UNIVERSIDADE DE BRASÍLIA
FACULDADE DE AGRONOMIA E MEDICINA VETERINÁRIA
PROGRAMA DE PÓS-GRADUAÇÃO EM AGRONOMIA

PEDOMETRIC MAPPING OF KEY TOPSOIL AND SUBSOIL ATTRIBUTES USING PROXIMAL AND REMOTE SENSING IN MIDWEST BRAZIL

RAÚL ROBERTO POPPIEL

TESE DE DOUTORADO EM AGRONOMIA

BRASÍLIA/DF

MARÇO/2020

UNIVERSIDADE DE BRASÍLIA
FACULDADE DE AGRONOMIA E MEDICINA VETERINÁRIA
PROGRAMA DE PÓS-GRADUAÇÃO EM AGRONOMIA

PEDOMETRIC MAPPING OF KEY TOPSOIL AND SUBSOIL ATTRIBUTES USING PROXIMAL AND REMOTE SENSING IN MIDWEST BRAZIL

RAÚL ROBERTO POPPIEL

ORIENTADOR: Prof^{a}. Dra. MARILUSA PINTO COELHO LACERDA
CO-ORIENTADOR: Prof. Titular JOSÉ ALEXANDRE MELO DEMATTÊ

TESE DE DOUTORADO EM AGRONOMIA

BRASÍLIA/DF

MARÇO/2020

UNIVERSIDADE DE BRASÍLIA
FACULDADE DE AGRONOMIA E MEDICINA VETERINÁRIA
PROGRAMA DE PÓS-GRADUAÇÃO EM AGRONOMIA

PEDOMETRIC MAPPING OF KEY TOPSOIL AND SUBSOIL ATTRIBUTES USING PROXIMAL AND REMOTE SENSING IN MIDWEST BRAZIL

RAÚL ROBERTO POPPIEL

TESE DE DOUTORADO SUBMETIDA AO PROGRAMA DE PÓS-GRADUAÇÃO EM AGRONOMIA, COMO PARTE DOS REQUISITOS NECESSÁRIOS À OBTENÇÃO DO GRAU DE DOUTOR EM AGRONOMIA

Aprovada por:

Marilusa P. C. Lacerda

MARILUSA PINTO COELHO LACERDA (Orientadora)
Professora Doutora (FAV - Universidade de Brasília), e-mail: marilusa@unb.br

EDSON EYJI SANO (Membro externo)
Pesquisador Doutor (Embrapa Cerrados), e-mail: edson.sano@embrapa.br

PAULO ROBERTO MENESES (Membro externo)
Professor Doutor (IG - Universidade de Brasília), e-mail: pmeneses@unb.br

ELPÍDIO INÁCIO FERNANDES FILHO (Membro externo)
Professor Doutor (Depto. Solos - Universidade Federal de Viçosa), e-mail: elpidio@ufv.br

Brasília/DF, 13 de março de 2020

REFERÊNCIA BIBLIOGRÁFICA

POPPIEL, R. R. **Pedometric mapping of key topsoil and subsoil attributes using proximal and remote sensing in Midwest Brazil.** Faculdade de Agronomia e Medicina Veterinária, Universidade de Brasília- Brasília, 2019; 105p. (Tese de Doutorado em Agronomia).

CESSÃO DE DIREITOS

NOME DO AUTOR: Raúl Roberto Poppiel

TÍTULO DA TESE DE DOUTORADO: Pedometric mapping of key topsoil and subsoil attributes using proximal and remote sensing in Midwest Brazil. GRAU: Doutor ANO: 2020

É concedida à Universidade de Brasília permissão para reproduzir cópias desta tese de doutorado e para emprestar e vender tais cópias somente para propósitos acadêmicos e científicos. O autor reserva outros direitos de publicação e nenhuma parte desta tese de doutorado pode ser reproduzida sem autorização do autor.

Raúl Roberto Poppiel

CPF: 703.559.901-05

Email: raulpoppiel@gmail.com

Poppiel, Raúl Roberto

Pedometric mapping of key topsoil and subsoil attributes using proximal and remote sensing in Midwest Brazil/ Raúl Roberto Poppiel. -- Brasília, 2020.

105f. : il.

Tese (Doutorado em Agronomia) - Universidade de Brasília, Faculdade de Agronomia e Medicina Veterinária, 2019.

Orientador: Prof^{a}. Dra. Marilusa Pinto Coelho Lacerda

Co-orientador: Prof. Titular José Alexandre Melo Demattê

Bibliografia

1. Gênese, Levantamento e Classificação de Solos. 2. Geoprocessamento aplicado às Ciências Agrárias. 3. Mapeamento digital de solos. I. Poppiel, Raúl Roberto. II. Universidade de Brasília. Faculdade de Agronomia e Medicina Veterinária. Doutorado em Agronomia. III. Título. IV. Marilusa, Marilusa Pinto Coelho. Doutor

I THANK

The life, Brazil and everyone around me.

I DEDICATE

To my dear mother, Liliana,
my source of inspiration that calms everything,
to my family and the memories of my grandparents.

I OFFER

To the academic community of Brazil.

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my advisor Profª. Marilusa P.C. Lacerda and co-advisor Prof. José A.M. Demattê for the continuous support on my research, motivation and immense knowledge.

Besides my advisors, I would like to thank the rest of my thesis examining committee: Dr. Edson E. Sano, Prof. Paulo R. Meneses, Prof. Elpídio I. Fernandes Filho and Prof. Tairone P. Leão for their insightful comments and ideas.

I am grateful to my partner Luis Fernando P. Pimentel, for all your affection, patience and emotional support along this intense period. I am also grateful to my family who have supported me along my life… I love you all!

I thank the Federal District Research Foundation – FAPDF, who funded my PhD scholarship, and the São Paulo Research Foundation – FAPESP, who funded this research through the grant numbers: 2014/22262-0, 2016/26124-6 and 2016/01597-9.

I am also grateful to the Faculty of Agronomy and Veterinary Medicine of the University of Brasilia (FAV/UnB), and to the Department of Soil Science of the "Luiz de Queiroz" College of Agriculture from University of São Paulo (LSO/ESALQ/USP), for the opportunity to develop my PhD. Particularly to my professors at the UnB, I am thankful for the valuable knowledge.

A very special gratitude goes out to all my colleagues and friends from the Geotechnologies in Soil Science group (https://esalqgeocis.wixsite.com/english/team) at the ESALQ/USP, and the Geoprocessing and Pedomorphology Laboratory at the FAV/UnB. I am not naming anyone, because you all have always supported me. It was great to share experiences (academic and non-academic) with all of you during the last four years.

This accomplishment would not have been possible without you all. Thank you!

SUMMARY

GENERAL ABSTRACT ---------- **1**
1. INTRODUCTION ---------- **3**
2. HYPOTHESIS ---------- **5**
3. OBJECTIVES ---------- **5**
3.1. GENERAL OBJECTIVE ---------- 5
3.2. SPECIFIC OBJECTIVES ---------- 5
4. GENERAL FRAMEWORK ---------- **5**
4.1. WORKING STEPS ---------- 5
4.2. THE THEMATIC PROJECT THAT SUPPORTS THIS THESIS ---------- 7
5. LITERATURE REVIEW ---------- **9**
5.1. WHAT IS PEDOMETRICS AND DIGITAL SOIL MAPPING? ---------- 9
5.2. RECENT ADVANCES IN PEDOMETRIC MAPPING ---------- 9
5.3. WHAT IS GOOGLE EARTH ENGINE? ---------- 10
5.4. GOOGLE EARTH ENGINE FOR PEDOMETRIC MAPPING ---------- 12
5.5. MACHINE LEARNING FOR PEDOMETRIC MAPPING ---------- 14
5.6. WHY PERFORM PEDOMETRIC MAPPING OF SOIL ATTRIBUTES? ---------- 15
6. REFERENCES ---------- **17**

CHAPTER 1 — MAPPING AT 30 M RESOLUTION OF PHYSICAL AND CHEMICAL SOIL ATTRIBUTES AT MULTIPLE DEPTHS IN MIDWEST BRAZIL ---------- **24**
1. INTRODUCTION ---------- **25**
2. MATERIAL AND METHODS ---------- **27**
2.1. STUDY AREA AND SOIL DATA ---------- 27
2.2.PREPARING ENVIRONMENTAL COVARIATES ---------- 29
2.2.1. Climate Data ---------- 31
2.2.2. Relief and Geology Data ---------- 31
2.2.3 Landsat-Derived Data ---------- 31
2.3. RANDOM FOREST (RF) REGRESSION ---------- 34
2.3.1 Calibration and Model Tuning ---------- 35
2.3.2 Validation and Variable Importance ---------- 35
2.3.3 Prediction of Continuous Soil Attributes ---------- 36
3. RESULTS ---------- **36**
3.1. SUMMARY AND RELATIONSHIPS BETWEEN SOIL ATTRIBUTES ---------- 36
3.2. SYNTHETIC SOIL IMAGE (SYSI) AND SYNTHETIC VEGETATION IMAGE (SYVI) ---------- 38
3.3. MODEL ASSESSMENTS ---------- 40
3.4. BEST PREDICTORS ---------- 42
3.5. SOIL MAPS AT MULTIPLE DEPTHS ---------- 43
4. DISCUSSION ---------- **45**
4.1. SOIL DATA ---------- 45

4.2. MACHINE LEARNING--45
4.3. INTERPRETATION OF COVARIATE LAYERS --46
4.4. RELIABILITY AND INTERPRETATION OF SOIL MAPS -------------------------------------48
4.5. POSSIBLE APPLICATIONS OF THE MAPS ---50
5. CONCLUSIONS AND FINAL CONSIDERATIONS ------------------------------------50
6. LIST OF ABBREVIATIONS---51
7. APPENDIX A ---51
8. REFERENCES ---57

