, Adjusted Correlation Index; CCC, Lin's Concordance Correlation Coefficient.

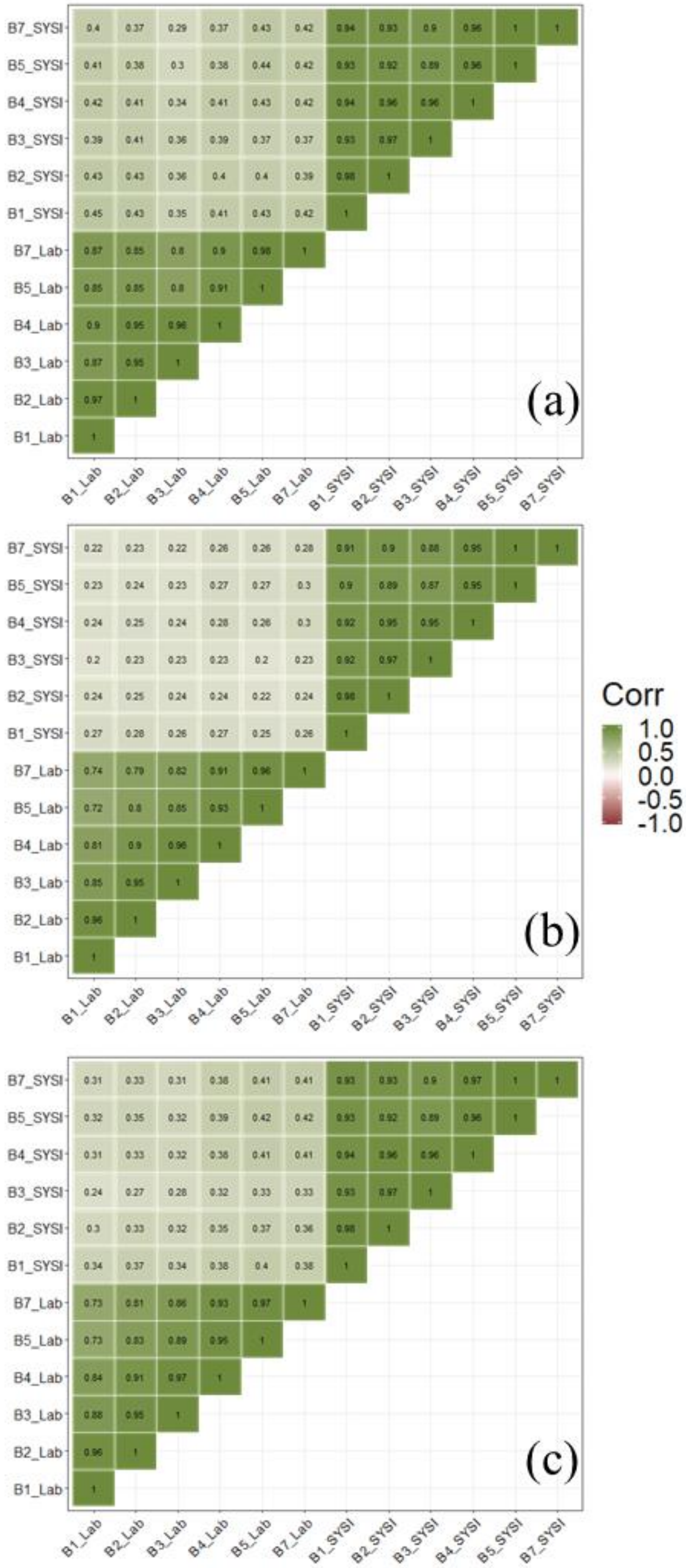


Fig. S1. Pearson's correlation coefficient between the laboratory convolved soil spectra and the Synthetic Soil Image's bands at 0 – 20 (a), 40 – 60 (B), and 80 – 100 cm depth (c).

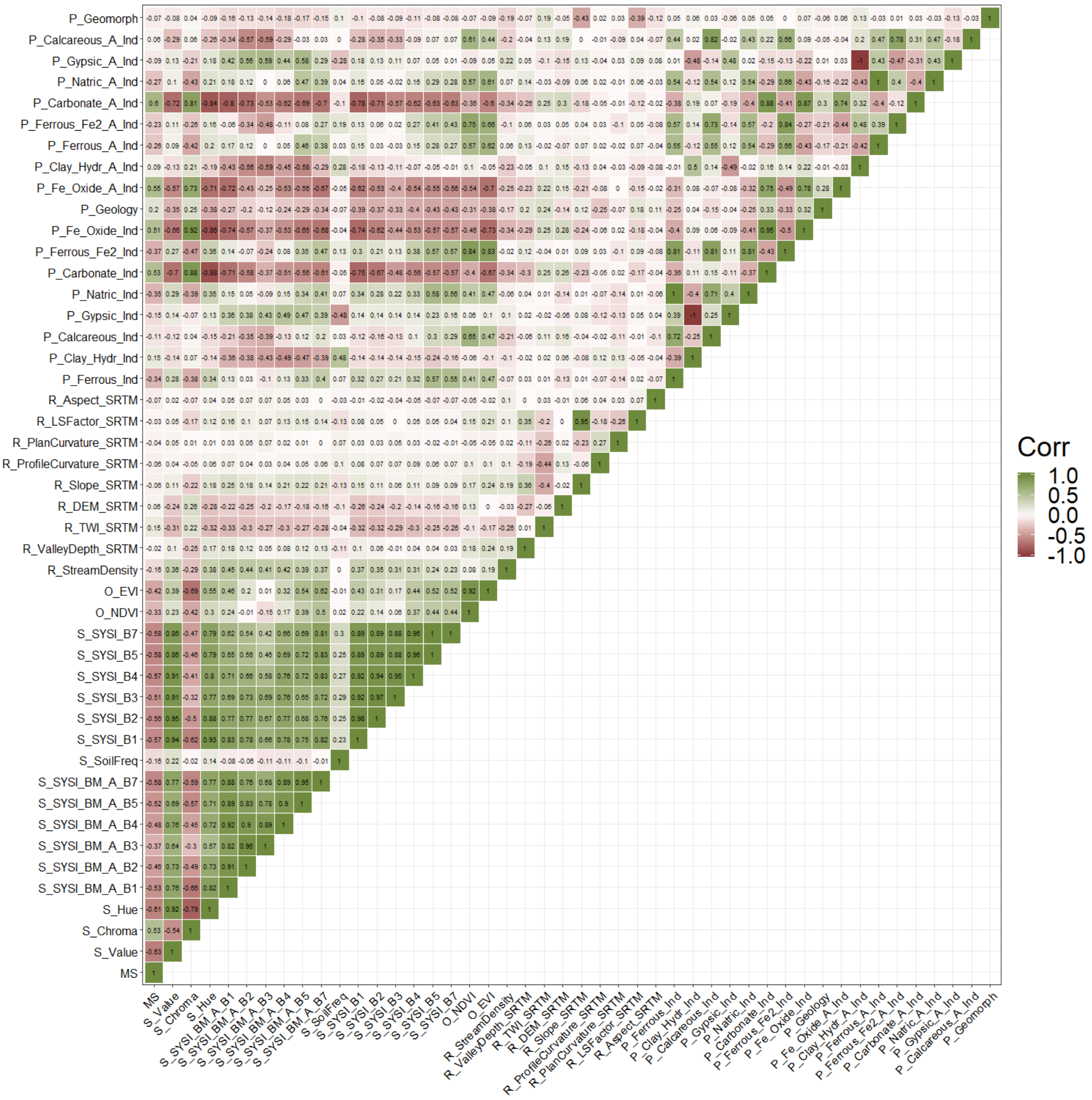


Fig. S2. Pearson's correlation coefficient between the magnetic susceptibility (χ) with the environmental covariates at 0 – 20 cm depth.

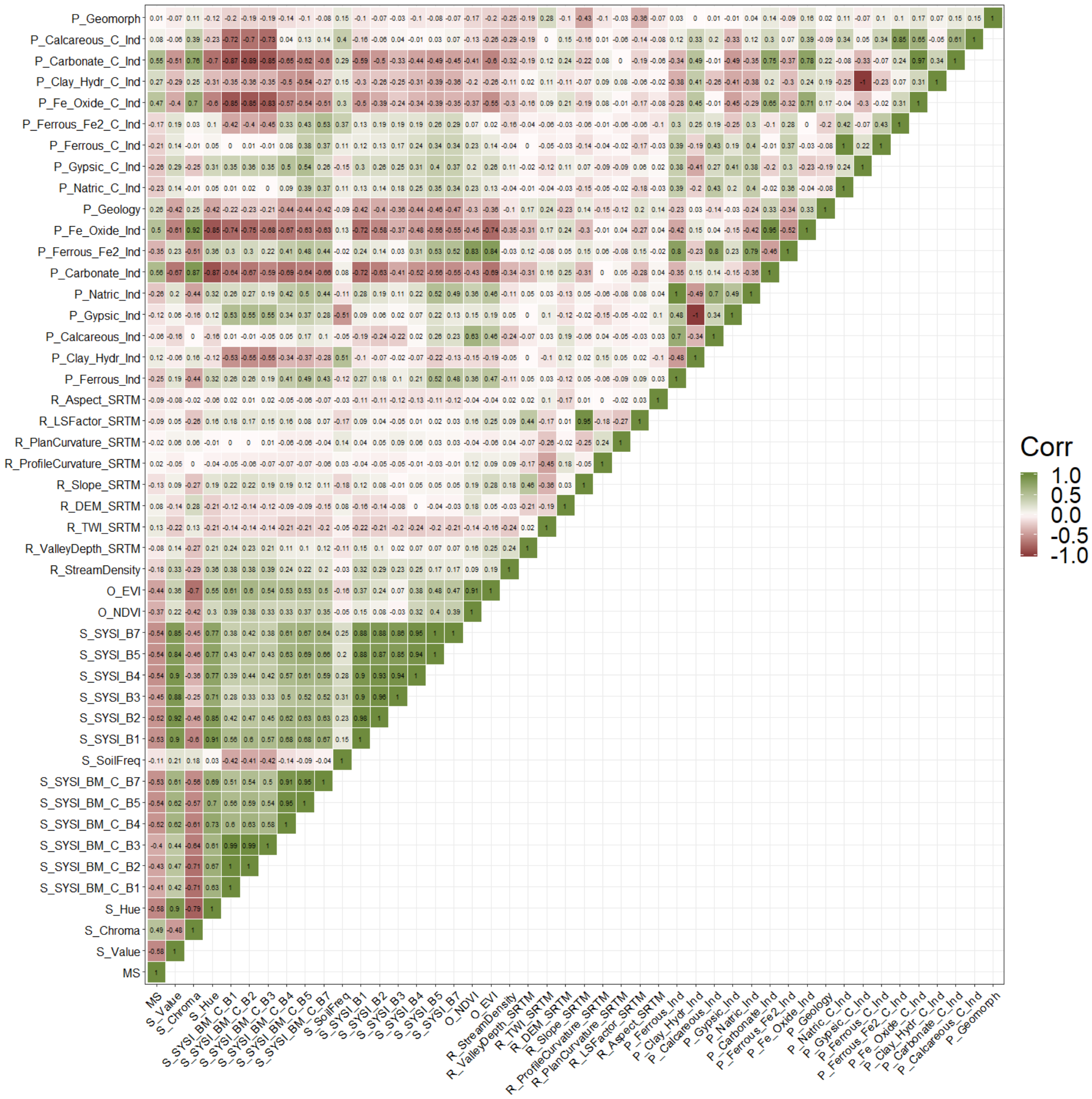


Fig. S3. Pearson's correlation coefficient between the magnetic susceptibility (χ) with the environmental covariates at 80 – 100 cm depth.

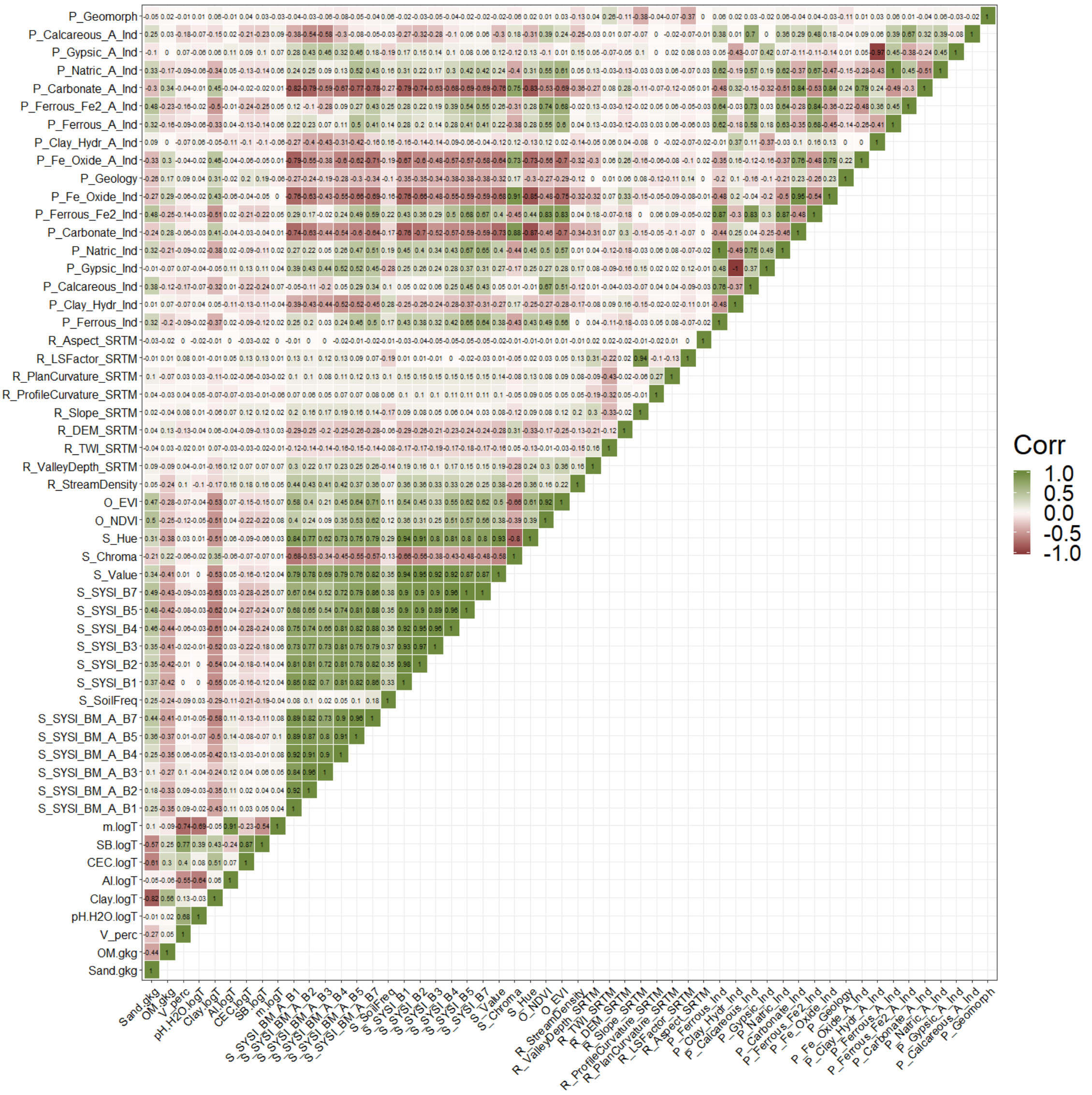


Fig. S4. Pearson's correlation coefficient between the soil attributes with the environmental covariates at 0 – 20 cm depth.

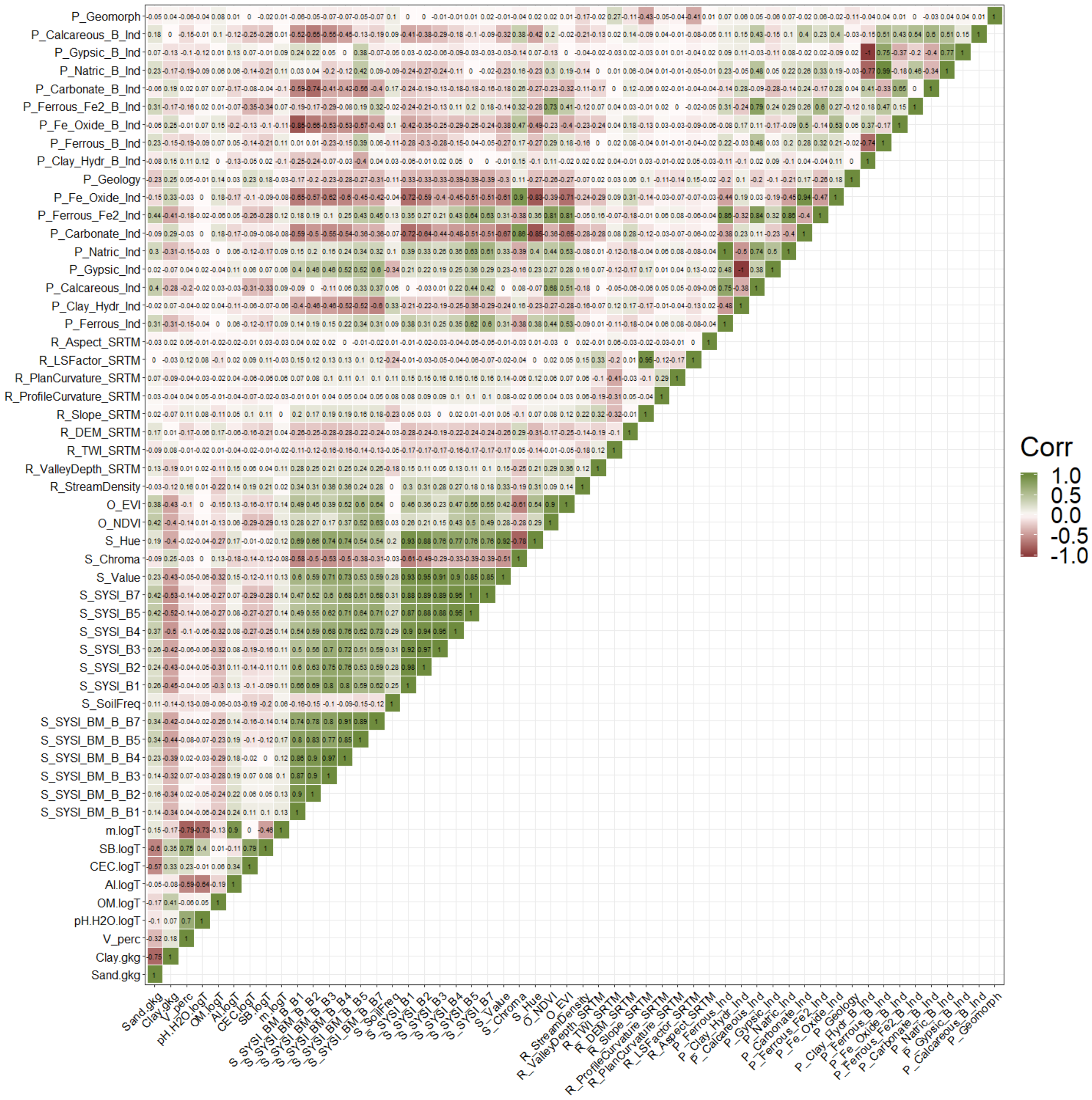


Fig. S5. Pearson's correlation coefficient between the soil attributes with the environmental covariates at 40 – 60 cm depth.

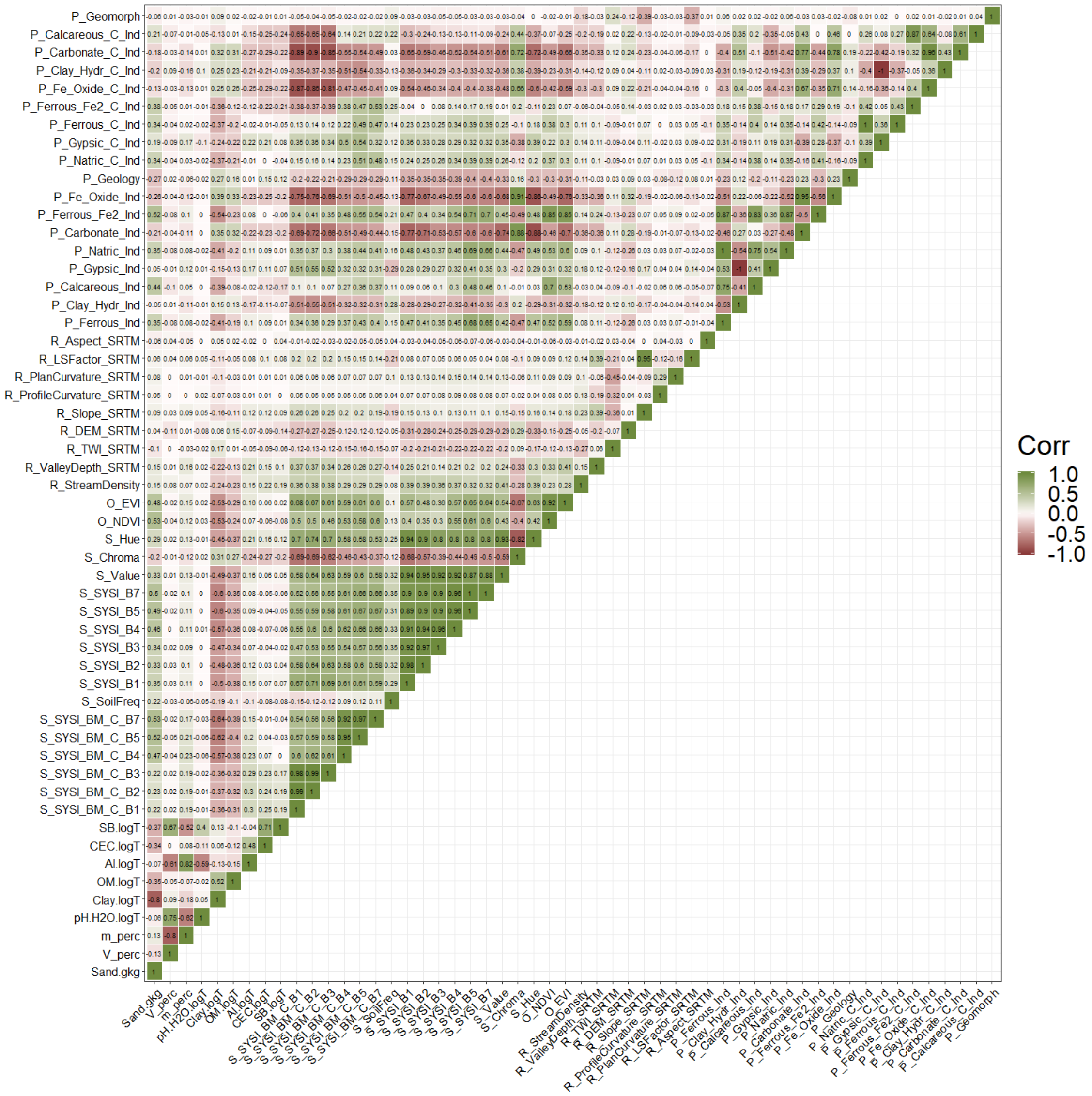


Fig. S6. Pearson’s correlation coefficient between the soil attributes with the environmental covariates at 80 – 100 cm depth.

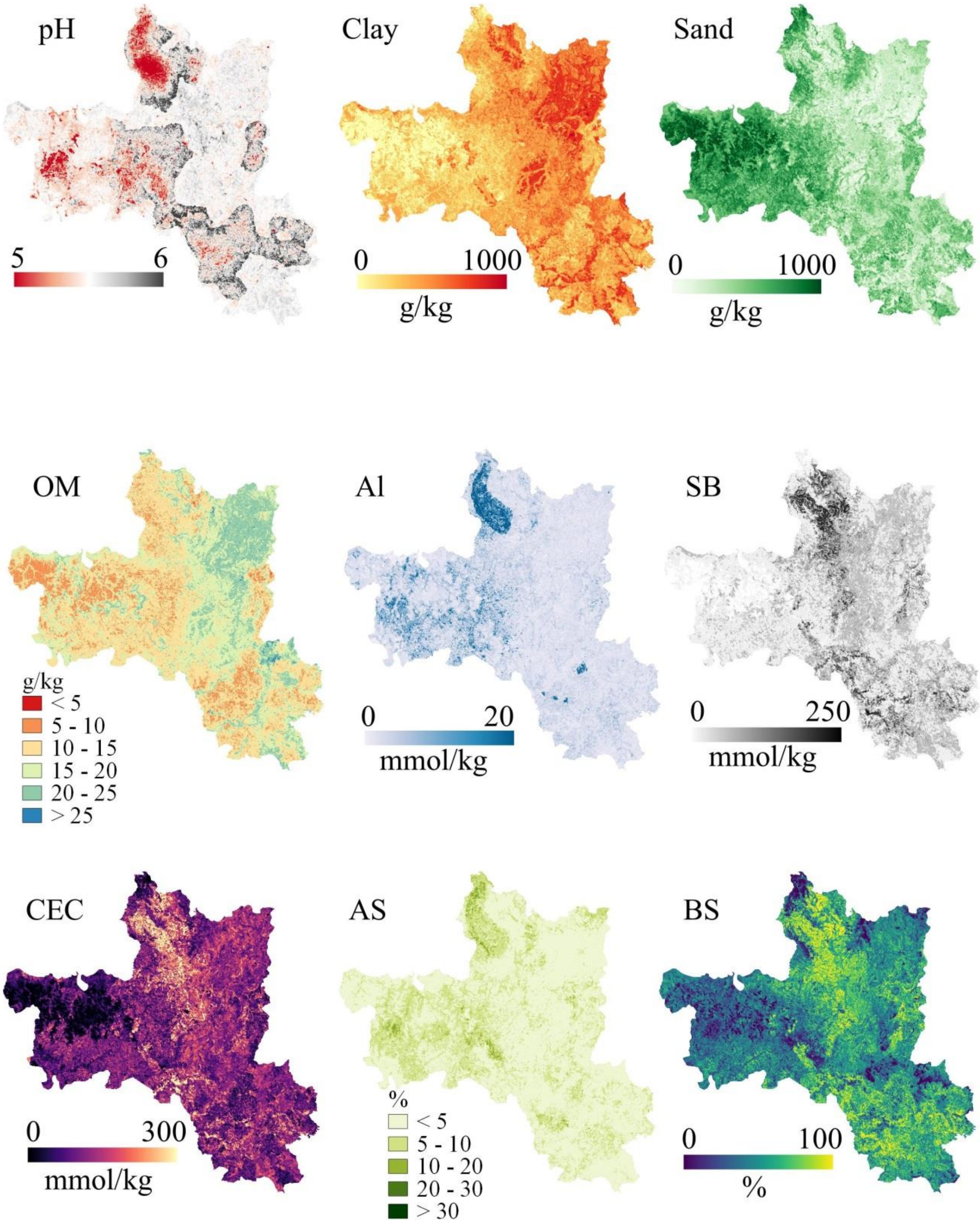


Fig. S7. Final predicted maps of the best machine learning algorithms performed using the Synthetic Soil Image's and the Best Synthetic Soil Image's bands for pH in water, clay, sand, and organic matter (OM) contents, exchangeable Al3+ (Al), sum of bases (SB), cation exchange capacity (CEC), aluminium saturation (AS), and base saturation (BS) at 0 – 20 cm depths in the study area.

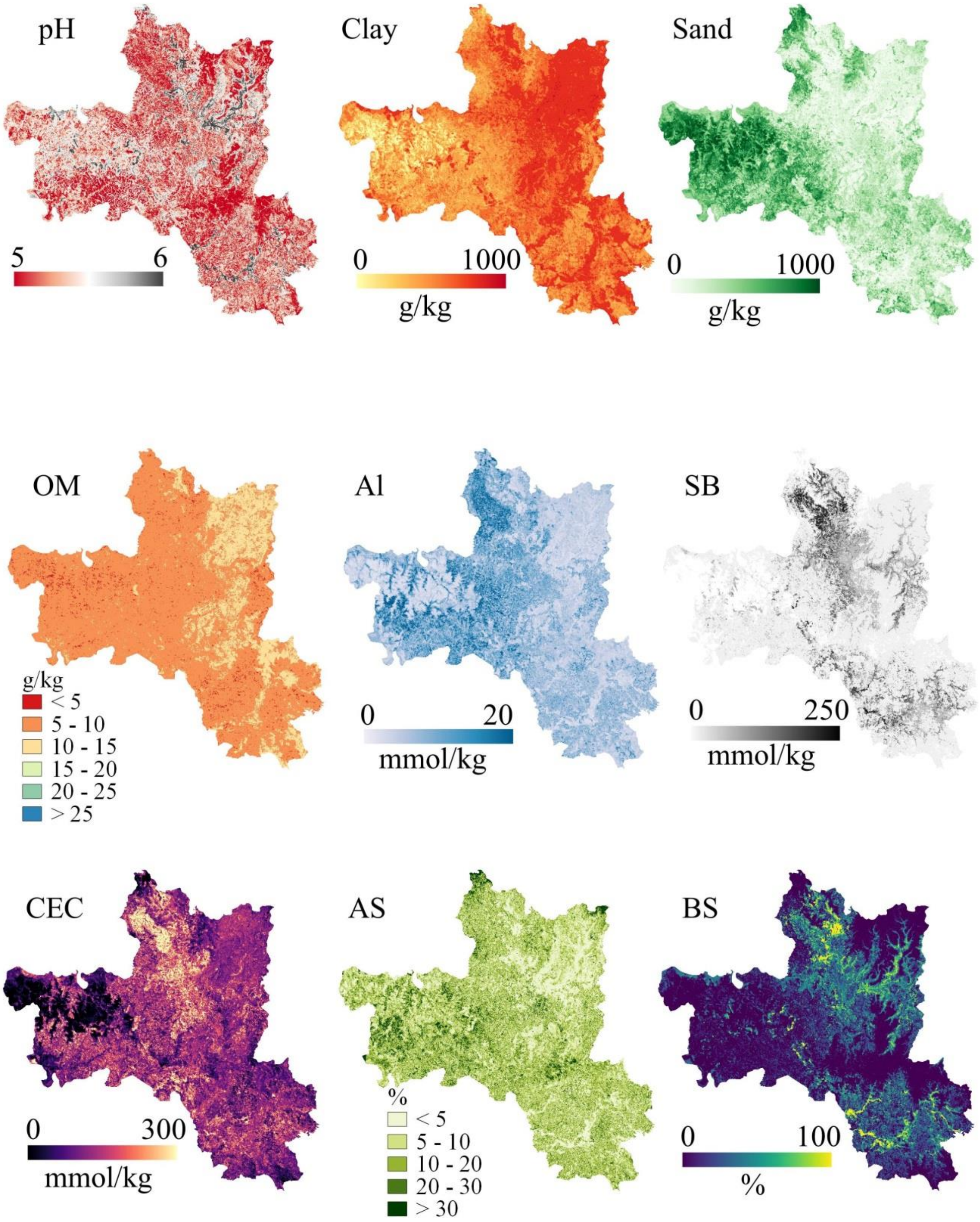


Fig. S8. Final predicted maps of the best machine learning algorithms performed using the Synthetic Soil Image's and the Best Synthetic Soil Image's bands for pH in water, clay, sand, and organic matter (OM) contents, exchangeable Al3+ (Al), sum of bases (SB), cation exchange capacity (CEC), aluminium saturation (AS), and base saturation (BS) at 40 – 60 cm depths in the study area.

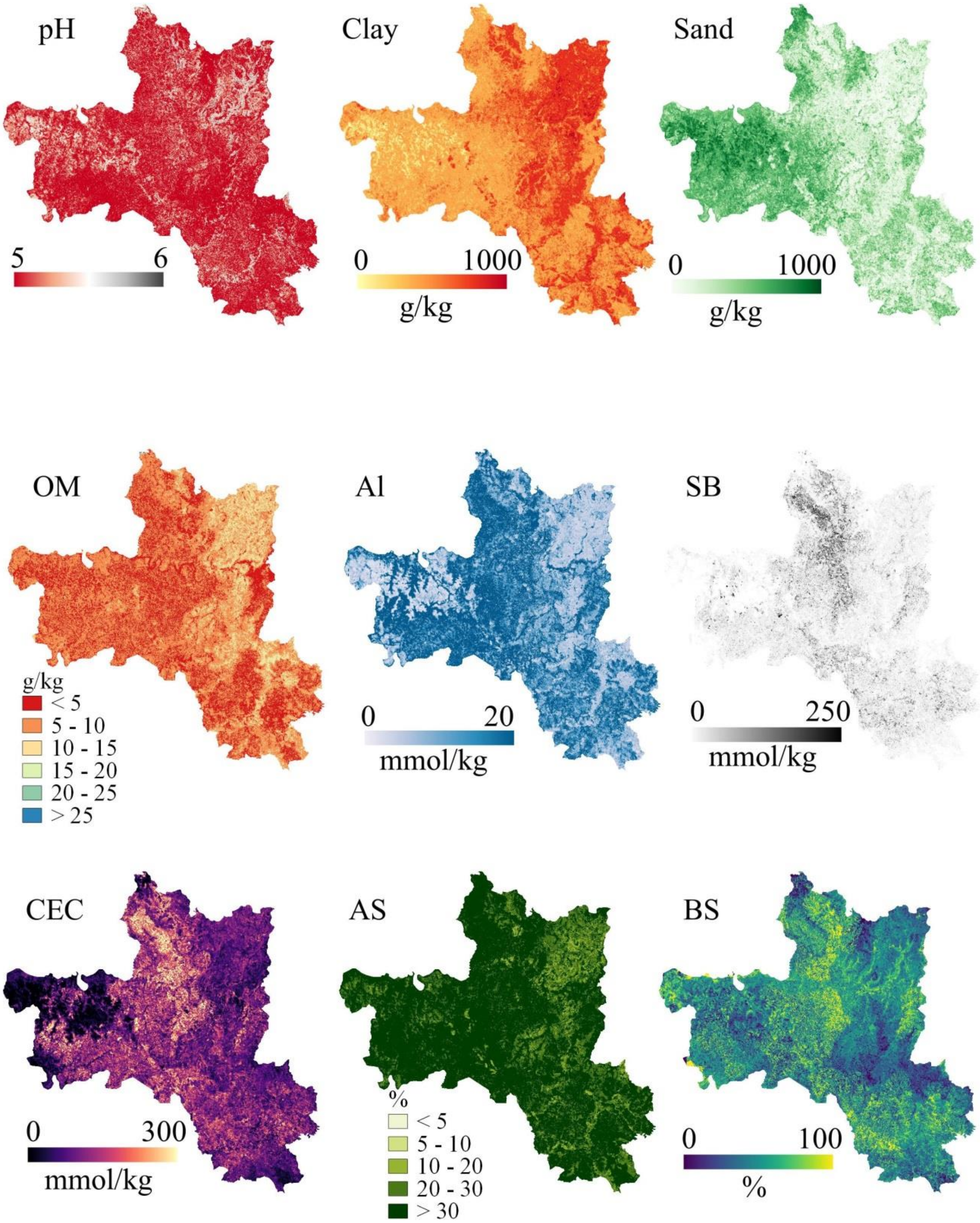


Fig. S9. Final predicted maps of the best machine learning algorithms performed using the Synthetic Soil Image's and the Best Synthetic Soil Image's bands for pH in water, clay, sand, and organic matter (OM) contents, exchangeable Al3+ (Al), sum of bases (SB), cation exchange capacity (CEC), aluminium saturation (AS), and base saturation (BS) at 80 – 100 cm depths in the study area.

REFERENCES

Angelini, M.E., Heuvelink, G.B.M., 2018. Including spatial correlation in structural equation modelling of soil properties. Spat. Stat. 25, 35–51. https://doi.org/10.1016/J.SPASTA.2018.04.003

Barrn, V., Torrent, J., 2002. Evidence for a simple pathway to maghemite in Earth and Mars soils. Geochim. Cosmochim. Acta 66, 2801–2806. https://doi.org/10.1016/S0016-7037(02)00876-1

Behrens, T., Schmidt, K., Ramirez-Lopez, L., Gallant, J., Zhu, A.X., Scholten, T., 2014. Hyper-scale digital soil mapping and soil formation analysis. Geoderma 213, 578–588. https://doi.org/10.1016/j.geoderma.2013.07.031

Ben-Dor, E., Banin, A., 1995. Quantitative analysis of convolved Thematic Mapper spectra of soils in the visible near-infrared and shortwave-infrared spectral regions (0·4–2·5 μm). Int. J. Remote Sens. 16, 3509–3528. https://doi.org/10.1080/01431169508954643

Bennett, J.M., McBratney, A., Field, D., Kidd, D., Stockmann, U., Liddicoat, C., Grover, S., Bennett, J.M., McBratney, A., Field, D., Kidd, D., Stockmann, U., Liddicoat, C., Grover, S., 2019. Soil Security for Australia. Sustainability 11, 3416. https://doi.org/10.3390/su11123416

Bonfatti, B.R., Demattê, J.A.M., Marques, K.P.P., Poppiel, R.R., Rizzo, R., Mendes, W. de S., Silvero, N.E.Q., Safanelli, J.L., 2020. Digital mapping of soil parent material in a heterogeneous tropical area. Geomorphology 107305. https://doi.org/10.1016/j.geomorph.2020.107305

Bonfatti, B.R., Hartemink, A.E., Giasson, E., Tornquist, C.G., Adhikari, K., 2016. Digital mapping of soil carbon in a viticultural region of Southern Brazil. https://doi.org/10.1016/j.geoderma.2015.07.016

Breiman, L., 2001. Random Forests.

Buol, S.W., Southard, R.J., Graham, R.C., McDaniel, P.A., 2011. Soil genesis and classification, 6th ed. John Wiley & Sons, Ltd.

Camargo, O.A., Moniz, A.C., Jorge, J.A., Valadares, J.M.A.S., 2009. Métodos de Análise Química, Mineralógica e Física de Solos do Instituto Agronômico de Campinas. Campinas.