CHAPTER 2 — SOIL COLOR AND MINERALOGY MAPPING USING PROXIMAL AND REMOTE SENSING IN MIDWEST BRAZIL---62
1. INTRODUCTION--63
2. MATERIALS AND METHODS---66
2.1. STUDY AREA AND SOIL DATA --66
2.2. REFLECTANCE TO SOIL COLOR --67
2.3. REFLECTANCE TO SOIL MINERALOGY --67
2.3.1. Spectral processing---67
2.3.2. Key spectral bands for mineral quantification--------------------------------68
2.4. ENVIRONMENTAL COVARIATES --68
2.5. SOIL MODELLING BY RANDOM FOREST (RF) ---69
2.5.1. Model tuning---70
2.5.2. Model performance--70
2.5.3. Covariates importance ---70
2.5.1. Soil Mapping ---71
3. RESULTS --72
3.1. SOIL ATTRIBUTES DERIVED FROM SPECTRA --72
3.2. PERFORMANCE OF SPATIAL MODELS--74
3.3. RELEVANCE OF COVARIATES--75
3.4. DIGITAL MAPS OF THE SOIL SURFACE AND SUBSURFACE ---------------------------77
3.4.1. Gridded Munsell soil color---77
3.4.2. Spatial patterns of the main minerals in studied soils ---------------------79
4. DISCUSSION ---82
4.1. RELATIONSHIPS BETWEEN SOIL COLOR AND MINERALOGY ----------------------------82
4.2. USE OF REGRESSION MODELS FOR MAPPING SOIL PROPERTIES---------------------84
4.3 INFLUENCE OF ENVIRONMENTAL PREDICTORS IN SOIL COLOR AND MINERALOGY PATTERNS--85
4.4. COMPARISON WITH LEGACY DATA AND MAPS--------------------------------------88
5. CONCLUSIONS--90
6. APPENDIX A ---91
7. REFERENCES ---93

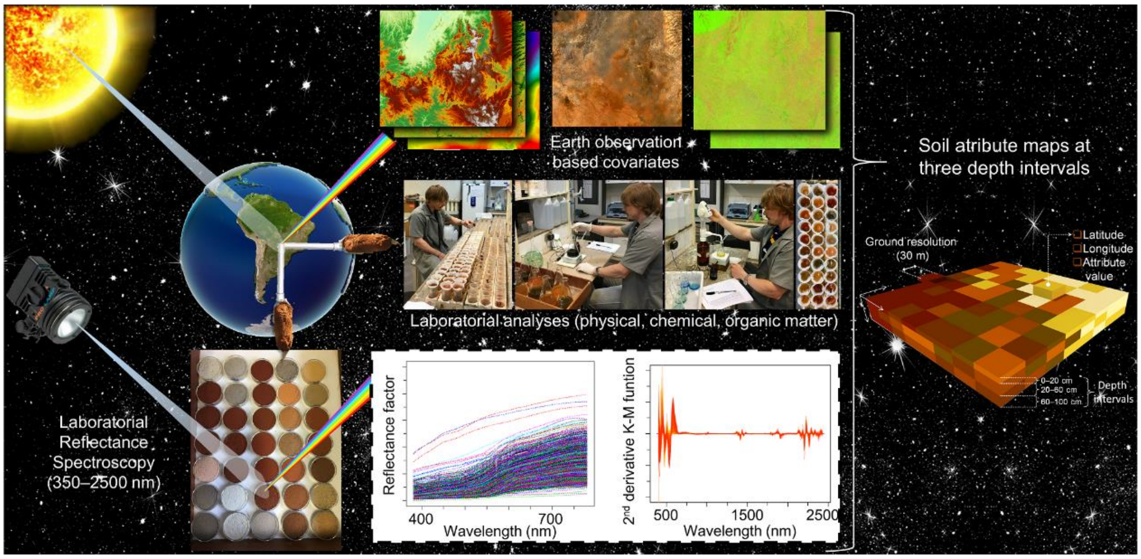
CHAPTER 3 — 250 M-GRIDDED SOIL TEXTURE AT MULTIPLE DEPTHS OF MIDWEST BRAZIL-----99
1. TABLE OF SPECIFICATIONS ----- 100
2. VALUE OF THE DATA----- 101
3. DATA----- 101
4. MATERIALS AND METHODS----- 103
5. REFERENCES ----- 104

GENERAL CONCLUDING REMARKS AND FUTURE OUTLOOK----- 105

General Abstract

The Midwest region in Brazil has the largest and most recent agricultural frontier in the country, where there is no currently detailed soil information to support the land use intensification. Producing large-extent digital soil maps is resource intensive. We aimed to use pedometric techniques coupled with proximal and remote sensing data to produce digital maps with 30 m resolution of key soil attributes at topsoil and subsoil for 851,000 km^2 of Midwest Brazil. For mapping key soil attributes we used multi-resolution covariates based on Earth observations: we produced composites of bare topsoil reflectance and potential natural vegetation reflectance using Landsat time series, which were coupled with terrain attributes, geologic and climate variables to capture short and long-range soil spatial patterns. We acquired soil data from observations at 0–20, 20–60 and 60–100 cm (rooting) depth intervals containing soil attributes, which are commonly used (as key criteria) for soil interpretations: clay, silt and sand, organic matter, pH, aluminum and base saturation. We also determined both the soil color in Munsell notation and the relative abundance of minerals in soil (hematite, goethite, kaolinite, gibbsite and 2:1 clay minerals) from laboratorial reflectance spectra (350–2500 nm). We fitted and validated optimal models for the spatial patterns of each soil attribute at topsoil and subsoil using Random Forest regression and 10-fold cross validation in R software. We identified the covariates that were most relevant to describe the soil spatial patterns in the study area. We mapped the spatial distribution of soil attributes at 30 m resolution for the 0–20, 20–60 and 60–100 cm depth intervals using the optimized models and Google Earth Engine. We made publicly available for download (GeoTIFF) at 250 m resolution the predicted soil maps of clay, silt and sand of the study area. We concluded that physical and chemical soil attributes, as well as soil color and mineralogy derived from spectra at multiple depth intervals, can be mapped using Earth observations data and machine learning methods with good performance.

Keywords: *Soil Science, big data, data mining, machine learning, cloud-computing*


Earth observation based covariates
Soil attribute maps at three depth intervals
Ground resolution (30 m)
Latitude
Longitude
Attribute value
Laboratorial analyses (physical, chemical, organic matter)
0–20 cm
20–60 cm
60–100 cm
Depth intervals
Laboratorial Reflectance Spectroscopy (350–2500 nm)
Reflectance factor
400
700
Wavelength (nm)
2nd derivative K-M funtion
500
1500
2500
Wavelength (nm)

GENERAL GRAPHICAL ABSTRACT

RESUMO GERAL (PORTUGUESE)

A região Centro-Oeste do Brasil tem a maior e mais recente fronteira agrícola do país, onde atualmente não há informações detalhadas sobre o solo para apoiar a intensificação do uso do solo. A produção de mapas de solo digitais de grandes extensões é intensiva em recursos. O principal objetivo desta pesquisa foi usar técnicas pedométricas acopladas a dados de sensoriamento proximal e remoto para produzir mapas digitais com resolução de 30 m dos principais atributos do solo em superfície e subsuperfície para 850.000 km^2 do Centro-Oeste do Brasil. Para mapear os principais atributos do solo, utilizamos covariáveis multi-resolução baseados em dados de observações da Terra: produzimos imagens compostas de refletância do solo exposto e de refletância da vegetação natural potencial usando séries temporais Landsat, que foram acoplados com atributos do terreno, variáveis geológicas e climáticas para capturar padrões espaciais do solo de curto e longo alcance. Adquirimos dados do solo a partir de observações em intervalos de profundidade (enraizamento) de 0–20, 20–60 e 60–100 cm, contendo atributos do solo que são comumente usados (como critério chave) para interpretações do solo: argila, silte e areia, matéria orgânica, pH, saturação de bases e de alumínio. Também determinamos a cor do solo em notação de Munsell e a abundância relativa de minerais no solo (hematita, goetita, caulinita, gibbsita e minerais de argila 2:1) a partir de espectros de laboratório (350–2500 nm). Foram ajustados e validados modelos ótimos para os padrões espaciais de cada atributo do solo em superfície e subsuperfície, usando regressão Random Forest e validação cruzada no software R. Identificamos as covariáveis mais relevantes que descreveram os padrões espaciais do solo na área de estudo. Mapeamos a distribuição espacial dos atributos do solo com resolução de 30 m para os intervalos de profundidade de 0-20, 20-60 e 60-100 cm usando os modelos otimizados e a plataforma Google Earth Engine. Disponibilizamos publicamente para consulta (GeoTIFF), com resolução de 250 m, os mapas de solo preditos de argila, silte e areia da área de estudo. Concluímos que atributos físicos e químicos do solo, assim como também a cor e a mineralogia do solo derivados de espectros de refletância, obtidos em múltiplos intervalos de profundidade, podem ser mapeados usando dados de observação da Terra e métodos de aprendizagem de máquinas com bom desempenho.

Palavras-chave: *Ciência do solo, big data, mineração de dados, aprendizado de máquina, computação em nuvem.*

1. INTRODUCTION

The word "soil," like many common words, has several meanings. In its traditional meaning, soil is the natural medium for the growth of land plants, whether or not it has discernible soil horizons (SOIL SURVEY STAFF, 2014). Soil is a natural body comprised of solids (minerals and organic matter), liquid, and gases varying in three dimension, that form as a result of the interaction of soil forming factors (JENNY, 1941). The repeating patterns formed by the soil across landscape allow soil scientists to develop quantitative models for digital soil mapping (DSM) (MCBRATNEY et al., 2003).

The soil plays essential role for natural and anthropic ecosystems (BÜNEMANN et al., 2018). Reliable spatial soil information can improve natural capital assessment, becoming important for food production, in large countries or emerging economies where the major demographic growth is expected (UNITED NATIONS, 2019). Conventional soil mapping is resource intensive and takes several years to perform adequate maps for large extents. This can be observed for Brazil, which is covered by soil class maps with small scales, mostly developed by Brazilian government institutions using RADAMBRASIL (Radar on Amazon and Brazil) project data (1:1,000,000 or nominally 2 km) (MCBRATNEY et al., 2003). In this case, such maps are not capable of supporting any decision making at regional or local scales.