Carré, F., McBratney, A.B., Minasny, B., 2007. Estimation and potential improvement of the quality of legacy soil samples for digital soil mapping. Geoderma. https://doi.org/10.1016/j.geoderma.2007.01.018

Castro-Franco, M., Córdoba, M.A., Balzarini, M.G., Costa, J.L., 2018. A pedometric technique to delimitate soil-specific zones at field scale. Geoderma 322, 101–111. https://doi.org/10.1016/J.GEODERMA.2018.02.034

Chen, S., Rao, P., 2008. Land degradation monitoring using multi-temporal Landsat TM/ETM data in a transition zone between grassland and cropland of northeast China. Int. J. Remote Sens. 29, 2055–2073. https://doi.org/10.1080/01431160701355280

Cracknell, A.P., 2007. Introduction to Remote Sensing, Second. ed. CRC Press, Boca Raton. https://doi.org/10.1201/b13575

Crucil, G., Castaldi, F., Aldana-Jague, E., van Wesemael, B., Macdonald, A., Van Oost, K., Crucil, G., Castaldi, F., Aldana-Jague, E., van Wesemael, B., Macdonald, A., Van Oost, K., 2019. Assessing the Performance of UAS-Compatible Multispectral and Hyperspectral Sensors for Soil Organic Carbon Prediction. Sustainability 11, 1889. https://doi.org/10.3390/su11071889

de Carvalho Junior, W., Lagacherie, P., da Silva Chagas, C., Calderano Filho, B., Bhering, S.B., 2014. A regional-scale assessment of digital mapping of soil attributes in a tropical hillslope environment. Geoderma 232–234, 479–486. https://doi.org/10.1016/j.geoderma.2014.06.007

De Jong, E., Pennock, D.J., Nestor, P.A., 2000. Magnetic susceptibility of soils in different slope positions in Saskatchewan, Canada. Catena 40, 291–305. https://doi.org/10.1016/S0341-8162(00)00080-1

de Souza Bahia, A.S.R., Marques, J., La Scala, N., Pellegrino Cerri, C.E., Camargo, L.A., 2017. Prediction and Mapping of Soil Attributes using Diffuse Reflectance Spectroscopy and Magnetic Susceptibility. Soil Sci. Soc. Am. J. 81, 1450–1462. https://doi.org/10.2136/sssaj2017.06.0206

Dearing, J., 1999. Environmental Magnetic Susceptibility: using the Bartington MS2 System, 1999. Chi Publ. Kenilworth, Engl. 43.

Demattê, J.A.M., 2016. From Profile Morphometrics to Digital Soil Mapping. pp. 383–399. https://doi.org/10.1007/978-3-319-28295-4_24

Demattê, J.A.M., Dotto, A.C., Paiva, A.F.S., Sato, M. V., Dalmolin, R.S.D., de Araújo, M. do S.B., da Silva, E.B., Nanni, M.R., ten Caten, A., Noronha, N.C., Lacerda, M.P.C., de Araújo Filho, J.C., Rizzo, R., Bellinaso, H., Francelino, M.R., Schaefer, C.E.G.R., Vicente, L.E., dos Santos, U.J., de Sá Barretto Sampaio, E. V., Menezes, R.S.C., de Souza, J.J.L.L., Abrahão, W.A.P., Coelho, R.M., Grego, C.R., Lani, J.L., Fernandes, A.R., Gonçalves, D.A.M., Silva, S.H.G., de Menezes, M.D., Curi, N., Couto, E.G., dos Anjos, L.H.C., Ceddia, M.B., Pinheiro, É.F.M., Grunwald, S., Vasques, G.M., Marques Júnior, J., da Silva, A.J., Barreto, M.C. de V., Nóbrega, G.N., da Silva, M.Z., de Souza, S.F., Valladares, G.S., Viana, J.H.M., da Silva Terra, F., Horák-Terra, I., Fiorio, P.R., da Silva, R.C., Frade Júnior, E.F., Lima, R.H.C., Alba, J.M.F., de Souza Junior, V.S., Brefin, M.D.L.M.S., Ruivo, M.D.L.P., Ferreira, T.O., Brait, M.A., Caetano, N.R., Bringhenti, I., de Sousa Mendes, W., Safanelli, J.L., Guimarães, C.C.B., Poppiel, R.R., e Souza, A.B., Quesada, C.A., do Couto, H.T.Z., 2019. The Brazilian Soil Spectral Library (BSSL): A general view, application and challenges. Geoderma 354, 113793. https://doi.org/10.1016/j.geoderma.2019.05.043

Demattê, J.A.M., Fongaro, C.T., Rizzo, R., Safanelli, J.L., 2018. Geospatial Soil Sensing System (GEOS3): A powerful data mining procedure to retrieve soil spectral reflectance from satellite images. Remote Sens. Environ. 212, 161–175. https://doi.org/10.1016/j.rse.2018.04.047

Demattê, J.A.M., Rizzo, R., Botteon, V.W., 2015. Pedological mapping through integration of digital terrain models spectral sensing and photopedology. Rev. Ciência Agronômica 46, 669–678. https://doi.org/10.5935/1806-6690.20150053

Demattê, J.A.M., Safanelli, J.L., Poppiel, R.R., Rizzo, R., Silvero, N.E.Q., Mendes, W. de S., Bonfatti, B.R., Dotto, A.C., Salazar, D.F.U., Mello, F.A. de O., Paiva, A.F. da S., Souza, A.B., Santos, N.V. dos, Maria Nascimento, C., Mello, D.C. de, Bellinaso, H., Gonzaga Neto, L., Amorim, M.T.A., Resende, M.E.B. de, Vieira, J. da S., Queiroz, L.G. de, Gallo, B.C., Sayão, V.M., Lisboa, C.J. da S., 2020. Bare Earth's Surface Spectra as a Proxy for Soil Resource Monitoring. Sci. Rep. 10, 4461. https://doi.org/10.1038/s41598-020-61408-1

Demattê, J.A.M., Sayão, V.M., Rizzo, R., Fongaro, C.T., 2017. Soil class and attribute dynamics and their relationship with natural vegetation based on satellite remote sensing. Geoderma 302, 39–51. https://doi.org/10.1016/j.geoderma.2017.04.019

Detwiler, R.P., 1986. Land use change and the global carbon cycle: the role of tropical soils. Biogeochemistry 2, 67–93. https://doi.org/10.1007/BF02186966

Dharumarajan, S., Kalaiselvi, B., Suputhra, A., Lalitha, M., Hegde, R., Singh, S.K., Lagacherie, P., 2020. Digital soil mapping of key GlobalSoilMap properties in Northern Karnataka Plateau. Geoderma Reg. https://doi.org/10.1016/j.geodrs.2019.e00250

Fongaro, C., Demattê, J., Rizzo, R., Lucas Safanelli, J., Mendes, W., Dotto, A., Vicente, L., Franceschini, M., Ustin, S., 2018. Improvement of Clay and Sand Quantification Based on a Novel Approach with a Focus on Multispectral Satellite Images. Remote Sens. 10, 21. https://doi.org/10.3390/rs10101555

Friedrich, G., Marker, A., Kanig, M., 1992. Heavy mineral surveys in exploration of lateritic terrain, in: Handbook of Exploration Geochemistry. Elsevier Science B.V., pp. 483–498. https://doi.org/10.1016/B978-0-444-89095-5.50024-9

Gallo, B.C., Demattê, J.A.M., Rizzo, R., Safanelli, J.L., Mendes, W. de S., Lepsch, I.F., Sato, M. V., Romero, D.J., Lacerda, M.P.C., 2018. Multi-temporal satellite images on topsoil attribute quantification and the relationship with soil classes and geology. Remote Sens. 10, 1571. https://doi.org/10.3390/rs10101571

Gianola, D., Okut, H., Weigel, K.A., Rosa, G.J.M., 2011. Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. BMC Genet. 12, 87. https://doi.org/10.1186/1471-2156-12-87

Gobin, A., Campling, P., Deckers, J., Feyen, J., 2000. Integrated Toposequence Analyses to combine local and scientific knowledge systems. Geoderma 97, 103–123. https://doi.org/10.1016/S0016-7061(00)00029-X

Godinho Silva, S.H., Poggere, G.C., de Menezes, M.D., Carvalho, G.S., Guilherme, L.R.G., Curi, N., 2016. Proximal sensing and digital terrain models applied to digital soil mapping and modeling of Brazilian Latosols (Oxisols). Remote Sens. 8, 614. https://doi.org/10.3390/rs8080614

Gray, J.M., Bishop, T.F.A., 2019. Mapping change in key soil properties due to climate change over south-eastern Australia. Soil Res. 57, 467–481. https://doi.org/10.1071/SR18139

Gray, J.M., Bishop, T.F.A., Wilford, J.R., 2016. Lithology and soil relationships for soil modelling and mapping. CATENA 147, 429–440. https://doi.org/10.1016/j.catena.2016.07.045

Hartemink, A.E.E., Zhang, Y., Bockheim, J.G.G., Curi, N., Silva, S.H.G.H.G., Grauer-Gray, J., Lowe, D.J.J., Krasilnikov, P., 2020. Soil horizon variation: A review, in: Advances in Agronomy. Academic Press Inc., pp. 125–185. https://doi.org/10.1016/bs.agron.2019.10.003

Hengl, T., Heuvelink, G.B.M., Stein, A., 2004. A generic framework for spatial prediction of soil variables based on regression-kriging. Geoderma 120, 75–93. https://doi.org/10.1016/j.geoderma.2003.08.018

Hengl, T., Nussbaum, M., Wright, M.N., Heuvelink, G.B.M., Gräler, B., 2018. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ e5518. https://doi.org/10.7717/peerj.5518

IGC, 2018. Geographic and Cartographic Institute of Sao Paulo, Revista do Instituto Geológico.

INMET, 2020. The Brazilian National Institute of Meteorology [WWW Document]. URL http://www.inmet.gov.br

Jordanova, N., 2016. Soil Magnetism: Applications in Pedology, Environmental Science and Agriculture, Soil Magnetism: Applications in Pedology, Environmental Science and Agriculture.

Keskin, H., Grunwald, S., 2018. Regression kriging as a workhorse in the digital soil mapper's toolbox. Geoderma 326, 22–41. https://doi.org/10.1016/j.geoderma.2018.04.004

Knotters, M., Brus, D.J., Oude Voshaar, J.H., 1995. A comparison of kriging, co-kriging and kriging combined with regression for spatial interpolation of horizon depth with censored observations. Geoderma 67, 227–246. https://doi.org/10.1016/0016-7061(95)00011-C

Kuhn, M., 2008. Building Predictive Models in R Using the caret Package. J. Stat. Softw. 28, 1–26. https://doi.org/10.18637/jss.v028.i05

Lal, R., 2015. Restoring Soil Quality to Mitigate Soil Degradation. Sustainability 7, 5875–5895. https://doi.org/10.3390/su7055875

Lane, P.W., 2002. Generalized linear models in soil science. Eur. J. Soil Sci. 53, 241–251. https://doi.org/10.1046/j.1365-2389.2002.00440.x

Lepsch, I.F., Menk, J.R.F., Oliveira, J.B., 1994. Carbon storage and other properties of soils under agriculture and natural vegetation in São Paulo State, Brazil. Soil Use Manag. 10, 34–42. https://doi.org/10.1111/j.1475-2743.1994.tb00455.x

Li, N., Zare, E., Huang, J., Triantafilis, J., 2018. Mapping Soil Cation-Exchange Capacity using Bayesian Modeling and Proximal Sensors at the Field Scale. Soil Sci. Soc. Am. J. 82, 1203–1216. https://doi.org/10.2136/sssaj2017.10.0356

Loiseau, T., Chen, S., Mulder, V.L., Román Dobarco, M., Richer-de-Forges, A.C., Lehmann, S., Bourennane, H., Saby, N.P.A., Martin, M.P., Vaudour, E., Gomez, C., Lagacherie, P., Arrouays, D., 2019. Satellite data integration for soil clay content modelling at a national scale. Int. J. Appl. Earth Obs. Geoinf. 82, 101905. https://doi.org/10.1016/J.JAG.2019.101905

Lourenço, A.M., Sequeira, E., Sant'Ovaia, H., Gomes, C.R., 2014. Magnetic, geochemical and pedological characterisation of soil profiles from different environments and geological backgrounds near Coimbra, Portugal. Geoderma 213, 408–418. https://doi.org/10.1016/j.geoderma.2013.07.035

Lu, Y., Si, B., Li, H., Biswas, A., 2019. Elucidating controls of the variability of deep soil bulk density. Geoderma 348, 146–157. https://doi.org/10.1016/j.geoderma.2019.04.033

Maher, B.A., 1998. Magnetic properties of modern soils and quaternary loessic paleosols: Paleoclimatic implications. Palaeogeogr. Palaeoclimatol. Palaeoecol. 137, 25–54. https://doi.org/10.1016/S0031-0182(97)00103-X

Malone, B.P., McBratney, A.B., Minasny, B., 2018. Description and spatial inference of soil drainage using matrix soil colours in the Lower Hunter Valley, New South Wales, Australia. PeerJ 6, e4659. https://doi.org/10.7717/peerj.4659

Malone, B.P., McBratney, A.B., Minasny, B., 2013. Spatial Scaling for Digital Soil Mapping. Soil Sci. Soc. Am. J. 77, 890. https://doi.org/10.2136/sssaj2012.0419

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

McBratney, A., Field, D.J., Koch, A., 2014. The dimensions of soil security. Geoderma 213, 203–213. https://doi.org/10.1016/j.geoderma.2013.08.013

McBratney, A.B., Odeh, I.O.A.A., Bishop, T.F.A.A., Dunbar, M.S., Shatar, T.M., 2000. An overview of pedometric techniques for use in soil survey. Geoderma 97, 293–327. https://doi.org/10.1016/S0016-7061(00)00043-4

McKenzie, N.J., Austin, M.P., 1993. A quantitative Australian approach to medium and small scale surveys based on soil stratigraphy and environmental correlation. Geoderma 57, 329–355. https://doi.org/10.1016/0016-7061(93)90049-Q

Mendes, W. de S., Medeiros Neto, L.G., Demattê, J.A.M., Gallo, B.C., Rizzo, R., Safanelli, J.L., Fongaro, C.T., 2019. Is it possible to map subsurface soil attributes by satellite spectral transfer models? Geoderma 343, 269–279. https://doi.org/10.1016/j.geoderma.2019.01.025

Meyer, H., Reudenbach, C., Wöllauer, S., Nauss, T., 2019. Importance of spatial predictor variable selection in machine learning applications – Moving from data reproduction to spatial prediction. Ecol. Modell. 411. https://doi.org/10.1016/j.ecolmodel.2019.108815

Minasny, B., Malone, B.P., McBratney, A.B., Angers, D.A., Arrouays, D., Chambers, A., Chaplot, V., Chen, Z.S., Cheng, K., Das, B.S., Field, D.J., Gimona, A., Hedley, C.B., Hong, S.Y., Mandal, B., Marchant, B.P., Martin, M., McConkey, B.G., Mulder, V.L., O'Rourke, S., Richer-de-Forges, A.C., Odeh, I., Padarian, J., Paustian, K., Pan, G., Poggio, L., Savin, I., Stolbovoy, V., Stockmann, U., Sulaeman, Y., Tsui, C.C., Vågen, T.G., van Wesemael, B., Winowiecki, L., 2017. Soil carbon 4 per mille. Geoderma. https://doi.org/10.1016/j.geoderma.2017.01.002

Minasny, B., McBratney, A.B., 2016. Digital soil mapping: A brief history and some lessons. Geoderma 264, 301–311. https://doi.org/10.1016/j.geoderma.2015.07.017

Nampak, H., Pradhan, B., Mojaddadi Rizeei, H., Park, H.-J., 2018. Assessment of land cover and land use change impact on soil loss in a tropical catchment by using multitemporal SPOT-5 satellite images and Revised Universal Soil Loss Equation model. L. Degrad. Dev. https://doi.org/10.1002/ldr.3112

Odgers, N.P., McBratney, A.B., Minasny, B., 2011. Bottom-up digital soil mapping. II. Soil series classes. Geoderma. https://doi.org/10.1016/j.geoderma.2011.03.013

Oliveira, J.B., Prado, H., 1989. Carta Pedológica Semi-detalhada do Estado de São Paulo: Quadrícula de Piracicaba. Folha SF-23-Y-A-IV. Instituto Agronômico de Campinas, Campinas.

Padarian, J., Minasny, B., McBratney, A.B., 2017. Chile and the Chilean soil grid: A contribution to GlobalSoilMap. Geoderma Reg. 9, 17–28. https://doi.org/10.1016/J.GEODRS.2016.12.001

Poggio, L., Gimona, A., Spezia, L., Brewer, M.J., 2016. Bayesian spatial modelling of soil properties and their uncertainty: The example of soil organic matter in Scotland using R-INLA. Geoderma 277, 69–82. https://doi.org/10.1016/J.GEODERMA.2016.04.026

Poppiel, R.R., Lacerda, M.P.C., Demattê, J.A.M., Oliveira, M.P., Gallo, B.C., Safanelli, J.L., 2019a. Pedology and soil class mapping from proximal and remote sensed data. Geoderma 348, 189–206. https://doi.org/10.1016/j.geoderma.2019.04.028

Poppiel, R.R., Lacerda, M.P.C., Safanelli, J.L., Rizzo, R., Oliveira, M.P., Novais, J.J., Demattê, J.A.M., 2019b. Mapping at 30 m resolution of soil attributes at multiple depths in midwest Brazil. Remote Sens. 11. https://doi.org/10.3390/rs11242905

Pouladi, N., Møller, A.B., Tabatabai, S., Greve, M.H., 2019. Mapping soil organic matter contents at field level with Cubist, Random Forest and kriging. Geoderma 342, 85–92. https://doi.org/10.1016/j.geoderma.2019.02.019

Preetz, H., Altfelder, S., Igel, J., 2008. Tropical Soils and Landmine Detection-An Approach for a Classification System. Soil Sci. Soc. Am. J. 72, 151–159. https://doi.org/10.2136/sssaj2007.0065

QGIS Development Team, 2020. QGIS geographic information system. Open source geospatial foundation project.

Quinlan, J.R. (John R., Ross, J., 1993. C4.5 : programs for machine learning. Morgan Kaufmann Publishers.

R Development Core Team, R., 2020. R: A Language and Environment for Statistical Computing.

Ramos, P.V., Dalmolin, R.S.D., Marques Júnior, J., Siqueira, D.S., De Almeida, J.A., Moura-Bueno, J.M., 2017. Magnetic susceptibility of soil to differentiate soil environments in southern Brazil. Rev. Bras. Cienc. do Solo 41. https://doi.org/10.1590/18069657rbcs20160189

Regmi, N.R., Rasmussen, C., 2018. Predictive mapping of soil-landscape relationships in the arid Southwest United States. Catena 165, 473–486. https://doi.org/10.1016/j.catena.2018.02.031

Rizzo, R., Medeiros, L.G., Mello, D.C. de, Marques, K.P.P., Mendes, W. de S., Quiñonez Silvero, N.E., Dotto, A.C., Bonfatti, B.R., Demattê, J.A.M., 2020. Multi-temporal bare surface image associated with transfer functions to support soil classification and mapping in southeastern Brazil. Geoderma 361, 114018. https://doi.org/10.1016/j.geoderma.2019.114018

Rudorff, B.F.T., de Aguiar, D.A., da Silva, W.F., Sugawara, L.M., Adami, M., Moreira, M.A., 2010. Studies on the rapid expansion of sugarcane for ethanol production in São Paulo state (Brazil) using Landsat data. Remote Sens. 2, 1057–1076. https://doi.org/10.3390/rs2041057

Rutgers, M., van Leeuwen, J.P., Vrebos, D., van Wijnen, H.J., Schouten, T., de Goede, R.G.M., Rutgers, M., van Leeuwen, J.P., Vrebos, D., van Wijnen, H.J., Schouten, T., de Goede, R.G.M., 2019. Mapping Soil Biodiversity in Europe and the Netherlands. Soil Syst. 3, 39. https://doi.org/10.3390/soilsystems3020039

Santos, H.G., Jacomine, P.K.T., Anjos, L.H.C. dos, Oliveira, V.A. de, Lumbreras, J.F., Coelho, M.R., Almeida, J.A., Araújo Filho, J.C. de, Oliveira, J.B., Cunha, T.J.F., 2018. Sistema Brasileiro de Classificação de Solos, 5 ed. ed. Embrapa, Brasília - DF.

Sayão, V.M., Demattê, J.A.M., Bedin, L.G., Nanni, M.R., Rizzo, R., 2018. Satellite land surface temperature and reflectance related with soil attributes. Geoderma 325, 125–140. https://doi.org/10.1016/j.geoderma.2018.03.026

Schaetzl, R.J., Anderson, S., 2005. Soils genesis and geomorphology. Cambridge University Press, New York, NY.

Shahbazi, F., Hughes, P., McBratney, A.B., Minasny, B., Malone, B.P., 2019. Evaluating the spatial and vertical distribution of agriculturally important nutrients — nitrogen, phosphorous and boron — in North West Iran. CATENA 173, 71–82. https://doi.org/10.1016/j.catena.2018.10.005

Silva, S.H.G., Menezes, M.D. de, Owens, P.R., Curi, N., 2016. Retrieving pedologist's mental model from existing soil map and comparing data mining tools for refining a larger area map under similar environmental conditions in Southeastern Brazil. Geoderma 267. https://doi.org/10.1016/j.geoderma.2015.12.025

Silvero, N.E.Q., Demattê, J.A.M., Amorim, M.T.A., Santos, N.V. dos, Rizzo, R., Safanelli, J.L., Poppiel, R.R., Mendes, W. de S., Bonfatti, B.R., 2021. Soil variability and quantification based on Sentinel-2 and Landsat-8 bare soil images: A comparison. Remote Sens. Environ. 252, 112117. https://doi.org/10.1016/j.rse.2020.112117

Silvero, N.E.Q., Siqueira, D.S., Coelho, R.M., Da Costa Ferreira, D., Marques, J., 2019. Protocol for the use of legacy data and magnetic signature on soil mapping of São Paulo Central West, Brazil. Sci. Total Environ. 693, 133463. https://doi.org/10.1016/j.scitotenv.2019.07.269

Sindayihebura, A., Ottoy, S., Dondeyne, S., Van Meirvenne, M., Van Orshoven, J., 2017. Comparing digital soil mapping techniques for organic carbon and clay content: Case study in Burundi's central plateaus. CATENA 156, 161–175. https://doi.org/10.1016/j.catena.2017.04.003

Siqueira, D.S., Marques, J., Pereira, G.T., Barbosa, R.S., Teixeira, D.B., Peluco, R.G., 2014. Sampling density and proportion for the characterization of the variability of Oxisol attributes on different materials. Geoderma 232–234, 172–182. https://doi.org/10.1016/j.geoderma.2014.04.037

Soil Survey Staff, 2014. Keys to soil taxonomy, 12th ed. U.S. Department of Agriculture Handbook.

Tajik, S., Ayoubi, S., Zeraatpisheh, M., 2020. Digital mapping of soil organic carbon using ensemble learning model in Mollisols of Hyrcanian forests, northern Iran. Geoderma Reg. 20, e00256. https://doi.org/10.1016/j.geodrs.2020.e00256

Teixeira, D.D.B., Marques, J., Siqueira, D.S., Vasconcelos, V., Carvalho, O.A., Martins, É.S., Pereira, G.T., 2018. Mapping units based on spatial uncertainty of magnetic susceptibility and clay content. CATENA 164, 79–87. https://doi.org/10.1016/j.catena.2017.12.038

Ticknor, J.L., 2013. A Bayesian regularized artificial neural network for stock market forecasting. Expert Syst. Appl. 40, 5501–5506. https://doi.org/10.1016/j.eswa.2013.04.013

Tien Bui, D., Pradhan, B., Lofman, O., Revhaug, I., Dick, O.B., 2012. Landslide susceptibility assessment in the Hoa Binh province of Vietnam: A comparison of the Levenberg–Marquardt and Bayesian regularized neural networks. Geomorphology 171–172, 12–29. https://doi.org/10.1016/j.geomorph.2012.04.023

Torrent, J., Liu, Q.S., Barrón, V., 2010. Magnetic susceptibility changes in relation to pedogenesis in a Xeralf chronosequence in northwestern Spain. Eur. J. Soil Sci. 61, 161–173. https://doi.org/10.1111/j.1365-2389.2009.01216.x

Triantafilis, J., Kerridge, B., Buchanan, S.M., 2009. Digital soil-class mapping from proximal and remotely sensed data at the field level. Agron. J. https://doi.org/10.2134/agronj2008.0112

USGS, 2018. USGS EROS Archive - Digital Elevation - Shuttle Radar Topography Mission (SRTM) 1 Arc-Second Global [WWW Document]. Earth Resour. Obs. Sci. Cent. https://doi.org/10.5066/F7PR7TFT

Vasques, G.M., Coelho, M.R., Dart, R.O., Oliveira, R.P., Teixeira, W.G., 2016. Mapping soil carbon, particle-size fractions, and water retention in tropical dry forest in Brazil. Pesqui. Agropecuária Bras. 51, 1371–1385. https://doi.org/10.1590/s0100-204x2016000900036

Venables, W.N., Ripley, B.D., 2002. Generalized Linear Models. pp. 183–210. https://doi.org/10.1007/978-0-387-21706-2_7

Vincent, S., Lemercier, B., Berthier, L., Walter, C., 2018. Spatial disaggregation of complex Soil Map Units at the regional scale based on soil-landscape relationships. Geoderma 311, 130–142. https://doi.org/10.1016/J.GEODERMA.2016.06.006

3. A NOVEL FRAMEWORK TO ESTIMATE SOIL MINERALOGY USING SOIL SPECTROSCOPY

ABSTRACT

Attempts on quantifying soil mineral consist of having an observation measured using traditional laboratory soil analysis. However, developments in interpreting and analysing the visible and near-infrared (VNIR) diffuse reflectance have allowed quantifying some soil minerals. In this study, we aimed to implement a novel framework using VNIR spectroscopy to quantify the main soil minerals and evaluated the application of digital soil mapping framework to spatialise those soil minerals. The soil spectra database comprised 2,701 observations in the spectral range of 350 – 2,500 nm at 0 – 20, 40 – 60, and 80 – 100 cm depths. The soil mineral bands in the VNIR spectra were selected based on the literature and in the strong maxima and minima of the second-derivative curves of the soil mineral standards using the Savitzky-Golay method. We proposed an estimative and conversion of the measurement unit of the soil minerals in weight percentages to g kg^{-1} based on clay content. For this procedure, we randomly selected 185 samples out of 2701 available at 0 – 20 cm depth and sent to traditional laboratory analyses to calibrate the final estimative. Therefore, a constant factor was determined to estimate mineral content in soils. For the digital soil mapping procedure, it was used the 2701 samples which were split into 80% and 20% for calibration and validation of the models for each of the nine minerals. This study showed that our proposed novel framework using VNIR spectroscopy and clay content, to estimate soil minerals are promising.

Keywords: remote sensing; digital soil mapping; clay minerals; near-infrared; proximal sensing

3.1. INTRODUCTION

Knowing mineral properties and their contributions to soil formation help to explain how plants growing mechanisms change over different soils, water filtration and partitioning behaviour over and under soils, and how potential toxic elements immobilisation and contamination work in soils (Churchman & Lowe, 2012). The soil minerals' occurrence, alteration and formation take place by physical and/or chemical weathering processes of the bedrocks or coarse particles, which could be influenced by climate, relief and time (Jenny, 1941; Viscarra Rossel, 2011; Zhao et al., 2018). Those processes provide primary and secondary minerals in soils, namely clay minerals or phyllosilicates, which is an indicator of soil formation (Churchman & Lowe, 2012; Dokuchaev, 1883; Jenny, 1941; Omran, 2017). The clay minerals strongly contribute to the soil chemical and physical characteristics. As described by Kämpf et al. (2012), the Brazilian Oxisols have an average of 73% of phyllosilicates (kaolinite), 14.5% of iron and 12.5% of aluminium oxides, and other phyllosilicates in the clay fraction. Winters and Simonson (1951), explaining the properties of subsoils for crop production and management. The authors pointed out that clay minerals are the main definers of soil characteristics at different depths. Essentially, most of the soils are constituted by phyllosilicates (i.e. kaolinite, montmorillonite, muscovite, illite, and chlorite), oxides (i.e. goethite, haematite/hematite, and gibbsite), and carbonate such as calcite (Churchman & Lowe, 2012; Weaver & Pollard, 1973; Winters & Simonson, 1951).

The most frequently analytical method to characterise soil minerals is the X-ray Powder Diffraction (XRD). This method can provide fair information about mineral chemical composition quantitatively, but mainly qualitatively (Bish & Plötze, 2011) and has been extensively used in soil science (L. A. Camargo, Marques, & Pereira, 2013; Carvalho, Nunes, & Coelho, 2017; Scheinost, Chavernas, Barrón, & Torrent, 1998; Xu et al., 2013; Zhang et al., 2016). Despite some improvements in this method, quantitative analysis using XRD is a laborious and demanding process (Bish & Plötze, 2011; Fang et al., 2018). Another aspect is the chemical extraction treatments of the samples, which are destructive, not environmentally friendly analyse and can interfere with the interpretation of the real soil status. The XRD method does not apply to large scale soil investigations.

The diffuse reflectance spectroscopy (i.e. visible and near-infrared, VNIR) uses wavelengths of 350-2500 nm to provide qualitative information and prediction of soil physical, chemical and biological characteristics (Ben-Dor & Banin, 1995; Brown, Shepherd, Walsh, Dewayne Mays, & Reinsch, 2006; José A M Demattê, Araújo, Fiorio, Fongaro,

& Nanni, 2015; Grunwald, Yu, & Xiong, 2018; Viscarra Rossel et al., 2016). The advantages of VNIR consist on a fast scan, cost-efficient (minimum area, portable sensors, field reading), and non-destructive (environmentally friendly source) methodology (Brown et al., 2006; Fang et al., 2018; Viscarra Rossel et al., 2016). This technique has been playing as a replacement for future laboratory soil analysis and XRD method. Brown et al. (2006), characterising soils using VNIR, predicted clay minerals such as montmorillonite and kaolinite finding close values to using XRD. Generally, the soil mineral is identified by the spectral features at a specific wavelength. For example, goethite and haematite are characterised around at 415 and 445 nm, and 535 and 580 nm, respectively (Scheinost et al., 1998). This identification is based on spectral morphology, which is the peak intensity, band surfaces, absorbance valley position and asymmetries. It allows to create indices to characterise the soil mineralogy such as those described in Madeira et al. (1997), and Terra et al. (2015).

Attempts on quantifying soil mineral consist of having an observation measured using traditional laboratory soil analysis. However, developments in interpreting and analysing VNIR diffuse reflectance have allowed quantifying some soil minerals. Fernandes et al. (2004), quantifying iron oxides in Brazilian oxisols by VNIR spectroscopy, proposed regression equations that the inputs are the spectral reflectance at the specific band to quantify hematite and goethite. They found correlation values varying between 0.46 and 0.94. The procedure involves applying the second derivative of the Kubelka-Munk (K-M) function in the original spectra. This differential equation permits a hyperbolic solution, which shows strongly and wide superposed bands at different wavelengths likewise the raw spectra. However, obtaining the derivative of the K-M function curves enhance the resolution of sharply defined features. Examples of the application and limitations of the K-M function to estimate iron oxides and kaolinite can be found in Barron and Torrent (1986), Jepson (1988), and Scheinost et al. (1998). The second-derivative enhances slight concavities and convexities of the original spectrum and presents a much narrower bandwidth (Kosmas, Curi, Bryant, & Franzmeier, 1984). The weak absorption on the K-M function curves become strong minima and maxima in the second-derivative curves. More recently, Mathian et al. (2018) proposed an approach to identify and quantify semi-quantitatively phyllosilicate minerals in laterite saprolites using VNIR diffuse reflectance second derivative method. It was found detection limit, in the existence of sizeable quantities of lateritic kaolinite, at values from 5-10 wt.% of the total clay content using the second derivative. The demand to derive detailed soil information has increased in the last decades to better manage land-use and sustainably increase food production. The Digital Soil Mapping (DSM) became the easiest and feasible approach to achieve such demand. The DSM basis was solemnised in the *scorpan* model by (McBratney, Mendonça Santos, & Minasny, 2003), and it considers the model of soil formation established by Jenny (1941). Mainly, the spatial prediction of soil is performed using only stochastic or deterministic models, or combining them with field observations, tacit knowledge and environmental and/or RS data in DSM. Furthermore, (Ma, Minasny, Malone, & Mcbratney, 2019) demonstrated the link between DSM and Pedology showing thriving applications of DSM in spatializing soil attributes, classes and profiles by integrating tacit knowledge and data-driven.

In this study, we aimed to implement a novel framework using VNIR spectroscopy to quantify the main soil minerals like kaolinite, montmorillonite, muscovite, illite, chlorite, goethite, haematite/hematite, gibbsite and calcite. Besides, we evaluated the application of DSM framework in the estimated values of those soil minerals using as environmental covariates a Synthetic Soil Image (SYSI), which represents bare soil areas from 1985 to 2019 at the soil surface, a Best Synthetic Soil Image (i.e. predicted from SYSI at 80 – 100 cm depth), a 3D drainage network and the Digital Elevation Model through the Random Forest algorithm.

3.2. MATERIAL AND METHODS

3.2.1. Study area

The study area covers about 2,574 km^2 and contains eight cities in the São Paulo State, Brazil. There are two defined seasons, dry winters and rainy summers, annual average temperature ranging from 20° to 22.5°C, and annual rainfall between 1,200 and 1,400 mm. The common topographic characteristics are rolling uplands and undulating hills with altitudes ranging from 450 to 950 m. The main land-use is predominantly agriculture (e.g. sugarcane and pastureland) with no-till and till farming, which implicates to have soil revolved up to 60 cm depth along the year before planting some crops. Geologically, there are a great diversity of parent materials such as siltstones, tillites, varvites, conglomerates, sandstones, limestones, siltstones, flint, dolomite, siltite, pyrombetuminosite, shales, diabase and basalt. This great diversity of parent materials and variation of topography (plain to strong rolling) gave the region contrasting minerals.

3.2.2. Soil spectra data

The soil spectra database comprises 2,701 observations in the spectral range of 350 – 2,500 nm at 0 – 20, 40 – 60, and 80 – 100 cm depths. These spectra were acquired using the Fieldspec 3 sensor (Analytical Spectral Devices, Boulder, Colorado, USA) with a spectral resolution of 1 nm in the laboratory. The soil samples were air-dried for 48h at 45°C, sieved (< 2 mm), and placed on petri dishes. The sensor was positioned vertically at 8 cm from the platform, spotted the energy reflected from two 50-W halogen lamps with no-collimated beam to the petri dishes. These lamps were positioned 35 cm from the platform at a zenith angle of 30° and three measurements for each sample were performed turning the petri dishes 90° between the sensor's reading intervals. A Spectralon was used as a white reference of ~100% reflectance for calibration. Afterwards, the average of the three readings and the white reference reflectance was used to calculate the final reflectance factor for each sample. Furthermore, the soil mineral spectra standard was retrieved from the USGS Spectral Library Version 7 (Kokaly et al., 2017). We selected the nine most common soil minerals (Table 1).

Table 1. List of the soil mineral spectra retrieved from the USGS Spectral Library Version 7.

Mineral	Type	Sample ID	Spectrometer
Goethite	Hydroxide	GDS134	FieldSpec3 standard resolution
Haematite	Oxide	HS45.3	FieldSpec3 standard resolution
Gibbsite	Hydroxide	HS423.2B	FieldSpec3 standard resolution
Kaolinite	Phyllosilicate	KGa-2	FieldSpec4 high-resolution next generation
Montmorillonite	Phyllosilicate	SAz-1	FieldSpec4 high-resolution next generation
Muscovite	Phyllosilicate	GDS113	FieldSpec4 high-resolution next generation
Illite	Phyllosilicate	IMt-1.a	FieldSpec4 high-resolution next generation
Chlorite	Phyllosilicate	HS179.2B	FieldSpec3 standard resolution
Calcite	Carbonate	GDS304	FieldSpec3 standard resolution

3.2.3. Processing spectral data

Qualitative evaluation of the raw spectra of most minerals presents sparse information on the optical features. One way to reveal this information is performing mathematical transformation and the most useful is given by the Kubelka-Munk theory (Barron & Torrent, 1986). The reflectance is stated as a function of the reflectance over a background and thickness of the layer, and the absorption and scattering are constants (Equation 1). The Kubelka-Munk (K-M) function at any wavelength is:

$$\frac{K}{S} = \frac{(1 - R_\infty)^2}{2R_\infty} = \theta \qquad (1)$$

where K and S are respectively the absorption and scattering coefficients, the θ is the remission or K-M function, and the R_∞ is the limiting reflectance.

This differential equation permits a hyperbolic solution, which shows strongly and wide superposed bands at different wavelengths, likewise the raw spectra. However, obtaining the derivative of the K-M function curves enhance the resolution of sharply defined features. Examples of the application of the K-M function to estimate iron oxides and kaolinite can be found in Barron and Torrent (1986), and Jepson (1988). The second-derivative enhances slight concavities and convexities of the original spectrum and presents a much narrower bandwidth (Kosmas et al., 1984). Yet, the weak absorption on the K-M function curves become strong minima and maxima in the second-derivative curves (Fig. 1). As spectra are acquired stepwise, a smoothing procedure must be performed for calculation of successive derivatives. One of the foremost algorithms of smoothing spectra was developed by Savitzky and Golay (Savitzky & Golay, 1964). This method consists of using a set of contiguous data points to fit a polynomial curve as described in Torrent and Barron (2015). As pointed out by Scheinost et al. (1998), and Silva et al. (2020), the second-derivative of the K-M function curves has slightly smaller detection sensitivity than X-ray diffraction for quantifying minerals in soils.

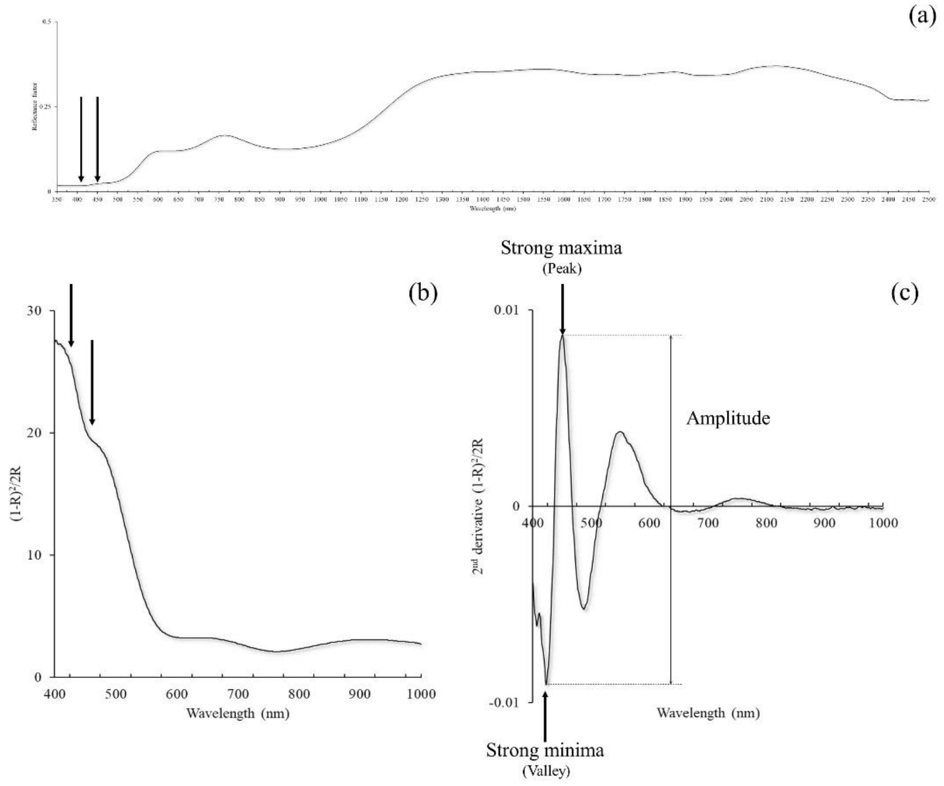


Fig. 1. Exemplifying spectral processing of goethite. Original spectra (a), Kubelka-Munk (K-M) function (b), and second-derivative of the K-M function using the Savitzky-Golay method (c).

3.2.4. Mineral quantification

The soil mineral bands in the Vis-NIR-SWIR spectra were selected based on the literature, but mainly in the strong maxima and minima of the second-derivative curves of the soil mineral standards (Table 2). The soil minerals selected in this study were: goethite (Gt), haematite/hematite (Hem), gibbsite (Gbs), kaolinite (Kln), montmorillonite (Mnt), muscovite (Ms), illite (Ill), chlorite (Chl), and calcite (Cal). The mineral abbreviations are in according to those recommended by the International Union of Geological Sciences (Siivola & Schmid, 2007). The amplitude of each mineral was calculated and then normalised.

The mineral quantification was achieved by transforming the original spectra (Fig. 1a) into the K-M function curves (Fig. 1b) and calculating the second-derivative using the Savitzky-Golay method with a set of 35 points and a polynomial function of order 2 (Fig. 1c) in the R software (R Development Core Team, 2020) and the AlradSpectra (Dotto, Dalmolin, Caten, Gris, & Ruiz, 2019). This procedure identified the strong minima and maxima in the second-derivative curve, matching the positions of the absorption bands in the original spectra (Fig. 1c). Table 2 shows the spectral bands selected to calculate the amplitude. Afterwards, the amplitude values were normalised by dividing by the maximum value per each value and multiplying by 100. Thus, the final results were measured as relative weight percentages (wt.%) for each mineral.

Table 2. Descriptive position of the spectral bands for identification and characterisation of soil minerals according to mineral purity of USGS Spectral Library Version 7 and the scientific literature.

Type mineral	Abbrev.	Selected	References	
		$(\lambda_{min}/\lambda_{max})$	$(\lambda_{min}/\lambda_{max})$	
Goethite	Gt	(422/450)	(~415/~445)	(Scheinost et al., 1998)
Haematite	Hem	(535/575)	(~535/~580)	(Scheinost et al., 1998)
Gibbsite	Gbs	(2265/2285)	(2265/2295)	(Clark et al., 1990)
Kaolinite	Kln	(1415/2205)	(1395/2165), (1406/2194) and (1415/2210)	(Mathian et al., 2018)
Montmorillonite	Mnt	(1415/1885), (1900/2190) and (2207/2236)	(1400/1900), (1413/2207) and (2207/2236)	(Mathian et al., 2018; Mulder et al., 2013)
Muscovite	Ms	(1415/2190) and (2350/2406)	(1412/2197) and (2350/2450)	(Mathian et al., 2018)
Illite	Ill	(2205/2280)	(2200/2350)	(Mulder et al., 2013)
Chlorite	Chl	(2247/2296) and (2326/2360)	(1407/2259) and (1415/2351)	(Mathian et al., 2018)
Calcite	Cal	(2342/2367)	(2300/2350)	(Mulder et al., 2013)

Note: λ_{min}, the spectral band with minima value in nm; λ_{max}, the spectral band with maxima value in nm.

3.2.5. Estimation and Assessment of the mineral quantification

We proposed an estimative and conversion of the measurement unit of the soil minerals in weight percentages to g kg^{-1} based on clay content. For this procedure, we randomly selected 185 samples out of 2701 available at 0 – 20 cm depth and sent to laboratory analyses. It is of noteworthy we had not enough soil samples of all 2701 observations and financial resources for chemical analyses. That is why we selected only at the soil surface because it had more samples available and well-distributed in the study area. The samples were air-dried (48h at 45ºC) and sieved (< 2mm mesh). The clay content was determined using the densimeter method as described in Camargo et al. (2009). Whilst, the potassium (K) was analysed using the 3051A method (US EPA, 2007). This method is the microwave-assisted acid digestion (extraction or dissolution) of sediments, sludges, soils, and oils and an alternative to conventional heating with nitric acid (HNO_3) or hydrochloric and nitric acid ($HCl+HNO_3$) solutions.

Frequency data in the literature shows that muscovites are more common in soils and contain between 9 and 10% of total potassium (Weaver & Pollard, 1973). This percentage has been used to estimate the mica content in the clay fraction of the soil (Jackson, 1958; Kämpf et al., 2012). Such parameter was applied to estimate the mica content from the 185 samples (Equation 2). Then, we calculated the probable value of total clay content expected for each sample as in Equation 3. Thus, the factor that has to be multiplied by the clay content and the relative weight percentage of each mineral was defined (Equation 4) by selecting the median value (Fig. 2). This factor (W) could be used to calculate the estimate of the soil mineral content as long as the users follow the methodology proposed here. Further assessment of this factor was performed by using the regression equation estimated by Fernandes et al. (2004) to calculate the Hematite and Goethite contents based on soil spectra. Using the W factor of 0.3 (Fig. 2), we estimated the hematite and goethite contents (Equation 5). Then, we assessed the good of fitting (e.g. RMSE, R^2, CCC, and Bias) considering as observed values the estimative values based on Fernandes et al. (2004) work and predicted values our estimative.

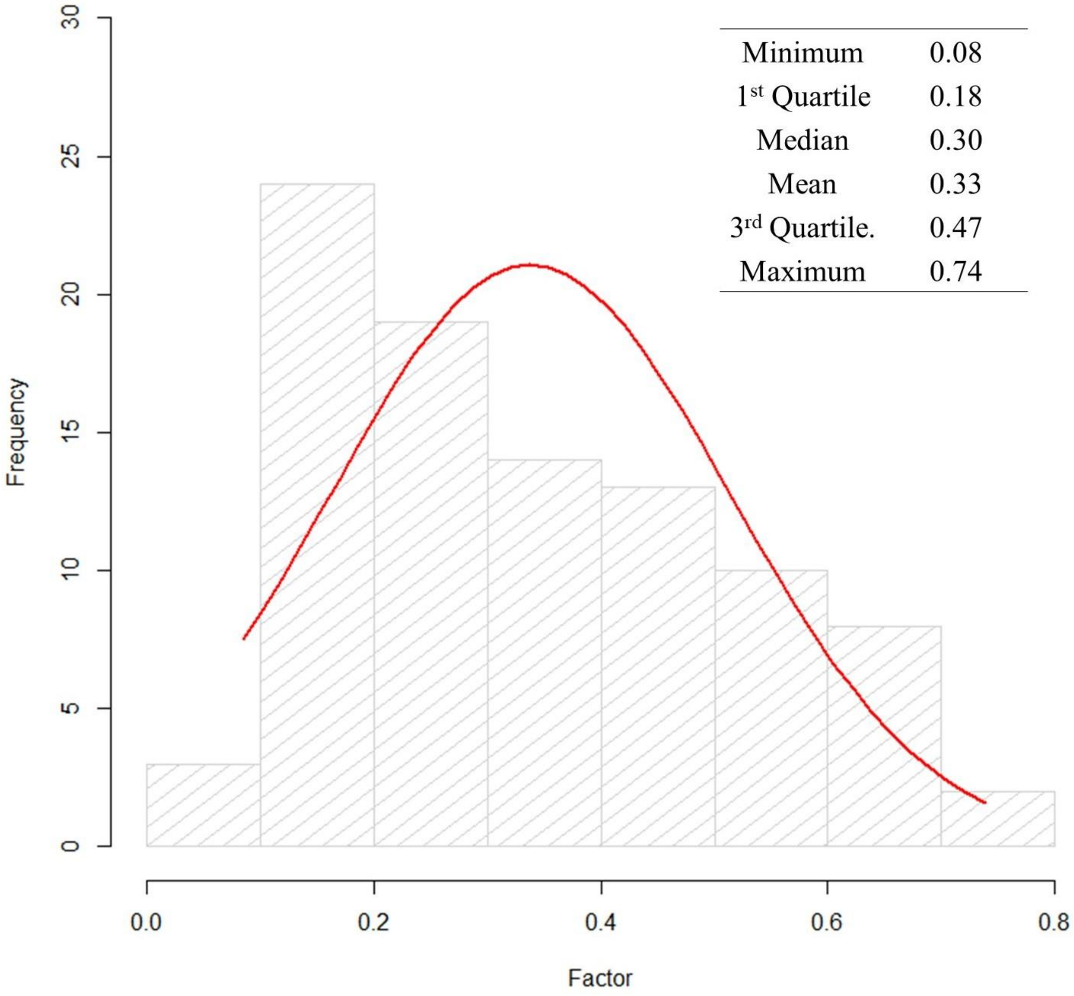
$$Ms_{g\ kg^{-1}} = 10 \times K \tag{2}$$

$$EClay_{g\ kg^{-1}} = \left(\frac{Ms_{g\ kg^{-1}}}{Ms_{wt\%}}\right) \times 100 \tag{3}$$

$$W = \left(\frac{EClay_{g\ kg^{-1}}}{Clay_{g\ kg^{-1}}}\right) \tag{4}$$

$$SM_{g\ kg^{-1}} = W(0.3) \times SM_{wt\%} \tag{5}$$

Where $Ms_{g\ kg^{-1}}$ is the estimated muscovite, K is the total potassium determined in the laboratory (1000g kg^{-1}), $EClay_{g\ kg^{-1}}$ is the estimated clay content, $Ms_{wt\%}$ is the muscovite calculated from the soil spectra, W is the estimated factor, $Clay_{g\ kg^{-1}}$ is the clay content determined in the laboratory, SM is the soil mineral of interest.



Minimum	0.08
1st Quartile	0.18
Median	0.30
Mean	0.33
3rd Quartile.	0.47
Maximum	0.74

Fig. 2. Histogram of the estimated factor for mineral quantification using Vis-NIR-SWIR spectra. Redline is the normal curve.

3.2.6. Digital Mapping

There was a total of 1008, 820, and 873 observations at 0 – 20, 40 – 60, and 80 – 100 cm depths. The 2701 samples were split into 80% and 20% for calibration and validation of the models for each of the nine minerals, respectively. The internal validation for each mineral was performed by using the calibration dataset and the external validation, which simulates field observations, was the 20% left out of the model. The environmental variables as proxies of the soil mineral formation into the digital soil mapping framework (McBratney et al., 2003) were the digital elevation model (DEM) retrieved from the Shuttle Radar Topography Mission at a resolution of 1 arc-second (USGS, 2018), the drainage density created using 3D digital aerophotographies, and the Synthetic Soil Image (SYSI) as described in Demattê et al. (2018, 2020). Mendes et al. (2019) mapping the soil attributes in the same study area, evaluated the original SYSI and the predicted SYSI (e.g. called Best Synthetic Soil Image – BSSI) from integrative hyper and multispectral approach. Thus, the authors concluded that the original SYSI could be used as a proxy (covariable) from soil depths above 60 cm and the BSSI for depths below 60 cm improving digital soil mapping predictive power. Based on those results, we used the original SYSI as environmental predictors at 0 – 20 and 40 – 60 cm depths, and the BSSI at 80 – 100 cm depth.

The chosen machine learning algorithm was the Random Forest (RF), which is categorised as an ensemble learning method. The RF split training samples into subsets and generates decision trees for each subset. Each new training subset used to build a decision tree, one third is randomly removed. This sample is called out-of-bag and the remaining samples (in-the-bag) are handled to build the decision tree. Out-of-bag samples are used to assess the model performance and select the training subset with higher accuracy. Thus, it was performed a grid search for optimal tuning hyperparameters and 100 interactions within the models by 10-fold repeated cross-validation method carried out five times for each mineral model. The hyperparameters selected is in Table 3. Each mineral modelling had different inputs in the hyperparameters *mTry*, *minimum node size*, and the *number of decision trees* at the three distinct soil depths. Further details on these hyperparameters are described in (Breiman, 2001; Cutler et al., 2007; Hengl, Nussbaum, Wright, Heuvelink, & Gräler, 2018).

Table 3. Hyperparamenters of random forest models for nine soil minerals.

Type	Depth	m_{try}	Min_{nodes}	n_{tree}
Goethite	0 – 20	5	26	5000
	40 – 60	7	27	1000
	80 – 100	7	29	1000
Haematite	0 – 20	5	29	1500
	40 – 60	7	24	1000
	80 – 100	7	25	1000
Gibbsite	0 – 20	6	25	1000
	40 – 60	6	10	2000
	80 – 100	7	21	1000
Kaolinite	0 – 20	7	15	1500
	40 – 60	7	17	1000
	80 – 100	2	26	1000
Montmorillonite	0 – 20	6	21	1500
	40 – 60	7	28	750
	80 – 100	7	26	1000
Muscovite	0 – 20	1	26	1500
	40 – 60	1	20	750
	80 – 100	7	24	1000
Illite	0 – 20	7	23	1000
	40 – 60	7	12	1000
	80 – 100	7	27	750
Chlorite	0 – 20	1	8	750
	40 – 60	1	26	1000
	80 – 100	1	20	1000
Calcite	0 – 20	2	27	750
	40 – 60	1	20	1500
	80 – 100	1	25	1000

3.2.7. Model evaluation

The metrics of model assessment used in this study were root mean squared error (RMSE), adjusted coefficient of determination (R^2_{adj}), concordance correlation coefficient (CCC), and bias. Each of these parameters explains the relationship between the predicted and observed values in distinct ways. The RMSE (Equation 5) explain how close the predicted values are to the real values by using the square root of the squares of the residuals, which sum up the degree of the residuals. The R^2_{adj} verifies the proportion of the variance of the covariates that affect the response variable by the approximated line of regression that the model enlightens (Equation 6). This metric allows digital soil mappers to compare model performances among distinct target variables. The Bias is calculated as the distance from the average prediction and observed values (Equation 7). This other validation metric fulfils one of the limitations left by the R^2_{adj}, showing the model bias. The last validation metric of model performance is the Lin's concordance correlation coefficient (CCC) (Equation 8). The CCC assesses the agreement between the predicted and observed values and so, it could be a more appropriate metric than R^2_{adj}.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{5}$$

$$R^2_{adj} = 1 - \frac{(SS_{res}/df_e)}{(SS_{tot}/df_t)} \tag{6}$$

$$Bias = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i) \tag{7}$$

$$CCC = \frac{2\rho\sigma_{pred}\sigma_{obs}}{\sigma_{pred}^2 + \sigma_{obs}^2 + \left(\mu_{pred} - \mu_{obs}\right)^2} \qquad (8)$$

Where *n*, y_i, and $\hat{y}_i$ are sample size, observed values, and predicted values of the response variable, respectively. SS_{res}, SS_{tot}, df_e, and df_t are respectively the sum of squares of the regression residual, the sum of the square of the total residual, the degrees of freedom of the estimated population error variance, and the degree of freedom of the estimated population variance of the dependent variable. σ_{pred}^2 and σ_{obs}^2 are the prediction and observation variances, respectively, μ_{pred} and μ_{obs} are the means of the predicted and observed values. ρ is the correlation coefficient between the predicted and observed values.

3.3. RESULTS AND DISCUSSION

3.3.1. Soil mineral spectra

The descriptive statistics are in Table 4. The negative values of amplitude were replaced by zero values. This was performed because whether the amplitude was negative, it meant there is not that mineral in soils. The normalisation was done as described in the methodology and the mineral content converted from spectral values to wt.%. The outliners were removed. Grouping each soil mineral into classes based on their content, in wt.%, it allowed us to identify their distinction on the vis-NIR-SWIR spectra (Fig. 3). The standards of the USGS Library were processed together with the original soil spectra. The literature reports Gt and Hem bands close to 415 and 445 nm, and 535 and 580, respectively. However, it was identified that their peak values were in bands 422 and 450 nm, and 535 and 575 nm. Some minerals had a slight difference in their specific band reported in the literature with those found in our study (Table 2). Thus, some bands were chosen based on the peaks from the spectral standards for each mineral not necessarily following the same bands in the literature.

Table 4. Descriptive statistic of the amplitude retrieved from the laboratory soil spectra for nine soil minerals.

	n	*Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum	SD
Agt	2701	0	0.00118	0.00243	0.00290	0.00408	0.05873	0.00255
Hem	2701	0	0.00035	0.00108	0.00233	0.00354	0.01273	0.00272
Gbs	2701	0	0.00000	0.00002	0.00013	0.00017	0.00293	0.00024
Kln	2701	0	0.00129	0.00219	0.00237	0.00316	0.00888	0.00139
Mnt	2701	0	0.00153	0.00221	0.00233	0.00295	0.00888	0.00114
Ms	2701	0	0.00047	0.00077	0.00086	0.00110	0.00473	0.00056
Ill	2701	0	0.00059	0.00100	0.00111	0.00147	0.00466	0.00069
Chl	2701	0	0.00000	0.00001	0.00007	0.00005	0.00334	0.00020
Cal	2701	0	0.00000	0.00001	0.00009	0.00010	0.00295	0.00021

Note. SD, standard deviation. *Negative values of amplitude were replaced by zero values considering no mineral's presence.

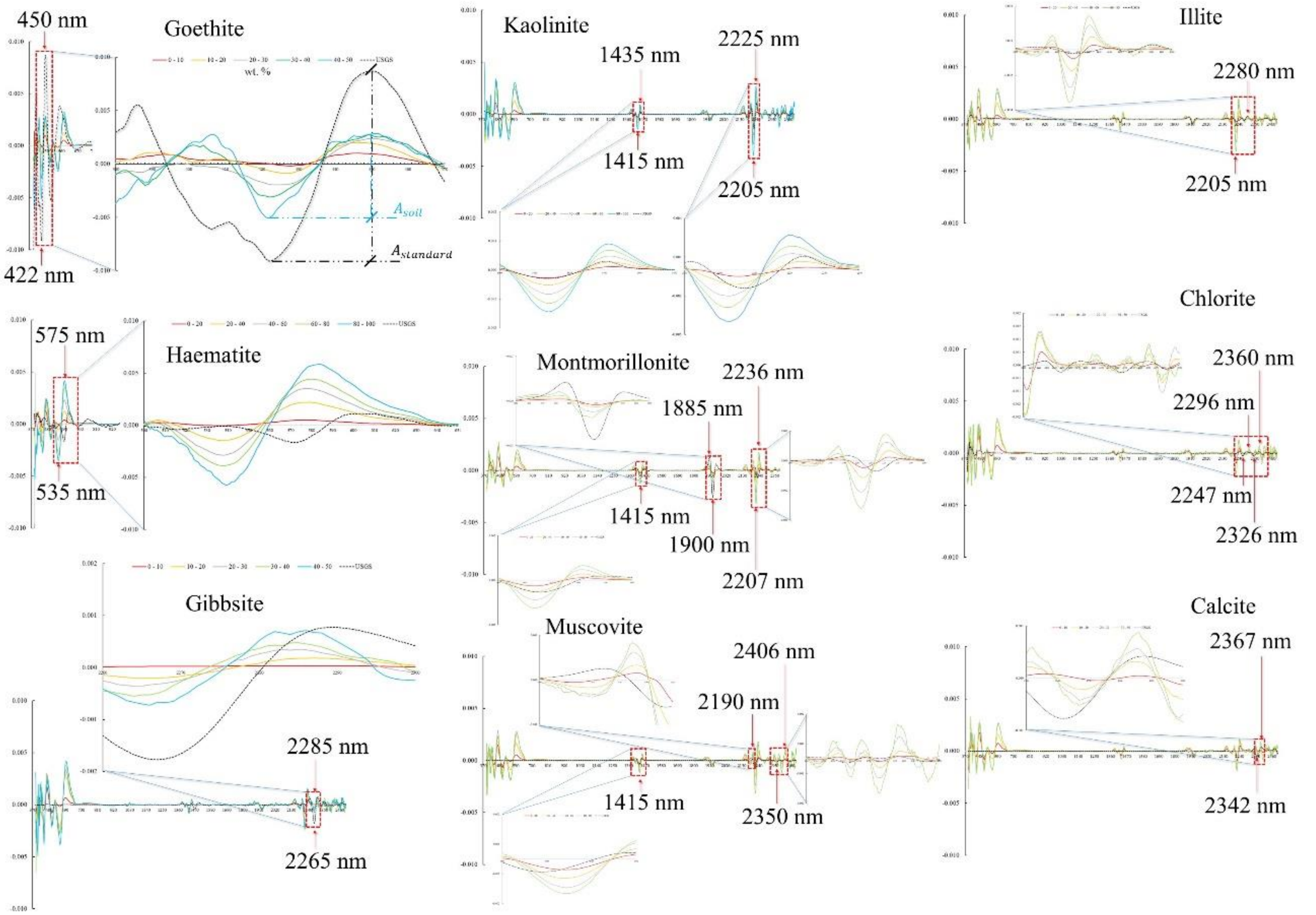


Fig. 3. Soil mineral spectral curves of the Second-Derivative using Savitzky-Golay with a set 35 points and a polynomial function of order 2. Blackline is the mineral standards from USGS Library.

3.3.2. Soil mineral content estimative

A soil mineral content estimative was suggested based on clay content (g kg^{-1}) from laboratory analysis. First of all, it was created an index, constant factor, using the observed values of potassium. As reported in the literature (Jackson, 1958; Kämpf et al., 2012; Weaver & Pollard, 1973), these values can be multiplied by 10 and the final values represent the actual content of muscovite in soils. With these values, it was computed, as explained in the methodology, the constant factor selecting the median (Fig. 2). Therefore, we found 0.3 as the constant factor to be multiplied by the clay content to print out the actual mineral content in soils. The muscovite content estimative using this factor presented similar values that the laboratory analysis (Fig. 4). As it is intrinsic to the term, we found an estimative to compute minerals content in soils using the vis-NIR-SWIR spectra. The assessment of this factor was performed by analysing the good of fitting (e.g. RMSE, R^2, CCC, and Bias) of the observed values which were estimated based on Fernandes et al. (2004) work and the predicted values from our estimative for Hem and Gt (Fig. 5). Without outliers, the total samples were 2619 and 2645 units for Hem and Gt, respectively. The Hem and Gt presented RMSE and R^2 values of 67.94 g kg^{-1} and 0.89, and 56.82 g kg^{-1} and 0.63, respectively. This proved the level of concordance between our methodology and that stated by those authors.

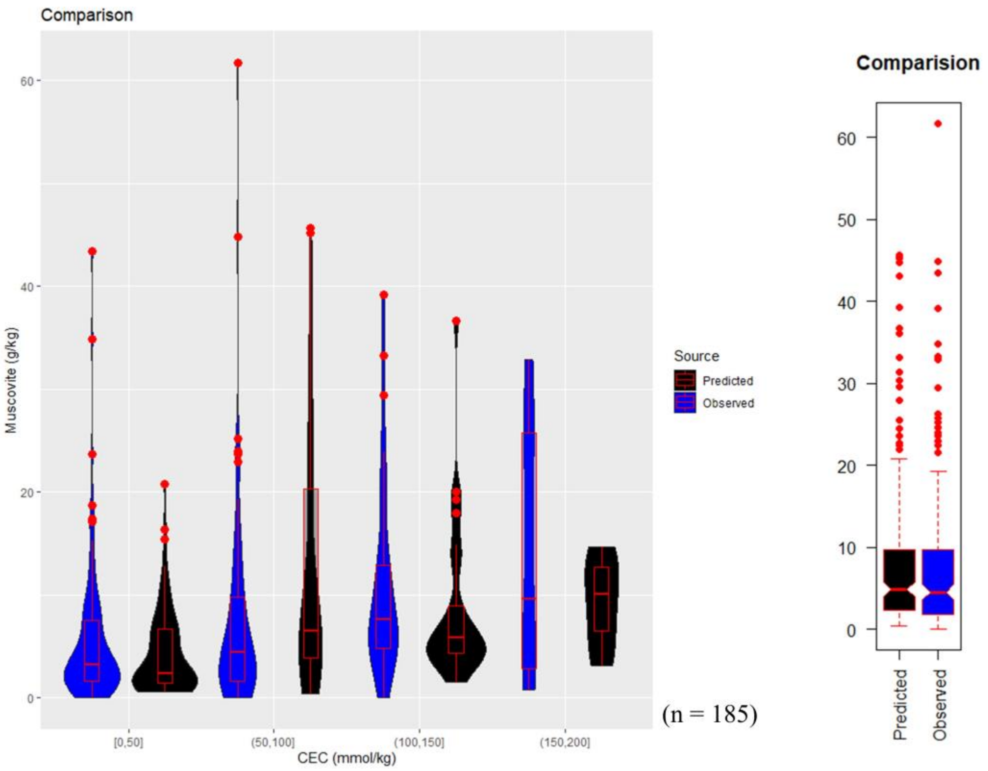


Fig. 4. Soil mineral content estimative of muscovite (predicted) and the total soil muscovite from laboratory analysis (observed).

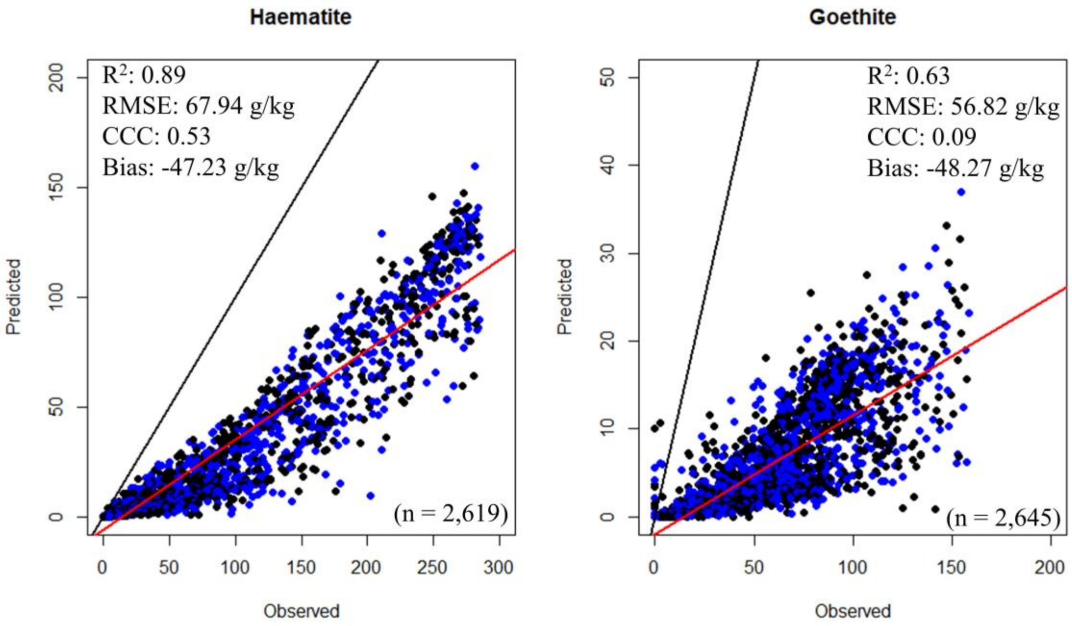


Fig. 5. Validating the soil mineral content estimative using the Hematite and Goethite minerals with observed values from the regression equation estimated method by Fernandes et al. (2004), and our estimated index.

3.3.3. Digital mineral mapping

The digital soil mapping is a framework that integrates field observations, algorithms, and remote sensing data to predict and create spatial information of soil attributes, properties and classes. This discipline of soil science was stated by McBratney et al. (2003). Herein, we create spatial information of the minerals content estimative in soils using the Random Forest algorithm with the environmental variables digital elevation model (DEM), drainage network, the Synthetic Soil Image (SYSI) at 0 – 20, and 40 – 60 cm depth, and the Best Synthetic Soil Image (BSSI) at 80 – 100 cm depth to predict those soil minerals. The initial assessment of the interaction between the soil minerals and the environmental variables was accessed by the Pearson correlation index (Fig. 6). Most of the correlation values showed a negative tendency among those variables and the response variable. Low correlation coefficients were found among DEM, Drainage and the soil minerals. Chlorite and Calcite presented low correlations with all environmental variables. The correlation coefficients ranged from 0.1 to -0.4 (Fig. 6a), 0.1 to -0.2 (Fig. 6b), and 0.0 to -0.50 (Fig. 6c) at 0 – 20, 40 – 60, and 80 – 100 cm depths, respectively. Strong negative correlation values were observed using the BSSI as a proxy of soil minerals (Fig. 6c).

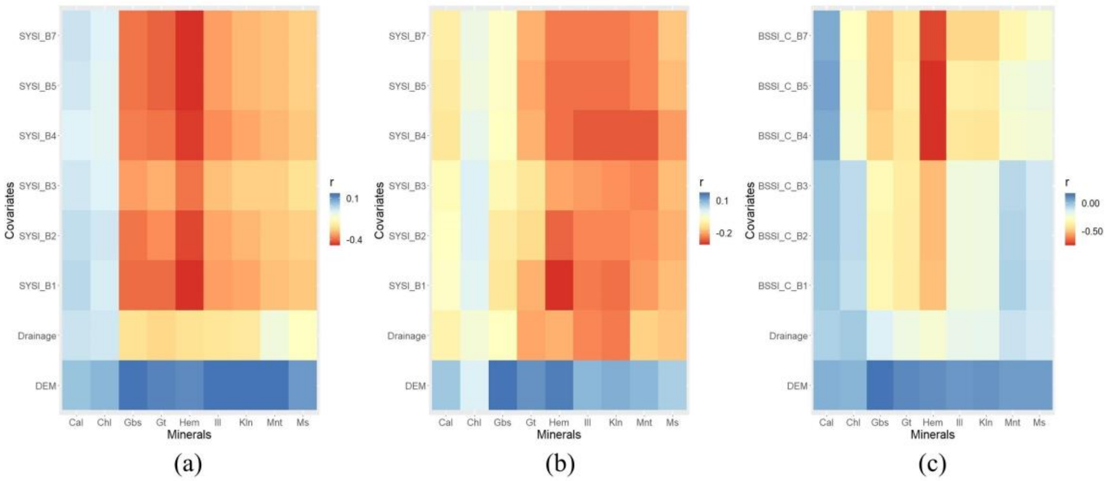


Fig. 6. Pearson correlation index ($p < 0.01$) between the soil minerals and the environmental variables used for Digital Mineral Mapping at 0 – 20 cm (a), 40 – 60 cm (b), and 80 – 100 cm (c) depths.

Table 5 displays the descriptive statistics of the minerals content estimative used as calibration (80%) and external validation (20%) of the models in the DSM framework in the three depths. The outliners were removed prompting different sets of observations for each mineral at each depth. The uncertainty inside the models was accessed by the RMSE and R^2 to choose the best-fitted model. The real predictive power of the models was accessed using the external validation dataset, which mimics the field observations (Table 6). In both cases, we could achieve better performances for soil subsurface (Fig. S1) rather than above 60 cm depth (Fig. S2 and S3). Even though the sampling density for soil subsurface was lower than above 60 cm depth, the models fitted better the predictions of the soil minerals. The only exception was the chlorite. Overall, the model did not fit well and some of the predictions presented lower correlation index, however, the RMSE values were satisfactory.

Table 5. Descriptive parameters of nine soil minerals (wt.%).

Type	Depth (cm)	Training	Validation	Total points	Skewness	Kurtosis	SD	CV
Goethite	0 – 20	796	196	992	0.79	-0.10	2.93	72.73
	40 – 60	640	156	796	0.71	-0.09	3.13	70.50
	80 – 100	640	158	798	0.43	-0.32	3.29	60.23
Haematite	0 – 20	778	192	970	1.32	0.52	17.87	121.84
	40 – 60	600	148	748	1.38	1.07	11.92	107.62
	80 – 100	651	160	811	0.95	-0.34	22.02	98.75
Gibbsite	0 – 20	758	188	946	1.34	0.64	3.21	137.95
	40 – 60	563	140	703	1.78	2.39	2.16	158.25
	80 – 100	613	152	765	1.23	0.47	4.92	129.66
Kaolinite	0 – 20	792	196	988	0.68	-0.18	10.23	54.97
	40 – 60	651	160	811	0.59	-0.33	15.37	53.33
	80 – 100	640	157	797	0.36	-0.26	13.03	40.72
Montmorillonite	0 – 20	789	196	985	0.45	-0.28	8.44	45.49
	40 – 60	648	160	808	0.24	-0.30	11.84	41.78
	80 – 100	636	156	792	0.47	0.02	9.81	32.05
Muscovite	0 – 20	783	192	975	0.59	-0.23	6.93	56.75
	40 – 60	632	156	788	0.46	-0.21	9.62	52.62
	80 – 100	631	156	787	0.19	0.00	9.29	44.74
Illite	0 – 20	783	192	975	0.80	0.00	9.26	56.49
	40 – 60	647	160	807	0.73	-0.12	14.03	55.21
	80 – 100	629	156	785	0.54	-0.09	11.88	42.25
Chlorite	0 – 20	698	173	871	1.98	2.95	0.74	182.50
	40 – 60	559	139	698	2.20	3.95	0.69	207.25
	80 – 100	564	140	704	1.69	1.93	1.04	163.62
Calcite	0 – 20	703	175	878	1.71	2.08	1.36	155.64
	40 – 60	575	142	717	1.79	2.22	2.30	166.54
	80 – 100	579	144	723	1.61	1.65	2.17	155.88

Table 6. Results of internal and external validation of the models for nine soil minerals (wt.%).

Type	Depth	RMSE$_{train}$	R$^2_{train}$	MAE$_{train}$	RMSE$_{val}$	R$^2_{val}$	CCC	Bias
Goethite	0 – 20	2.47	0.29	1.84	2.81	0.16	0.31	0.09
	40 – 60	2.91	0.14	2.29	3.22	0.10	0.19	0.14
	80 – 100	2.91	0.22	2.22	2.81	0.24	0.39	0.01
Haematite	0 – 20	12.97	0.47	9.37	12.81	0.54	0.68	-0.72
	40 – 60	11.42	0.09	8.75	11.70	0.17	0.23	0.53
	80 – 100	13.26	0.63	9.32	13.77	0.62	0.74	0.37
Gibbsite	0 – 20	2.56	0.36	1.85	2.59	0.32	0.51	0.03
	40 – 60	2.09	0.07	1.56	1.95	0.00	0.04	-0.17
	80 – 100	3.97	0.36	2.95	3.56	0.38	0.57	-0.06
Kaolinite	0 – 20	9.30	0.18	7.26	9.78	0.17	0.28	0.03
	40 – 60	14.19	0.15	11.40	14.37	0.09	0.19	0.34
	80 – 100	11.11	0.28	8.64	10.79	0.42	0.51	0.82
Montmorillonite	0 – 20	7.96	0.11	6.28	7.98	0.16	0.26	0.17
	40 – 60	11.39	0.08	9.18	11.90	0.04	0.11	-0.02
	80 – 100	8.56	0.24	6.71	8.83	0.27	0.38	0.42
Muscovite	0 – 20	6.52	0.12	5.10	6.86	0.07	0.16	0.05
	40 – 60	9.48	0.04	7.53	9.32	0.04	0.08	0.47
	80 – 100	8.91	0.09	6.82	9.30	0.08	0.16	0.17
Illite	0 – 20	8.45	0.17	6.54	7.84	0.19	0.35	-1.21
	40 – 60	13.26	0.11	10.57	12.84	0.13	0.20	0.00
	80 – 100	9.99	0.30	7.78	10.01	0.25	0.42	-0.47
Chlorite	0 – 20	0.73	0.03	0.53	0.74	0.07	0.10	0.02
	40 – 60	0.69	0.01	0.49	0.62	0.05	0.07	-0.02
	80 – 100	0.99	0.10	0.75	0.93	0.06	0.18	-0.17
Calcite	0 – 20	1.33	0.05	0.99	1.44	0.01	0.05	0.01
	40 – 60	2.29	0.02	1.73	2.00	0.02	0.06	-0.23
	80 – 100	2.10	0.06	1.61	1.90	0.09	0.17	-0.03

It was also retrieved the level of importance that each environmental variable had inside the model to predict each mineral (Fig. 7). The drainage network and DEM were the most relevant among the others predictors at 0 – 20 and 40 – 60 cm depths (Fig. 7a, b). Drainage influences reductive and oxidative soil processes altering iron oxides state (Bigham, Golden, Bowen, Buol, & Weed, 1978; Malone, McBratney, & Minasny, 2018). Whilst, the relief alter the dynamic of water and also interfere in the particle remove process (Terra, Demattê, & Viscarra Rossel, 2018). Nevertheless, this trend was not observed below 80 cm depth (Fig. 7c). As increase depth, the concentration of primary minerals increases as well (Ben-Dor et al., 2006; Melo, Corrêa, Maschio, Ribeiro, & Lima, 2003), and it could affect, positively or negatively, the absorption and reflectance processes in soil spectra.

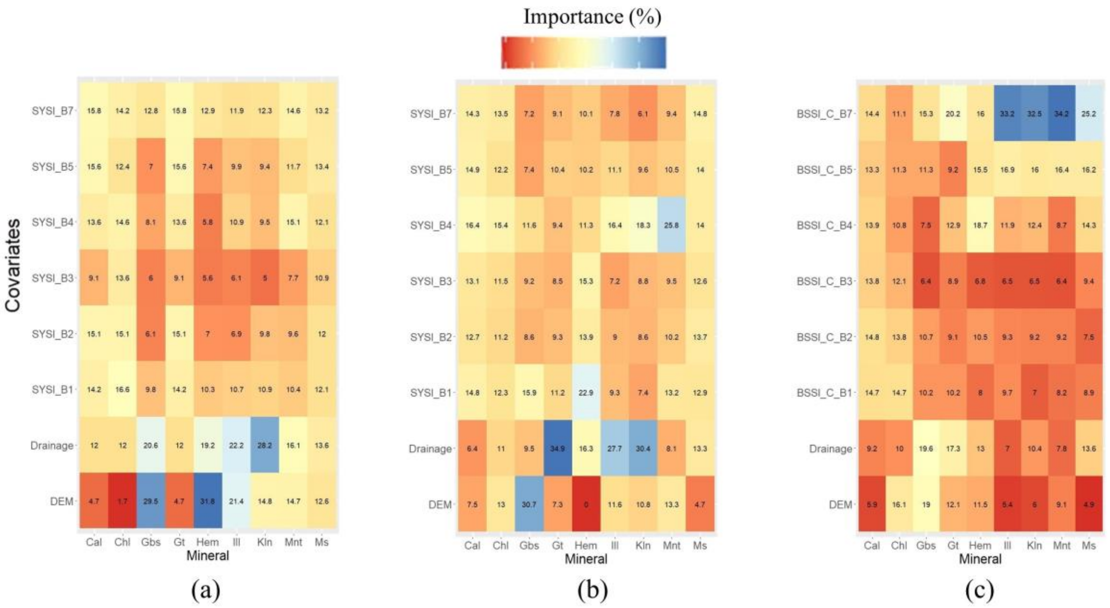


Fig. 7. Variable importance in predicting soil minerals using Random Forest algorithm at 0 – 20 (a), 40 – 60 (b), and 80 – 100 (c) cm depths.

3.3.4. Interpreting the predicted maps and their application

The predicted maps for the soil oxides (e.g. Hem, Gt, and Gbs), carbonate, 1:1 and 2:1 clay minerals are presented in the Fig. 8, 9, and 10. The minerals content increases as depth increase, which are in according to the principles of soil science because the minerals are close to the parent material (Buol, Southard, Graham, & McDaniel, 2011; Schaetzl & Anderson, 2005). The weathering of primary minerals forms most iron and aluminium oxides which are considered secondary minerals. The gibbsite is the most plentiful Al-oxide in soils and its formation is elevated by high precipitation in freely drainage sceneries, which allows the leaching of silica from the underlying parent material (Yokozeki, Watanabe, Sakata, & Otsuki, 2004), and by warm temperatures, that promote the weathering of primary minerals (Buol et al., 2011). Aspects of soil fertility in humid tropics are straitly related to weathering mechanism being kaolinite (1:1 phyllosilicate) with more proportion in those soils.