At the moment of this thesis, Gomes et al. (2019) performed one of the few studies on pedometric (or quantitative) mapping for a large extent in Brazil, where they mapped organic carbon stocks. When we searched for studies on soil color mapping, currently, Viscarra Rossel et al. (2010) performed one of the few works, where the authors also mapped iron oxides of Australian soil using reflectance spectra (350–2500 nm) and geostatistics. Conversely, we found some studies on soil mineralogy mapping from laboratorial spectra between 350 and 2500 nm wavelengths (MALONE et al., 2014; MULDER et al., 2013; ROSSEL et al., 2011; VISCARRA ROSSEL, 2011). Other studies (ROBERTS et al., 2019; DUCART et al., 2016; MADEIRA NETTO et al., 1997) used enhanced mineral mapping techniques to produce a thematic mineral map of soil using the spectral response of Landsat imagery.

Midwest Brazil has the largest and most recent agricultural frontier of the country (PARENTE et al., 2019), which contributes about 34% and more than 10% to the agricultural production and gross domestic product of the nation, respectively (IBGE, 2018). Nevertheless, the spatial patterns of soil attributes, such as physical, chemical, mineralogical and the color, under current conditions remains not mapped at fine resolution for this region and most Brazilian soils, both qualitatively and quantitatively. One reason for that might be the lack of dissemination of DSM knowledge to the community, because it is a multidisciplinary technique that involves soil knowledge, statistics, and mathematics applied to geoinformation science to understand soil variability across landscape (DALMOLIN et al., 2020). Another reason might be that in Brazilian repositories (SAMUEL-ROSA et al., 2020) there is a lack of mineralogical data, possible due to traditional methods are resource intensive (MULDER et al., 2013). Soil color in national datasets (SAMUEL-ROSA et al., 2020) also lacks or do not contain spatial referencing or was visually approximated (MARQUES et al., 2019). Furthermore, several key soil forming factors are still not fully represented by classical environmental covariates, being necessary to develop new covariates that provide improved proxies for describing topsoil and subsoil spatial patterns.

Advances in Earth observation (satellite images and products), digital elevation models and DSM frameworks, coupled with data mining and cloud-based computing, might be a solution to the lack of adequate soil data (HENGL et al., 2018). Satellite image provides measurements of topsoil reflectance, which are directly related to clay content, organic matter, mineralogy, moisture and soil color (STENBERG et al., 2010). Some studies has been shown that topsoil spectral patterns are related to the subsoil pattern variations and dynamic processes which occurs within the soil profile (MENDES et al., 2019; POPPIEL et al., 2018, 2019). The synergy between satellite images, especially bare soil composites, and auxiliary covariates (e.g. elevation, climate) coupled with machine learning for DSM was reported in several studies (MENDES et al., 2019; ROGGE et al., 2018; DEMATTÊ et al., 2018; FONGARO et al., 2018; HENGL et al., 2017; DIEK et al., 2017, 2016). These techniques reduce overall soil data production costs, combining a number of sciences (PADARIAN et al., 2019; DEMATTÊ et al., 2018; HENGL et al., 2018; MCBRATNEY et al., 2003). Furthermore, environmentally clean, quick and low cost techniques such as reflectance spectroscopy (350–2500 nm) was successfully used in pedometry to determine the color and mineralogy of soil (SIMON et al., 2019; MARQUES et al., 2019; RIZZO et al., 2016; SCHEINOST et al., 1998; MATTIKALLI, 1997; ESCADAFAL et al., 1988; FERNANDEZ et al., 1987). Random Forest (RF) is one of most popular algorithm in DSM, being employed in several soil mapping studies (AMIRIAN-CHAKAN et al., 2019; GOMES et al., 2019; HENGL et al., 2015, 2018; KESKIN et al., 2019; LOISEAU et al., 2019; NUSSBAUM et al., 2018M; A et al., 2017).

Revealing the spatial patterns of key soil attributes at multiple (rooting) depth increments might provide adequate information to account for the multi-functionality of soil in Midwest Brazil.

2. HYPOTHESIS

- Earth observation data based covariates can describe topsoil and subsoil spatial patterns over a large geographical extent in Midwest Brazil;
- Proximal soil sensing data (350–2500 nm) have potential to provide accurate information on soil color and mineralogy;
- Earth observation data based covariates and machine learning, coupled with soil observations can promote a favorable framework to produce accurate soil predictions.
- It is possible to map physical and chemical soil attributes, and also the soil color and mineralogy at three fixed (rooting) depth intervals (0–20, 20–60 and 80–100 cm) with 30 m resolution across the Midwest region in Brazil, using open source software.

3. OBJECTIVES

3.1. General objective

The main objective of this research was to use pedometric techniques coupled with proximal and remote sensing data to produce digital maps with 30 m resolution of key soil attributes at topsoil and subsoil for 851,000 km^2 of Midwest Brazil.

3.2. Specific objectives

I. To map physical and chemical soil attributes with 30 m resolution at multiple depth intervals in Midwest brazil;

II. To map the soil color and mineralogy using proximal and remote sensing data at three depth intervals in Midwest Brazil;

III. To make publicly available for download as integer GeoTIFF format at 250 m resolution the soil texture maps of Midwest Brazil.

4. GENERAL FRAMEWORK

To achieve the objectives of this research, we followed the working steps described below and represented in the general methodology flowchart.

4.1. Working steps

I. Assessment of the pedomorphogeological characteristics across different areas centralized over Goiás State, based on legacy data (maps and soil observations) and satellite images, for size definition of the study area.

II. Production of new covariates (composites) by Landsat data mining (30 m resolution) in Google Earth Engine for better representation of some soil forming factors: bare topsoil reflectance and potential natural vegetation reflectance. Interpolation of the gaps by ordinary kriging after fitting the semivariogram.

III. Acquisition of traditional covariates for soil mapping (relief and climate), and selection by the least correlated between them. Adjustment of the pixel size to 30 m.

IV. Exploration of soil datasets and acquisition of soil observations at topsoil and subsoil from Brazilian databases. Evaluation of the datasets to identify the need for new soil observations in the study area.

V. Visiting and collection sites planning according to pedomorphogeological characteristics of the region, adjusted with new detailed covariates.

VI. New soil samples preparation for traditional (wet) determination of physical and chemical attributes, and for reflectance measurements (spectroscopy) between 350 and 2500 nm, both in laboratory.

VII. Calculation of physical and chemical attributes values in their specific units, and reflectance spectra pre-processing for splice and noise bands removing.

VIII. Determination of the soil color in Munsell notation, and the relative abundance of minerals in soil, both from laboratorial spectra (350–2500 nm);

IX. Preparation of the final soil datasets by checking and aggregating data into depth intervals for soil modelling: physical and chemical soil dataset, and soil color and mineralogical dataset.

X. Fitting, optimization and validation of regression models for the spatial patterns of each soil attribute at topsoil and subsoil using Random Forest algorithm and 10-fold cross validation.

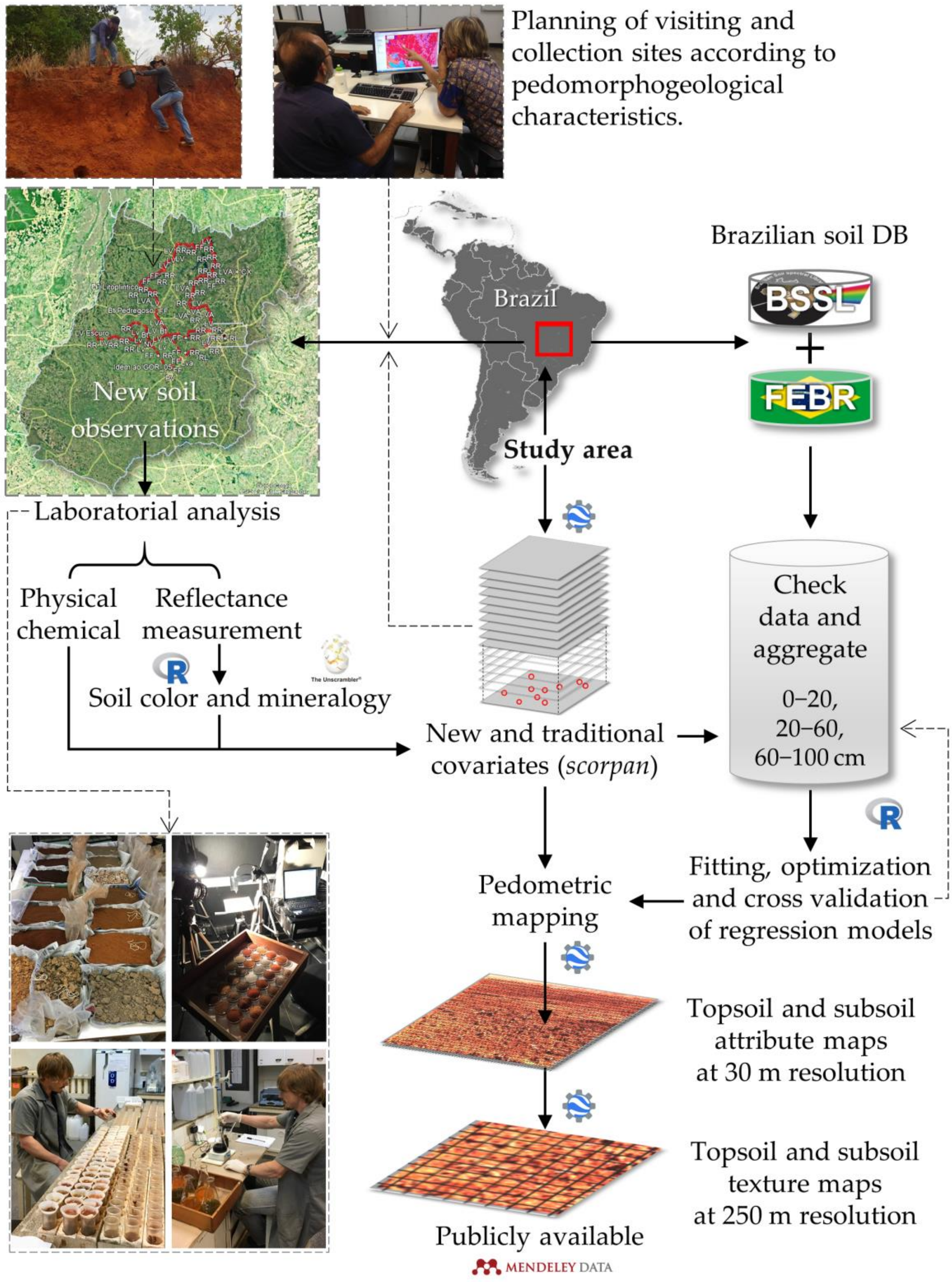
XI. Identification of the most relevant covariates for describing the soil spatial patterns in the study area.

XII. Mapping the spatial distribution of soil attributes at 30 m resolution for the 0–20, 20–60 and 60–100 cm depth intervals using optimized models in Google Earth Engine.

XIII. Clustering of lithologies using the maps of physical and chemical soil attributes, and verification of the spatial correspondence of predicted values with parent materials.

XIV. Verification of the spatial correspondence (using Pearson correlation) between the predicted maps with legacy soil observations acquired from a national dataset, and weathering degree and hue, both inferred from a legacy soil class map of the study area.

XV. Reduction of the spatial resolution at 250 m, tailing and making publicly available the predicted soil maps of clay, silt and sand for the study area.


Planning of visiting and collection sites according to pedomorphogeological characteristics.
Brazilian soil DB
Brazil
BSSL
+
FEBR
New soil observations
Study area
Laboratorial analysis
Physical chemical
Reflectance measurement
Soil color and mineralogy
The Unscrambler
New and traditional covariates (scorpan)
Check data and aggregate
0–20, 20–60, 60–100 cm
Fitting, optimization and cross validation of regression models
Pedometric mapping
Topsoil and subsoil attribute maps at 30 m resolution
Topsoil and subsoil texture maps at 250 m resolution
Publicly available
MENDELEY DATA

General Methodology Flowchart.

4.2. The thematic project that supports this thesis

This thesis is part of the thematic project entitled "*Geotechnologies on a Detailed Digital Soil Mapping and the Brazilian Soil Spectral Library: Development and Applications*", founded by the São Paulo Research Foundation (FAPESP), grant number 14/22262-0, and coordinated by Professor José Alexandre Melo Demattê from University of São Paulo (USP), College of Agriculture Luiz de Queiroz (ESALQ), Departamento of Soil Science, Piracicaba, SP. The project is co-coordinated by Professor Marcos Rafael Nanni from

Maringá State University (UEM), and by Professor Marilusa Pinto Coelho Lacerda from FAV/UnB.

The thematic project aims to act in many soil knowledge fields through geotechnologies at laboratorial, field, aerial and orbital acquisition levels. This will be organized into 6 projects as follows: 1) Brazilian Soil Spectral library; 2) Satellite images on soil mapping; 3) Geotechnologies for in-situ soil mapping; 4) Stratigraphy on soil mapping; 5) Geotechnology for digital soil mapping. The first theme aims the implementation of a Brazil's soil spectral library. Soil reflectance spectra will be obtained and organized from the broadest attainable regions. The objective is to develop a database with soil patterns via spectra and make accessible to the community, to support future studies.

The second theme regards to the development of a detailed soil map across the municipality of Piracicaba in São Paulo State, by the integration of geotechnologies. In order to accomplish this goal, many subprojects will be performed. The first will be the compartmentalization of stratigraphic surfaces according to their geological and geomorphological features. Composites from reflectance satellite images will be displayed in three dimensional software for observing and relating the bare topsoil to the alndscape. Hyperspectral images will be also acquired using a sensor with 620 spectral bands on a plane, providing detailed spectral patterns of soils. Studies on photopedology using 3D software will be performed.

Within the compartments, several transect points will be allocated along the study area, and soil samples will be collected with auger and from soil profiles. After fieldworks, soil samples will be prepared for determination of physical and chemical soil attributes, and for measurements of reflectance spectra, both in laboratory. The results will be used to produce a soil database. Three pilot areas within each compartment will be selected based on soil database and environmental variables, where a detailed in-situ soil characterization will be performed using diverse geotechnologies: field spectroradiometer, colorimeter, GPS, gammaspectrometer, computer and open source software. Such tools will provide information for soil interpretation at real time, allowing the pedologist decision.

At the final, an approach based on a mix of traditional and geotechnological methods will be applied for soil mapping. Two different products are expected as a result from the thematic project: a soil spectral library publicly available and a detailed soil map assisted by a new geotechnological methodology. In addition, detailed maps of soil attributes for different regions of Brazil are expected as results.

5. LITERATURE REVIEW

5.1. What is Pedometrics and Digital Soil Mapping?

Pedometrics is a new expression derived from the Greek words *pedos* (meaning soil) and *metron* (meaning measurement), defined by the Pedometric society (www.pedometrics.org) as "the application of mathematical and statistical methods for the study of the distribution and genesis of soils". Nevertheless, it will always intergrade to all areas of soil science and quantitative methods.

DSM is defined as "the creation and population of spatial soil information systems by the use of field and laboratory observational methods coupled with spatial and non-spatial soil inference systems" (LAGACHERIE et al., 2007). Other terminology has also been used or proposed, including: computer-assisted soil cartography, numerical soil cartography, pedometric mapping, environmental correlation, predictive soil mapping, or geographical extrapolation using models (MINASNY et al., 2016).

McBratney et al. (2003) formalized the DSM framework, which started prior to the 21st century, as *scorpan* based on Jenny's *clorp*t model (JENNY, 1941) of soil formation, where the acronym *scorpan* stands for soil (s), climate (c), organisms (o), relief (r), parent material (p), age or time (a) and spatial position (n). This updated equation provides a spatial model to express quantitatively the relationship between a soil attribute or class and environmental variables, for a given spatial location (WADOUX et al., 2020).

Pedometric mapping is generally characterized as a quantitative, (geo)statistical production of soil geoinformation, also referred to as the predictive soil mapping (SCULL et al., 2003) or DSM (MCBRATNEY et al., 2003), as it depends heavily on the use of information technologies. Pedometric mapping, however, specifically means that quantitative methods are used in the production of soil geoinformation. The most recent topics covered by Pedometrics include: analysis and modelling of spatial and temporal variation of soil attributes; multi-resolution data integration; soil-landscape modelling using digital terrain analysis; quantitative soil classification algorithms; soil genesis simulation; soil pattern analysis; design and evaluation of sampling schemes; incorporation of exhaustively sampled information (remote sensing) in soil mapping; precision agriculture applications, among others (HENGL et al., 2019).

5.2. Recent advances in Pedometric Mapping

In recent years, the advance of DSM was due to several factors, mainly the accessibility of Landsat images and digital elevation models, as well as to the availability of computing power to process big data, the development of data mining tools and progress in open source geospatial software, applications beyond geostatistics, and institutional rejuvenation with a new generation of soil scientists that were attracted by the spatial analysis of soils (MINASNY et al., 2016). Other important factor is the current strong demand for soil maps by agricultural interest (food, feed, fuel) that brought soils back onto the global research agenda (OMUTO et al., 2013).

The recent advances cited above agreed with Figure 1, where the number of citations from keyword search “digital soil mapping”, coupled with “Remote Sensing”, “Random Forest regression” and “Pedometrics” on scientific papers increased from the year 2010 to present, mainly due to advances in soil data availability (MINASNY et al., 2016), such as from Earth observations and the popularization of soil spectral libraries (DEMATTÊ et al., 2019; VISCARRA ROSSEL et al., 2016) to produce new soil information. According to McBratney et al. (2019), we are going into the global mapping era (2015 onwards), aiming to make a global digital soil map at fine resolution.

Although the term pedometrics was coined by McBratney in 1986, and that quantitative methods have been increasing ever since (MCBRATNEY et al., 2019), the number of citations from keyword “pedometrics” on papers still remains small (Figure 1), possibly due to a more recent dissemination of the term between the scientific community. In a bibliometric study of the composition of papers on soil science, from its inception in 1967 until 2001, Hartemink et al. (2001) showed that papers on pedometrics have risen from less than 3% in 1967 to around 18% of all papers in 2000. Furthermore, very few pedometrics-related papers in high impact journals on soil and Earth sciences describe the information technology used, like computer hardware, algorithms, sensors, models, etc. (ROSSITER, 2018). Thus, from the nature of the algorithms and datasets we can infer something about the information technology needed for the reported studies.

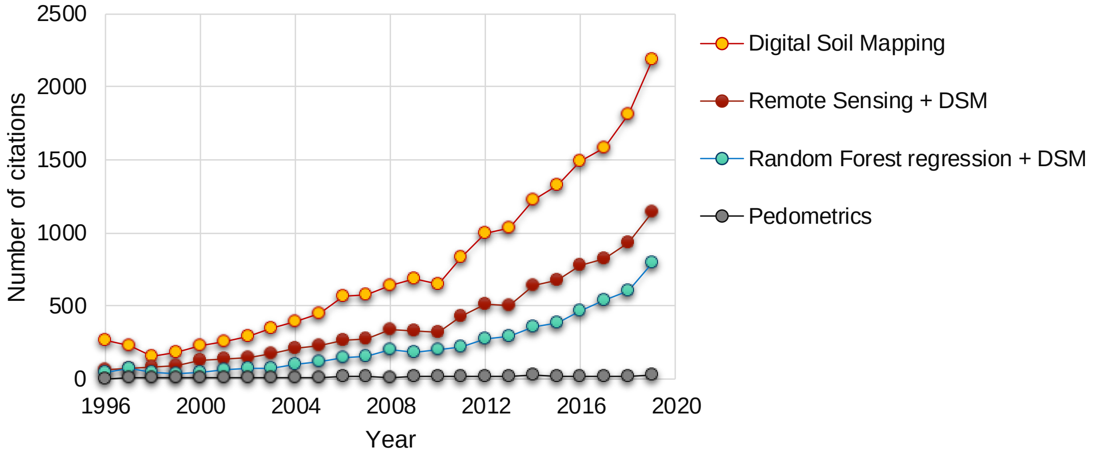


Figure 1. Number of citations from keyword search on scientific papers from the ScienceDirect database (data extracted in February 2020). DSM: digital soil mapping.