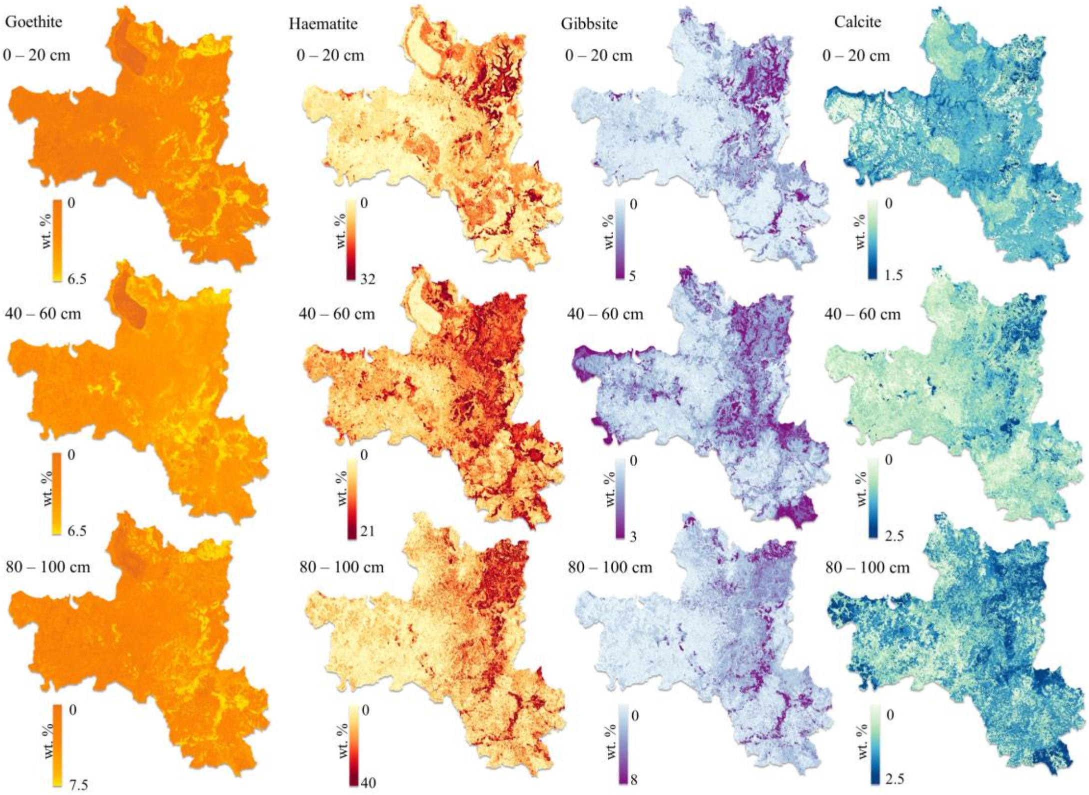


Fig. 8. Predicted maps of soil oxides and carbonate minerals at three soil depths.

As pointed out by Schwertmann and Herbillon (2015), Oxisols and Ultisols present unvaried mineralogy (hematite, goethite, gibbsite and kaolinite) in very weathered soils such as those in humid tropics. The oxides can arise as discrete particles in highly weathered environments and as coatings on mineral grains in moderated weathered ones. More rich in soils than the gibbsite is the goethite and haematite affecting soil colours brown to yellowish and grey that is common in an anaerobic environment. The occurrence of chlorides and carbonates are mainly as salt crusts in arid soils and as inherited from calcareous or formed in root zones, respectively. The first one is more soluble than the latter. Illites and muscovites are phyllosilicates of the mica group and part of the clay fraction highlighted by the potassium content into their chemical composition (Churchman, 2010).

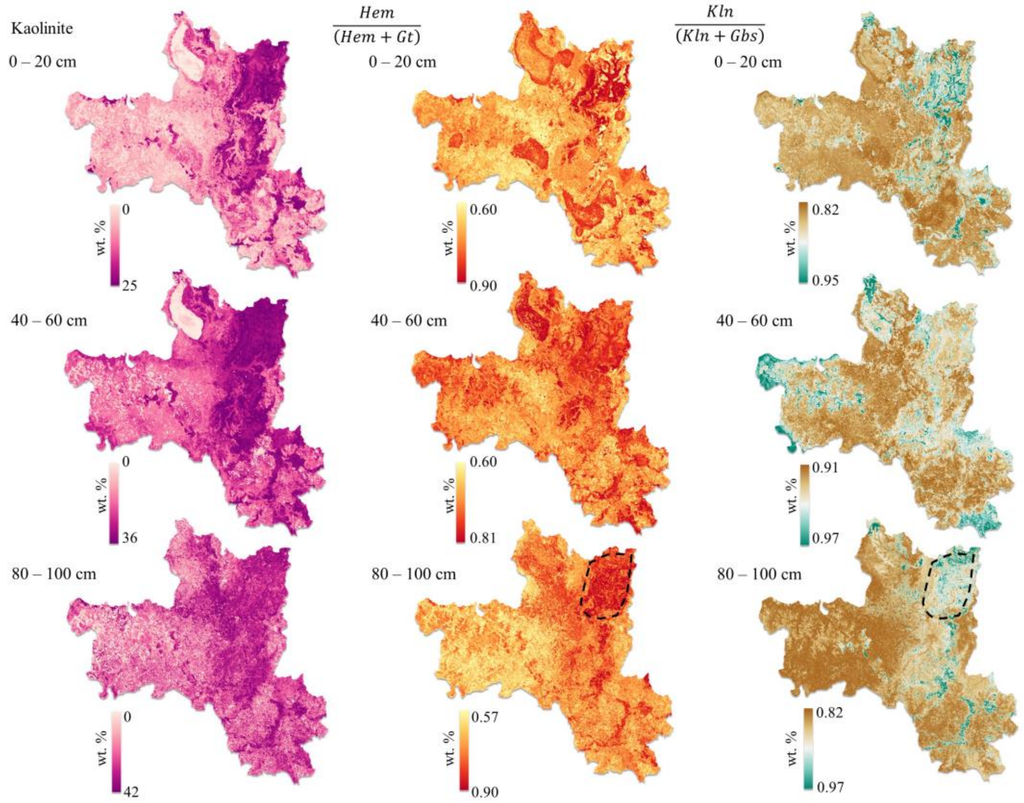


Fig. 9. Predicted maps of 1:1 clay mineral, hematite and goethite ratio, and kaolinite and gibbsite ratio at three soil depths. Blackline shows an area of Oxisols derived from basalt.

Igneous rocks are the source of micas and most of the phyllosilicates in soils. Micas similar to muscovites in soils are called illites when are found in clay fractions and biotite when in coarse fractions (Kämpf et al., 2012). Montmorillonites are also 2:1 phyllosilicates typified by a low layer charge and common in soils (Brown et al., 2006; Coyne et al., 1990; Dufréchou, Grandjean, & Bourguignon, 2015). Those regions with a high concentration of phyllosilicates and oxides are located in basalt and could indicate how it is interconnected to the parent material. The ratio of Hem and Gt, and Kln and Gbs show where they are more concentrated and help to access their dynamic in soils derived from sedimentary to igneous materials. As described in Gallo et al. (2018), quantifying topsoil attributes and the relationship with soil classes and geology, there is a great variation in soils and strong relationship with parent materials of this study area (black line in Fig. 9).

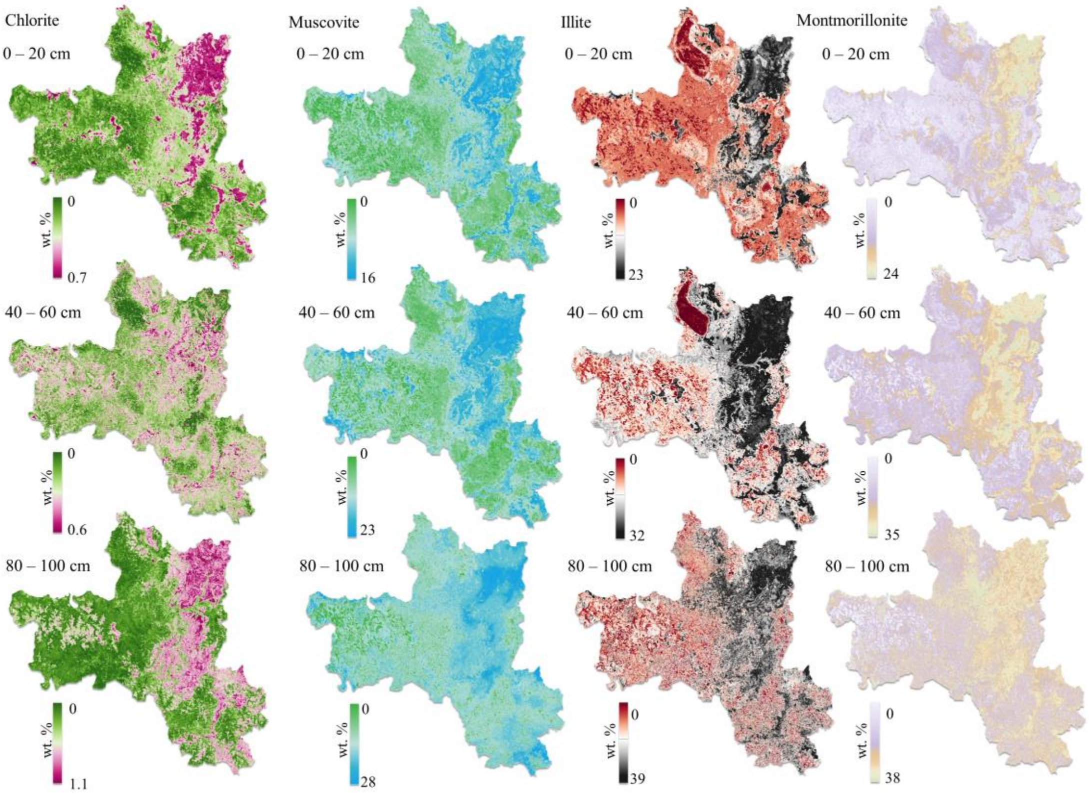


Fig. 10. Predicted maps of 2:1 clay minerals at three soil depths.

The mineral maps could enhance some previous and published studies in the area. As an example of this application, we selected the same area that reported in Vidal-Torrado and Lepsch (1999), which described the relationship between parent material and pedogenesis on a slope dominated by clayey oxidic soils (Fig. 11). The clay mineralogy of pedon 1 (P1 – clayey Dark-red Latosol/Rhodudox) was characterised as rich in kaolinite, gibbsite and illite. In the pedon 5 (Ultisol), those authors rose a question if there is mica in the form of illite above 500 cm depth. Thus, as in Fig. 11, we found the answer to that question without XRD and for a larger area including that one.

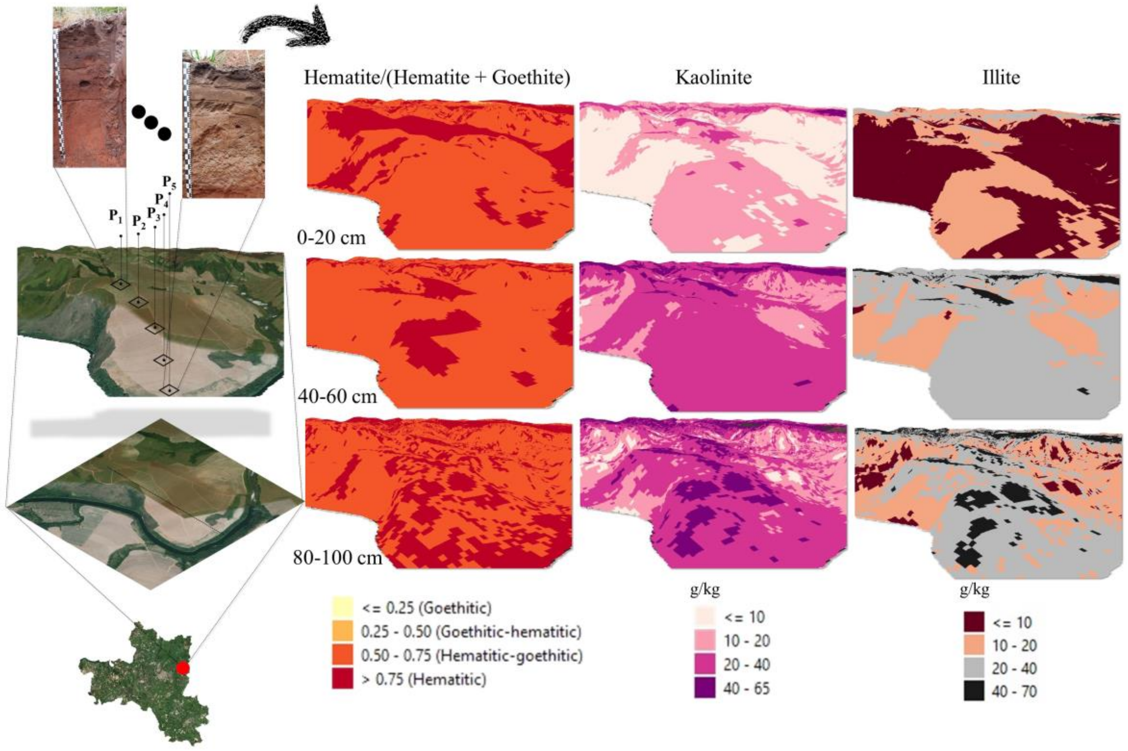


Fig. 11. Example of an application showing the spatial distribution of soil minerals in a former punctual soil survey inside the study area. The selected area and soil profiles (e.g. P$_1$, …, and P$_5$) are described in Vidal-Torrado and Lepsch (1999).

3.4. CONCLUSIONS

Attempts on quantifying soil mineral consist of having an observation measured using traditional laboratory soil analysis. However, developments in interpreting and analysing VNIR diffuse reflectance have allowed quantifying some soil minerals. This study showed that our proposed novel framework using VNIR spectroscopy and clay content, to quantify soil minerals in soils are feasible, non-destructive, quick and cost-efficient rather than traditional laboratory analysis and XRD. This methodology is promising to estimate soil mineralogy.

Furthermore, the application of DSM framework in the estimated values of those soil minerals using as environmental covariates a Synthetic Soil Image, a Best Synthetic Soil Image, a 3D drainage network and the Digital Elevation Model through the Random Forest algorithm proved to add spatialized soil mineral information into former investigations.

ACKNOWLEDGMENTS

This research is part of PhD's thesis of the first author, presented to the Graduate Program of Soil and Plant Nutrition at the "Luiz de Queiroz" College of Agriculture, University of São Paulo. We acknowledge the financial support of the São Paulo Research Foundation (FAPESP grant numbers 2016/26124-6 and 2014/22262-0). The authors also thank the members of the Geotechnologies in Soil Science Group (GeoCIS/GeoSS) (https:// esalqgeocis.wixsite.com/english).

APPENDIX A. SUPPLEMENTARY DATA

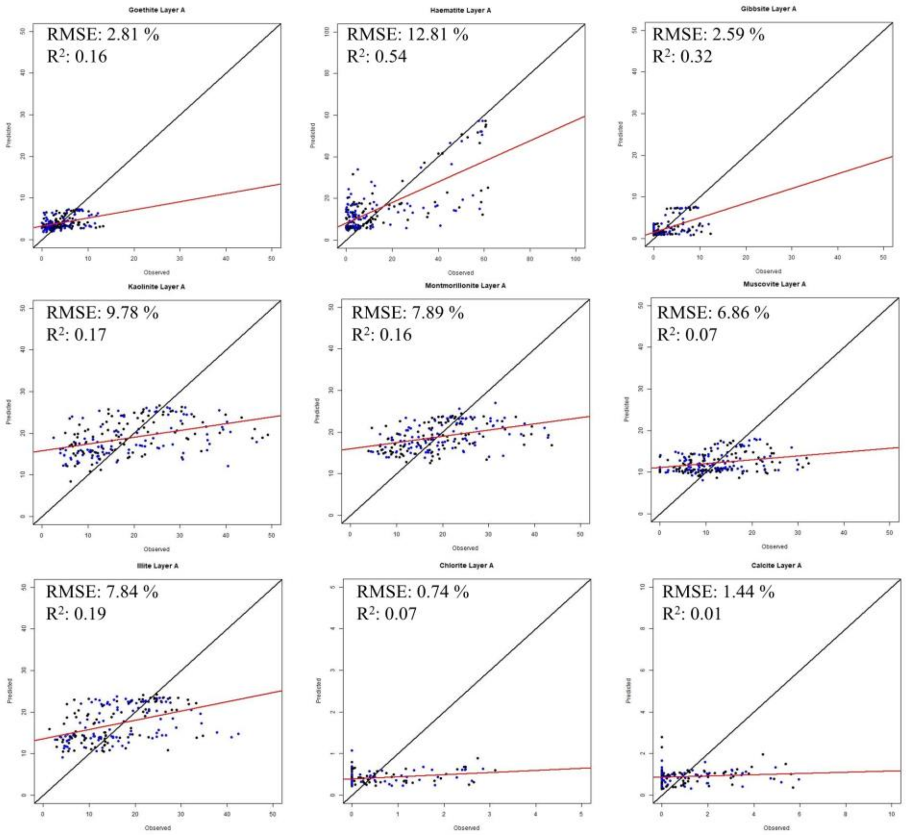


Fig. S1. Graphs of predicted and observed values (wt.%). of the soil minerals at 0 – 20 cm depth in the digital mineral mapping.

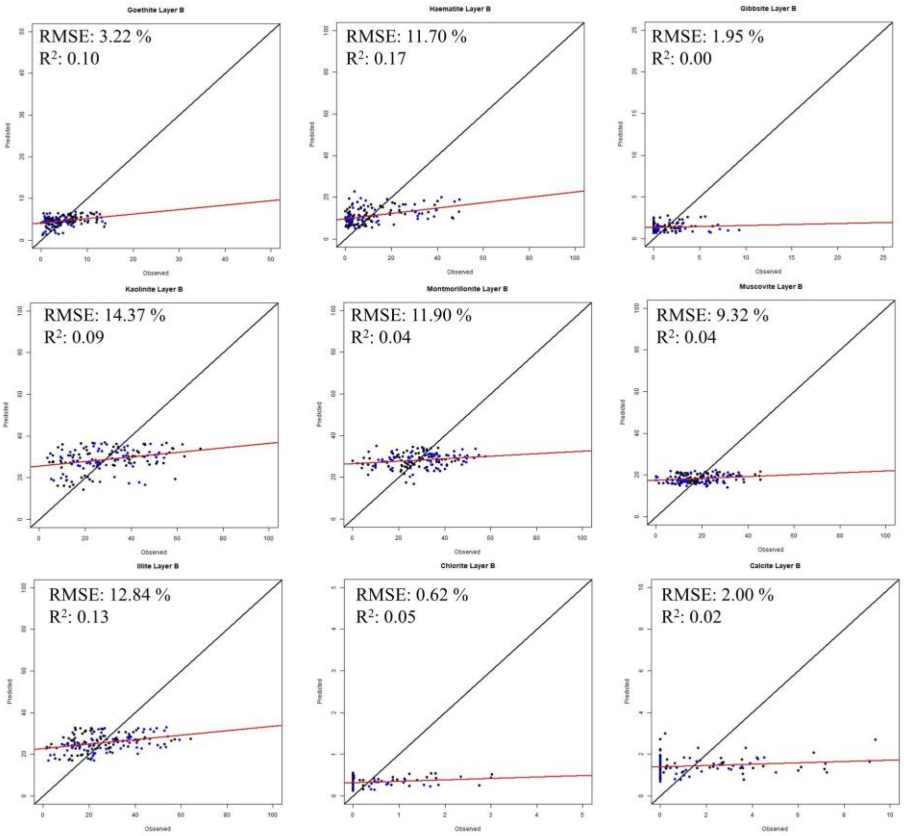


Fig. S2. Graphs of predicted and observed values (wt.%). of the soil minerals at 40 – 60 cm depth in the digital mineral mapping.

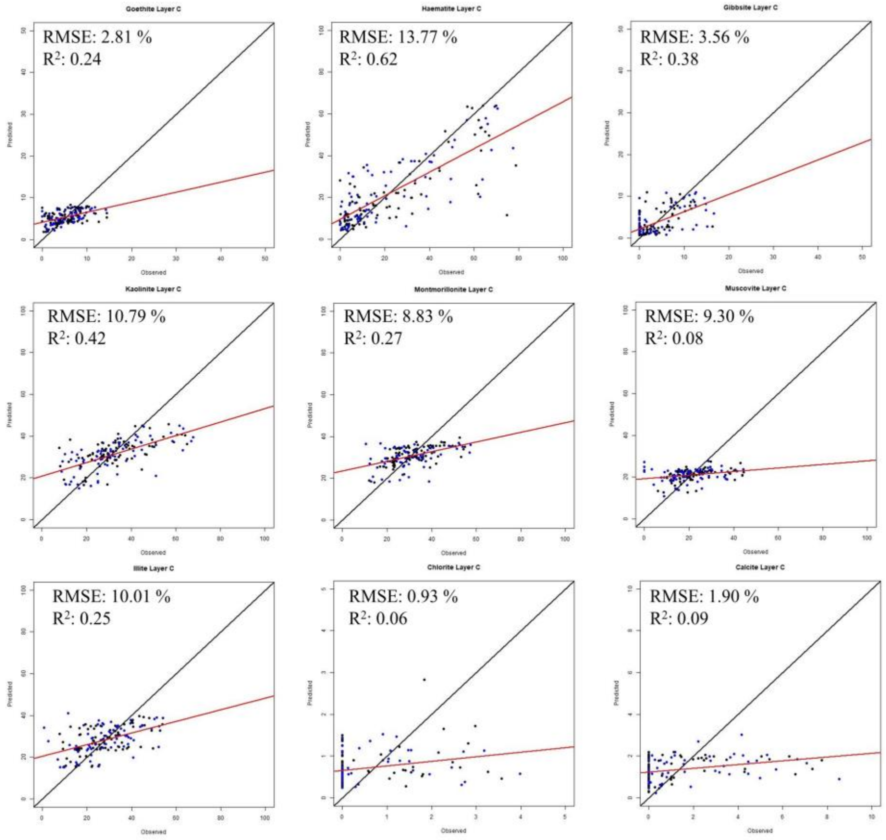


Fig. S3. Graphs of predicted and observed values (wt.%) of the soil minerals at 80 – 100 cm depth in the digital mineral mapping.

REFERENCES

Barron, V., & Torrent, J. (1986). Use of the Kubelka-Munk theory to study the influence of iron oxides on soil colour. *Journal of Soil Science*, *37*(4), 499–510. https://doi.org/10.1111/j.1365-2389.1986.tb00382.x

Ben-Dor, E., & Banin, A. (1995). Quantitative analysis of convolved Thematic Mapper spectra of soils in the visible near-infrared and shortwave-infrared spectral regions (0·4–2·5 μm). *International Journal of Remote Sensing*, *16*(18), 3509–3528. https://doi.org/10.1080/01431169508954643

Ben-Dor, E., Levin, N., Singer, A., Karnieli, A., Braun, O., & Kidron, G. J. (2006). Quantitative mapping of the soil rubification process on sand dunes using an airborne hyperspectral sensor. *Geoderma*, *131*(1–2), 1–21. https://doi.org/10.1016/j.geoderma.2005.02.011

Bigham, J. M., Golden, D. C., Bowen, L. H., Buol, S. W., & Weed, S. B. (1978). Iron Oxide Mineralogy of Well-drained Ultisols and Oxisols: I. Characterization of Iron Oxides in Soil Clays by Mössbauer Spectroscopy, X-ray Diffractometry, and Selected Chemical Techniques. *Soil Science Society of America Journal*, *42*(5), 816–825. https://doi.org/10.2136/sssaj1978.03615995004200050033x

Bish, D. L., & Plötze, M. (2011). X-ray Powder Diffraction with Emphasis on Qualitative and Quantitative Analysis in Industrial Mineralogy. In *Advances in the characterization of industrial minerals* (Vol. 9, pp. 35–76). European Mineralogical Union. https://doi.org/10.1180/EMU-notes.9.3

Breiman, L. (2001). *Random Forests* (Vol. 45). Retrieved from https://link.springer.com/content/pdf/10.1023/A:1010933404324.pdf

Brown, D. J., Shepherd, K. D., Walsh, M. G., Dewayne Mays, M., & Reinsch, T. G. (2006). Global soil characterization with VNIR diffuse reflectance spectroscopy. *Geoderma*, *132*(3–4), 273–290. https://doi.org/10.1016/j.geoderma.2005.04.025

Buol, S. W., Southard, R. J., Graham, R. C., & McDaniel, P. A. (2011). *Soil genesis and classification* (6th ed.). John Wiley & Sons, Ltd.

Camargo, L. A., Marques, J., & Pereira, G. T. (2013). Mineralogia da fração argila de um Argissolo em curvaturas do relevo. III - Variabilidade espacial. *Revista Brasileira de Ciencia Do Solo*, *37*(2), 295–306. https://doi.org/10.1590/S0100-06832013000200001

Camargo, O. A., Moniz, A. C., Jorge, J. A., & Valadares, J. M. A. S. (2009). *Métodos de Análise Química, Mineralógica e Física de Solos do Instituto Agronômico de Campinas*. Campinas.

Carvalho, A. M. G., Nunes, R. S., & Coelho, A. A. (2017). X-ray powder diffraction of high-Absorption materials at the XRD1 beamline off the best conditions: Application to (Gd, Nd)5Si4 compounds. *Powder Diffraction*, *32*(1), 10–14. https://doi.org/10.1017/S0885715616000646

Churchman, G. J. (2010). Is the geological concept of clay minerals appropriate for soil science? A literature-based and philosophical analysis. *Physics and Chemistry of the Earth*, *35*(15–18), 927–940. https://doi.org/10.1016/j.pce.2010.05.009

Churchman, G. J., & Lowe, D. J. (2012). Alteration, formation, and occurrence of minerals in soils. In P. Huang, Y. Li, & M. Sumner (Eds.), *Handbook of Soil Sciences Properties and Processes, Volume 1 - Properties and processes* (2nd ed., pp. 1–72). Boca Raton: CRC Press.

Coyne, L. M., Bishop, J. L., Scattergood, T., Banin, A., Carle, G., & Orenberg, J. (1990). Quantifying Iron and Surface Water in a Series of Variably Cation-Exchanged Montmorillonite Clays. In *Near-Infrared Correlation Spectroscopy* (pp. 407–429). American Chemical Society. https://doi.org/10.1021/bk-1990-0415.ch021

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random Forests for Classification in Ecology. *Ecology*, *88*(11), 2783–2792. https://doi.org/10.1890/07-0539.1

Demattê, José A. M., Safanelli, J. L., Poppiel, R. R., Rizzo, R., Silvero, N. E. Q., Mendes, W. de S., … Lisboa, C. J. da S. (2020). Bare Earth's Surface Spectra as a Proxy for Soil Resource Monitoring. *Scientific Reports*, *10*(1), 4461. https://doi.org/10.1038/s41598-020-61408-1

Demattê, José A M, Araújo, S. R., Fiorio, P. R., Fongaro, C. T., & Nanni, M. R. (2015). Espectroscopia VIS-NIR-SWIR na avaliação de solos ao longo de uma topossequência em Piracicaba (SP). *Revista Ciencia Agronomica*, *46*(4), 679–688. https://doi.org/10.5935/1806-6690.20150054

Demattê, José Alexandre Melo, Fongaro, C. T., Rizzo, R., & Safanelli, J. L. (2018). Geospatial Soil Sensing System (GEOS3): A powerful data mining procedure to retrieve soil spectral reflectance from satellite images. *Remote Sensing of Environment*, *212*, 161–175. https://doi.org/10.1016/j.rse.2018.04.047

Dokuchaev, V. V. (1883). Russian chernozem. Selected works of V.V. Dokuchaev. *Israel Program for Scientific Translations, Vol. I*, Translated in 1967.

Dotto, A. C., Dalmolin, R. S. D., Caten, A. ten, Gris, D. J., & Ruiz, L. F. C. (2019). AlradSpectra: a Quantification Tool for Soil Properties Using Spectroscopic Data in R. *Revista Brasileira de Ciência Do Solo*, *43*. https://doi.org/10.1590/18069657rbcs20180263

Dufréchou, G., Grandjean, G., & Bourguignon, A. (2015). Geometrical analysis of laboratory soil spectra in the short-wave infrared domain: Clay composition and estimation of the swelling potential. *Geoderma*. https://doi.org/10.1016/j.geoderma.2014.12.014

Fang, Q., Hong, H., Zhao, L., Kukolich, S., Yin, K., & Wang, C. (2018). Visible and Near-Infrared Reflectance Spectroscopy for Investigating Soil Mineralogy: A Review. *Journal of Spectroscopy*, *2018*, 1–14. https://doi.org/10.1155/2018/3168974

Fernandes, R. B. A., Barrón, V., Torrent, J., & Fontes, M. P. F. (2004). Quantificação de óxidos de ferro de latossolos brasileiros por espectroscopia de refletância difusa. *Revista Brasileira de Ciencia Do Solo*, *28*(2), 245–257. https://doi.org/10.1590/s0100-06832004000200003

Gallo, B. C., Demattê, J. A. M., Rizzo, R., Safanelli, J. L., Mendes, W. de S., Lepsch, I. F., … Lacerda, M. P. C. (2018). Multi-temporal satellite images on topsoil attribute quantification and the relationship with soil classes and geology. *Remote Sensing*, *10*(10), 1571. https://doi.org/10.3390/rs10101571

Grunwald, S., Yu, C., & Xiong, X. (2018). Transferability and Scalability of Soil Total Carbon Prediction Models in Florida, USA. *Pedosphere*, *28*(6), 856–872. https://doi.org/10.1016/S1002-0160(18)60048-7

Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B. M., & Gräler, B. (2018). Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, (8), e5518. https://doi.org/10.7717/peerj.5518

Jackson, M. L. (1958). *Soil Chemical Analysis*. Englewood Cliffs, N.J: Prentice-Hall.

Jenny, H. (1941). *Factors of soil formation : a system of quantitative pedology*. New York: McGraw-Hill.

Jepson, W. B. (1988). Structural Iron in Kaolinites and in Associated Ancillary Minerals. In *Iron in Soils and Clay Minerals* (pp. 467–536). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-009-4007-9_15

Kämpf, N., Marques, J. J., & Curi, N. (2012). Mineralogia de Solos Brasileiros. In J. C. Ker, N. Curi, C. E. G. R. Schaefer, & P. Vidal-Torrado (Eds.), *Pedologia - Fundamentos* (1ª, p. 343). Viçosa-MG: Sociedade Brasileira de Ciência do Solo.

Kokaly, R. F., Clark, R. N., Swayze, G. A., Livo, K. E., Hoefen, T. M., Pearson, N. C., … Klein, A. J. (2017). *USGS Spectral Library Version 7: U.S. Geological Survey Data Series 1035*. https://doi.org/10.3133/ds1035

Kosmas, C. S., Curi, N., Bryant, R. B., & Franzmeier, D. P. (1984). Characterization of Iron Oxide Minerals by Second-Derivative Visible Spectroscopy. *Soil Science Society of America Journal*, *48*(2), 401–405. https://doi.org/10.2136/sssaj1984.03615995004800020036x

Ma, Y., Minasny, B., Malone, B. P., & Mcbratney, A. B. (2019). Pedology and digital soil mapping (DSM). *European Journal of Soil Science*, *70*(2), 216–235. https://doi.org/10.1111/ejss.12790

Madeira, J., Bedidi, A., Cervelle, B., Pouget, M., & Flay, N. (1997). Visible spectrometric indices of hematite (Hm) and goethite (Gt) content in lateritic soils: The application of a Thematic Mapper (TM) image for soil-mapping in Brasilia, Brazil. *International Journal of Remote Sensing*, *18*(13), 2835–2852. https://doi.org/10.1080/014311697217369

Malone, B. P., McBratney, A. B., & Minasny, B. (2018). Description and spatial inference of soil drainage using matrix soil colours in the Lower Hunter Valley, New South Wales, Australia. *PeerJ*, *6*, e4659. https://doi.org/10.7717/peerj.4659

Mathian, M., Hebert, B., Baron, F., Petit, S., Lescuyer, J. L., Furic, R., & Beaufort, D. (2018). Identifying the phyllosilicate minerals of hypogene ore deposits in lateritic saprolites using the near-IR spectroscopy second derivative methodology. *Journal of Geochemical Exploration*. https://doi.org/10.1016/j.gexplo.2017.11.019

McBratney, A. . B., Mendonça Santos, M. . L., & Minasny, B. (2003). On digital soil mapping. *Geoderma*, *117*(1–2), 3–52. https://doi.org/10.1016/S0016-7061(03)00223-4

Melo, V. F., Corrêa, G. F., Maschio, P. A., Ribeiro, A. N., & Lima, V. C. (2003). Importância das espécies minerais no potássio total da fração argila de solos do Triângulo Mineiro. *Revista Brasileira de Ciência Do Solo*, *27*(5), 09–10. https://doi.org/10.1590/s0100-06832003000500005

Mendes, W. de S., Medeiros Neto, L. G., Demattê, J. A. M., Gallo, B. C., Rizzo, R., Safanelli, J. L., & Fongaro, C. T. (2019). Is it possible to map subsurface soil attributes by satellite spectral transfer models? *Geoderma*, *343*, 269–279. https://doi.org/10.1016/j.geoderma.2019.01.025

Omran, E. S. E. (2017). Rapid prediction of soil mineralogy using imaging spectroscopy. *Eurasian Soil Science*, *50*(5), 597–612. https://doi.org/10.1134/S106422931705012X

R Development Core Team, R. R: A Language and Environment for Statistical Computing (2020). Retrieved from https://www.r-project.org

Savitzky, A., & Golay, M. J. E. (1964). Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry*, *36*(8), 1627–1639. https://doi.org/10.1021/ac60214a047

Schaetzl, R. J., & Anderson, S. (2005). *Soils genesis and geomorphology*. New York, NY: Cambridge University Press.

Scheinost, A. C., Chavernas, A., Barrón, V., & Torrent, J. (1998). Use and limitations of second-derivative diffuse reflectance spectroscopy in the visible to near-infrared range to identify and quantify Fe oxide minerals in soils. *Clays and Clay Minerals*, *46*(5), 528–536. https://doi.org/10.1346/CCMN.1998.0460506

Schwertmann, U., & Herbillon, A. J. (2015). Some Aspects of Fertility Associated with the Mineralogy of Highly Weathered Tropical Soils. In *Myths and science of soils of the tropics* (pp. 47–59). https://doi.org/10.2136/sssaspecpub29.c4

Siivola, J., & Schmid, R. (2007). List of Mineral abbreviations. *IUGS Subcommission on the Systematics of Metamorphic Rocks*, 1–14.

Silva, L. S., Marques Júnior, J., Barrón, V., Gomes, R. P., Teixeira, D. D. B., Siqueira, D. S., & Vasconcelos, V. (2020). Spatial variability of iron oxides in soils from Brazilian sandstone and basalt. *Catena*, *185*, 104258. https://doi.org/10.1016/j.catena.2019.104258

Terra, F. S., Demattê, J. A. M., & Viscarra Rossel, R. A. (2015). Spectral libraries for quantitative analyses of tropical Brazilian soils: Comparing vis–NIR and mid-IR reflectance data. *Geoderma*, *255–256*, 81–93. https://doi.org/10.1016/j.geoderma.2015.04.017

Terra, F. S., Demattê, J. A. M., & Viscarra Rossel, R. A. (2018). Proximal spectral sensing in pedological assessments: vis–NIR spectra for soil classification based on weathering and pedogenesis. *Geoderma*, *318*, 123–136. https://doi.org/10.1016/j.geoderma.2017.10.053

Torrent, J., & Barron, V. (2015). Diffuse reflectance spectroscopy of iron oxides. In P. Somasundaran (Ed.), *Encyclopedia of Surface and Colloid Science* (3rd Editio, pp. 1731–1739). CRC Press. https://doi.org/10.1081/E-ESCS3-120000047

US EPA. (2007). Test Methods for Evaluating Solid Waste, Physical/Chemical Methods. *EPA Publication*, *IV*(February), 1–30. https://doi.org/10.1017/CBO9781107415324.004

USGS. (2018). USGS EROS Archive - Digital Elevation - Shuttle Radar Topography Mission (SRTM) 1 Arc-Second Global. https://doi.org/10.5066/F7PR7TFT

Vidal-Torrado, P., & Lepsch, I. F. (1999). Relações material de origem / solo e pedogênese em uma seqüência de solos predominantemente argilosos e Latossólicos sobre psamitos na depressão periférica Paulista: Paulo State Peripheral Depression, southeastern Brazil. *Revista Brasileira de Ciência Do Solo*, *23*(2), 357–369. https://doi.org/10.1590/S0100-06831999000200019

Viscarra Rossel, R. A. (2011). Fine-resolution multiscale mapping of clay minerals in Australian soils measured with near infrared spectra. *Journal of Geophysical Research: Earth Surface*, *116*(4). https://doi.org/10.1029/2011JF001977

Viscarra Rossel, R. A., Behrens, T., Ben-Dor, E., Brown, D. J., Demattê, J. A. M., Shepherd, K. D., … Ji, W. (2016). A global spectral library to characterize the world's soil. *Earth-Science Reviews*, *155*(July), 198–230. https://doi.org/10.1016/j.earscirev.2016.01.012

Weaver, C. E. (Charles E., & Pollard, L. D. (Lin D. (1973). *The chemistry of clay minerals*. Elsevier Scientific Pub. Co.