5.3. What is Google Earth Engine?

Google Earth Engine (GEE), established towards the end of 2010, is a cloud-based platform for planetary-scale geospatial analysis that brings Google's massive computational capabilities to bear on a variety of high-impact societal issues (GORELICK et al., 2017). The main components of GEE are 1) Datasets: a petabyte-scale archive of publicly available remotely sensed imagery and other terrain, climatic, geophysical data available at https://developers.google.com/earth-engine/datasets; 2) Compute power: Google's computational infrastructure optimized for parallel processing (high-performance) of geospatial data; 3) API: Application Program Interfaces (focused on JavaScript) for making requests to the Earth Engine servers, where documentation (Docs) contains links to sections or pages about important data types (or objects) such as `ee.Image()`, `ee.ImageCollection()`, `ee.Feature()`, `ee.FeatureCollection()`, `ee.Geometry()`, etc., and methods (or functions) such as `filter()`, `clip()`, `ee.Algorithms.Terrain()`, `Export.image.toDrive()`, etc. Generally, a method is applied to an object, as follow `ee.Image().clip()`. And 4) Code Editor: an online Integrated Development Environment (IDE) for rapid prototyping and visualization of complex spatial analyses using the JavaScript API, available at https://code.earthengine.google.com. All this information and much more details can be found in the GEE User Guides (https://developers.google.com/earth-engine). JavaScript is an interpreted programming language that is most well-known as the scripting language for Web pages (wikipedia.org).

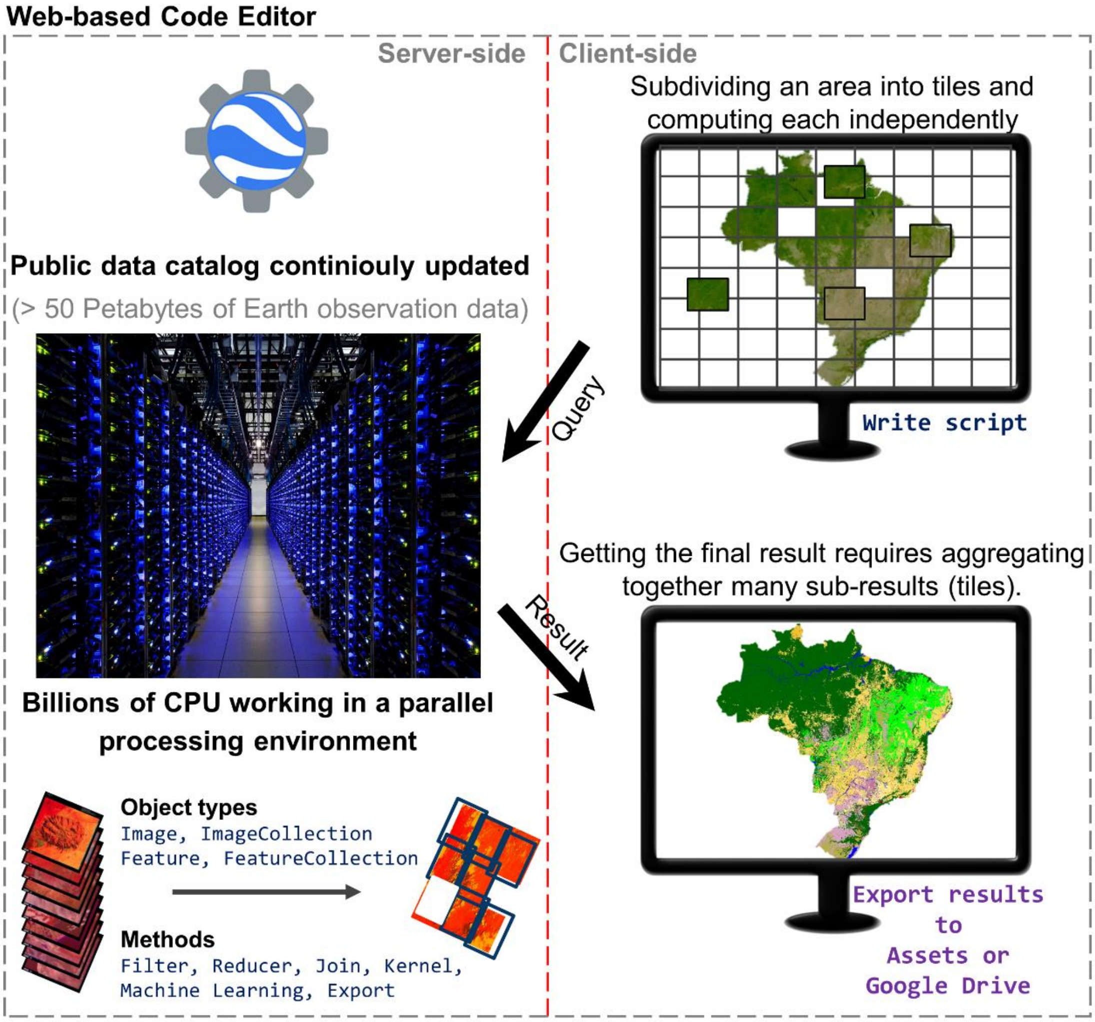

Web-based Code Editor
Server-side
Client-side
Subdividing an area into tiles and computing each independently
Public data catalog continiouly updated
(> 50 Petabytes of Earth observation data)
Query
Write script
Getting the final result requires aggregating together many sub-results (tiles).
Result
Billions of CPU working in a parallel processing environment
Object types
Image, ImageCollection
Feature, FeatureCollection
Methods
Filter, Reducer, Join, Kernel, Machine Learning, Export
Export results to Assets or Google Drive

Figure 2. A simplified system architecture diagram of the Google Earth Engine Code Editor. Adapted from Earth Engine educational resources available at https://developers.google.com/earth-engine/edu.

Gorelick et al. (2017) and Padarian et al. (2015) described "Cloud computing" as if you have access to a supercomputer designed for geospatial analysis. All the hard work is done on Google servers, and to get the result is low bandwidth consuming. How does it work? The code you write in the Code Editor using JavaScript programming language (client-side) gets turned into an object representing the set of instructions which is then sent to Google (server-side) for processing (Figure 2). The requested analysis is then run in parallel on many computers or central processing units (CPUs). What you get back in your browser (or monitor) is only what you request, for example a statistic or chart printed to the console or small RGB tiles to display on the map.

5.4. Google Earth Engine for pedometric mapping

At the moment, most pedometric mappings take place on local computers using local processors, which becomes resource intensive because require massive amounts of processing power and memory for spatial big data analyses (HENGL et al., 2019). Kumar et al. (2018) found that, of the total number of research papers published between 2010 and 2017 using GEE, less than 3% were applied to DSM, as observed by the red line in Figure 3. Nevertheless, GEE is an interesting new platform which could be implemented in a routine DSM workflow, since it is specifically designed to manage large volumes of gridded information (rasters), which are used to represent soil forming factors (PADARIAN et al., 2015). The major advantage of using GEE is that many rasters are already available at catalogue, making easier the process of collecting data, and the parallel nature of its algorithms, which considerably accelerates the computation times (GORELICK et al., 2017). The user can also upload his own dataset (raster or vectorial) to the assets with 250 Gigabytes of space, and couple them with GEE datasets for processing.

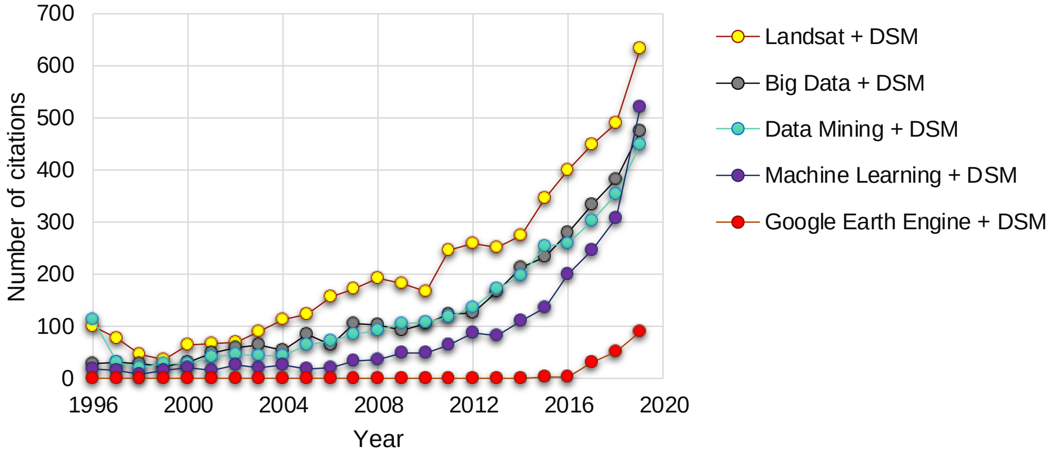


Figure 3. Number of citations from keyword search on scientific papers from the ScienceDirect database (data extracted in February 2020). DSM: digital soil mapping.

GEE Code Editor is accessed and controlled through an internet-accessible API and an associated web-based IDE that enables rapid prototyping and visualization of results (Figure 4). Users can access and analyze data from the public catalog as well as their own private data (stored in assets) using a library of operators provided by the Code Editor. These operators are implemented in a large automatic parallelization that enables global-scale analyses such as that by Hansen et al. (2013), Pekel et al. (2016) and Murray et al. (2019). Padarian et al. (2015) explored the feasibility of using this platform for DSM by presenting two soil mapping examples over the contiguous United States. The authors specified that, independent of the extent and aim, pedometric mapping usually has a standard workflow, which can be successfully implemented and improved within GEE:

a) Environmental variables:
    - Acquisition: many covariates are already available in the GEE catalogues, which can be imported into the code editor and directly apply calculation of derivatives, up or downscale to a target grid resolution, spatial interpolation, filtering, subset and stack rasters, etc.
    - Development: based on previous studies (DEMATTÊ et al., 2018; DIEK et al., 2017; ROBERTS et al., 2019; ROGGE et al., 2018), we highlight the strong potential that GEE has for big data mining (GORELICK et al., 2017), and thus to obtain new spatially continuous soil predictors based Earth observation data for their use as soil forming factor proxies in pedometric mappings of large geographical extents.

b) Sampling: the user can upload points to their assets, using tables (.csv) or vectors (shapefiles), overlay them with covariates for sampling data, and thus prepare a value matrix.

c) Modeling: the available algorithm within GEE covers both prediction of class (categorical) and continuous variables (soil attributes), such as linear regression, regression kriging, neural networks and Random Forests. The platform is not flexible enough to build and tuning models. Nevertheless, in our experience, sampled data (described in item b) can be exported and used within a geostatistical software, like R, to optimize the models. Then, tuned hyperparameters from optimized models can be used within GEE algorithms for spatial predictions.