Winters, E., & Simonson, R. W. (1951). The Subsoil. In *Advances in Agronomy* (Vol. 3, pp. 1–92). Academic Press. https://doi.org/10.1016/S0065-2113(08)60366-1

Xu, J., PENG, B., Yu, C., Yang, G., Tang, X., Tan, C., … Xiao, M. (2013). Geochemistry of soils derived from black shales in the Ganziping mine area, western Hunan, China. *Environmental Earth Sciences*, *70*(1), 175–190. https://doi.org/10.1007/s12665-012-2114-0

Yokozeki, K., Watanabe, K., Sakata, N., & Otsuki, N. (2004). Modeling of leaching from cementitious materials used in underground environment. *Applied Clay Science*, *26*(1–4), 293–308. https://doi.org/10.1016/j.clay.2003.12.027

Zhang, Z. Y., Huang, L., Liu, F., Wang, M. K., Fu, Q. L., & Zhu, J. (2016). Characteristics of clay minerals in soil particles of two Alfisols in China. *Applied Clay Science*, *120*, 51–60. https://doi.org/10.1016/j.clay.2015.11.018

Zhao, L., Hong, H., Liu, J., Fang, Q., Yao, Y., Tan, W., … Algeo, T. J. (2018). Assessing the utility of visible-to-shortwave infrared reflectance spectroscopy for analysis of soil weathering intensity and paleoclimate reconstruction. *Palaeogeography, Palaeoclimatology, Palaeoecology*, *512*, 80–94. https://doi.org/10.1016/j.palaeo.2017.07.007

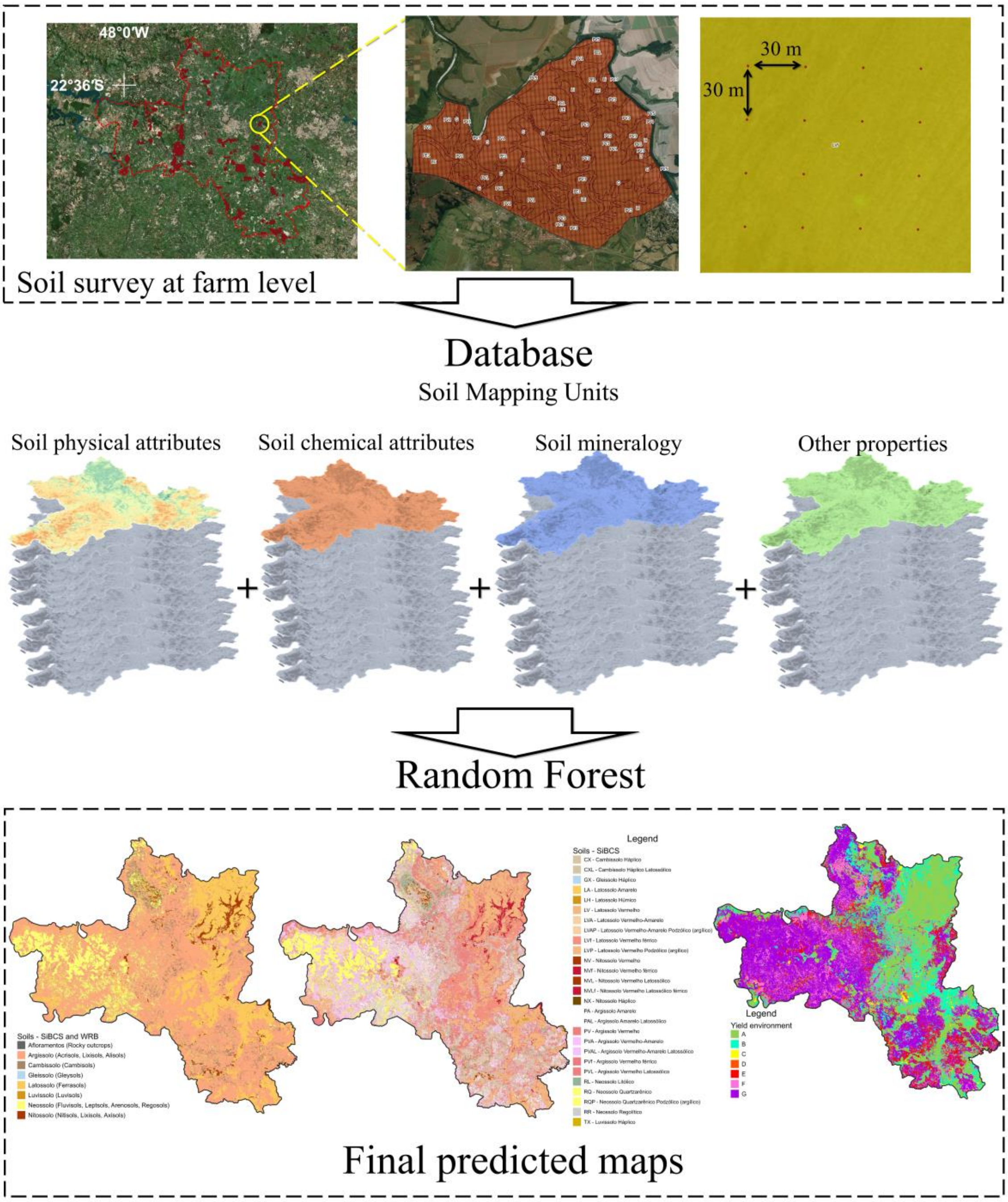
4. APPLYING THE DIGITAL SOIL MAPPING PRODUCTS TO PREDICT THE SOIL TYPES IN BRAZIL

ABSTRACT

The civilisation lives in a world of maps and soil maps are vital at regional and farm levels to achieve best management agricultural practices. Soil is the substrate for plant growth and vital to the fulfilment of the food demand. However, the cartographic scale of those soil maps, which for the best management agricultural practice (BMAP) have to be the most detailed as possible are scarce. Therefore, the objectives of this research were to (i) present the potentiality of using the digital soil mapping (DSM) products such as soil chemical, physical, indices, mineralogy, and properties to extrapolate former soil survey maps at 1:20000 scale; (ii) create the digital yield environment for sugarcane based on the DSM products; and (iii) evaluate qualitatively the predict soil maps and relationship with former research and the predicted yield environment. The region of interest (ROI) covers eight cities and almost 2,598 km^2 in the São Paulo state, Brazil. The soil survey at farm level conducted covered almost 86.52 km^2, which is ~3.33% of the total area (96.67% of the unmapped area). Thus, we created a point grid (centroid) with the same spatial resolution (30 m) of the rasters used as covariates for soil mapping unit (SMU) predictions. Such grid intended to retrieve the representative soil mapping unit of each geometric polygon. It was retrieved 117,413 points representing twenty-seven soil mapping units of seven soil orders at a first categorical level according to the Brazilian Classification System and seven yield environment for sugarcane production. The prediction of the SMUs and their respectively soil orders were performed using the random forest machine learning regression method. The level of association between the SMUs and yield environments was 0.34 ($p<0.01$) by the Cramer's V coefficient displaying a very strong relationship. The digital yield environment for sugarcane based on the DSM products was created and an qualitatively evaluation of the predict soil maps and relationship with former research showed that our findings and framework could attend the need for soil maps at regional and farm levels to achieve best management agricultural practices.

Keywords: pedometrics; sugarcane yield; digital soil mapping; remote sensing.

Graphical Abstract


48°0'W
22°36'S
30 m
30 m
Soil survey at farm level
Database
Soil Mapping Units
Soil physical attributes
Soil chemical attributes
Soil mineralogy
Other properties
+
+
+
Random Forest
Soils - SiBCS and WRB
Afloramentos (Rocky outcrops)
Argissolo (Acrisols, Lixisols, Alisols)
Cambissolo (Cambisols)
Gleissolo (Gleysols)
Latossolo (Ferrasols)
Luvissolo (Luvisols)
Neossolo (Fluvisols, Leptsols, Arenosols, Regosols)
Nitossolo (Nitisols, Lixisols, Axisols)
Legend
Soils - SiBCS
CX - Cambissolo Háplico
CXL - Cambissolo Háplico Latossólico
GX - Gleissolo Háplico
LA - Latossolo Amarelo
LH - Latossolo Húmico
LV - Latossolo Vermelho
LVA - Latossolo Vermelho-Amarelo
LVAP - Latossolo Vermelho-Amarelo Podzólico (argílico)
LVf - Latossolo Vermelho férrico
LVP - Latossolo Vermelho Podzólico (argílico)
NV - Nitossolo Vermelho
NVf - Nitossolo Vermelho férrico
NVL - Nitossolo Vermelho Latossólico
NVLf - Nitossolo Vermelho Latossólico férrico
NX - Nitossolo Háplico
PA - Argissolo Amarelo
PAL - Argissolo Amarelo Latossólico
PV - Argissolo Vermelho
PVA - Argissolo Vermelho-Amarelo
PVAL - Argissolo Vermelho-Amarelo Latossólico
PVf - Argissolo Vermelho férrico
PVL - Argissolo Vermelho Latossólico
RL - Neossolo Litólico
RQ - Neossolo Quartzarênico
RQP - Neossolo Quartzarênico Podzólico (argílico)
RR - Neossolo Regolítico
TX - Luvissolo Háplico
Legend
Yield environment
A
B
C
D
E
F
G
Final predicted maps

4.1. INTRODUCTION

The soil is known in Soil Science as part of the landscape and the self-organised complex natural system wherein all abiotic and biotic processes take place, which it is also the substrate for the forest, crop, and pasture growth. Acting so, the fulfilment of the food demand cannot be achieved without better knowledge of soil formation (e.g. inputs, translocation/movement, transformation/change, and outputs), genesis, and spatial variability. As stated by Jenny (1941), the five main interacting factors that affect soil formation are climate, organisms (e.g. including human activities), relief, parent material, and time. These soil formation factors are responsible for the soil genesis or pedogenesis (Buol et al., 2011). Soil genesis is the roundabout that connects soil mineralogy, chemistry, physics, climatology, geology, anthropology, geography, biology, and agriculture aiming the soil quality (Norfleet et al., 2003). The pedogenesis encompasses specific soil-forming processes such as gleization, podsolization, lateralization, plinthization, carbonization, salinization, sodification, turbation, and paludization. Understanding these processes enable to classify soil and determine which are the limitation and advantage for plant growth and productivity.

The soil classification system and maps are the final step of a soil survey grouping soils by similar attributes and/or properties, which it makes more accessible to policy-makers, farmers, and scientific community. Each country around the world or most of them has its classification system. However, the main challenge is to make a unified classification system or a more comprehensive/simplified soil classification. Another point is the cartographic scale of those soil maps, which for the best management agricultural practice (BMAP) have to be the most detailed as possible. Hartemink et al. (2013), reviewing this topic, pointed out that our civilisation lives in a world of maps and soil maps are vital at regional and farm levels to achieve BMAP. For instance, according to Embrapa (2020) and Polidoro et al. (2017), up to 5% (~425,000 out of 8.5 million km^2) of the Brazilian territory has soil maps at ≥ 1:100,000 scale while the United States of America has mapped almost their entire territory at 1:20,000 – 1:40,000 scale.

The soil maps at a detailed scale can be used to determine capability groups and/or yield environment for agricultural purposes. The capability groupings were conceived by Klingebiel (1958), and associated soil classes into eight capability classes, which described their potentiality for plant growth and soil conservation management. Acting similar to this approach, it was created the yield environment for sugarcane production in Brazil by Demattê and Demattê (2009) associating eight yield environment with soil mapping units, evapotranspiration, and sugarcane tons per hectare (STH). Each of these classes is related to specific estimative of sugarcane production. Those two approaches showed how the detailed soil map is vital to provide valuable information for agriculture. However, none of that valuable information has been available on a large scale for policy-makers, farmers, and the scientific community. A feasible, fast, and recent framework and discipline in Soil Science that can help to change those scenarios (e.g. scarce soil map and capability groupings/yield environment at detailed scale) is the digital soil mapping.

The Digital soil Mapping (DSM) framework was stated by McBratney et al. (2003), and consists of using quantitative models (e.g. stochastic, deterministic, and hybrid methods) from field soil georeferenced samples interrelated with landscape variables (e.g. remote sensing data) to predict soil attributes, properties, and classes considering the soil formation factors established by Jenny (1941). A couple of reviews in this topic has been done since then (Grunwald, 2010; Hartemink et al., 2020; Ma et al., 2019; Mendonça-Santos et al., 2006; Odgers et al., 2011; Rossiter, 2018; Sanchez et al., 2009). Most of the studies in digital soil mapping predict the soil attributes (Chen et al., 2020; Gallo et al., 2018; Li et al., 2018; Padarian et al., 2017), properties (César de Mello et al., 2020; Dharumarajan et al., 2020; Odeha et al., 1994; Poppiel et al., 2020; Steinberg et al., 2016) or classes (Debella-Gilo and Etzelmüller, 2009; Demattê et al., 2017; Poppiel et al., 2019a, 2019b; Wolski et al., 2017; Zeng et al., 2017) using as predictors remote sensing data, which retrieve the main landscape patterns. Another example of soil class predictions is the

Disaggregation and Harmonization of Soil Map Units Through Resampled Classification Trees (DSMART) algorithm enables to remap conventional soil maps at detailed scale as described in Odgers et al. (2014) who developed the DSMART. However, it is rarely seen in DSM the soil attributes and/or properties as predictors to model soil classes. This question rise because if the final predicted attributes or properties play in the same way that the traditional laboratory analyses to determine soil classes punctually, why not use those predict soil attributes and properties as predictors to model the soil types. Exploiting that it will enhance the potentiality of DSM which consists in spatialize soil data easily, detailed, and low-costs.

Therefore, the objectives of this research were to (i) present the potentiality of using the DSM products such as soil chemical, physical, indices, mineralogy, and properties to extrapolate former soil survey maps at 1:20000 scale; (ii) create the digital yield environment for sugarcane based on the DSM products; and (iii) evaluate qualitatively the predict soil maps and relationship with former research and the predicted yield environment.

4.2. MATERIAL AND METHODS

4.2.1. Description of the region of interest (ROI)

The study area covers eight cities and almost 2,598 square kilometres in the São Paulo state, Brazil (Fig. 1a). We chose the limits of such area based on the regional administrative criteria (IBGE, 2019), agricultural impacts (Machado et al., 2017), and importance (Gallo et al., 2018). The economic activity is predominantly sugarcane with no-till and till farming along the year (Rudorff et al., 2010). The total sugarcane cultivated area for the crop year 2013/2014 was 1,221.04 square kilometres (Fig. 1b), which is approximately 47% of the total study area.

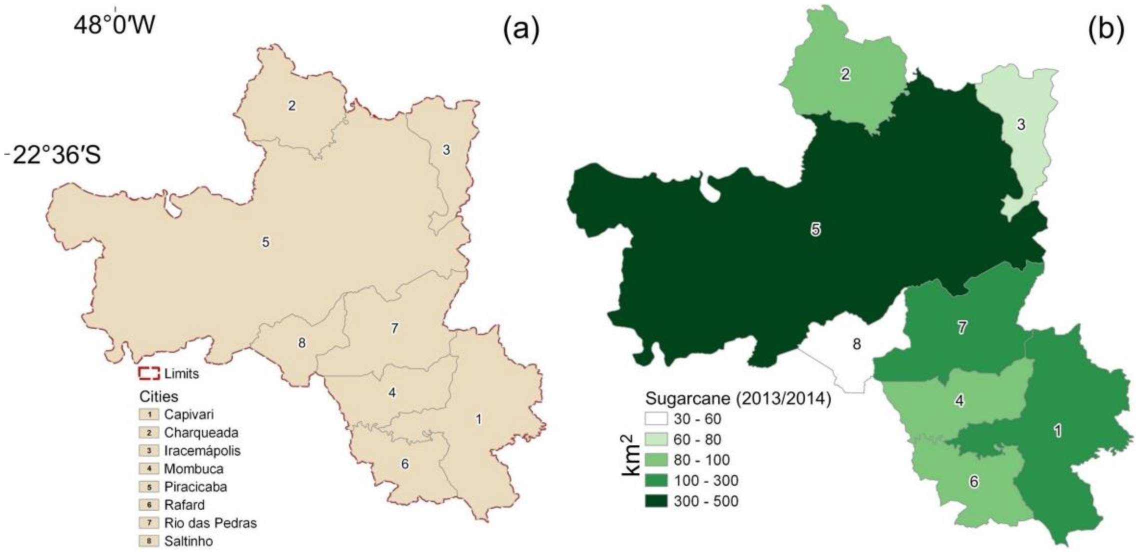


Fig. 1. Study area (a) and sugarcane cultivated area by municipalities (b).

Dry winters and rainy summers are the defined weather seasons with annual average rainfall and temperature ranging from 1200 to 1400 mm and 20 to 23ºC, respectively. The area covers 500 m in elevation (450 – 950 m) with gentle slopes, undulating hills, and rolling uplands, and has parent materials such as siltstones, tillites, varvites, conglomerates, sandstones, shales, limestones, dolomite, flint, diabase, and basalt.

4.2.2. Data

The soil survey at farm level conducted in the ROI covered almost 86.52 km^2, which is ~3.33% of the total area. This means there were around 96.67% of the unmapped area at the same map scale. Thus, we created a point grid (centroid) with the same spatial resolution (30 m) of the rasters used as covariates for soil mapping unit (SMU) predictions (Fig. 2). Some polygons had more points inside than others. Such grid intended to retrieve the representative soil mapping unit of each geometric polygon. It was retrieved 117,413 points representing twenty-seven soil mapping units of seven soil orders at a first categorical level according to the Brazilian Classification System (Table 1) and seven yield environment (Demattê and Demattê, 2009). The latter is defined by the joint of two or more soil mapping units with equal yield capacity based on the soil-climate-plant interaction and local characteristics. This is similar to the capability groupings described in Klingebiel (1958) and adopted by the Natural Resources Conservation Service of the United States Department of Agriculture (USDA-NRCS).

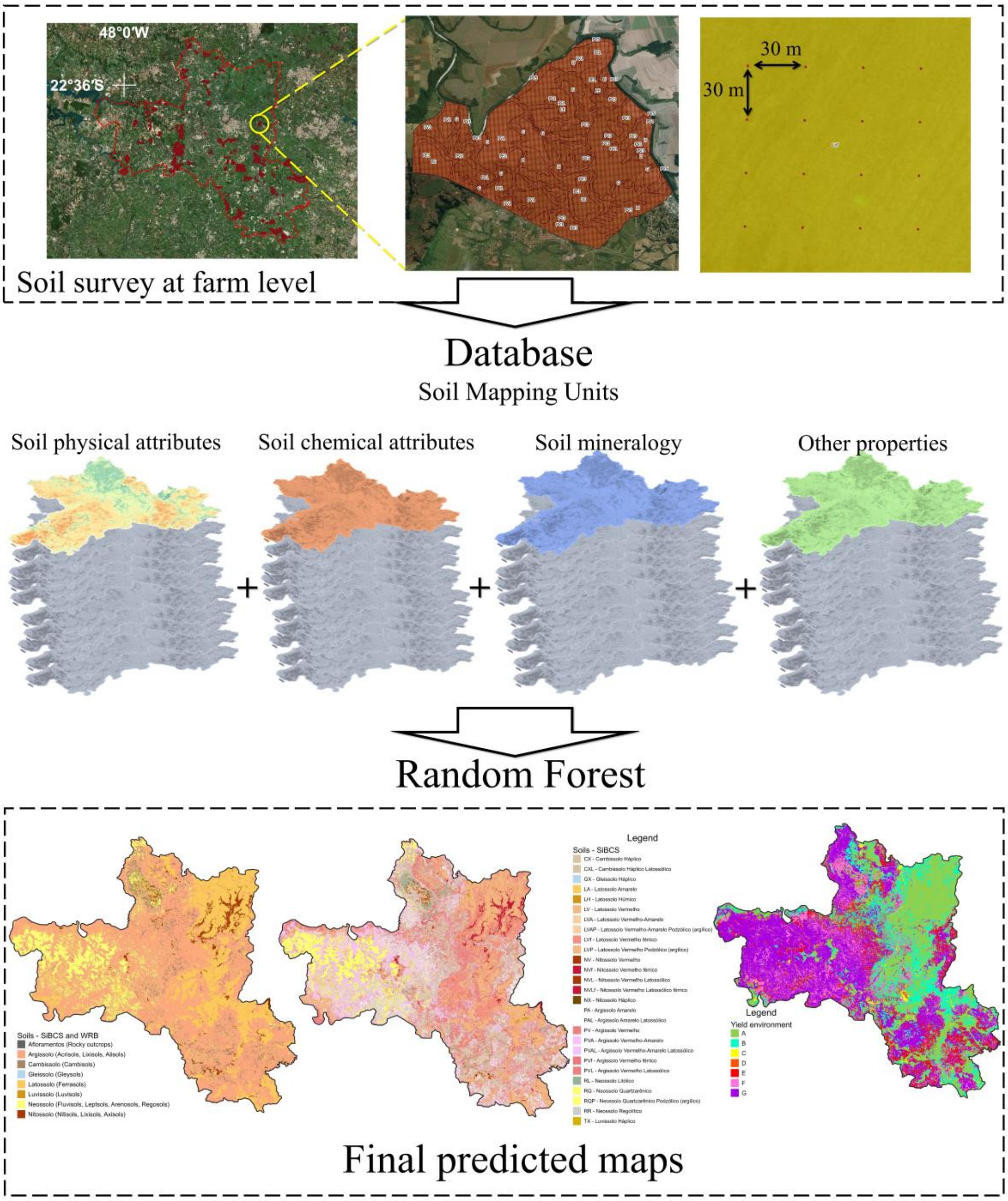

48°0'W
22°36'S
30 m
30 m
Soil survey at farm level
Database
Soil Mapping Units
Soil physical attributes
Soil chemical attributes
Soil mineralogy
Other properties
+
+
+
Random Forest
Soils - SiBCS and WRB
Afloramentos (Rocky outcrops)
Argissolo (Acrisols, Lixisols, Alisols)
Cambissolo (Cambisols)
Gleissolo (Gleysols)
Latossolo (Ferrasols)
Luvissolo (Luvisols)
Neossolo (Fluvisols, Leptsols, Arenosols, Regosols)
Nitossolo (Nitisols, Lixisols, Axisols)
Legend
Soils - SiBCS
CX - Cambissolo Háplico
CXL - Cambissolo Háplico Latossólico
GX - Gleissolo Háplico
LA - Latossolo Amarelo
LH - Latossolo Húmico
LV - Latossolo Vermelho
LVA - Latossolo Vermelho-Amarelo
LVAP - Latossolo Vermelho-Amarelo Podzólico (argílico)
LVf - Latossolo Vermelho férrico
LVP - Latossolo Vermelho Podzólico (argílico)
NV - Nitossolo Vermelho
NVf - Nitossolo Vermelho férrico
NVL - Nitossolo Vermelho Latossólico
NVLf - Nitossolo Vermelho Latossólico férrico
NX - Nitossolo Háplico
PA - Argissolo Amarelo
PAL - Argissolo Amarelo Latossólico
PV - Argissolo Vermelho
PVA - Argissolo Vermelho-Amarelo
PVAL - Argissolo Vermelho-Amarelo Latossólico
PVf - Argissolo Vermelho férrico
PVL - Argissolo Vermelho Latossólico
RL - Neossolo Litólico
RQ - Neossolo Quartzarênico
RQP - Neossolo Quartzarênico Podzólico (argílico)
RR - Neossolo Regolítico
TX - Luvissolo Háplico
Legend
Yield environment
A
B
C
D
E
F
G
Final predicted maps

Fig. 2. Scheme of the methodology and results.

Table 1. Description and correspondence among mapping unit, the Brazilian Soil Classification System (SiBCS) (Santos et al., 2018), World Reference Base (WRB) (IUSS Working Group WRB, 2014), and Soil Taxonomy (Soil Survey Staff, 2014). Number of samples for each soil mapping unit (N).

Mapping unit	Soil group (1st Letter)	SiBCS	WRB	Soil taxonomy	Mapping unit (other letters)	Colour/Parent material/Other character	N
CX	C	Cambissolo	Cambisol	Udepts	X	Haplic	9713
CXL	C				XL	Haplic and Latosolic	23
GX	G	Gleissolo	Gleisol	Aqualf	X	Haplic	202
LA	L	Latossolo	Ferralsol	Udox	A	Yellow	1762
LH	L				H	Humic	652
LV	L				V	Red	27065
LVA	L				VA	Red-Yellow	944
LVAP	L				VAP	Red-Yellow and Argilic	16
LVf	L				Vf	Red and ferric	10707
LVP	L				VP	Red and Latosolic	547
NV	N	Nitossolo	Nitisol	Udalf, Udult	V	Red	1234
NVf	N				Vf	Red and ferric	3870
NVL	N				VL	Red and Latosolic	709
NVLf	N				VLf	Red, Latosolic and ferric	710
NX	N				X	Haplic	84
PA	P	Argissolo	Lixisol		A	Yellow	7607
PAL	P				AL	Yellow and Latosolic	40
PV	P				V	Red	15187
PVA	P				VA	Red-Yellow	12850
PVAL	P				VAL	Red-Yellow and Latosolic	230
PVf	P				Vf	Red and ferric	97
PVL	P				VL	Red and Latosolic	3014
RL	R	Neossolo	Leptsol	Lithic Udorthent/Psamments	L	Lithic	2997
RQ	R		Arenosols	Quartzipsamment	Q	Quartzenic	10735
RQP	R		Arenosols	Quartzipsamment	QP	Quartzenic and Argilic	2317
RR	R		Regosols	Psamment, Orthent	R	Regolitic	3790
TX	T	Luvissolo	Luvisols	Alfisol, Aridisol	X	Haplic	311

The digital soil mapping framework uses a series of ancillary variables that represent quantitatively and spatially the soil formation factors as described in McBratney et al. (2003). In this sense, our ancillary variables were the soil chemical, physical, and mineralogical attributes plus other properties such as magnetic susceptibility and free iron (Fig. 2). The response variable, in this case, was the soil mapping units. The soil constituents were predicted in Chapters 1, 2, and 3, wherein further information about the methodology can be found, for the same ROI. It was selected the best-fitted models from those studies comprising nine soil minerals, three soil physical and six soil chemical attributes, three indices calculated from other rasters, and two soil properties (Table 2). At total, we had 64 predictors.

Table 2. Selected digital soil mapping products in the study area from Chapters 1; 2; and 3.

Soil Constituents		Abbrev.	Depth (cm)	RMSE	R^2	CCC
Mineralogy	Goethite (g kg^{-1})	Gt	0 – 20	2.81	0.16	0.31
			40 – 60	3.22	0.10	0.19
			80 – 100	2.81	0.24	0.39
	Hematite (g kg^{-1})	Hem	0 – 20	12.81	0.54	0.68
			40 – 60	11.70	0.17	0.23
			80 – 100	13.77	0.62	0.74
	Gibbsite (g kg^{-1})	Gbs	0 – 20	2.59	0.32	0.51
			40 – 60	1.95	0.00	0.04
			80 – 100	3.56	0.38	0.57
	Kaolinite (g kg^{-1})	Kln	0 – 20	9.78	0.17	0.28
			40 – 60	14.37	0.09	0.19
			80 – 100	10.79	0.42	0.51
	Chlorite (g kg^{-1})	Chl	0 – 20	0.74	0.07	0.10
			40 – 60	0.62	0.05	0.07
			80 – 100	0.93	0.06	0.18
	Calcite (g kg^{-1})	Cal	0 – 20	1.44	0.01	0.05
			40 – 60	2.00	0.02	0.06
			80 – 100	1.90	0.09	0.17
	Illite (g kg^{-1})	Ill	0 – 20	7.84	0.19	0.35
			40 – 60	12.84	0.13	0.20
			80 – 100	10.01	0.25	0.42
	Muscovite (g kg^{-1})	Ms	0 – 20	6.86	0.07	0.16
			40 – 60	9.32	0.04	0.08
			80 – 100	9.30	0.08	0.16
	Montmorillonite (g kg^{-1})	Mnt	0 – 20	7.98	0.16	0.26
			40 – 60	11.90	0.04	0.11
			80 – 100	8.83	0.27	0.38
Physical	Clay (g kg^{-1})	-	0 – 20	119.33	0.58	0.72
			40 – 60	129.74	0.44	0.59
			80 – 100	118.12	0.63	0.76
	Sand (g kg^{-1})	-	0 – 20	177.58	0.46	0.62
			40 – 60	192.80	0.43	0.57
			80 – 100	178.55	0.46	0.62
	Soil Organic Matter (g kg^{-1})	SOM	0 – 20	7.77	0.35	0.48
			40 – 60	4.51	0.24	0.34
			80 – 100	3.97	0.46	0.63
Chemical	Soluble Al^{3+} (mmol$_c$ kg^{-1})	Al	0 – 20	10.17	0.08	0.21
			40 – 60	21.03	0.13	0.10
			80 – 100	28.16	0.20	0.22
	Cation Exchange Capacity (mmol$_c$ kg^{-1})	CEC	0 – 20	36.12	0.42	0.61
			40 – 60	58.74	0.30	0.41
			80 – 100	70.09	0.29	0.39
	Sum of Bases (mmol$_c$ kg^{-1})	SB	0 – 20	32.98	0.42	0.51
			40 – 60	46.02	0.37	0.37
			80 – 100	40.33	0.17	0.25
	Aluminium Saturation (%)	AS	0 – 20	18.64	0.13	0.10
			40 – 60	28.63	0.05	0.11
			80 – 100	24.72	0.21	0.30
	Base Saturation (%)	BS	0 – 20	19.19	0.12	0.21
			40 – 60	22.11	0.16	0.27
			80 – 100	21.79	0.09	0.16
	pH	-	0 – 20	0.55	0.08	0.11
			40 – 60	0.62	0.04	0.10
			80 – 100	0.57	0.10	0.17
Indices	Hematite/ (Hematite + Goethite)	Hem/ (Hem+Gt)	0 – 20	Hematite / (Hematite + Goethite)		
			40 – 60			
			80 – 100			
	Kaolinite/ (Kaolinite + Gibbsite)	Kln/ (Kln+Gbs)	0 – 20	Kaolinite / (Kaolinite + Gibbsite)		
			40 – 60			
			80 – 100			
	B textural horizon	Bt	Clay (40 – 60 cm)/ Clay (0 – 20 cm)			
Properties	Magnetic Susceptibility (10^{-8} m^3 kg^{-1})	χ	0 – 20	313.30	0.29	0.53
			80 – 100	212.80	0.66	0.65
	Free Iron (g kg^{-1})	FI	0 – 20	24.98	0.84	0.69*

*Coefficient Of Efficiency

4.2.3. Modelling and evaluation

Random Forest (RF) models are decision tree algorithms that deal with reducing the variance. The RF is widely and well-posed in the literature for digital soil mapping purposes (Gray et al., 2016; Møller et al., 2019;

Nussbaum et al., 2018). In this sense, it was applied to the RF machine learning regression method, which is categorised as an ensemble learning method. The RF split training samples into subsets and generates decision trees for each subset. Each new training subset used to build a decision tree, one third is randomly removed. This sample is called out-of-bag and the remaining samples (in-the-bag) are handled to build the decision tree. Out-of-bag samples are used to assess the model performance and select the training subset with higher accuracy. Thus, it was performed a grid search for optimal tuning using the "caret" R package (Kuhn, 2008) by 10-fold repeated cross-validation method carried out five times and the best accuracy and kappa selected (Table 3).

Table 3. Random forest parameters of calibration and internal validation.

	Number of Factors	Ntree	Mtry	Accuracy	Kappa
Soil Orders	7	500	33	0.86	0.81
Soil Mapping Units	27	500	33	0.84	0.82
Yield Environment	7	500	33	0.86	0.83
Soil Legacy (IAC 1989)	42	500	33	0.58	0.52

Number of trees (Ntree); Number of variables available for splitting at each tree node (Mtry).

The default metrics for multi-class classification used to initially assess model performances were the accuracy and kappa coefficient (Eq. 1). The latter was proposed by Cohen (1960), and is based on a confusion matrix that has the upper limit of 1 only when there is an entire agreement between the predicted and observed samples.

$$Kappa = \frac{P_o - P_e}{1 - P_e} \tag{1}$$

where P_o is the proportion of correctly classified sites, and P_e is the probability of random agreement.

The Cohen's kappa is more useful on multi-class classification wherein there is an imbalance in the classes. The accuracy coefficient is more likely to access a binary classification rather than various classes classification. Another important result of the RF is the mtry which makes explicit how many predictors is needed to outperform the predictions. For example, in our study, 33 out of 64 predictors were enough to outperform the soil mapping unit predictions. Furthermore, we analysed one former study of five soil profiles from 1999 (Vidal-Torrado and Lepsch, 1999) in the ROI and compared to our predictions, and evaluated three other soil profiles in the field. Another index of consistency between two categorical variables is the Cramer's V coefficient (Liebetrau, 1983; Rees, 2008), which measures the level of association based on Chi-squared test similar to Pearson's correlation with values varying from 0 to 1. Usually, the interpretation of this index is as follows: very strong (> 0.25), strong (0.15 – 0.25), moderate (0.10 – 0.15), weak (0.05 – 0.10), and no or very weak (0.05 – 0.00)(Akoglu, 2018). This index was used to evaluate the level of association between the SMUs and yield environments at $p < 0.01$.

4.3. RESULTS

4.3.1. Soil mapping units by soil orders

The information of the sixty-four soil attributes, properties and indices was retrieved by using the 117,413 points from twenty-seven soil mapping units (SMUs) of seven soil orders at a first categorical level according to the Brazilian Classification System (Santos et al., 2018)(Table 1). We grouped the descriptive statistics of those SMUs into three categories as soil chemical attributes, soil properties (e.g. free iron and magnetic susceptibility), indices and physical attributes, and soil mineralogy. To sum up the data, we chose to describe the soil orders instead of each SMUs. The soil chemical attributes are presented in Table 4. As the soil components were predicted based on increment core samples at 0 – 20, 40 – 60, and 80 – 100 cm depths, we concentrated the quantitative description only in the deepest layer because it is the less revolved in agriculture areas. The chemical attributes are not diagnostic to classify soils in the 1st categorical level of the Brazilian Classification System, but related to soil fertility without considering taxonomy. Analysing the C layer, there is no significant difference among soil orders in the ROI. However, SMUs in the Latossolo (L), Nitossolo (N), Neossolo (R), Luvissolo (T), Argissolo (P), Cambissolo (C), and Gleissolo (G) orders presented median values of soluble Al3+ between 5 and 14 mmolc kg-1, CEC from 62.4 to 89.6 mmolc kg-1, SB betwixt 19 and 29 mmolc kg-1, AS from 25.3 to 44%, BS between 36 and 43%, and constant pH (5.2). The lowest values of Al3+, CEC, and AS were found for soils in the L and N orders, and the highest for the G order. SB and BS values were higher for soils in the C (SB = 28.7 mmolc kg-1; BS = 40.4%), G (SB = 27.6 mmolc kg-1; BS = 39.7%), N (SB = 25.7 mmolc kg-1; BS = 42.9%), and P (SB = 24.4 mmolc kg-1; BS = 39.9%) orders. The lowest results were found in the L (SB = 19.8 mmolc kg-1; BS = 37.7%), R (SB = 19.8 mmolc kg-1; BS = 37.7%), and T (SB = 20.5 mmolc kg-1; BS = 36.8%) orders. The most differentiating soil chemical attributes among soils were AS and CEC (Fig. 3a).

Table 4. Descriptive statistics of soil chemical attributes soluble Al^{3+}, cation exchange capacity (CEC), sum of bases (SB), aluminium saturation (AS), base saturation (BS), and pH Argissolo (P), Latossolo (L), Neossolo (R), Cambissolo (C), Gleissolo (G), Nitossolo (N), and Luvissolo (L) soil orders (1st categorical level) of SiBCS.

		Al^{3+}			CEC			SB			AS			BS			pH		
		A_1	B_1	C_1	A_1	B_1	C_1	A_1	B_1	C_1	A_1	B_1	C_1	A_1	B_1	C_1	A_1	B_1	C_1
						mmol$_c$ kg^{-1}							%					Dimensionless	
	1Q	1.1	3.6	6.9	52.1	51.7	52.4	29.0	23.4	17.8	2.4	5.7	31.3	56.4	44.1	36.3	5.5	5.3	5.1
P	M	1.7	5.1	11.3	66.3	70.4	71.1	36.8	29.4	24.4	3.4	9.5	37.3	59.9	48.4	39.9	5.5	5.3	5.2
	3Q	2.8	7.5	16.9	85.5	92.2	100.8	47.0	37.1	32.8	4.7	14.8	43.5	63.4	53.1	43.6	5.6	5.4	5.3
	SD	3.2	3.9	10.3	28.1	31.8	38.2	15.4	13.5	13.1	2.1	8.2	9.4	5.4	7.1	5.5	0.1	0.2	0.1
	1Q	1.0	2.9	4.4	62.5	56.3	51.1	32.6	22.3	15.3	2.1	5.7	24.8	54.5	39.7	33.5	5.5	5.2	5.1
L	M	1.5	3.9	6.1	74.7	69.1	63.7	39.1	26.6	19.8	2.9	9.6	30.5	58.0	43.9	37.7	5.5	5.3	5.2
	3Q	2.1	5.2	9.5	87.7	82.8	79.5	45.9	32.1	26.1	4.2	15.8	38.0	61.1	48.4	41.7	5.6	5.4	5.3
	SD	2.5	2.9	6.2	21.3	25.1	31.1	11.7	10.9	10.9	1.8	10.0	9.6	5.7	7.6	6.3	0.1	0.2	0.2
	1 Q	1.3	2.8	3.4	33.0	27.4	28.2	16.4	8.6	8.6	3.5	7.4	33.5	49.6	37.0	33.7	5.4	5.2	5.1
R	M	2.2	4.7	9.5	57.0	60.5	64.4	27.9	23.9	19.8	4.7	11.6	39.3	54.5	44.3	37.5	5.5	5.3	5.2
	3Q	3.8	7.6	16.3	78.6	91.3	101.3	37.4	33.6	30.7	6.3	18.9	46.4	59.2	49.7	41.2	5.5	5.4	5.3
	SD	7.7	4.4	18.2	30.0	39.4	46.5	18.2	18.1	15.6	2.6	11.4	9.6	7.3	9.5	5.8	0.1	0.2	0.2
	1 Q	1.1	4.0	9.7	55.6	57.4	61.7	31.0	26.3	21.3	2.2	5.4	33.8	57.6	46.3	36.7	5.5	5.3	5.1
C	M	1.8	5.7	13.8	72.1	78.0	83.1	42.2	33.3	28.7	3.4	9.1	39.4	61.8	51.0	40.4	5.6	5.4	5.2
	3Q	3.0	8.1	18.8	94.2	103.8	113.3	55.5	44.7	39.0	5.3	14.4	44.7	65.8	55.4	44.3	5.6	5.5	5.3
	SD	11.9	3.8	11.6	33.3	38.3	44.4	19.3	18.7	16.4	2.8	9.5	8.9	6.6	8.2	6.2	0.1	0.2	0.1
	1 Q	1.3	4.7	12.1	58.1	71.5	71.3	36.3	29.2	21.3	3.2	6.9	36.5	57.4	45.4	34.2	5.5	5.3	5.1
G	M	1.9	6.0	15.6	66.1	89.9	89.6	40.7	33.4	27.6	4.1	10.4	40.9	60.6	51.4	39.7	5.6	5.4	5.2
	3Q	3.1	7.2	19.5	84.7	114.3	119.6	46.4	41.2	35.9	4.9	16.0	45.9	63.6	55.4	44.7	5.6	5.4	5.3
	SD	4.3	4.3	13.8	30.8	29.6	37.2	17.7	14.5	14.0	2.0	6.5	7.8	5.5	6.5	6.3	0.1	0.1	0.1
	1 Q	0.9	2.2	3.4	67.6	64.6	51.8	39.6	27.9	18.9	1.5	3.1	19.3	58.3	44.9	37.7	5.5	5.3	5.2
N	M	1.3	2.9	5.2	80.3	75.6	65.6	45.1	32.3	25.7	2.2	6.1	25.3	61.5	50.5	42.9	5.6	5.4	5.3
	3Q	1.9	4.0	9.8	100.0	89.9	81.7	54.7	41.9	36.0	3.1	10.9	32.8	65.0	57.3	49.1	5.6	5.6	5.5
	SD	4.2	3.2	6.5	31.4	34.2	47.8	18.3	19.9	20.8	1.7	5.9	9.5	5.4	8.5	7.7	0.1	0.2	0.2
	1Q	2.0	3.8	7.4	55.3	47.7	49.2	24.3	21.7	17.5	4.3	8.3	37.1	47.2	44.5	34.5	5.4	5.2	5.1
T	M	3.0	5.6	10.2	67.4	65.7	62.4	30.8	26.8	20.5	6.4	12.7	44.0	50.9	47.0	36.8	5.5	5.3	5.2
	3Q	4.2	9.8	17.9	80.6	84.6	99.5	36.5	29.7	24.7	8.7	19.6	51.6	58.6	49.2	39.8	5.6	5.4	5.3
	SD	4.4	6.5	12.7	17.8	24.7	42.1	7.9	7.2	5.8	3.3	15.9	9.8	6.3	4.3	3.9	0.1	0.1	0.1

A_1 (0 – 20 cm); B_1 (40 – 60 cm); C_1 (80 – 100 cm); 1Q (1st Quartile); M (Median); 3Q (3rd Quartile); RMSE (Standard Deviation).