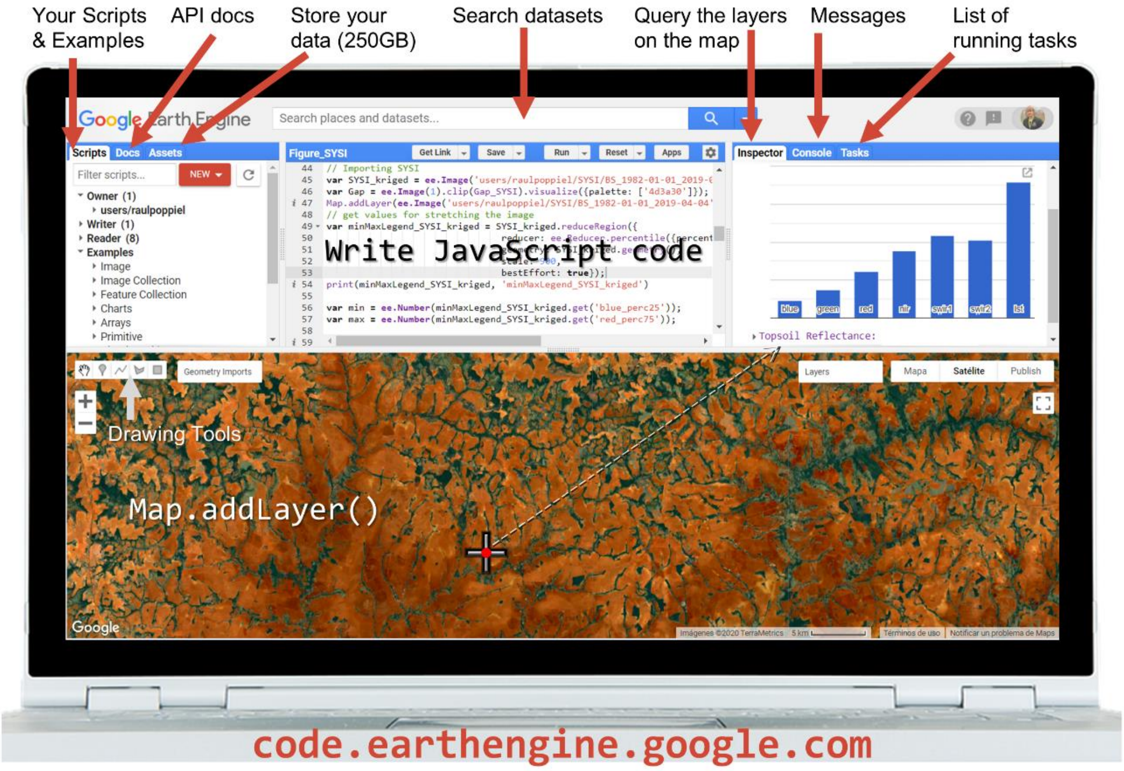
d) Mapping: spatial predictions are virtually made by tiles (sub-areas) in a parallel process for the whole extent. This is the major advantage of GEE, bringing speed gain to this step. According to Padarian et al. (2015), DSM in GEE was 40–100 times faster than DSM using desktop workstation.

e) Uncertainty: unfortunately, techniques for uncertainty estimating are not implemented in GEE, and there is no straightforward way to program it at the current development stage. GEE is in active development and constantly being updated.

f) Displaying: GEE uses the scale specified by the output (zoom level) to determine the appropriate level of the image pyramid to use as input by aggregating 2 x 2 blocks of 4 pixels. That is, each tile is always 256 x 256 pixels, and it is recalculated with every change in zoom level for visualization.

g) Map layout: code editor allows to design the traditional layouts by coding each element: frame, scale and gradient bars, legend, north, text and map background.

h) Distribution: the code, the soil observations used to obtain the maps, and also the results obtained and stored in assets can be shared with other researcher. Results can also be exported to Google cloud storage (and share as a web map service) or Google Drive (for downloading).



```
// Importing SYSI
var SYSI_kriged = ee.Image('users/raulpoppiel/SYSI/BS_1982-01-01_2019-0
var Gap = ee.Image(1).clip(Gap_SYSI).visualize({palette: ['4d3a30']});
Map.addLayer(ee.Image('users/raulpoppiel/SYSI/BS_1982-01-01_2019-04-04'
// get values for stretching the image
var minMaxLegend_SYSI_kriged = SYSI_kriged.reduceRegion({
                          reducer: ee.Reducer.percentile({percent
                          geometry: SYSI_kriged.geometry(
                          scale: 900,
                          bestEffort: true});
print(minMaxLegend_SYSI_kriged, 'minMaxLegend_SYSI_kriged')

var min = ee.Number(minMaxLegend_SYSI_kriged.get('blue_perc25'));
var max = ee.Number(minMaxLegend_SYSI_kriged.get('red_perc75'));
```

Figure 4. Google Earth Engine Code Editor interface and their components. API: Application Program Interfaces (focused on JavaScript) for making requests to the Earth Engine servers. Adapted from a screenshot by the author.

5.5. *Machine learning for pedometric mapping*

Machine learning (ML) was implemented in the 1990s as a tool for DSM (LAGACHERIE, 2008). ML techniques refer to a large class of non-linear data-driven algorithms employed primarily for data mining and pattern recognition purposes, and now frequently used for regression and classification tasks in all fields of science. ML algorithms do not make an assumption of the observations' distribution, unlike geostatistical methods, where transformation of the original values is often required to satisfy the assumptions. ML algorithms can also handle a large number of cross-correlated covariates as predictor.

In parallel, there has been a tremendous increase in the production and availability of regional and global soil databases. For example, the Soil and Terrain Digital Database (OLDEMAN et al., 1993), the World Soil Information Service (BATJES et al., 2017) and the World Spectral Library (VISCARRA ROSSEL et al., 2016) at global extent, and the Free Brazilian Repository for Open Soil Data (SAMUEL-ROSA et al., 2020) and the The Brazilian Soil Spectral Library (DEMATTÊ et al., 2019) at national extent, among others. Additionally, numerous spatially exhaustive *scorpan* covariates are available at global extent (mostly available within GEE data catalogue at https://developers.google.com/earth-engine/datasets). Conventional regression techniques seem, to some extent, outdated to deal with the increased complexity of soil datasets. This justifies the increasing use of machine learning algorithms for digital soil mapping (WADOUX et al., 2020).

There is a large number of studies using ML for mapping soil attributes from local, regional to global extent. For example, Pouladi et al. (2019) make a quantitative map over a 10 ha field in Denmark while Hengl et al. (2017) produce quantitative and categorical maps for the whole world. Wadoux et al. (2020) found that most of studies (70%) only predict topsoil attributes, while a smaller number of works mapped soil attributes at topsoil and subsoil (ARROUAYS et al., 2014; HENGL et al., 2015; VISCARRA ROSSEL et al., 2015; GOMES et al., 2019).

ML algorithms have been successfully applied for pedometric mapping of various soil attributes, such as soil organic carbon concentration (GOMES et al., 2019; HENDERSON et al., 2005; POULADI et al., 2019; SIEWERT, 2018), soil texture (FONGARO et al., 2018; LIU et al., 2020; LOISEAU et al., 2019), bulk density (VISCARRA ROSSEL et al., 2015), pH (DHARUMARAJAN et al., 2017), or cation exchange capacity (FORKUOR et al., 2017). ML also was used for mapping nutrients such as nitrogen (FORKUOR et al., 2017; VISCARRA ROSSEL et al., 2015), phosphorus (HENGL et al., 2017; VISCARRA ROSSEL et al., 2015), potassium, calcium or magnesium (HENGL et al., 2017).

Environmental covariates represent soil forming factors, where studies with ML used from less than five (PADARIAN et al., 2019) to more than 100 predictor (HENGL et al., 2017; RAMCHARAN et al., 2018). Some examples of studies using multi-resolution covariates for mapping with ML algorithms were performed by Behrens et al. (2010), Miller et al. (2015) and Behrens et al. (2018). For instance, Miller et al. (2015) employed 412 covariates, several of which were derived from the aggregation of terrain attributes from a fine elevation map (grid cell size of 2 m X 2 m).

5.6. Why perform pedometric mapping of soil attributes?

The soil covers the Earth, and its attributes vary spatially and temporally in a sometimes continuous, sometimes discrete and sometimes complex or random way (MCBRATNEY et al., 2003). The attributes can either be direct or indirect (proximally or remotely sensed, including humanly sensed) measurements of physical, chemical, biological, morphological and mineralogical soil features in the field or on samples taken back to the laboratory or multivariate soil classes derived from them (MCBRATNEY et al., 2018). The word “attribute” is commonly used in soil science because the features are *attributed* to the soil and to their formation (SOIL SURVEY STAFF, 2014).