The magnetic susceptibility (χ) and the free iron contents (FI) were used to characterise the soil mapping units. Moreover, three indices were calculated such as the hematite and goethite ratio, the kaolinite and gibbsite ratio and the B textural horizon. The soil physical attributes clay, sand and soil organic matter were also analysed. Those soil properties, indices, and physical attributes are described in Table 5. For the χ, we had data at 0 – 20 and 80 – 100 cm depths only. Analysing the deepest layer, N (χ = 552) and L (χ = 324) showed the highest values followed by P, G, R, C, and T. These last five soil orders had χ values betwixt 22 and 30. For the FI, we only had values from the soil surface (0 – 20 cm). High contents of FI were found in N (FI = 106.7 g kg^{-1}) and L (FI = 81.9 g kg^{-1}) orders. The other soil orders had median values from 10.5 to 25.4 g kg^{-1} being the T order the lowest. The same patterns were described for clay and sand contents. Assessing the 80 – 100 cm depth, clayey soils were those classified into N and L orders and sandy soils into R and T orders. C, G, and P presented clay contents ranging from 230.3 to 234.1 g kg^{-1}. N and L had SOM contents of 9.0 and 8.4 g kg^{-1}, whilst the other soil orders had a constant value of 5.3 g kg^{-1}. At the deepest layer, SOM could differentiate N and L from others soil orders (Fig. 3a), as well as, MS, clay, sand, and FI (0 – 20 cm) contents (Fig. 3b). Among SMUs in the seven soil orders presented none significant difference for the median values of the soil indices (Fig. 3c).

Table 5. Descriptive statistics of magnetic susceptibility (χ), hematite and goethite ratio (Hem/Hem+Gt), kaolinite and gibbsite ratio (Kln/Kln+Gt), Bt horizon, free iron (FI), clay, sand, and soil organic matter (SOM) contents for Argissolo (P), Latossolo (L), Neossolo (R), Cambissolo (C), Gleissolo (G), Nitossolo (N), and Luvissolo (L) soil orders (1st categorical level) of SiBCS.

		χ		Hem/(Hem+Gt)			Kln/(Kln+Gbs)			Bt	FI	Clay			Sand			SOM		
		A_1	C_1	A_1	B_1	C_1	A_1	B_1	C_1	B_1/ A_1	A_1	A_1	B_1	C_1	A_1	B_1	C_1	A_1	B_1	C_1
		10^{-8}m^3 kg^{-1}		Dimensionless								g kg^{-1}								
	1Q	11.3	14.6	0.7	0.7	0.7	0.9	0.9	0.9	1.3	11.0	127.3	199.1	188.9	478.6	397.2	390.1	10.9	6.6	4.6
P	M	26.6	29.6	0.7	0.7	0.7	0.9	1.0	0.9	1.5	20.1	164.8	263.1	230.3	596.5	507.7	506.3	13.0	7.4	5.6
	3Q	74.6	58.8	0.8	0.7	0.8	0.9	1.0	1.0	1.7	37.3	229.0	326.5	294.6	679.9	607.6	602.2	15.9	8.6	6.9
	SD	424.0	147.4	0.1	0.0	0.1	0.0	0.0	0.0	0.3	30.4	75.7	91.0	83.1	132.8	140.1	130.8	3.4	1.6	1.9
	1Q	53.1	55.9	0.7	0.7	0.7	0.9	0.9	0.9	1.2	27.1	188.0	280.6	254.3	412.3	354.9	327.1	15.1	8.4	6.4
L	M	280.1	324.0	0.7	0.7	0.8	0.9	0.9	0.9	1.4	81.9	252.6	360.0	345.3	500.4	442.2	416.0	17.8	10.1	8.4
	3Q	750.4	755.7	0.8	0.8	0.8	0.9	1.0	0.9	1.6	108.2	325.0	439.9	435.9	618.5	554.0	538.8	20.1	11.5	10.4
	SD	1088.5	519.6	0.1	0.0	0.1	0.0	0.0	0.0	0.3	41.9	105.1	109.9	116.5	129.5	130.8	129.0	3.8	2.0	2.8
	1 Q	6.4	12.9	0.7	0.7	0.6	0.9	0.9	0.9	1.3	9.0	91.9	137.5	140.2	620.9	548.9	521.7	9.4	6.1	4.4
R	M	17.4	26.2	0.7	0.7	0.7	0.9	1.0	0.9	1.5	10.5	120.9	182.7	175.5	738.8	682.0	656.0	10.9	7.1	5.3
	3Q	30.1	51.4	0.8	0.7	0.7	0.9	1.0	1.0	1.7	14.8	152.6	241.6	212.3	834.1	806.0	756.8	12.8	7.8	6.3
	SD	608.2	185.2	0.1	0.0	0.1	0.0	0.0	0.0	0.3	18.3	58.1	81.3	69.7	152.9	175.6	157.3	2.9	1.4	1.6
	1 Q	14.2	13.6	0.7	0.7	0.7	0.9	0.9	0.9	1.3	12.2	129.9	196.3	186.9	436.2	356.4	363.2	10.7	6.5	4.4
C	M	25.1	24.6	0.7	0.7	0.7	0.9	1.0	0.9	1.5	23.9	172.4	262.5	232.0	545.1	456.1	460.1	12.4	7.2	5.3
	3Q	61.4	45.1	0.8	0.7	0.8	0.9	1.0	1.0	1.6	38.4	235.5	327.4	293.9	640.6	584.8	566.4	14.8	8.2	6.4
	SD	603.2	217.6	0.1	0.1	0.1	0.0	0.0	0.0	0.3	29.7	73.5	90.1	78.4	134.6	141.1	126.5	3.2	1.4	1.8
	1 Q	18.9	19.0	0.7	0.7	0.7	0.9	1.0	0.9	1.4	19.0	150.9	260.5	201.6	470.4	407.7	407.3	13.0	6.8	4.4
G	M	30.9	29.3	0.7	0.7	0.7	0.9	1.0	1.0	1.6	25.4	167.9	286.2	234.1	567.8	468.9	478.9	14.2	7.7	5.3
	3Q	46.4	44.4	0.7	0.7	0.7	0.9	1.0	1.0	1.8	31.2	206.7	309.3	264.4	613.4	507.0	555.0	15.8	8.6	6.0
	SD	96.2	48.2	0.0	0.0	0.1	0.0	0.0	0.0	0.3	19.9	47.8	41.6	60.8	98.8	86.9	105.8	2.1	1.2	1.4
	1 Q	270.6	242.6	0.7	0.7	0.8	0.9	0.9	0.9	1.2	89.3	255.2	367.7	336.4	364.4	318.1	284.4	17.3	9.0	7.4
N	M	671.2	552.0	0.8	0.7	0.8	0.9	0.9	0.9	1.3	106.7	303.4	420.2	406.8	415.0	364.5	339.5	19.2	10.3	9.0
	3Q	1987.8	1187.4	0.8	0.8	0.8	0.9	1.0	0.9	1.5	116.7	399.2	497.8	470.3	480.9	425.3	414.9	21.5	11.2	10.7
	SD	2473.7	779.2	0.1	0.0	0.1	0.0	0.0	0.0	0.2	31.9	103.4	84.8	108.5	81.9	79.4	95.6	3.5	1.7	2.8
	1Q	5.2	13.8	0.7	0.7	0.7	0.9	1.0	0.9	1.3	8.7	104.3	169.5	172.1	643.0	578.1	542.6	10.2	6.6	4.7
T	M	10.4	22.6	0.8	0.7	0.7	0.9	1.0	1.0	1.5	10.5	126.1	204.4	189.3	723.2	674.5	595.8	11.9	7.2	5.3
	3Q	16.4	31.6	0.8	0.7	0.7	0.9	1.0	1.0	1.8	13.8	152.9	235.7	213.0	773.3	732.6	646.2	13.8	7.8	6.0
	SD	7.8	11.5	0.0	0.0	0.0	0.0	0.0	0.0	0.3	5.0	38.9	45.4	38.1	93.8	94.0	85.9	2.8	0.9	1.0

A_1 (0 – 20 cm); B_1 (40 – 60 cm); C_1 (80 – 100 cm); 1Q (1st Quartile); M (Median); 3Q (3rd Quartile); SD (Standard Deviation).

Nine soil minerals were used as predictors of the SMUs such as muscovite (Ms), gibbsite (Gbs), chlorite (Chl), calcite (Cal), illite (Ill), kaolinite (Kln), montmorillonite (Mnt), hematite (Hem), and goethite (Gt) at 0 – 20, 40 – 60, and 80 – 100 cm depths. The descriptive statistics for the three layers are in Table 6. We concentrated the descriptive analysis at 80 – 100 cm depth because it is the less revolved soil layer. The most hematitic soils were those into the N and L orders with 26.1 and 18.4 g kg^{-1} each. The other SMUs grouped into the R, T, G, P, and C presented median values of Hem ranging from 4.9 to 7.4 g kg^{-1}. Iron and aluminium oxides were higher in the SMUs which belonged to N and L orders. The kaolinite, 1:1 clay mineral, contents were between 14 and 45 g kg^{-1}. The 2:1 phyllosilicates presented values ranging from 14 - 41 g kg^{-1} for Ill, 14 - 42 g kg^{-1} for Mnt, and 14 – 30 g kg^{-1} for Ms. Chl and Cal displayed values ranging from 0.2 to 2.2 g kg^{-1}. The Ill, Mnt, and Ms contents allowed to distinguish most of the main SMUs into their orders (Fig. 3d).

Table 6. Descriptive statistics of soil minerals muscovite (Ms), gibbsite (Gbs), chlorite (Chl), calcite (Cal), illite (Ill), kaolinite (Kln), montmorillonite (Mnt), hematite (Hem), and goethite (Gt) for Argissolo (P), Latossolo (L), Neossolo (R), Cambissolo (C), Gleissolo (G), Nitossolo (N), and Luvissolo (L) soil orders (1st categorical level) of SiBCS.

		Gt			Hem			Gbs			Kln			Ill			Cal			Mnt			Chl			Ms		
		A_1	B_1	C_1	A_1	B_1	C_1	A_1	B_1	C_1	A_1	B_1	C_1	A_1	B_1	C_1	A_1	B_1	C_1	A_1	B_1	C_1	A_1	B_1	C_1	A_1	B_1	C_1
															g kg^{-1}													
	1Q	1.1	2.4	2.4	2.5	4.9	4.4	0.4	0.6	0.6	5.6	15.3	15.5	5.0	14.3	13.1	0.3	0.7	0.7	5.8	15.3	15.7	0.1	0.2	0.2	3.9	10.2	10.9
P	M	1.6	3.4	3.2	4.1	7.6	7.4	0.6	1.0	1.2	7.9	21.7	21.1	6.8	20.5	17.7	0.5	1.0	1.2	7.9	21.2	21.1	0.2	0.3	0.3	5.5	14.1	14.5
	3Q	2.4	4.6	4.5	7.6	11.2	13.2	1.0	1.6	2.3	12.1	30.0	29.8	10.4	26.7	26.5	0.7	1.5	1.6	11.9	28.6	29.3	0.3	0.3	0.5	8.0	18.3	19.9
	SD	1.4	1.8	1.9	5.9	6.0	9.3	0.9	0.8	1.9	5.9	11.5	11.7	5.5	9.2	10.9	0.3	0.5	0.6	5.3	9.8	10.2	0.1	0.1	0.3	3.4	6.3	6.8
	1Q	2.0	3.8	3.9	5.0	9.3	9.1	0.8	1.3	1.6	9.8	25.2	24.0	8.3	24.0	21.0	0.5	1.2	1.5	9.8	24.9	24.2	0.2	0.2	0.4	6.9	16.8	16.6
L	M	3.5	5.9	5.9	10.5	16.4	18.4	1.6	2.0	2.9	17.3	36.7	35.6	15.5	32.0	32.4	0.8	1.8	2.0	16.5	33.8	34.4	0.4	0.3	0.8	10.2	22.1	24.0
	3Q	5.0	7.2	7.9	15.8	22.4	31.3	2.7	2.6	6.1	22.8	45.8	49.4	20.7	39.4	45.1	1.1	2.2	2.4	21.5	41.2	45.5	0.5	0.5	1.1	13.4	28.0	31.1
	SD	2.2	2.1	2.7	11.4	8.4	17.8	1.6	0.9	3.6	8.6	13.7	15.3	7.9	10.7	14.8	0.5	0.9	0.7	7.7	11.3	13.1	0.2	0.2	0.4	5.1	7.8	9.6
	1 Q	0.8	1.4	1.7	2.0	3.6	3.5	0.3	0.4	0.5	3.9	8.5	9.5	3.3	9.6	8.5	0.2	0.4	0.5	3.8	8.8	10.0	0.1	0.1	0.2	2.7	6.8	7.1
R	M	1.0	2.1	2.2	2.5	4.9	4.9	0.4	0.6	0.7	5.3	13.6	14.3	4.8	12.7	12.2	0.3	0.7	0.7	5.5	13.9	14.5	0.1	0.2	0.2	3.6	9.2	10.0
	3Q	1.4	3.0	3.0	3.8	6.6	7.0	0.6	0.8	1.1	7.2	18.9	19.6	6.4	17.5	16.3	0.5	1.0	1.0	7.5	19.5	19.8	0.2	0.3	0.3	4.9	12.6	13.1
	SD	0.9	1.3	1.3	3.3	3.6	7.5	0.4	0.5	1.6	3.8	8.3	9.1	3.5	6.8	8.3	0.2	0.5	0.4	3.8	7.9	8.5	0.1	0.1	0.2	2.4	4.8	5.2
	1 Q	1.1	2.3	2.2	2.8	4.7	4.3	0.4	0.6	0.7	5.9	14.9	14.9	5.2	14.0	12.6	0.3	0.7	0.6	6.2	14.8	15.2	0.1	0.2	0.2	4.1	9.9	10.6
C	M	1.7	3.3	3.0	4.8	7.3	7.4	0.6	0.9	1.3	8.4	20.9	20.6	7.2	19.6	17.4	0.5	1.1	1.0	8.6	21.3	20.8	0.2	0.3	0.3	5.6	13.6	14.2
	3Q	2.6	4.5	4.4	7.8	10.8	13.1	1.0	1.3	2.3	12.5	29.4	29.3	10.4	26.3	26.4	0.7	1.6	1.5	12.6	28.1	29.3	0.3	0.4	0.5	8.1	17.8	19.5
	SD	1.3	1.8	1.9	5.1	5.9	9.5	0.7	0.6	1.9	5.7	11.1	11.4	5.2	8.8	10.6	0.3	0.6	0.6	5.2	9.7	9.8	0.1	0.1	0.3	3.2	6.1	6.3
	1 Q	1.3	3.3	2.6	2.9	7.3	5.1	0.5	0.8	0.7	6.5	21.0	16.9	5.7	20.2	14.7	0.5	1.0	0.9	6.7	21.4	17.4	0.1	0.2	0.2	5.0	14.4	12.0
G	M	1.5	3.7	3.2	4.1	8.6	6.9	0.6	1.0	1.0	7.9	24.5	21.4	6.8	22.4	17.7	0.6	1.2	1.1	8.2	24.3	21.6	0.2	0.3	0.3	5.6	16.1	14.5
	3Q	2.1	4.2	3.8	6.3	10.8	11.7	0.8	1.2	2.2	11.2	27.9	26.8	9.0	24.5	24.3	0.7	1.5	1.4	10.9	26.8	26.7	0.2	0.3	0.4	6.9	17.7	17.6
	SD	0.9	1.0	1.4	2.6	3.1	7.6	0.7	0.3	1.6	4.1	5.9	9.2	3.6	5.1	8.6	0.2	0.3	0.4	3.8	4.5	7.8	0.1	0.1	0.2	2.0	2.9	4.9
	1 Q	3.0	6.0	5.6	10.4	16.2	18.7	1.4	1.7	2.8	16.5	38.1	34.8	14.5	32.1	31.0	0.8	1.7	1.8	15.8	35.5	33.4	0.3	0.3	0.7	9.9	22.8	22.4
N	M	4.3	7.0	7.0	15.1	20.4	26.1	2.4	2.3	4.3	20.9	43.9	44.7	19.1	38.1	40.3	1.0	2.1	2.2	19.9	40.2	41.9	0.5	0.4	1.0	12.5	26.6	27.8
	3Q	5.7	7.7	8.1	26.5	23.6	44.0	3.6	2.9	7.1	26.4	51.3	53.2	24.4	44.9	50.6	1.4	2.8	2.6	25.1	43.7	49.2	0.7	0.6	1.2	17.2	31.2	32.7
	SD	1.8	1.8	2.2	15.2	6.0	20.4	2.1	0.8	3.3	8.7	11.3	12.3	7.6	9.3	14.3	0.5	1.0	0.6	7.5	7.7	11.3	0.3	0.2	0.5	5.6	6.3	10.4
	1Q	0.8	1.9	2.2	2.7	4.4	4.5	0.3	0.5	0.6	5.3	13.2	14.4	4.7	13.1	12.7	0.2	0.6	0.6	5.5	12.7	14.4	0.1	0.2	0.2	3.1	8.4	10.0
T	M	1.0	2.4	2.5	3.7	5.5	5.7	0.4	0.7	0.8	6.5	16.9	16.9	5.6	15.4	14.7	0.3	0.8	0.8	6.5	15.8	17.2	0.1	0.2	0.2	4.0	10.4	11.5
	3Q	1.4	2.9	3.0	5.2	6.7	6.8	0.6	0.8	1.2	7.6	19.5	20.9	6.4	18.4	18.6	0.5	0.9	1.1	7.5	19.8	20.8	0.2	0.2	0.3	5.1	12.7	13.8
	SD	0.4	0.7	0.9	1.7	2.2	3.5	0.2	0.2	0.8	1.8	4.8	6.0	1.4	4.4	5.9	0.2	0.2	0.5	1.6	4.8	5.5	0.0	0.0	0.1	1.4	2.9	4.2

A_1 (0 – 20 cm); B_1 (40 – 60 cm); C_1 (80 – 100 cm); 1Q (1st Quartile); M (Median); 3Q (3rd Quartile); SD (Standard Deviation).

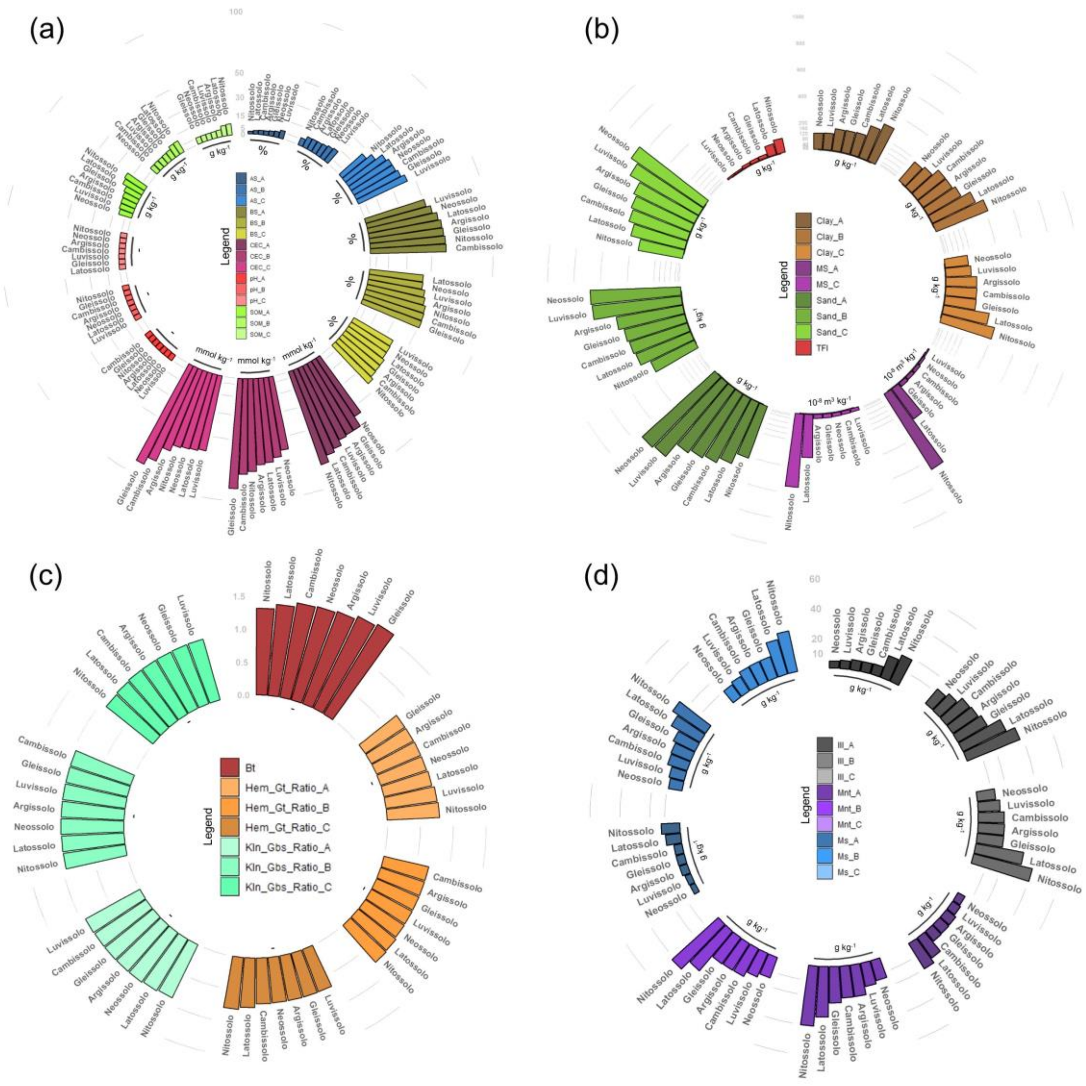


Fig. 3. Median values of the initial dataset for the seven soil orders according the Brazilian Soil Classification System (SiBCS) grouped by soil chemical attributes (a), soil properties and physical attributes (b), soil indices (c), and main soil mineralogy (d) at 0 – 20 (_A), 40 – 60 (_B), and 80 – 100 cm (_C) depths.

The major soil chemical, physical and mineralogical components, as well as, soil indices and properties median values were analysed for the 27 SMUs in the ROI (Fig. 4). The most significant chemical attribute to differentiate SMUs was the cation exchange capacity (Fig. 4a). The other chemical attributes such as AS, BS, and pH did not play an important role in separating the SMUs. The SOM had more content in red and ferric Nitisols (NVf) rather than among others. The soil physical attributes (e.g. clay and sand), and soil properties (e.g. FI and MS) proved to be vital players splitting up the SMUs (Fig. 4b). However, soil indices such as B textural horizon, Hem/(Hem+Gt), and Kln/(Kln+Gbs) could not quantitatively disaggregate well the SMUs (Fig. 4c). It does not mean that those indices are not important for soil classification, but they were not significant in the ROI. Conversely, the 2:1 phyllosilicates showed that could be quantitatively essential splitting up the SMUs (Fig. 4d). Generally, the SMUs were well described by the soil attributes, indices and properties selected from predictive maps using the DSM framework.

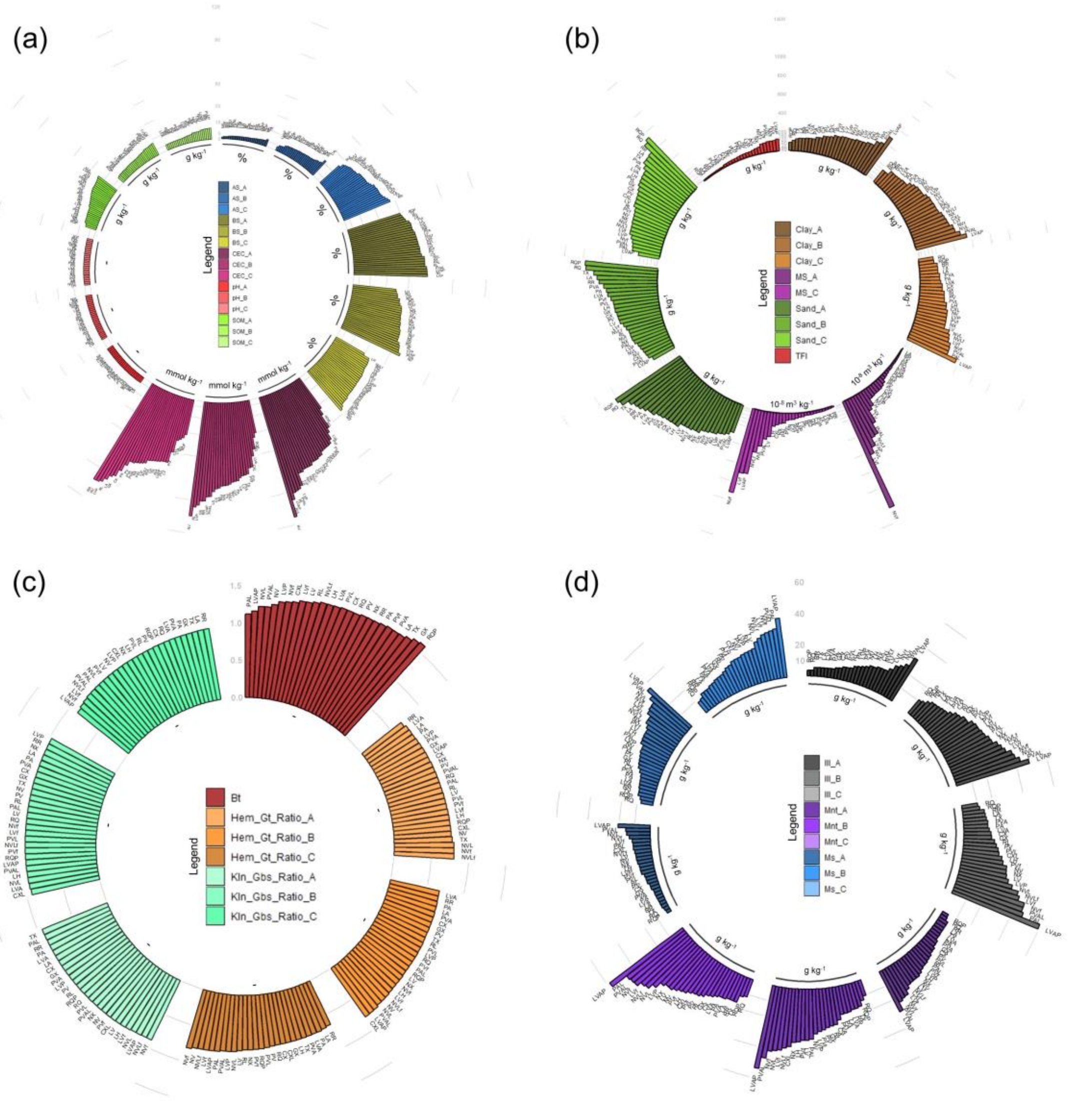


Fig. 4. Median values of the initial dataset for twenty-seven soil mapping units grouped by soil chemical attributes (a), soil properties and physical attributes (b), soil indices (c), and main soil mineralogy (d) at 0 – 20 (_A), 40 – 60 (_B), and 80 – 100 cm (_C) depths.

4.3.2. Yield environment

As it was done for the SMUs, it was performed for the yield environment. This yield environment was based on the same SMUs and used the sugarcane yield with soil fertility levels creating an index from A to G, where A is the best and G is the worst. Intrinsic to yield environment is the real evapotranspiration and sugarcane tons per hectare (STH), which delineated the thresholds for those yield environment (Demattê and Demattê, 2009). For example, yield environment A has 95 – 100 STH, LVf (Red ferric Ferralsol) SMU, and 5 mm day^{-1} evapotranspiration. However, such data is not easy to find and deal, that is why we used that previous information to analyse and propose the yield environment based on soil properties, attributes, indices, and mineralogy.

Assessing the soil chemical aspects related to each yield environment at the deepest depth (Table 7), soluble Al^{3+} median values ranged from 5.2 to 12.8 mmol$_c$ kg^{-1}. CEC and SB presented values between 65 – 76, and 20 – 25 mmol$_c$ kg^{-1}, respectively. AS and BS displayed median values from 26.4 – 40.6, and 37.6 – 41.1%. The pH values were still around 5.2, not displaying significant differences among yield environments. Aluminium saturation at 80 – 100 cm and cation exchange capacity at 0 – 20 cm depth were the foremost soil chemical attributes distinguishing the seven yield environments (Fig. 5a). Considering the less revolved soil layer, AS is the soil chemical attributes to take into account classifying the yield environments.

Table 7. Descriptive statistics of soil chemical attributes soluble Al^{3+}, cation exchange capacity (CEC), sum of bases (SB), aluminium saturation (AS), base saturation (BS), and pH for each yield environment A, B, C, D, and E.

		Al^{3+}			CEC			SB			AS			BS			pH		
		A$_1$	B$_1$	C$_1$	A$_1$	B$_1$	C$_1$	A$_1$	B$_1$	C$_1$	A$_1$	B$_1$	C$_1$	A$_1$	B$_1$	C$_1$	A$_1$	B$_1$	C$_1$
		mmol$_c$ kg^{-1}									%						Dimensionless		
A	1Q	0.9	2.5	3.7	67.9	63.3	53.4	38.2	25.9	18.1	1.8	3.9	21.1	57.3	43.7	37.5	5.5	5.3	5.2
	M	1.3	3.2	5.2	80.2	73.6	65.4	44.2	30.8	23.0	2.5	6.9	26.4	60.4	47.7	41.1	5.6	5.4	5.3
	3Q	1.9	4.3	8.0	94.0	86.1	79.4	50.9	37.3	30.3	3.4	11.3	32.7	63.3	54.1	45.6	5.6	5.5	5.4
	SD	2.7	2.2	5.0	23.7	24.6	32.7	13.6	14.3	14.1	1.5	6.7	8.4	5.3	8.2	6.4	0.1	0.2	0.2
B	1Q	1.0	3.1	5.0	64.3	59.4	54.2	35.0	23.1	16.1	2.2	6.0	26.4	55.4	39.8	33.6	5.5	5.2	5.1
	M	1.4	4.0	6.8	75.5	71.0	65.4	40.0	27.2	20.5	3.0	9.7	31.1	58.8	43.8	37.6	5.5	5.3	5.2
	3Q	2.1	5.5	11.0	89.9	86.7	84.2	47.6	33.6	28.0	4.0	15.3	36.9	62.3	49.5	41.9	5.6	5.4	5.3
	SD	2.3	2.6	6.5	25.9	28.5	35.5	14.2	12.6	13.3	1.6	9.6	7.9	5.8	7.9	6.0	0.1	0.2	0.2
C	1Q	1.2	3.2	4.9	55.3	48.6	48.4	27.4	19.4	14.3	2.3	6.0	27.9	52.4	38.8	33.0	5.5	5.2	5.1
	M	1.7	4.3	8.4	68.2	65.2	63.1	37.0	26.2	20.5	3.3	10.1	35.0	58.6	44.9	38.1	5.5	5.3	5.2
	3Q	2.4	5.9	13.1	89.3	87.9	88.5	47.3	35.6	29.1	5.2	16.1	42.5	63.2	51.8	42.6	5.6	5.4	5.3
	SD	3.6	4.2	12.2	29.6	33.4	37.1	16.8	16.5	13.7	2.2	10.0	9.9	7.6	9.7	7.2	0.1	0.2	0.2
D	1Q	1.0	3.6	6.9	55.7	50.5	51.8	30.3	21.3	16.1	2.3	6.2	31.4	56.0	42.0	35.0	5.5	5.2	5.1
	M	1.6	4.9	10.6	67.0	66.9	69.6	38.6	29.1	24.3	3.2	10.5	36.7	60.2	47.3	39.4	5.5	5.3	5.2
	3Q	2.4	6.7	15.5	83.6	88.1	98.5	48.9	38.0	34.4	4.4	16.5	42.1	64.1	52.8	43.3	5.6	5.4	5.3
	SD	2.7	4.0	7.0	27.9	36.4	37.9	16.2	18.4	16.0	2.0	8.7	8.5	6.3	8.9	6.6	0.1	0.2	0.2
E	1Q	1.3	4.0	7.3	47.6	49.5	49.3	26.7	22.5	16.5	2.4	6.5	31.5	55.3	44.3	36.8	5.5	5.3	5.2
	M	2.1	5.6	11.4	63.3	69.7	72.3	35.2	28.7	24.5	3.6	10.4	37.7	59.5	48.8	40.1	5.5	5.3	5.2
	3Q	3.3	7.8	16.9	80.9	91.7	99.9	46.1	37.3	33.4	5.2	15.9	44.0	63.4	52.7	43.4	5.6	5.4	5.3
	SD	3.1	3.6	15.5	30.4	35.1	41.2	17.7	15.9	15.4	2.3	8.6	9.7	6.6	6.7	5.3	0.1	0.1	0.1
F	1Q	1.3	4.0	7.6	50.4	49.7	51.8	26.7	22.7	16.7	2.7	6.4	34.2	54.9	42.8	35.0	5.5	5.2	5.1
	M	2.1	5.9	12.8	65.5	70.7	75.6	34.0	28.6	24.1	4.0	10.6	40.1	58.5	47.1	38.7	5.5	5.3	5.2
	3Q	3.8	9.1	18.5	84.0	95.8	108.1	43.5	35.5	32.8	5.6	17.1	46.4	62.0	51.4	42.5	5.6	5.4	5.3
	SD	7.9	4.3	11.0	27.0	34.1	42.0	16.2	15.0	13.5	2.4	10.1	9.4	5.9	7.7	6.0	0.1	0.2	0.2
G	1Q	1.3	3.4	5.4	39.2	34.5	33.1	20.2	13.8	10.8	3.2	7.1	35.3	51.6	39.1	33.9	5.4	5.2	5.1
	M	2.1	5.1	11.4	59.0	61.4	67.6	29.9	25.4	21.2	4.4	11.2	40.6	56.1	46.0	37.7	5.5	5.3	5.2
	3Q	3.6	7.8	17.3	76.4	89.7	101.3	37.4	33.0	30.5	6.0	18.3	47.1	60.1	50.5	41.2	5.6	5.4	5.3
	SD	7.6	4.1	15.5	26.8	36.3	44.8	14.7	14.3	14.2	2.6	11.0	8.7	6.7	8.5	5.3	0.1	0.2	0.1

A$_1$ (0 – 20 cm); B$_1$ (40 – 60 cm); C$_1$ (80 – 100 cm); 1Q (1st Quartile); M (Median); 3Q (3rd Quartile); SD (Standard Deviation).

The evaluation of the χ, FI, Bt horizon, kaolinite and hematite indices, and soil physical attributes are in Table 8. The soil properties and physical attributes were excellent qualifiers of the yield environment (Fig. 5b). Nevertheless, there was no relevant values variation into the indices (Fig. 5c). Yield environment classified as A, B, C, D, E, F, and G presented χ median values of 543.4, 245.3, 86.9, 40.8, 24.5, 22.0, and 21.6, respectively. The FI values were 102.0, 84.2, 40.9, 25.5, 14.9, 13.6, and 10.9 g kg^{-1}. Moreover, the clay content was 400.5, 349.4, 274.4, 255.4, 210.9, 206.7, and 184.1 g kg^{-1} for A, B, C, D, E, F, and G yield environments. SOM values presented no such difference among the seven yield environments (Fig. 5a). The trend is true for the sandier the soil the less productive it will be.

Table 8. Descriptive statistics of magnetic susceptibility (χ), hematite and goethite ratio (Hem/Hem+Gt), kaolinite and gibbsite ratio (Kln/Kln+Gt), Bt horizon, free iron (FI), clay, sand, and soil organic matter (SOM) contents for each yield environment A, B, C, D, and E.