In contrast to the quartz sand coarse soil fraction, the finer soil fraction of the highly weathered soil mantle of Midwest Brazil (SCHAEFER et al., 2008) is dominated by minerals that protect soil organic matter from mineralization by microorganisms through sorption and/or entrapment of organic matter in small microaggregates (BALDOCK et al., 2000; CHENU et al., 2006). Among these minerals, sesquioxides such as gibbsite and goethite, have a greater affinity for organic matter than clay minerals, thanks to their large specific surface area (KAISER et al., 2003). These minerals also constrain the nutrient status in soils with low pH through their ability to remove inorganic anions, such as phosphate or nitrate, from solution, forming inner or outer sphere complexes (SCHWERTMANN et al., 1989). Together with these very stable sesquioxides, kaolinite 1:1 clay, is the most abundant mineral of the fine soil fraction in highly weathered soils (KAISER et al., 2003). Despite being less reactive than sesquioxides, kaolinite can present an important anion exchange capacity at low pH (MELO et al., 2001), contributing to lower extractable phosphorus concentrations traditionally used as an indicator of readily plant accessible P (GÉRARD, 2016; MCGRODDY et al., 2008). Thus, the same clays and sesquioxides that lead to greater soil organic matter storage also occlude P into less accessible forms (RUTTENBERG et al., 2011), reducing its mobility and making uptake of P resources from the soil more difficult. Phosphorus is an essential element determining plant growth and productivity (MALHOTRA et al., 2018), that was found to be spatially correlated with other yield-limiting factors (soil attributes) (NAWAR et al., 2017).

In a recent scientific report, Soong et al. (2020) demonstrates how variation in soil attributes that retain carbon and nutrients can help to explain variation in tropical forest growth and mortality (Figure 5). The authors observed strong positive relationships between soil attributes (soil texture and mineralogy) and forest dynamics of growth and mortality across tropical forests in a phosphorus-poor region of the Guiana Shield in South America. Average tree growth increased from 0.81 to 2.1 mm yr^{-1} along a soil texture gradient from 0 to 67% clay, and increasing metal-oxide content. Topsoil (30 cm depth) organic carbon stocks ranged from 30 to 118 tons C ha^{-1}, phosphorus content ranged from 7 to 600 mg kg^{-1} soil, and the relative abundance of arbuscular mycorrhizal fungi ranged from 0 to 50%, all positively correlating with soil clay and sesquioxides content. In contrast, already low extractable phosphorus content decreased from 4.4 to <0.02 mg kg^{-1} in soil with increasing clay content. A greater prevalence of arbuscular mycorrhizal fungi in more clayey forests that had higher tree growth and mortality, but not biomass, indicates that despite the greater investment in nutrient uptake required, soils with higher clay content may actually serve to sustain high tree growth in tropical forests by avoiding phosphorus losses from the ecosystem.

Findings as cited are very important to soil scientists, because it supports that continuous and quantitative soil attribute maps could be used to predict (or to indicate) which soils are expected to be responsive to nutrient additions based on their soil attributes. In other words, to indicate the best soil management or places for food production.

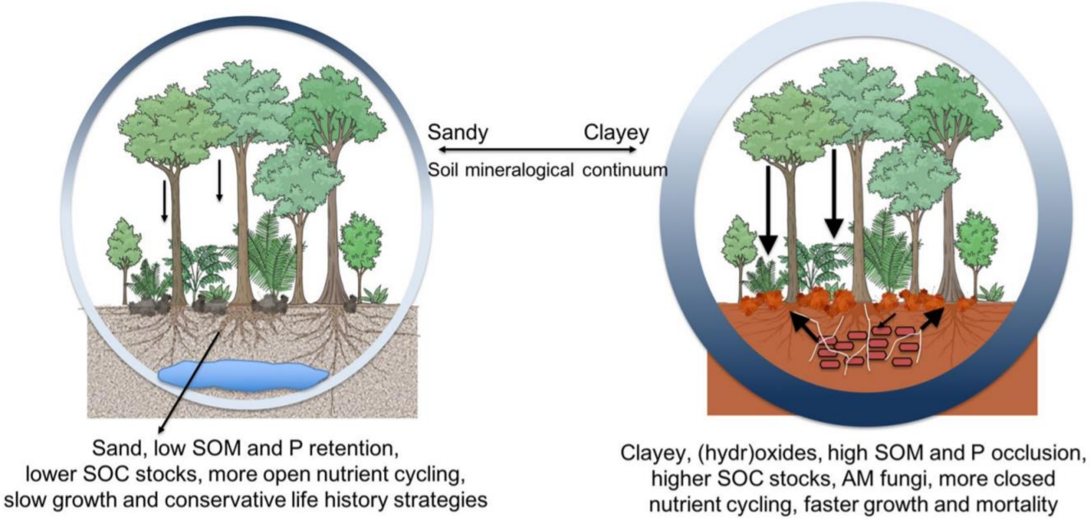

Sandy
Clayey
Soil mineralogical continuum
Sand, low SOM and P retention, lower SOC stocks, more open nutrient cycling, slow growth and conservative life history strategies
Clayey, (hydr)oxides, high SOM and P occlusion, higher SOC stocks, AM fungi, more closed nutrient cycling, faster growth and mortality

Figure 5. A simplified conceptual model of the influence of soil properties on tree growth and mortality, but not biomass, across phosphorus-depleted tropical forests. From Soong et al. (2020). Both forests have the same aboveground biomass, but different turnover rates and soil properties. At the sandy end of the soil continuum are forests with slower (narrower) nutrient cycling due to greater nutrient retention in the aboveground biomass (dark blue) based on slower growth, greater longevity, lower quality litter. At the other end of the spectrum are forests where the greater capacity of clay and (hydr)oxide-rich soils to retain phosphorus and organic matter support faster (wider) nutrient cycling forests. At clayey sites, nutrient recycling via decomposition (dark blue) is supported by a greater relative abundance of arbuscular mycorrhizal (AM) fungi.

6. REFERENCES

AMIRIAN-CHAKAN, A.; MINASNY, B.; TAGHIZADEH-MEHRJARDI, R.; AKBARIFAZLI, R.; DARVISHPASAND, Z.; KHORDEHBIN, S. Some practical aspects of predicting texture data in digital soil mapping. **Soil and Tillage Research**, v. 194, p. 104289, 2019.

ARROUAYS, D.; GRUNDY, M. G.; HARTEMINK, A. E.; HEMPEL, J. W.; HEUVELINK, G. B. M.; HONG, S. Y.; LAGACHERIE, P.; LELYK, G.; MCBRATNEY, A. B.; MCKENZIE, N. J.; MENDONCA-SANTOS, M. D. L.; MINASNY, B.; MONTANARELLA, L.; ODEH, I. O. A.; SANCHEZ, P. A.; THOMPSON, J. A.; ZHANG, G.-L. GlobalSoilMap: Toward a Fine-Resolution Global Grid of Soil Properties. **Advances in Agronomy**, [s.l: s.n.]. v. 125, p. 93–134.

BALDOCK, J. .; SKJEMSTAD, J. . Role of the soil matrix and minerals in protecting natural organic materials against biological attack. **Organic Geochemistry**, v. 31, n. 7–8, p. 697–710, Jul. 2000.

BATJES, N. H.; RIBEIRO, E.; OOSTRUM, A. VAN; LEENAARS, J.; HENGL, T.; MENDES DE JESUS, J. WoSIS: providing standardised soil profile data for the world. **Earth System Science**

Data, v. 9, n. 1, p. 1–14, 17 Jan. 2017.

BEHRENS, T.; SCHMIDT, K.; MACMILLAN, R. A.; VISCARRA ROSSEL, R. A. Multi-scale digital soil mapping with deep learning. **Scientific Reports**, 2018.

BEHRENS, T.; ZHU, A.-X.; SCHMIDT, K.; SCHOLTEN, T. Multi-scale digital terrain analysis and feature selection for digital soil mapping. **Geoderma**, v. 155, n. 3–4, p. 175–185, Mar. 2010.

BÜNEMANN, E. K.; BONGIORNO, G.; BAI, Z.; CREAMER, R. E.; DEYN, G. DE; GOEDE, R. DE; FLESKENS, L.; GEISSEN, V.; KUYPER, T. W.; MÄDER, P.; PULLEMAN, M.; SUKKEL, W.; GROENIGEN, J. W. VAN; BRUSSAARD, L. Soil quality – A critical review. **Soil Biology and Biochemistry**, v. 120, p. 105–125, 2018.

CHENU, C.; PLANTE, A. F. Clay-sized organo-mineral complexes in a cultivation chronosequence: revisiting the concept of the "primary organo-mineral complex." **European Journal of Soil Science**, v. 57, n. 4, p. 596–607, Aug. 2006.

DALMOLIN, R. S. D.; MOURA-BUENO, J. M.; SAMUEL-ROSA, A.; FLORES, C. A. How is the learning process of digital soil mapping in a diverse group of land use planners? **Revista Brasileira de Ciência do Solo**, v. 44, 2020.

DEMATTÊ, J. A. M.; DOTTO, A. C.; PAIVA, A. F. S.; SATO, M. V; DALMOLIN, R. S. D.; ARAÚJO, M. DO S. B.; SILVA, E. B.; NANNI, M. R.; CATEN, A. TEN; NORONHA, N. C.; LACERDA, M. P. C.; ARAÚJO FILHO, J. C.; RIZZO, R.; BELLINASO, H.; FRANCELINO, M. R.; SCHAEFER, C. E. G. R.; VICENTE, L. E.; SANTOS, U. J.; SÁ BARRETTO SAMPAIO, E. V DE; MENEZES, R. S. C.; 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.; MENEZES, M. D.; CURI, N.; COUTO, E. G.; ANJOS, L. H. C.; CEDDIA, M. B.; PINHEIRO, É. F. M.; GRUNWALD, S.; VASQUES, G. M.; MARQUES JÚNIOR, J.; SILVA, A. J.; BARRETO, M. C. DE V.; NÓBREGA, G. N.; SILVA, M. Z.; SOUZA, S. F.; VALLADARES, G. S.; VIANA, J. H. M.; SILVA TERRA, F. DA; HORÁK-TERRA, I.; FIORIO, P. R.; SILVA, R. C.; FRADE JÚNIOR, E. F.; LIMA, R. H. C.; ALBA, J. M. F.; 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.; SOUSA MENDES, W.; SAFANELLI, J. L.; GUIMARÃES, C. C. B.; POPPIEL, R. R.; SOUZA, A. B.; QUESADA, C. A.; COUTO, H. T. Z. The Brazilian Soil Spectral Library (BSSL): A general view, application and challenges. **Geoderma**, v. 354, p. 113793, 2019.