		χ		Hem/(Hem+Gt)			Kln/(Kln+Gbs)			Bt	FI	Clay			Sand			SOM		
		A_1	C_1	A_1	B_1	C_1	A_1	B_1	C_1	B_1/ A_1	A_1	A_1	B_1	C_1	A_1	B_1	C_1	A_1	B_1	C_1
		$10^{-8}m^3 kg^{-1}$					Dimensionless								g kg^{-1}					
A	1Q	173.8	166.4	0.7	0.7	0.7	0.9	0.9	0.9	1.2	51.0	237.9	351.9	309.5	376.4	322.4	297.5	16.5	8.8	6.9
	M	541.3	543.4	0.8	0.7	0.8	0.9	1.0	0.9	1.4	102.0	294.0	414.7	400.5	429.1	378.2	351.6	19.0	10.2	8.6
	3Q	1334.6	1072.3	0.8	0.8	0.8	0.9	1.0	0.9	1.6	115.7	384.7	482.4	472.3	505.5	444.8	449.1	21.3	11.5	10.5
	SD	1615.5	637.3	0.1	0.0	0.1	0.0	0.0	0.0	0.3	38.0	107.1	92.8	111.2	99.8	92.0	107.0	3.5	1.9	2.7
B	1Q	66.7	49.3	0.7	0.7	0.7	0.9	0.9	0.9	1.3	33.6	200.2	296.3	266.1	419.0	349.9	331.4	14.5	8.0	6.1
	M	229.8	245.3	0.8	0.7	0.8	0.9	1.0	0.9	1.4	84.2	256.3	359.1	349.4	487.1	417.0	394.9	17.6	9.9	8.1
	3Q	601.6	564.9	0.8	0.8	0.8	0.9	1.0	0.9	1.6	103.8	310.8	422.8	419.8	569.2	497.7	494.4	19.4	11.4	10.1
	SD	927.9	389.2	0.1	0.1	0.1	0.0	0.0	0.0	0.3	38.7	84.1	87.7	97.4	106.1	106.0	109.3	3.5	2.1	2.8
C	1Q	22.4	25.4	0.7	0.7	0.7	0.9	0.9	0.9	1.3	18.4	155.9	249.8	220.4	443.7	389.7	369.1	12.3	7.1	5.2
	M	97.3	86.9	0.7	0.7	0.7	0.9	1.0	0.9	1.4	40.9	220.2	303.8	274.4	566.4	507.0	491.8	15.8	8.7	7.0
	3Q	254.9	275.9	0.8	0.8	0.8	0.9	1.0	1.0	1.6	79.2	266.5	360.2	343.4	650.5	587.4	587.1	18.1	10.5	9.5
	SD	1037.8	391.9	0.1	0.1	0.1	0.0	0.0	0.0	0.3	37.0	85.2	100.1	96.8	141.0	143.4	132.1	3.8	2.2	2.8
D	1Q	18.5	20.4	0.7	0.7	0.7	0.9	0.9	0.9	1.3	16.1	156.2	242.2	212.7	480.1	386.1	395.8	12.1	7.0	4.9
	M	40.7	40.8	0.7	0.7	0.7	0.9	1.0	0.9	1.5	25.5	189.6	285.2	255.4	583.9	489.4	489.8	14.4	8.1	6.1
	3Q	116.5	90.6	0.8	0.7	0.8	0.9	1.0	1.0	1.7	46.6	240.1	334.0	304.8	648.0	585.0	571.4	16.7	9.4	7.8
	SD	404.5	243.0	0.1	0.1	0.1	0.0	0.0	0.0	0.3	28.4	63.4	74.9	72.7	118.4	124.8	112.9	3.2	1.8	2.2
E	1Q	10.7	12.7	0.7	0.7	0.6	0.9	1.0	0.9	1.3	10.4	114.7	179.7	172.1	511.4	401.0	410.8	10.4	6.5	4.5
	M	21.0	24.5	0.7	0.7	0.7	0.9	1.0	1.0	1.5	14.9	149.2	234.3	210.9	623.2	525.5	541.7	12.0	7.1	5.3
	3Q	40.9	44.7	0.8	0.7	0.7	0.9	1.0	1.0	1.7	27.6	194.6	292.7	257.8	713.1	643.0	640.0	14.0	7.9	6.4
	SD	733.0	108.4	0.1	0.0	0.1	0.0	0.0	0.0	0.3	22.7	63.7	79.1	70.3	145.4	160.7	143.3	2.8	1.3	1.7
F	1Q	6.4	11.4	0.7	0.7	0.6	0.9	1.0	0.9	1.3	9.9	115.8	178.5	174.1	538.6	450.0	438.3	10.6	6.6	4.5
	M	15.8	22.0	0.7	0.7	0.7	0.9	1.0	1.0	1.5	13.6	145.0	225.4	206.7	642.3	569.3	555.4	12.1	7.4	5.4
	3Q	45.6	44.1	0.8	0.7	0.7	0.9	1.0	1.0	1.7	27.8	197.0	291.9	259.4	718.9	656.6	635.6	15.2	8.5	6.6
	SD	571.6	188.5	0.1	0.1	0.1	0.0	0.0	0.0	0.3	23.5	64.0	80.5	73.9	131.4	146.3	133.9	3.2	1.5	1.9
G	1Q	6.2	11.5	0.7	0.7	0.6	0.9	0.9	0.9	1.3	9.3	100.0	148.7	151.3	606.7	534.6	498.5	9.6	6.2	4.3
	M	14.6	21.6	0.7	0.7	0.7	0.9	1.0	1.0	1.5	10.9	125.3	190.7	184.1	698.1	638.2	611.4	11.1	7.0	5.1
	3Q	25.3	38.6	0.8	0.7	0.7	0.9	1.0	1.0	1.7	14.8	154.1	242.4	217.8	800.5	768.9	717.6	12.9	7.8	6.1
	SD	226.3	61.0	0.1	0.1	0.1	0.0	0.0	0.0	0.3	11.9	49.0	70.6	60.2	139.0	156.7	148.0	2.7	1.3	1.4

A_1 (0 – 20 cm); B_1 (40 – 60 cm); C_1 (80 – 100 cm); 1Q (1st Quartile); M (Median); 3Q (3rd Quartile); SD (Standard Deviation).

The soil mineralogy for the seven yield environment is presented in Table 9. The Hem contents were more relevant splitting the yield environment instead of Gt contents. The Al oxide represented by Gbs contents varied from 0.7 to 4.2 g kg^{-1}. The 1:1 phyllosilicate median values were between 15 and 44 g kg^{-1}. The Cal and Chl showed values up to 2.1 and 1.0, respectively. The Ill and Ms median values ranged from 12.9 - 39.6 g kg^{-1} and 10.6 – 27.8 g kg^{-1} by the less productive yield environment to the most productive one. The 2:1 clay mineral Mnt presented values of 40.9, 34.9, 26.5, 24.4, 18.4, 18.8, and 15.4 g kg^{-1} for A, B, C, D, E, F, and G yield environment, respectively. There was no difference between the values for D and E yield environment. Most of the soil minerals estimated and used to split the seven yield environment worked satisfactory (Fig. 5d).

Table 9. Descriptive statistics of soil minerals muscovite (Ms), gibbsite (Gbs), chlorite (Chl), calcite (Cal), illite (Ill), kaolinite (Kln), montmorillonite (Mnt), hematite (Hem), and goethite (Gt) for each yield environment A, B, C, D, and E.

		Gt			Hem			Gbs			Kln			Ill			Cal			Mnt			Chl			Ms		
		A_1	B_1	C_1	A_1	B_1	C_1	A_1	B_1	C_1	A_1	B_1	C_1	A_1	B_1	C_1	A_1	B_1	C_1	A_1	B_1	C_1	A_1	B_1	C_1	A_1	B_1	C_1
															g kg^{-1}													
	1Q	3.0	5.4	5.2	8.3	14.2	14.6	1.2	1.7	2.4	15.3	34.9	31.0	13.5	30.4	28.0	0.7	1.7	1.7	15.0	32.0	30.2	0.3	0.3	0.6	9.3	21.2	21.0
A	M	4.3	6.8	6.9	13.4	19.3	24.9	2.2	2.3	4.2	20.8	42.8	43.3	19.0	37.5	39.6	0.9	2.1	2.1	20.0	38.5	40.9	0.5	0.4	1.0	12.0	26.4	27.8
	3Q	5.6	7.6	8.5	21.8	24.0	40.1	3.5	2.9	7.2	26.4	50.2	52.9	24.2	43.8	50.0	1.3	2.7	2.6	24.8	43.5	49.1	0.7	0.6	1.2	16.4	31.1	33.7
	SD	2.1	1.8	2.5	13.9	7.3	19.2	2.0	0.9	3.5	8.7	12.0	13.8	7.9	9.4	14.5	0.5	1.0	0.7	7.6	9.4	12.1	0.3	0.2	0.5	5.4	6.8	9.8
	1Q	2.2	4.2	4.0	6.1	10.0	9.8	0.8	1.3	1.8	11.1	27.0	25.4	9.4	24.8	22.0	0.6	1.4	1.5	10.9	25.5	25.3	0.3	0.3	0.5	7.4	16.8	17.4
B	M	3.3	5.7	5.8	11.1	17.0	17.7	1.6	2.0	2.8	16.7	36.7	35.9	14.7	31.6	32.4	0.8	1.8	2.0	15.9	33.9	34.9	0.4	0.3	0.8	10.1	22.1	24.0
	3Q	4.6	7.0	7.6	15.1	22.4	27.8	2.5	2.7	5.3	21.0	44.6	47.3	18.9	38.2	42.7	1.0	2.1	2.4	19.6	40.7	44.0	0.5	0.4	1.0	12.6	26.6	29.9
	SD	1.9	1.9	2.5	9.3	7.7	14.9	1.3	0.9	3.4	7.1	11.5	13.6	6.6	8.8	12.6	0.3	0.6	0.7	6.3	9.9	11.6	0.2	0.2	0.4	4.1	6.3	7.8
	1Q	1.4	3.1	3.0	3.6	6.9	6.7	0.6	0.9	1.0	7.6	19.8	19.3	6.5	18.7	16.5	0.5	1.0	1.0	7.7	20.1	19.4	0.2	0.2	0.3	5.2	13.3	13.6
C	M	2.5	4.8	4.5	7.8	12.1	10.8	1.1	1.6	1.9	13.2	30.7	26.7	11.9	27.3	23.4	0.7	1.5	1.6	12.8	28.8	26.5	0.3	0.3	0.5	8.3	18.2	18.2
	3Q	3.9	6.0	6.1	11.5	16.6	19.7	1.7	2.0	3.3	18.1	36.5	36.9	16.5	31.7	33.4	0.9	1.8	2.0	17.1	34.0	35.1	0.4	0.4	0.8	10.7	21.6	23.8
	SD	1.8	2.1	2.4	8.0	7.4	14.2	1.0	0.8	2.9	7.3	12.8	13.8	6.9	9.8	12.7	0.3	0.6	0.7	6.5	10.9	11.6	0.2	0.1	0.3	4.0	6.9	7.6
	1Q	1.4	3.0	2.8	3.4	6.6	6.1	0.5	0.8	0.9	7.1	19.6	18.4	6.1	18.9	15.7	0.5	1.0	0.9	7.2	19.5	18.5	0.2	0.2	0.3	5.1	12.9	12.8
D	M	1.9	3.8	3.6	5.1	9.0	9.1	0.8	1.1	1.6	9.6	25.0	24.1	8.2	23.1	20.8	0.6	1.2	1.4	9.6	24.7	24.4	0.2	0.3	0.4	6.7	16.5	16.6
	3Q	2.8	4.9	5.0	8.6	13.2	14.7	1.3	1.8	2.4	13.1	31.4	31.2	11.7	28.0	28.3	0.7	1.6	1.8	13.5	29.8	30.5	0.3	0.3	0.6	9.0	19.3	20.7
	SD	1.2	1.6	1.8	4.6	5.3	9.4	0.7	0.6	1.9	5.0	9.5	10.7	4.5	7.5	9.9	0.3	0.6	0.6	4.6	8.4	9.1	0.1	0.1	0.3	2.8	5.2	6.0
	1Q	1.0	2.0	2.1	2.3	4.5	3.8	0.4	0.6	0.6	5.0	13.0	13.2	4.3	12.5	11.4	0.3	0.6	0.6	5.1	13.4	13.9	0.1	0.2	0.2	3.4	9.2	9.7
E	M	1.4	3.0	2.8	3.3	6.5	6.4	0.5	0.8	0.9	6.8	18.8	18.3	6.0	17.8	15.5	0.5	0.9	0.9	7.0	18.9	18.4	0.2	0.2	0.3	4.8	12.4	12.6
	3Q	1.9	3.9	3.7	5.4	8.9	9.9	0.7	1.0	1.6	9.6	25.1	24.6	8.3	22.8	21.3	0.6	1.2	1.3	9.5	24.2	24.7	0.2	0.3	0.4	6.4	16.1	16.5
	SD	1.0	1.4	1.5	3.8	4.3	7.9	0.6	0.5	1.5	4.4	9.0	10.1	4.0	7.3	9.3	0.2	0.5	0.5	4.1	7.8	8.8	0.1	0.1	0.2	2.6	5.0	5.6
	1Q	1.0	1.9	2.1	2.2	4.1	3.6	0.4	0.5	0.6	5.2	12.8	13.4	4.6	12.3	11.3	0.3	0.6	0.5	5.3	13.4	14.2	0.1	0.2	0.2	3.4	8.9	9.7
F	M	1.3	2.8	2.8	3.4	6.0	6.1	0.5	0.8	0.9	6.9	17.7	18.6	6.0	16.7	15.5	0.4	0.9	0.8	7.0	18.2	18.8	0.1	0.2	0.2	4.6	11.9	12.7
	3Q	2.0	3.9	3.8	5.7	9.1	9.8	0.8	1.3	1.8	9.8	25.0	24.6	8.3	23.8	21.6	0.6	1.3	1.4	9.9	25.2	24.9	0.2	0.3	0.4	6.8	16.4	16.8
	SD	1.1	1.5	1.6	3.9	4.5	9.0	0.5	0.6	1.9	4.7	9.5	10.5	4.3	8.1	9.9	0.3	0.6	0.6	4.5	8.5	9.2	0.1	0.1	0.3	2.9	5.3	5.9
	1Q	0.8	1.5	1.8	2.1	3.6	3.4	0.3	0.5	0.5	4.2	9.2	10.0	3.6	10.2	8.8	0.2	0.4	0.5	4.2	9.7	10.8	0.1	0.2	0.2	2.9	7.1	7.6
G	M	1.0	2.2	2.3	2.6	4.9	4.9	0.4	0.6	0.7	5.6	14.5	15.2	5.0	13.4	12.9	0.3	0.7	0.7	5.7	14.5	15.4	0.1	0.2	0.2	3.8	9.6	10.6
	3Q	1.4	3.1	3.0	3.9	6.7	7.0	0.6	0.9	1.2	7.3	19.7	20.3	6.4	18.2	16.9	0.5	1.0	1.0	7.5	19.6	20.5	0.2	0.3	0.3	4.9	12.7	13.6
	SD	0.7	1.2	1.1	2.4	3.1	4.7	0.4	0.4	1.1	3.2	7.5	8.1	2.8	5.9	7.2	0.2	0.4	0.4	3.1	6.9	7.6	0.1	0.1	0.2	1.8	4.2	4.9

A_1 (0 – 20 cm); B_1 (40 – 60 cm); C_1 (80 – 100 cm); 1Q (1st Quartile); M (Median); 3Q (3rd Quartile); SD (Standard Deviation).

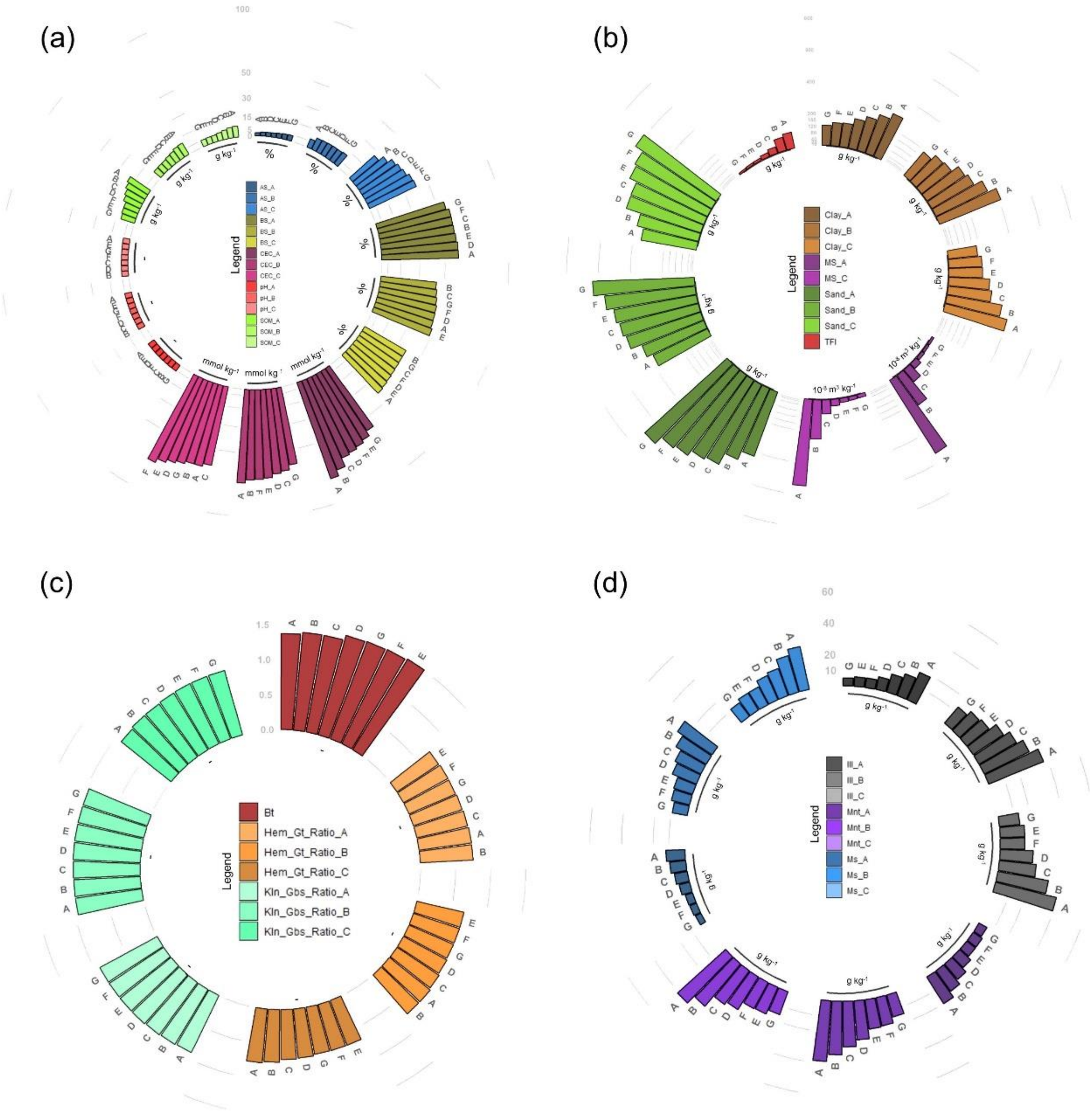


Fig. 5. Median values of the initial dataset for seven yield environment (A, B, C, D, E, and F) grouped by soil chemical attributes (a), soil properties and physical attributes (b), soil indices (c), and main soil mineralogy (d) at 0 – 20 (_A), 40 – 60 (_B), and 80 – 100 cm (_C) depths.

4.3.3. Prediction of soil mapping units and yield environment

The prediction of the SMUs and their respectively soil orders were performed using the parameters of random forest machine learning regression method available in the "caret" R package as described in the methodology. For the three response variables, it was necessary 33 variables for splitting at each tree node (Table 3). To avoid unbiased evaluation of the models, we did not leave out validation dataset and main assess the accuracy by retrieving a formal study in the same study area plus some soil profiles. Analysing the variable importance for soil order predictions, there was a certain variation of which ancillary variables had to be chosen for modelling (Table 10).

Table 10. Variable importance in percentage from Random Forest for the soil orders.

	Argissolo	Cambissolo	Gleissolo	Latossolo	Luvissolo	Neossolo	Nitossolo
Hem_A	**52**	30	12	37	13	24	27
Hem_B	33	30	9	30	13	15	21
Hem_C	42	22	6	20	4	13	15
Gt_A	45	24	8	32	14	15	26
Gt_B	**50**	32	8	31	14	14	38
Gt_C	42	24	7	26	7	12	20
Gbs_A	**52**	36	11	31	17	20	20
Gbs_B	35	29	18	33	17	28	33
Gbs_C	46	30	5	22	7	17	19
Kln_A	49	35	8	22	10	12	18
Kln_B	32	24	7	19	9	12	16
Kln_C	39	24	5	19	4	13	17
Chl_A	35	26	8	40	13	19	24
Chl_B	**79**	50	9	45	12	41	35
Chl_C	**62**	33	7	26	11	24	20
Cal_A	**59**	36	8	41	17	13	30
Cal_B	**58**	34	11	45	14	14	34
Cal_C	37	26	13	35	13	20	25
Ill_A	20	19	9	19	10	12	12
Ill_B	34	20	10	16	9	16	16
Ill_C	36	25	4	18	5	13	16
Ms_A	37	25	5	26	14	17	18
Ms_B	34	24	11	22	15	13	16
Ms_C	37	24	6	29	5	11	21
Mnt_A	34	29	9	21	11	12	16
Mnt_B	47	36	8	19	12	12	24
Mnt_C	39	24	5	20	5	11	16
H/Gt_A	**100**	**73**	20	**76**	21	40	**66**
H/Gt_B	**57**	46	14	**79**	23	**58**	46
H/Gt_C	**53**	35	4	33	7	28	32
K/Gb_A	**86**	**67**	18	**66**	29	43	37
K/Gb_B	**85**	**59**	20	**79**	31	**53**	**62**
K/Gb_C	**60**	36	11	37	12	42	27
Bt	**63**	34	9	**55**	14	40	32
χ _A	37	32	23	32	42	45	28
χ _C	34	33	21	**89**	26	29	**63**
FI_A	27	25	15	36	27	19	22
Clay_A	49	31	5	35	12	13	19
Clay_B	42	28	9	29	11	11	20
Clay_C	39	23	6	22	7	13	18
Sand_A	42	33	14	48	18	18	26
Sand_B	**62**	48	14	**51**	21	23	36
Sand_C	**51**	34	9	29	10	16	20
SOM_A	42	30	13	39	18	15	23
SOM_B	**70**	**56**	25	**59**	41	45	23
SOM_C	37	17	11	39	10	26	21
Al_A	**70**	**52**	7	**60**	10	45	39
Al_B	**78**	45	12	**54**	21	26	32
Al_C	45	36	23	**68**	20	29	39
CEC_A	**67**	37	10	42	10	18	35
CEC_B	47	38	19	**60**	21	19	40
CEC_C	**51**	34	17	**63**	20	37	47
AS_A	**88**	**66**	18	**94**	23	**58**	**57**
AS_B	**67**	41	6	48	8	41	32
AS_C	**77**	41	10	49	22	44	22
SB_A	**58**	40	13	45	17	28	45
SB_B	**65**	**57**	17	43	24	44	39
SB_C	33	23	13	41	11	20	34
BS_A	**64**	45	11	**55**	31	31	37
BS_B	**65**	**54**	15	**81**	18	44	**52**
BS_C	**56**	37	7	36	8	32	32
pH_A	**82**	**73**	37	**88**	49	**68**	**54**
pH_B	**57**	43	7	45	15	46	30
pH_C	**68**	41	8	38	10	36	29

A (0 – 20 cm); B (40 – 60 cm); C (80 – 100 cm); Hematite/ (Hematite + Goethite) (H/Gt); Kaolinite/ (Kaolinite + Gibbsite) (K/Gb).

The patterns of forty-nine predictors were relevant to model the soil orders. At 0 – 20 cm depth, the most relevant ancillary variables were Hem and Kln indices, pH, soluble Al^{3+}, SB, CEC, BS, AS, sand and SOM contents, Hem, Gt, Gbs, Kln, Chl, and Cal contents, MS and FI contents. At 40 – 60 cm depth, the Hem and Kln indices, pH, soluble Al^{3+}, SB, CEC, BS, AS, sand and SOM contents, Gt, Gbs, Cal, Chl, Ms, Ill, and Mnt contents were more important. At 80 – 100 cm, the indices of Bt horizon, Hem and Kln, and pH, soluble Al^{3+}, SB, CEC, BS, AS, sand and SOM contents, Chl and Cal contents and χ contents were the most relevant. It is of note that those predictors randomly selected by the random forest algorithm are not necessarily the only vital soil data needed to best-predicted soil orders. We just intended here to show how those variables played inside the model as it is for the SMUs and yield environment predictions. Other models could show better performance using all sixty-four predictors, but we did not test them in this specific study as the random forest is the most used algorithm in digital soil mapping.

In the region of interest, seven soil orders plus rocky outcrops class were predicted classified according to the Brazilian Soil Classification System and their correspondence with the World Reference Base (Fig. 6). The predominant soil orders were Latossolo and Argissolo, followed by Neossolo, Nitossolo, Cambissolo, Luvissolo, Gleissolo, and Afloramentos. The cartographic scale of this soil map is 1:20000, which is considered detailed mapping.

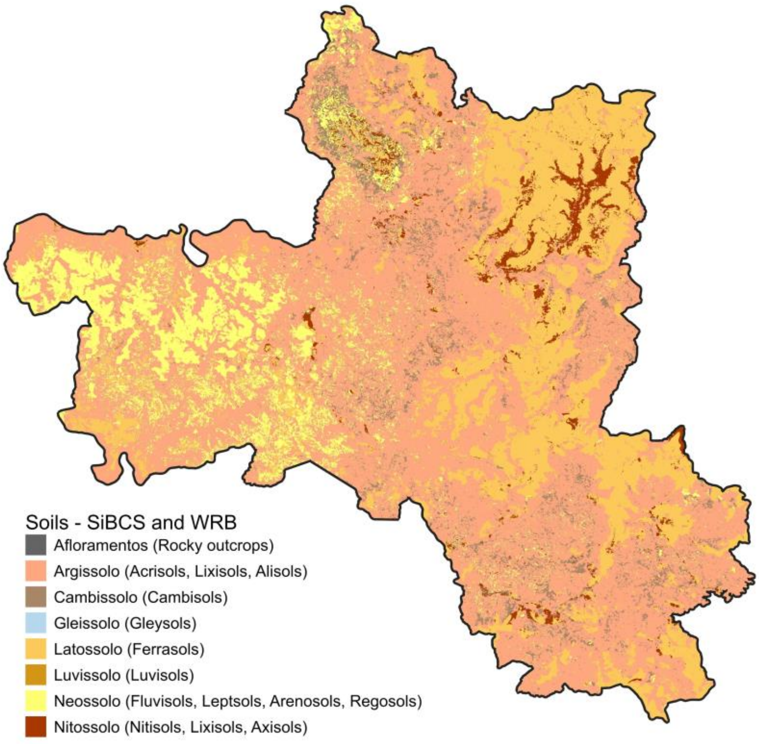

Soils - SiBCS and WRB
Afloramentos (Rocky outcrops)
Argissolo (Acrisols, Lixisols, Alisols)
Cambissolo (Cambisols)
Gleissolo (Gleysols)
Latossolo (Ferrasols)
Luvissolo (Luvisols)
Neossolo (Fluvisols, Leptsols, Arenosols, Regosols)
Nitossolo (Nitisols, Lixisols, Axisols)

Fig. 6. Predicted soil orders according to the Brazilian Soil Classification System with their correspondence to World Reference Base.

We had 27 soil mapping units within those seven soil orders (Table 11). The main predictors modelling the SMUs at 0 – 20 cm depth were Hem and Kln indices, pH, soluble Al^{3+}, SB, CEC, BS, AS, sand, clay and SOM contents, Hem, Gt, Gbs, Kln, Chl, Cal, Ill, and Mnt contents, χ and FI contents. Assessing the 40 – 60 cm depth, the Hem and Kln indices, pH, soluble Al^{3+}, SB, CEC, BS, AS, sand and SOM contents, Hem, Gt, Gbs, Kln, Chl, Cal, Ill, Ms, and Mnt contents displayed as most valuable predictors. At the deepest layer, the most relevant ancillary variables were Bt horizon, Hem and Kln indices, pH, soluble Al^{3+}, SB, CEC, BS, AS, sand, clay and SOM contents, Gt, Chl, and Cal contents, and χ contents. Thus, fifty-six out of sixty-four predictors were expressive to model the SMUs. Fig. 7 presents the final predicted map of the soil mapping units. The CX, CXL, LVA, LVAP, LVf, PV, PVf, NVLf, NVf, and RQ are the most frequent soil mapping units in the region of interest.

Table 11. Variable importance in percentage from Random Forest for the soil mapping units.

Attrib.	CX	CXL	GX	LA	LH	LV	LVA	LVAP	LVf	LVP	NV	NVf	NVL	NVLf	NX	PA	PAL	PV	PVA	PVAL	PVf	PVL	RL	RQ	RQP	RR	TX
Hem_A	42	4	14	31	12	40	18	0	25	23	22	18	22	18	9	42	7	47	46	9	13	35	28	31	19	35	17
Hem_B	39	8	8	18	18	39	16	6	23	23	22	25	17	17	8	28	6	34	28	15	14	26	29	24	11	24	15
Hem_C	33	3	7	13	8	24	10	3	14	15	11	19	5	12	3	26	7	38	31	8	11	26	17	15	15	22	6
Gt_A	39	7	10	23	14	**54**	15	6	36	22	22	33	26	28	8	33	7	**54**	43	17	14	34	30	23	13	24	19
Gt_B	42	6	10	19	17	**58**	19	9	29	23	19	38	32	24	10	37	5	50	36	12	16	32	26	18	13	28	16
Gt_C	37	3	8	12	8	32	13	7	28	16	16	28	10	16	4	31	8	40	36	16	12	30	21	16	11	24	6
Gbs_A	39	7	13	16	13	43	21	2	27	21	20	17	20	19	7	39	6	48	43	10	16	30	30	27	14	26	18
Gbs_B	**74**	15	31	28	28	**63**	**56**	8	37	**61**	42	33	35	20	15	**61**	17	**84**	**63**	24	38	**63**	**63**	36	37	34	36
Gbs_C	37	3	6	13	9	28	12	3	16	14	12	18	8	12	3	31	6	40	38	8	13	28	20	21	18	22	8
Kln_A	49	5	10	23	12	35	14	4	26	18	19	21	18	17	7	43	5	49	**52**	12	12	35	34	18	20	32	14
Kln_B	22	6	8	16	11	25	13	10	28	18	17	16	15	16	7	19	6	28	24	18	14	23	23	17	11	18	12
Kln_C	33	3	6	14	6	24	8	4	19	14	11	16	8	10	2	26	7	31	29	8	11	24	16	13	12	23	7
Chl_A	28	6	9	27	11	33	13	2	27	17	20	17	17	16	7	28	7	30	38	9	11	26	22	19	12	26	13
Chl_B	**65**	5	13	37	16	**70**	25	3	31	28	24	38	27	27	13	**63**	7	**72**	**56**	12	18	48	36	20	25	40	16
Chl_C	45	6	11	18	10	42	25	6	15	29	14	20	14	17	7	43	9	**53**	47	9	15	37	27	27	25	36	17
Cal_A	25	7	12	27	15	40	18	5	27	23	25	23	24	19	9	33	8	34	40	11	13	33	20	25	13	37	20
Cal_B	**53**	7	15	23	31	**51**	23	4	**56**	27	35	32	23	26	12	45	8	**62**	48	15	25	44	39	27	11	32	18
Cal_C	30	5	13	21	11	30	20	3	25	23	18	21	15	18	8	33	8	35	32	14	14	26	26	20	21	24	14
Ill_A	26	5	12	16	14	26	11	6	20	21	13	16	18	15	9	17	8	26	23	15	14	20	19	16	11	18	11
Ill_B	22	7	13	21	12	27	16	3	15	22	15	29	13	12	11	22	8	27	24	13	14	22	29	17	15	22	14
Ill_C	33	4	6	11	7	21	9	2	15	15	11	16	6	10	4	25	7	30	29	8	10	22	14	13	12	17	6
Ms_A	18	5	9	15	10	28	10	5	22	16	15	15	17	13	5	17	7	20	22	12	11	19	17	14	13	15	12
Ms_B	23	8	12	16	11	22	16	8	22	20	14	12	11	11	10	21	8	25	25	14	14	20	18	14	12	16	17
Ms_C	33	3	8	14	7	28	12	4	19	15	13	18	11	13	6	22	9	37	31	8	10	26	20	14	13	22	7
Mnt_A	23	5	12	26	12	31	16	4	23	18	14	20	16	14	9	22	5	37	32	9	12	28	27	17	25	29	15
Mnt_B	45	6	11	21	22	33	16	11	50	21	22	33	18	31	10	38	6	48	39	16	15	39	29	17	10	31	15
Mnt_C	32	3	7	11	7	21	9	4	17	16	12	17	8	14	3	26	6	31	27	6	10	26	16	15	10	22	6
H/Gt_A	**73**	7	22	48	21	**100**	43	5	47	**53**	41	42	31	**51**	11	**82**	5	**87**	**93**	20	20	**73**	**67**	**50**	41	**93**	31
H/Gt_B	**60**	13	19	29	25	**85**	35	4	44	36	31	**57**	38	34	17	**53**	15	**77**	**60**	26	29	**69**	38	37	30	37	31
H/Gt_C	41	5	6	16	10	35	13	3	25	16	21	24	11	19	6	32	5	**52**	31	10	13	32	21	26	15	24	7
K/Gb_A	**68**	9	23	32	23	**91**	37	3	47	37	37	39	28	30	10	**63**	14	**96**	**79**	16	27	**56**	**61**	47	38	40	31
K/Gb_B	**73**	14	21	27	21	**76**	36	6	**59**	41	36	46	43	33	13	**76**	11	**99**	**81**	26	25	**72**	48	46	**52**	41	38
K/Gb_C	49	6	10	19	13	49	17	3	22	17	16	28	13	15	4	43	7	**58**	**55**	7	16	38	29	37	30	28	11
Bt	47	4	10	19	11	**58**	20	5	32	21	19	28	20	18	5	46	13	**52**	**53**	13	13	35	24	31	16	33	16
χ_A	41	11	28	**67**	24	38	31	10	21	38	33	28	24	19	17	**59**	12	39	46	17	28	44	43	**56**	32	**71**	**52**
χ_C	35	5	24	27	27	**56**	33	13	38	30	32	40	36	31	14	33	14	41	33	22	38	35	37	32	34	26	27
FI_A	39	10	21	26	38	**68**	30	19	**88**	49	**60**	**51**	**52**	**96**	20	33	13	43	37	45	30	**56**	35	32	31	30	33
Clay_A	30	5	8	18	12	36	13	4	22	17	16	14	19	14	7	27	10	30	32	12	12	26	17	18	17	28	14
Clay_B	23	5	9	17	9	22	19	3	18	15	16	15	13	12	8	19	6	28	26	9	14	20	17	15	12	18	12
Clay_C	39	4	8	14	9	21	11	3	17	17	12	18	8	10	4	30	8	37	35	10	11	26	21	14	15	28	8
Sand_A	44	11	17	25	20	**62**	26	8	46	33	21	33	30	32	13	40	15	47	39	25	22	47	36	21	29	26	25
Sand_B	**54**	10	16	20	14	**54**	27	8	46	34	31	38	22	27	13	49	10	**60**	39	22	17	50	43	24	15	24	19
Sand_C	44	5	11	24	10	31	18	6	23	20	15	25	11	13	5	45	9	38	44	5	12	36	24	16	23	34	11
SOM_A	**62**	8	15	31	14	**51**	24	3	28	45	24	34	27	20	10	**61**	7	**74**	**60**	16	17	49	46	20	15	48	21

SOM_B	49	7	19	**62**	48	**52**	32	6	25	37	22	17	16	16	16	**64**	12	**51**	**66**	16	39	**56**	31	33	29	**81**	41
SOM_C	31	6	9	17	14	43	13	3	35	20	12	17	21	16	5	38	4	42	46	18	15	35	18	28	23	37	9
Al_A	47	3	11	22	22	**55**	17	4	33	22	17	35	11	18	6	47	4	**58**	**57**	14	16	38	32	45	24	43	10
Al_B	**50**	8	15	33	16	**64**	22	2	38	26	18	36	20	20	8	**59**	6	**66**	**73**	16	17	43	30	28	27	40	24
Al_C	**51**	9	23	30	14	**75**	30	1	**56**	34	20	45	22	29	11	**51**	9	39	**67**	11	20	**53**	31	35	32	41	26
CEC_A	49	5	11	26	20	**53**	14	3	34	25	25	43	27	24	10	49	4	**56**	49	12	19	42	28	24	23	44	12
CEC_B	45	7	16	27	22	**51**	20	3	37	29	25	40	21	23	9	38	8	41	37	14	16	42	32	23	13	40	20
CEC_C	**53**	10	18	26	17	**72**	**53**	4	36	31	28	43	23	21	8	**61**	9	**55**	**57**	17	18	**52**	35	28	18	35	25
AS_A	**73**	6	25	19	18	**89**	42	3	**58**	37	31	48	43	31	19	**77**	9	**81**	**78**	14	27	**64**	**51**	47	35	38	27
AS_B	43	7	9	25	20	50	20	4	31	21	15	35	15	17	5	42	4	**68**	48	9	16	38	25	37	20	32	13
AS_C	50	4	12	24	18	**55**	23	0	39	21	16	27	11	20	6	**51**	5	**58**	**58**	10	15	40	33	42	24	34	21
SB_A	42	7	14	26	21	47	28	5	39	29	36	33	22	27	12	44	9	**52**	48	14	24	39	45	24	26	44	17
SB_B	**63**	9	20	38	27	**50**	32	8	37	36	43	40	25	28	18	**61**	16	**54**	**60**	24	32	**55**	**54**	**51**	28	48	28
SB_C	31	7	10	19	12	44	19	3	28	17	21	27	16	16	8	32	2	30	36	11	12	30	18	22	19	30	11
BS_A	48	7	11	31	27	**50**	22	8	28	22	20	35	30	22	11	**51**	8	49	**56**	16	19	48	27	45	24	39	33
BS_B	**70**	9	20	28	28	**72**	29	13	49	40	45	**52**	42	43	15	**52**	16	**74**	**70**	35	27	**63**	**53**	37	26	48	20
BS_C	43	7	8	18	30	42	18	3	28	13	26	34	26	18	6	42	7	**53**	48	11	16	40	21	33	23	29	12
pH_A	**77**	9	26	**71**	29	**98**	37	6	**54**	**57**	**51**	38	33	36	22	**85**	8	**91**	**71**	23	30	**80**	44	**58**	**52**	**76**	40
pH_B	**52**	5	8	31	20	47	19	2	39	23	18	33	32	26	13	46	5	**55**	**53**	16	17	40	31	44	24	38	15
pH_C	**51**	5	8	24	18	37	16	1	23	16	16	24	20	20	7	46	2	**55**	**50**	10	14	40	31	29	13	32	9

A (0 – 20 cm); B (40 – 60 cm); C (80 – 100 cm); Hematite/ (Hematite + Goethite) (H/Gt); Kaolinite/ (Kaolinite + Gibbsite) (K/Gb).

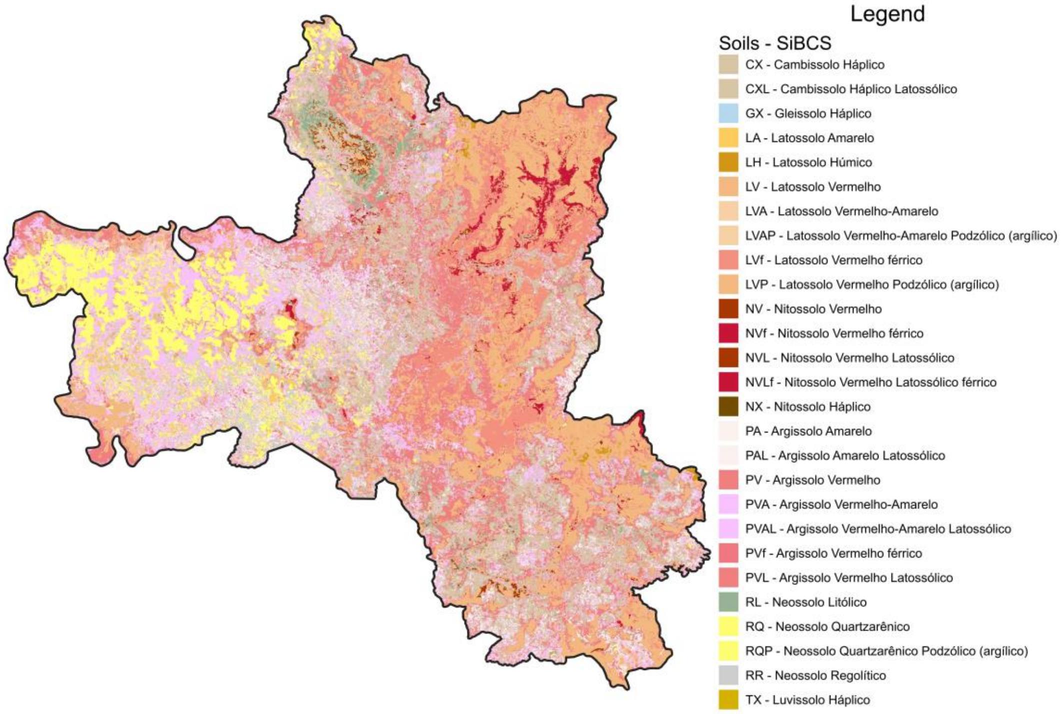

Legend
Soils - SiBCS
CX - Cambissolo Háplico
CXL - Cambissolo Háplico Latossólico
GX - Gleissolo Háplico
LA - Latossolo Amarelo
LH - Latossolo Húmico
LV - Latossolo Vermelho
LVA - Latossolo Vermelho-Amarelo
LVAP - Latossolo Vermelho-Amarelo Podzólico (argílico)
LVf - Latossolo Vermelho férrico
LVP - Latossolo Vermelho Podzólico (argílico)
NV - Nitossolo Vermelho
NVf - Nitossolo Vermelho férrico
NVL - Nitossolo Vermelho Latossólico
NVLf - Nitossolo Vermelho Latossólico férrico
NX - Nitossolo Háplico
PA - Argissolo Amarelo
PAL - Argissolo Amarelo Latossólico
PV - Argissolo Vermelho
PVA - Argissolo Vermelho-Amarelo
PVAL - Argissolo Vermelho-Amarelo Latossólico
PVf - Argissolo Vermelho férrico
PVL - Argissolo Vermelho Latossólico
RL - Neossolo Litólico
RQ - Neossolo Quartzarênico
RQP - Neossolo Quartzarênico Podzólico (argílico)
RR - Neossolo Regolítico
TX - Luvissolo Háplico

Fig. 7. Predicted soil mapping units in the region of interest, which represents the adapted fifth categorical level of the Brazilian Soil Classification System.

The yield environment classes from the most productive (A) to the less productive (G) needed forty-four out of sixty-four predictors to fit well the model (Table 12). The most important variables to build the yield environment at the three depths were Bt horizon, Hem and Kln indices, pH, soluble Al^{3+}, SB, CEC, BS, AS, sand, clay (0-20 cm) and SOM contents, Hem (0-20 cm), Gt (0-20 and 40-60 cm), Gbs (0-20 and 40-60 cm), Kln (0-20 cm), Chl (40-60 cm), and Cal (0-20 and 40-60 cm) contents, MS (0-20 and 80-100 cm) and FI (0-20 cm) contents.

Table 12. Variable importance in percentage from Random Forest for yield environment.

Attributes	A	B	C	D	E	F	G
Hem_A	29	27	31	33	35	41	35
Hem_B	17	20	22	22	20	19	16
Hem_C	16	17	25	27	24	30	14
Gt_A	23	35	37	35	27	43	24
Gt_B	29	29	37	37	36	39	23
Gt_C	26	24	27	33	28	34	15
Gbs_A	31	33	36	40	29	35	29
Gbs_B	43	43	47	**57**	**58**	**53**	43
Gbs_C	19	23	30	31	27	37	18
Kln_A	16	31	32	35	33	41	19
Kln_B	18	30	26	32	27	24	21
Kln_C	17	15	18	26	21	29	13
Chl_A	31	22	23	20	18	19	15
Chl_B	45	42	**51**	**57**	45	**62**	**60**
Chl_C	28	32	29	37	29	34	22
Cal_A	33	41	42	39	29	43	28
Cal_B	**50**	46	43	41	30	38	31
Cal_C	25	29	27	30	25	33	27
Ill_A	17	15	19	18	17	16	14
Ill_B	13	19	16	18	17	16	17
Ill_C	14	18	19	25	19	29	12
Ms_A	11	11	12	13	12	14	13
Ms_B	17	18	23	26	22	21	19
Ms_C	16	18	25	27	21	31	16
Mnt_A	16	26	22	32	29	37	22
Mnt_B	17	33	25	24	18	23	20
Mnt_C	16	15	19	27	18	28	14
H/Gt_A	**65**	**71**	**70**	**67**	**62**	**79**	**58**
H/Gt_B	41	**51**	**56**	46	41	46	48
H/Gt_C	15	30	33	36	32	39	33
K/Gb_A	46	**66**	**59**	**67**	**52**	**51**	47
K/Gb_B	**72**	**76**	**74**	**74**	**74**	**81**	**67**
K/Gb_C	25	36	35	38	37	50	38
Bt	38	45	38	40	36	50	44
χ_A	36	30	34	41	43	**54**	41
χ_C	48	44	42	40	41	45	**55**
FI_A	30	**76**	33	34	27	28	21
Clay_A	25	25	28	36	23	38	36
Clay_B	17	19	22	19	17	20	19
Clay_C	13	17	20	27	23	34	14
Sand_A	28	46	**52**	**51**	38	**51**	44
Sand_B	34	48	50	**51**	49	**54**	**56**
Sand_C	26	29	29	36	31	43	23
SOM_A	33	41	49	49	44	**60**	**51**
SOM_B	41	**64**	**55**	48	46	**59**	**55**
SOM_C	40	38	27	36	31	49	47
Al_A	49	**51**	36	39	35	**51**	46
Al_B	37	46	**55**	44	42	**56**	33
Al_C	43	47	**56**	49	43	**55**	32
CEC_A	**50**	**59**	37	47	34	**53**	28
CEC_B	42	**53**	40	42	38	44	27
CEC_C	48	**53**	45	46	43	47	38
AS_A	**65**	**68**	**68**	**61**	**53**	**67**	**65**
AS_B	37	44	40	43	39	50	**50**
AS_C	32	45	43	45	37	44	43
SB_A	44	**51**	39	49	33	**52**	24
SB_B	**51**	49	43	**58**	**53**	**66**	**50**
SB_C	47	40	23	38	28	45	22
BS_A	45	32	27	38	34	**51**	**53**
BS_B	**83**	**95**	49	**67**	**60**	**68**	46
BS_C	35	24	24	35	35	34	35
pH_A	**100**	**94**	**94**	**100**	**93**	**75**	**82**
pH_B	40	28	36	45	49	**55**	**64**
pH_C	23	24	27	35	31	46	47

A (0 – 20 cm); B (40 – 60 cm); C (80 – 100 cm); Hematite/ (Hematite + Goethite) (H/Gt); Kaolinite/ (Kaolinite + Gibbsite) (K/Gb).

The predicted yield environment map was well represented in the study area (Fig. 8). The yield environment classes display sugarcane productivity as very high (A, > 100 ton ha^{-1}), high (B, 90-100 ton ha^{-1}), medium/high (C, 85-90 ton ha^{-1}), medium (D, 80-85 ton ha^{-1}), low (E, 75-80 ton ha^{-1}), very low (F, 70-75 ton ha^{-1}), and extremely low (G, < 70 ton ha^{-1}). Most of the ROI had A, B, and G yield environment followed by the other classes.

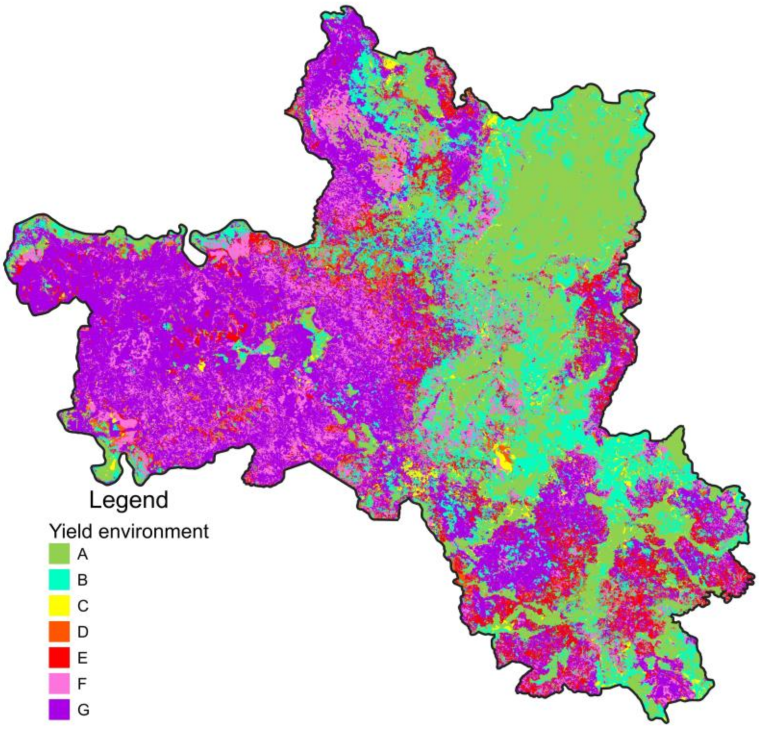


Fig. 8. Predicted map of yield environment in the region of interest.

4.3.4. Consistency between SMUs and Yield Environments

The level of association between the SMUs and yield environments was accessed using the Cramer's V coefficient, which generalises statistically contingency tables of varying sizes (Table 13). This coefficient value was 0.34 ($p < 0.01$) between those two categorical maps displaying a very strong relationship. The predicted yield environment in the region of interest had approximately 22.2, 16.5, 2.4, 4.2, 7.5, 13.3, and 33.5% of the pixels with A, B, C, D, E, F, and G yield environments, respectively. Therefore, most of the productive areas for sugarcane had very low yield.

Table 13. Consistency table of the relationship between soil mapping units (SMUs) and yield environments by predicted pixels in the region of interest.

SMUs	A	B	C	D	E	F	G
				%			
RL	0.142	0.103	0.005	0.007	0.018	0.662	0.608
CX	1.068	1.983	0.266	0.888	3.078	1.990	8.102
CXL	0.000	0.000	0.000	0.001	0.000	0.000	0.000
GX	0.000	0.002	0.000	0.002	0.010	0.001	0.008
LA	0.000	0.003	0.000	0.008	0.018	0.060	0.699
LH	0.079	0.111	0.010	0.001	0.000	0.000	0.000
LV	10.113	5.989	0.875	0.859	0.128	1.559	1.447
LVA	0.002	0.000	0.000	0.001	0.000	0.009	0.066
LVAP	0.000	0.000	0.000	0.000	0.000	0.000	0.000
LVf	3.918	1.769	0.087	0.001	0.000	0.000	0.008
LVP	0.007	0.038	0.001	0.009	0.001	0.001	0.001
NV	0.421	0.065	0.012	0.003	0.001	0.071	0.005
NVf	1.511	0.052	0.000	0.000	0.004	0.001	0.007
NVL	0.035	0.002	0.010	0.000	0.000	0.000	0.000
NVLf	0.004	0.038	0.000	0.000	0.000	0.000	0.000
NX	0.004	0.000	0.000	0.000	0.000	0.000	0.000
PA	0.069	0.298	0.066	0.360	0.996	1.652	2.520
PAL	0.000	0.000	0.000	0.001	0.000	0.000	0.000
PV	4.279	4.708	0.792	1.050	1.587	2.503	3.022
PVA	0.484	1.289	0.274	0.926	1.248	4.020	7.108
PVAL	0.006	0.008	0.038	0.001	0.000	0.000	0.000
PVf	0.005	0.000	0.000	0.000	0.000	0.000	0.000
PVL	0.037	0.107	0.013	0.127	0.004	0.037	0.023
RQ	0.006	0.007	0.008	0.001	0.119	0.369	7.440
RQP	0.012	0.005	0.001	0.002	0.302	0.051	0.544
RR	0.003	0.004	0.001	0.001	0.027	0.367	1.978
TX	0.003	0.000	0.010	0.000	0.000	0.004	0.003
Total	22.209	16.581	2.470	4.252	7.543	13.357	33.588
100% (N = 3,494,378 pixels of 30 x 30 m)							

The SMUs of LV, LVf, PV, NVf, and CX were associated with the A yield environment, which can provide very high sugarcane productivity. High yield environment can be displayed for LV, PV, CX, LVf, and PVA SMUs. Those and these SMUs are derived from igneous rocks, that is why they presented high and very high yield environments. Medium/high yield environment, which is then characterised by C, was related to SMUS of LV and PV. The PV, PVA, CX, and LV SMUs presented most of the D yield environment, that is 80-85 ton ha^{-1} of sugarcane. The CX, PV, and PVA SMUs were the foremost representatives in the E yield environment. While for the F yield environment, the highlighted soil mapping units were PVA, PV, CX, PA, and LV. Moreover, the G yield environment (extremely low sugarcane productivity) were found in the CX, RQ, PVA, PV, PA, and RR SMUs. Most of those SMUs are derived from sedimentary materials, which form sandy soils.

4.3.5. Qualitative evaluation

In the study area, we found a study of five soil profiles developed by Vidal-Torrado and Lepsch (1999) and evaluated qualitatively the similarity with the predicted SMUs (Fig. 9). These authors classified those five soil pits as LV, PVAL, and PVL which matched to our SMUs prediction. This shows the potentiality of spatializing the soil information even though using the dataset from outside that area. Thus, it means the digital soil mapping applying our approach could reach a reasonable level of certainty.

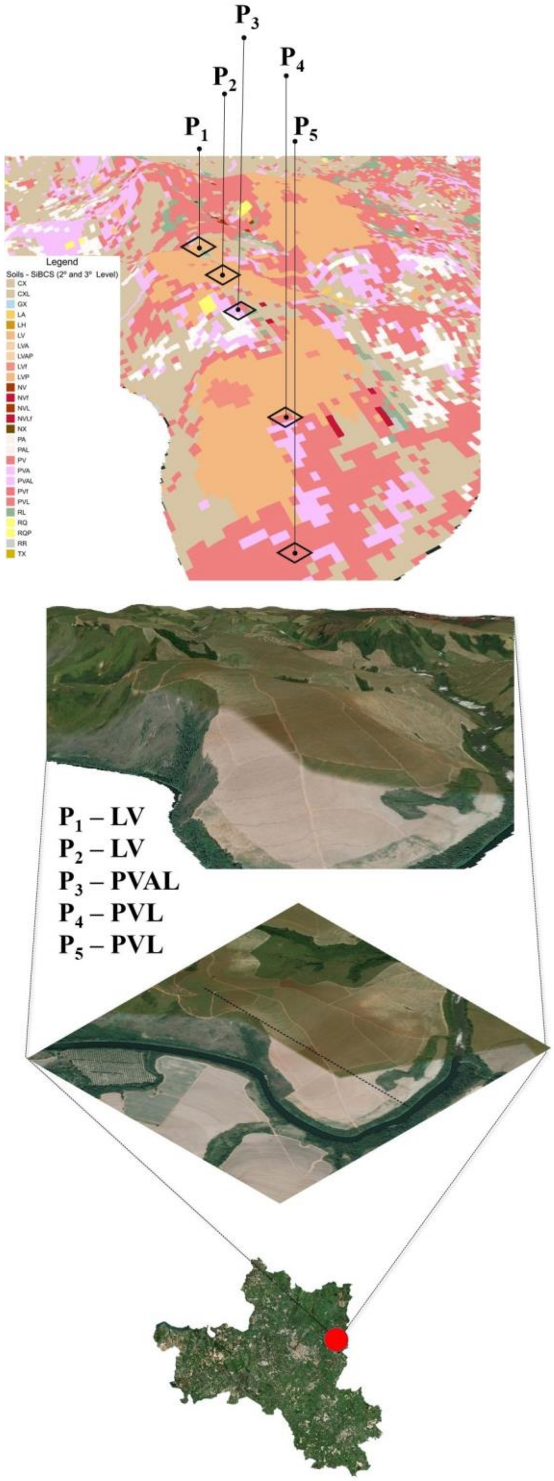


Fig. 9. Assessing the qualitative association between the predicted SMUs and a former soil survey in a site in the region of interest published by (Vidal-Torrado and Lepsch, 1999).

Furthermore, we acquired the field information of three soil profiles in the ROI to assess the predictive power of our approach in a different area (Fig. 10). The predicted SMUs corresponded to those in the field. Of course, more need to be done to improve more and more the quality of the methodology developed here in this study, but we proved that a satisfactory level of accuracy was reached.

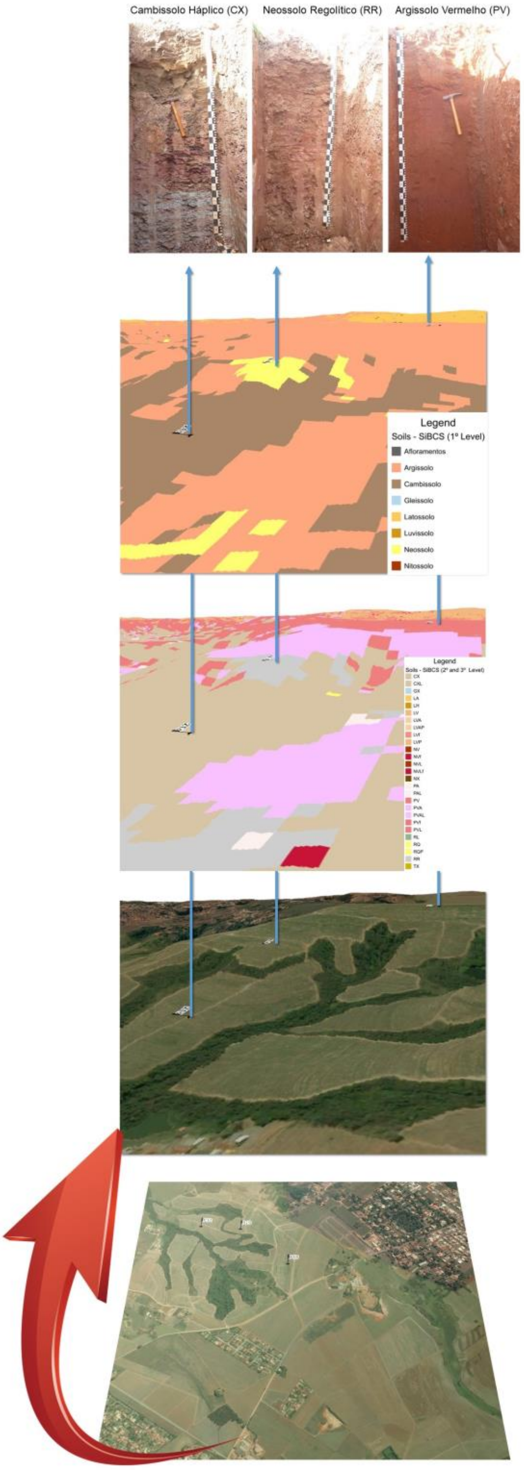


Fig. 10. Evaluating qualitatively the predicted SMUs by three soil profiles classified in the region of interest.

4.4. DISCUSSION

4.4.1. Soil mapping units' mineralogy

The magnetic susceptibility and the estimated soil mineralogy content of the Ill, Mnt, and Ms allowed distinguishing most of the main SMUs into their orders. Marconi (1974), analysing the main soil mineralogy of the soils in the same ROI, pointed out that those soils had the presence of minerals such as magnetite and muscovite.

Rodrigues and Marconi (1990), evaluating the mineralogy of sand fraction of LVf in the municipality of Piracicaba, highlighted the magnetite as the main mineral. Those studies corroborate with our findings pointing out the importance of the magnetic susceptibility to classify the soil units. This soil property is indirectly related to soil texture because soils that present high values of magnetic susceptibility will display high clay contents. The same trend applies to free iron content (Nandra, 1974), which also satisfactorily characterised the soil mapping units. Those two soil attributes play a vital role in ROI because of the diversity of parent materials such as siltstones, tillites, varvites, conglomerates, sandstones, shales, limestones, dolomite, flint, diabase, and basalt (Bonfatti et al., 2020; Teramoto et al., 2001). Another characteristic of the tropical soils in the ROI is the common presence of montmorillonite and kaolinite as revealed by Demattê et al. (2006). This fact underlies our findings in differentiating the SMUs.

4.4.2. Soil mapping units' fertility and yield environment

The yield environment combines the soil physical, chemical, and types into a single component to classify areas for sugarcane production. The Brazilian soils usually have high acidity, which affects plant growth depending on the Al^{3+} and H^{+} ions in the soil solution (Abreu et al., 2003). Soluble aluminium is the major factor that can affect plant growth and produce a toxic environment for root development. From all soil chemical attributes analysed in our study, considering the less revolved soil layer, aluminium saturation was the foremost attribute in classifying the yield environments and SMUs. (Cerri and Magalhães, 2012), analysing the correlation of Oxisols physical and chemical attributes with the sugarcane yield (e.g. yield environment), concluded that for Oxisols the physical and chemical attributes are weak to explain the yield variation. However, the Al^{3+} had a positive effect on the sugarcane yield. Those facts underlie our results and highlight the importance of this soil chemical attribute for yield environment in the ROI. Understanding the SMUs and their relationship with the yield environments allow farmers to better manage their areas allocating harvest blocks, for instance. The main challenging on this, according to Demattê and Demattê (2009), is to choose the suitable sugarcane variety for soils of low fertility as are most of the Brazilian soils.

4.4.3. Application and limitations

Our approach of using the Digital Soil Mapping products to predict the SMUs is similar to the Disaggregating and Harmonisation of Soil Map Units Through Resampled Classification Trees (DSMART) developed by Odgers et al. (2014) and extensively applied in some studies (Ellili-Bargaoui et al., 2020; Ellili Bargaoui et al., 2019; Møller et al., 2019; Vincent et al., 2018). The difference is that the DSMART uses as predictors the environmental variables necessary to predict the soil attributes in DSM, while our approach is the first attempt to use the predict soil attributes as predictors of the SMUs. That is the novelty of our study. Here, we just applied the Random Forest algorithm, which is the most tested in DSM and achieved reasonable results as it was verified by accessing other former studies in ROI and field observations. Other studies using decision tree C4.5 (Giasson et al., 2011) and logistic regression (Giasson et al., 2006) presented overall accuracy lower than our findings because they had not used the DSM products instead. We know that is not a perfect system yet. There are some limitations as the actual soil survey protocol developed by humans has. However, creating the synergy between the technology available and the tacit knowledge is a step forward to improve more and more the spatialisation and classification of soils.

4.5. CONCLUSION

In this study, we applied the DSM products as predictors of the SMUs and yield environment in a complex diversity area. Therefore, it was presented the potentiality of using the DSM products such as soil chemical, physical, indices, mineralogy, and properties to extrapolate former soil survey maps at 1:20000 scale. The digital yield environment for sugarcane based on the DSM products was created and a qualitatively evaluation of the predict soil maps and relationship with former research showed that our findings and framework could attend the need for soil maps at regional and farm levels to achieve best management agricultural practices.

ACKNOWLEDGMENTS

This research is part of PhD's thesis of the first author, presented to the Graduate Program of Soil and Plant Nutrition at the "Luiz de Queiroz" College of Agriculture, University of São Paulo. We acknowledge the financial support of the São Paulo Research Foundation (FAPESP grant numbers 2016/26124-6 and 2014/22262-0). The authors also thank the members of the Geotechnologies in Soil Science Group (GeoCIS/GeoSS) (https://esalqgeocis.wixsite.com/english).

REFERENCES

Abreu, C.H., Muraoka, T., Lavorante, A.F., 2003. Relações entre acidez e propriedades químicas de solos Brasileiros. Sci. Agric. 60, 337–343. https://doi.org/10.1590/S0103-90162003000200019

Akoglu, H., 2018. User's guide to correlation coefficients. Turkish J. Emerg. Med. https://doi.org/10.1016/j.tjem.2018.08.001

Bonfatti, B.R., Demattê, J.A.M., Marques, K.P.P., Poppiel, R.R., Rizzo, R., Mendes, W. de S., Silvero, N.E.Q., Safanelli, J.L., 2020. Digital mapping of soil parent material in a heterogeneous tropical area. Geomorphology 107305. https://doi.org/10.1016/j.geomorph.2020.107305

Buol, S.W., Southard, R.J., Graham, R.C., McDaniel, P.A., 2011. Soil genesis and classification, 6th ed. John Wiley & Sons, Ltd.

Cerri, D.G.P., Magalhães, P.S.G., 2012. Correlation of physical and chemical attributes of soil with sugarcane yield. Pesqui. Agropecu. Bras. 47, 613–620. https://doi.org/10.1590/S0100-204X2012000400018

César de Mello, D., Demattê, J.A.M., Silvero, N.E.Q., Di Raimo, L.A.D.L., Poppiel, R.R., Mello, F.A.O., Souza, A.B., Safanelli, J.L., Resende, M.E.B., Rizzo, R., 2020. Soil magnetic susceptibility and its relationship with naturally occurring processes and soil attributes in pedosphere, in a tropical environment. Geoderma 372, 114364. https://doi.org/10.1016/j.geoderma.2020.114364

Chen, S., Mulder, V.L., Heuvelink, G.B.M., Poggio, L., Caubet, M., Román Dobarco, M., Walter, C., Arrouays, D., 2020. Model averaging for mapping topsoil organic carbon in France. Geoderma 366, 114237. https://doi.org/10.1016/j.geoderma.2020.114237

Cohen, J., 1960. A Coefficient of Agreement for Nominal Scales. Educ. Psychol. Meas. 20, 37–46. https://doi.org/10.1177/001316446002000104

Debella-Gilo, M., Etzelmüller, B., 2009. Spatial prediction of soil classes using digital terrain analysis and multinomial logistic regression modeling integrated in GIS: Examples from Vestfold County, Norway 77, 8–18.

Demattê, J.A.M., Sayão, V.M., Rizzo, R., Fongaro, C.T., 2017. Soil class and attribute dynamics and their relationship with natural vegetation based on satellite remote sensing. Geoderma 302, 39–51. https://doi.org/10.1016/j.geoderma.2017.04.019

Demattê, J.A.M., Sousa, A.A., Alves, M.C., Nanni, M.R., Fiorio, P.R., Campos, R.C., 2006. Determining soil water status and other soil characteristics by spectral proximal sensing. Geoderma 135, 179–195. https://doi.org/10.1016/j.geoderma.2005.12.002

Demattê, J.L.I., Demattê, J.A.M., 2009. Ambientes de produção como estratégia de manejo na cultura da cana-de-açúcar. Informações Agronômicas 10–18.

Dharumarajan, S., Kalaiselvi, B., Suputhra, A., Lalitha, M., Hegde, R., Singh, S.K., Lagacherie, P., 2020. Digital soil mapping of key GlobalSoilMap properties in Northern Karnataka Plateau. Geoderma Reg. https://doi.org/10.1016/j.geodrs.2019.e00250

Ellili-Bargaoui, Y., Malone, B.P., Michot, D., Minasny, B., Vincent, S., Walter, C., Lemercier, B., 2020. Comparing three approaches of spatial disaggregation of legacy soil maps based on the Disaggregation and Harmonisation of Soil Map Units Through Resampled Classification Trees (DSMART) algorithm. SOIL 6, 371–388. https://doi.org/10.5194/soil-6-371-2020

Ellili Bargaoui, Y., Walter, C., Michot, D., Saby, N.P.A., Vincent, S., Lemercier, B., 2019. Validation of digital maps derived from spatial disaggregation of legacy soil maps. Geoderma 356, 113907. https://doi.org/10.1016/j.geoderma.2019.113907

Embrapa, 2020. Pronasolos - Programa Nacional de Solos do Brasil [WWW Document]. URL https://www.embrapa.br/en/pronasolos (accessed 7.22.20).

Gallo, B.C., Demattê, J.A.M., Rizzo, R., Safanelli, J.L., Mendes, W. de S., Lepsch, I.F., Sato, M. V., Romero, D.J., Lacerda, M.P.C., 2018. Multi-temporal satellite images on topsoil attribute quantification and the relationship with soil classes and geology. Remote Sens. 10, 1571. https://doi.org/10.3390/rs10101571

Giasson, E., Clarke, R.T., Inda, A.V., Merten, G.H., Tornquist, C.G., 2006. Digital soil mapping using multiple logistic regression on terrain parameters in southern Brazil. Sci. Agric. 63, 262–268. https://doi.org/10.1590/S0103-90162006000300008

Giasson, E., Sarmento, E.C., Weber, E., Flores, C.A., Hasenack, H., 2011. Árvores de decisão para o mapeamento digital de solos em encostas basálticas subtropicais. Sci. Agric. 68, 167–174. https://doi.org/10.1590/S0103-90162011000200006

Gray, J.M., Bishop, T.F.A., Wilford, J.R., 2016. Lithology and soil relationships for soil modelling and mapping. CATENA 147, 429–440. https://doi.org/10.1016/j.catena.2016.07.045

Grunwald, S., 2010. Current State of Digital Soil Mapping and What Is Next, in: Digital Soil Mapping. Springer Netherlands, Dordrecht, pp. 3–12. https://doi.org/10.1007/978-90-481-8863-5_1

Hartemink, A.E., Krasilnikov, P., Bockheim, J.G., 2013. Soil maps of the world. Geoderma 207–208, 256–267. https://doi.org/10.1016/J.GEODERMA.2013.05.003

Hartemink, A.E.E., Zhang, Y., Bockheim, J.G.G., Curi, N., Silva, S.H.G.H.G., Grauer-Gray, J., Lowe, D.J.J., Krasilnikov, P., 2020. Soil horizon variation: A review, in: Advances in Agronomy. Academic Press Inc., pp. 125–185. https://doi.org/10.1016/bs.agron.2019.10.003

IBGE, 2019. Digital Municipal Network of the Political - Brazilian Administrative Division [WWW Document]. Brazilian Insitute Geogr. Stat. URL https://www.ibge.gov.br/geociencias/organizacao-do-territorio/15774-malhas.html?=&t=o-que-e (accessed 6.2.20).

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

Klingebiel, A.A., 1958. Soil Survey Interpretation-Capability Groupings. Soil Sci. Soc. Am. J. 22, 160–163. https://doi.org/10.2136/sssaj1958.03615995002200020019x

Kuhn, M., 2008. Building Predictive Models in R Using the caret Package. J. Stat. Softw. 28, 1–26. https://doi.org/10.18637/jss.v028.i05

Li, N., Zare, E., Huang, J., Triantafilis, J., 2018. Mapping Soil Cation-Exchange Capacity using Bayesian Modeling and Proximal Sensors at the Field Scale. Soil Sci. Soc. Am. J. 82, 1203–1216. https://doi.org/10.2136/sssaj2017.10.0356

Liebetrau, A.M., 1983. Measures of Association, 1st ed, Sage Publications. SAGE Publications, Inc., Iowa.

Ma, Y., Minasny, B., Malone, B.P., Mcbratney, A.B., 2019. Pedology and digital soil mapping (DSM). Eur. J. Soil Sci. 70, 216–235. https://doi.org/10.1111/ejss.12790

Machado, P.G., Walter, A., Picoli, M.C., João, C.G., 2017. Potential impacts on local quality of life due to sugarcane expansion: a case study based on panel data analysis. Environ. Dev. Sustain. 19, 2069–2092. https://doi.org/10.1007/s10668-016-9823-6

Marconi, A., 1974. Mineralogia de solos das séries Paredão Vermelho, Ribeirão Claro e Saltinho, do município de Piracicaba, SP. An. da Esc. Super. Agric. Luiz Queiroz 31, 403–418. https://doi.org/10.1590/s0071-12761974000100031

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

Mendonça-Santos, M.L.L., Santos, H.G. dos G., dos Santos, H.G., Santos, H.G. dos G., 2006. Chapter 3 The State of the Art of Brazilian Soil Mapping and Prospects for Digital Soil Mapping, in: Developments in Soil Science. Elsevier, pp. 39–601. https://doi.org/10.1016/S0166-2481(06)31003-3

Møller, A.B., Malone, B., Odgers, N.P., Beucher, A., Iversen, B.V., Greve, M.H., Minasny, B., 2019. Improved disaggregation of conventional soil maps. Geoderma 341, 148–160. https://doi.org/10.1016/J.GEODERMA.2019.01.038

Nandra, S.S., 1974. Free iron oxide content of a tropical soil. Plant Soil 40, 453–456. https://doi.org/10.1007/BF00011532

Norfleet, M.L., Ditzler, C.A., Puckett, W.E., Grossman, R.B., Shaw, J.N., 2003. Soil Quality and Its Relationship To Pedology. Soil Sci. 168, 149–155. https://doi.org/10.1097/01.ss.0000058887.60072.07

Nussbaum, M., Spiess, K., Baltensweiler, A., Grob, U., Keller, A., Greiner, L., Schaepman, M.E., Papritz, A., 2018. Evaluation of digital soil mapping approaches with large sets of environmental covariates. SOIL 4, 1–22. https://doi.org/10.5194/soil-4-1-2018

Odeha, I.O. a., McBratney, A.B., Chittleborough, D.J., 1994. Spatial prediction of soil properties from landform attributes derived from a digital elevation model. Geoderma 63, 197–214. https://doi.org/10.1016/0016-7061(94)90063-9

Odgers, N.P., McBratney, A.B., Minasny, B., 2011. Bottom-up digital soil mapping. I. Soil layer classes. Geoderma. https://doi.org/10.1016/j.geoderma.2011.03.014

Odgers, N.P., Sun, W., McBratney, A.B., Minasny, B., Clifford, D., 2014. Disaggregating and harmonising soil map units through resampled classification trees. Geoderma 214–215, 91–100. https://doi.org/10.1016/j.geoderma.2013.09.024

Padarian, J., Minasny, B., McBratney, A.B., 2017. Chile and the Chilean soil grid: A contribution to GlobalSoilMap. Geoderma Reg. 9, 17–28. https://doi.org/10.1016/J.GEODRS.2016.12.001

Polidoro, J.C., Mendonça-Santos, M. de L., Lumbreras, J.F., Coelho, M.R., Carvalho Filho, A. de, Motta, P.E.F. da, Carvalho Junior, W. de, Araujo Filho, J.C. de, Curcio, G.R., Correia, J.R., Martins, E. de S., Spera, S.T., Oliveira, S.R. de M., Bolfe, E.L., Manzatto, C. V., Tosto, S.G., Venturieri, A., Sa, I.B., Oliveira, V.A. de, Shinzato, E., Anjos, L.H.C. dos, Valladares, G.S., Ribeiro, J.L., Medeiros, P.S.C. de, Moreira, F.M. de S., Silva, L.S.L., SequinattO, L., Aglio, M.L.D., Dart, R. de O., DART, R. de O., 2017. Programa Nacional de Solos do Brasil (PronaSolos), 2016th ed. Embrapa Solos, Rio de Janeiro, RJ.

Poppiel, R.R., Lacerda, M.P.C., Demattê, J.A.M., Oliveira, M.P., Gallo, B.C., Safanelli, J.L., 2019a. Soil class map of the Rio Jardim watershed in Central Brazil at 30 meter spatial resolution based on proximal and remote sensed data and MESMA method. Data Br. https://doi.org/10.1016/j.dib.2019.104070

Poppiel, R.R., Lacerda, M.P.C., Demattê, J.A.M., Oliveira, M.P., Gallo, B.C., Safanelli, J.L., 2019b. Pedology and soil class mapping from proximal and remote sensed data. Geoderma 348, 189–206. https://doi.org/10.1016/j.geoderma.2019.04.028

Poppiel, R.R., Lacerda, M.P.C., Rizzo, R., Safanelli, J.L., Bonfatti, B.R., Silvero, N.E.Q., Demattê, J.A.M., 2020. Soil Color and Mineralogy Mapping Using Proximal and Remote Sensing in Midwest Brazil. Remote Sens. 12, 1197. https://doi.org/10.3390/rs12071197

Rees, W.G., 2008. Comparing the spatial content of thematic maps. Int. J. Remote Sens. 29, 3833–3844. https://doi.org/10.1080/01431160701852088

Rodrigues, E.M., Marconi, A., 1990. Mineralogia da fração areia de latossolo roxo do município de Piracicaba, SP. An. da Esc. Super. Agric. Luiz Queiroz 47, 221–232. https://doi.org/10.1590/s0071-12761990000100013

Rossiter, D.G., 2018. Past, present & future of information technology in pedometrics. Geoderma 324, 131–137. https://doi.org/10.1016/j.geoderma.2018.03.009

Rudorff, B.F.T., de Aguiar, D.A., da Silva, W.F., Sugawara, L.M., Adami, M., Moreira, M.A., 2010. Studies on the rapid expansion of sugarcane for ethanol production in São Paulo state (Brazil) using Landsat data. Remote Sens. 2, 1057–1076. https://doi.org/10.3390/rs2041057

Sanchez, P.A., Ahamed, S., Carre, F., Hartemink, A.E., Hempel, J., Huising, J., Lagacherie, P., McBratney, A.B., McKenzie, N.J., Mendonca-Santos, M. d. L., Minasny, B., Montanarella, L., Okoth, P., Palm, C.A., Sachs, J.D., Shepherd, K.D., Vagen, T.-G., Vanlauwe, B., Walsh, M.G., Winowiecki, L.A., Zhang, G.-L., 2009. Digital Soil Map of the World. Science (80-.). 325, 680–681. https://doi.org/10.1126/science.1175084

Santos, H.G., Jacomine, P.K.T., Anjos, L.H.C. dos, Oliveira, V.A. de, Lumbreras, J.F., Coelho, M.R., Almeida, J.A., Araújo Filho, J.C. de, Oliveira, J.B., Cunha, T.J.F., 2018. Sistema Brasileiro de Classificação de Solos, 5 ed. ed. Embrapa, Brasília - DF.

Steinberg, A., Chabrillat, S., Stevens, A., Segl, K., Foerster, S., 2016. Prediction of Common Surface Soil Properties Based on Vis-NIR Airborne and Simulated EnMAP Imaging Spectroscopy Data: Prediction Accuracy and Influence of Spatial Resolution. Remote Sens. 8, 613. https://doi.org/10.3390/rs8070613

Teramoto, E.R., Lepsch, I.F., Vidal-Torrado, P., 2001. Relações solo, superfície geomórfica e substrato geológico na microbacia do ribeirão marins (Piracicaba - SP). Sci. Agric. 58, 361–371. https://doi.org/10.1590/S0103-90162001000200021

Vidal-Torrado, P., Lepsch, I.F., 1999. Relações material de origem / solo e pedogênese em uma seqüência de solos predominantemente argilosos e Latossólicos sobre psamitos na depressão periférica Paulista: Paulo State Peripheral Depression, southeastern Brazil. Rev. Bras. Ciência do Solo 23, 357–369. https://doi.org/10.1590/S0100-06831999000200019

Vincent, S., Lemercier, B., Berthier, L., Walter, C., 2018. Spatial disaggregation of complex Soil Map Units at the regional scale based on soil-landscape relationships. Geoderma 311, 130–142. https://doi.org/10.1016/J.GEODERMA.2016.06.006

Wolski, M.S., Dalmolin, R.S.D., Flores, C.A., Moura-Bueno, J.M., Caten, A. ten, Kaiser, D.R., Wolski, M.S., Dalmolin, R.S.D., Flores, C.A., Moura-Bueno, J.M., Caten, A. ten, Kaiser, D.R., 2017. Digital soil mapping and its implications in the extrapolation of soil-landscape relationships in detailed scale. Pesqui. Agropecuária Bras. 52, 633–642. https://doi.org/10.1590/s0100-204x2017000800009

Zeng, R., Rossiter, D.G., Yang, F., Li, D.-C., Zhao, Y.-G., Zhang, G.-L., 2017. How accurately can soil classes be allocated based on spectrally predicted physio-chemical properties? Geoderma 303, 78–84. https://doi.org/10.1016/j.geoderma.2017.05.011