DEMATTÊ, J. A. M.; FONGARO, C. T.; RIZZO, R.; SAFANELLI, J. L. Geospatial Soil Sensing System (GEOS3): A powerful data mining procedure to retrieve soil spectral reflectance from satellite images. **Remote Sensing of Environment**, v. 212, p. 161–175, Jun. 2018.

DHARUMARAJAN, S.; HEGDE, R.; SINGH, S. K. Spatial prediction of major soil properties using Random Forest techniques - A case study in semi-arid tropics of South India. **Geoderma Regional**, v. 10, p. 154–162, 2017.

DIEK, S.; FORNALLAZ, F.; SCHAEPMAN, M.; JONG, R. DE. Barest Pixel Composite for Agricultural Areas Using Landsat Time Series. **Remote Sensing**, v. 9, n. 12, p. 1245, 2017.

DIEK, S.; SCHAEPMAN, M. E.; JONG, R. DE. Creating multi-temporal composites of airborne imaging spectroscopy data in support of digital soil mapping. **Remote Sensing**, 2016.

ESCADAFAL, R.; GIRARD, M. C.; DOMINIQUE, C. Modeling the relationships between Munsell soil color and soil spectral properties. **International Agrophysics**, v. 4, n. 3, p. 249–261, 1 Jan. 1988.

FERNANDEZ, R. N.; SCHULZE, D. G. Calculation of Soil Color from Reflectance Spectra. **Soil Science Society of America Journal**, v. 51, p. 1277–1282, 1987.

FONGARO, C.; DEMATTÊ, J.; RIZZO, R.; LUCAS SAFANELLI, J.; MENDES, W.; DOTTO, A.; VICENTE, L.; FRANCESCHINI, M.; USTIN, S. Improvement of Clay and Sand Quantification Based on a Novel Approach with a Focus on Multispectral Satellite Images. **Remote Sensing**, v.

10, n. 10, p. 1555, 27 Sep. 2018.

FORKUOR, G.; HOUNKPATIN, O. K. L.; WELP, G.; THIEL, M. 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**, v. 12, n. 1, p. e0170478, 23 Jan. 2017.

GÉRARD, F. Clay minerals, iron/aluminum oxides, and their contribution to phosphate sorption in soils — A myth revisited. **Geoderma**, v. 262, p. 213–226, Jan. 2016.

GOMES, L. C.; FARIA, R. M.; SOUZA, E. DE; VELOSO, G. V.; SCHAEFER, C. E. G. R.; FILHO, F. Modelling and mapping soil organic carbon stocks in Brazil. **Geoderma**, v. 340, p. 337–350, 2019.

GORELICK, N.; HANCHER, M.; DIXON, M.; ILYUSHCHENKO, S.; THAU, D.; MOORE, R. Google Earth Engine: Planetary-scale geospatial analysis for everyone. **Remote Sensing of Environment**, v. 202, p. 18–27, 2017.

HANSEN, M. C.; POTAPOV, P. V.; MOORE, R.; HANCHER, M.; TURUBANOVA, S. A.; TYUKAVINA, A.; THAU, D.; STEHMAN, S. V.; GOETZ, S. J.; LOVELAND, T. R.; KOMMAREDDY, A.; EGOROV, A.; CHINI, L.; JUSTICE, C. O.; TOWNSHEND, J. R. G. High-Resolution Global Maps of 21st-Century Forest Cover Change. **Science**, v. 342, n. 6160, p. 850–853, 15 Nov. 2013.

HARTEMINK, A. E.; MCBRATNEY, A. B.; CATTLE, J. A. Developments and trends in soil science: 100 volumes of Geoderma (1967–2001). **Geoderma**, v. 100, n. 3, p. 217–268, 2001.

HENDERSON, B. L.; BUI, E. N.; MORAN, C. J.; SIMON, D. A. P. Australia-wide predictions of soil properties using decision trees. **Geoderma**, v. 124, n. 3–4, p. 383–398, Feb. 2005.

HENGL, T.; HEUVELINK, G. B. M.; KEMPEN, B.; LEENAARS, J. G. B.; WALSH, M. G.; SHEPHERD, K. D.; SILA, A.; MACMILLAN, R. A.; MENDES DE JESUS, J.; TAMENE, L.; TONDOH, J. E. Mapping Soil Properties of Africa at 250 m Resolution: Random Forests Significantly Improve Current Predictions. **PLOS ONE**, v. 10, n. 6, p. e0125814, 25 Jun. 2015.

HENGL, T.; MACMILLAN, R. A. **Predictive Soil Mapping with R**. Wageningen, the Netherlands: OpenGeoHub foundation, 2019.

HENGL, T.; MENDES DE JESUS, J.; HEUVELINK, G. B. M.; RUIPEREZ GONZALEZ, M.; KILIBARDA, M.; BLAGOTIĆ, A.; SHANGGUAN, W.; WRIGHT, M. N.; GENG, X.; BAUER-MARSCHALLINGER, B.; GUEVARA, M. A.; VARGAS, R.; MACMILLAN, R. A.; BATJES, N. H.; LEENAARS, J. G. B.; RIBEIRO, E.; WHEELER, I.; MANTEL, S.; KEMPEN, B. SoilGrids250m: Global gridded soil information based on machine learning. **PLoS ONE**, v. 12, n. 2, p. e0169748, 16 Feb. 2017.

HENGL, T.; NUSSBAUM, M.; WRIGHT, M. N.; HEUVELINK, G. B. M.; GRÄLER, B. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. **PeerJ**, v. 6, p. e5518, 2018.

IBGE - INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Produção Agrícola Municipal [Municipal Agricultural Production]**. URL: https://sidra.ibge.gov.br/pesquisa/pam/tabelas. Acesso em: 29 sep. 2019.

JENNY, H. **Factors of soil formation: a system of quantitative pedology**. New York: Courier Dover Publications, 1941.

KAISER, K.; GUGGENBERGER, G. Mineral surfaces and soil organic matter. **European Journal of Soil Science**, v. 54, n. 2, p. 219–236, Jun. 2003.

KESKIN, H.; GRUNWALD, S.; HARRIS, W. G. Digital mapping of soil carbon fractions with machine learning. **Geoderma**, v. 339, p. 40–58, 2019.

KUMAR, L.; MUTANGA, O. Google Earth Engine Applications Since Inception: Usage, Trends, and Potential. **Remote Sensing**, v. 10, n. 10, p. 1509, 20 Sep. 2018.

LAGACHERIE, P. Digital soil mapping: A state of the art. *In*: HARTEMINK, A.; MCBRATNEY, A.; MENDONÇA-SANTOS, M. L. (Eds.). **Digital Soil Mapping with Limited Data**. [s.l.] Springer Netherlands, 2008. p. 3–14.

LAGACHERIE, P.; MCBRATNEY, A. B. Spatial soil information systems and spatial soil inference systems: perspectives for digital soil mapping. *In*: LAGACHERIE, P.; MCBRATNEY, A. B.; VOLTZ, M. (Eds.). **Digital soil mapping - an introductory perspective**. 31. ed., v. 31, 2007, p. 3–22.

LIU, F.; ZHANG, G.-L.; SONG, X.; LI, D.; ZHAO, Y.; YANG, J.; WU, H.; YANG, F. High-resolution and three-dimensional mapping of soil texture of China. **Geoderma**, v. 361, p. 114061, 2020.

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. Satellite data integration for soil clay content modelling at a national scale. **International Journal of Applied Earth Observation and Geoinformation**, v. 82, p. 101905, Oct. 2019.

MA, Y.; MINASNY, B.; WU, C. Mapping key soil properties to support agricultural production in Eastern China. **Geoderma Regional**, v. 10, p. 144–153, 2017.

MADEIRA NETTO, J. S.; BEDIDI, A.; CERVELLE, B.; POUGET, M.; FLAY, N. 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**, v. 18, n. 13, p. 2835–2852, 1997.

MALHOTRA, H.; VANDANA; SHARMA, S.; PANDEY, R. Phosphorus Nutrition: Plant Growth in Response to Deficiency and Excess. *In*: **Plant Nutrients and Abiotic Stress Tolerance**. Singapore: Springer Singapore, 2018. p. 171–190. URL: http://link.springer.com/10.1007/978-981-10-9044-8_7

MALONE, B. P.; HUGHES, P.; MCBRATNEY, A. B.; MINASNY, B. A model for the identification of terrons in the Lower Hunter Valley, Australia. **Geoderma Regional**, v. 1, p. 31–47, 2014.

MARQUES, K. P.; RIZZO, R.; DOTTO, A. C.; SOUZA, A. B.; MELLO, F. A.; NETO, L. G.; ANJOS, L. H. C.; DEMATTÊ, J. A. How qualitative spectral information can improve soil profile classification? **Journal of Near Infrared Spectroscopy**, p. 096703351882196, 3 Jan. 2019.

MATTIKALLI, N. M. Soil color modeling for the visible and near-infrared bands of Landsat sensors using laboratory spectral measurements. **Remote Sensing of Environment**, v. 59, n. 1, p. 14–28, 1997.

MCBRATNEY, A. B. **Introduction to pedometrics: a course of lectures, CSIRO Division of Soils Technical Memorandum 53/1986. (mimeo)**. Glen Osmond, Australia: CSIRO Division of Soils, 1986. URL: http://www.pedometrics.org/docs/Intro_to_pedometrics.pdf

MCBRATNEY, A. B.; MENDONCA SANTOS, M. L.; MINASNY, B. On digital soil mapping. **Geoderma**, v. 117, n. 1–2, p. 3–52, Nov. 2003.

MCBRATNEY, A. B.; MINASNY, B.; STOCKMANN, U. **Pedometrics**. Cham: Springer International Publishing, 2018.

MCBRATNEY, A.; GRUIJTER, J. DE; BRYCE, A. Pedometrics timeline. **Geoderma**, v. 338, p. 568–575, 2019.

MCGRODDY, M. E.; SILVER, W. L.; OLIVEIRA JR., R. C. DE; MELLO, W. Z. DE; KELLER, M. Retention of phosphorus in highly weathered soils under a lowland Amazonian forest ecosystem. **Journal of Geophysical Research: Biogeosciences**, v. 113, n. G4, 1 Dec. 2008.

MELO, V. F.; SINGH, B.; SCHAEFER, C. E. G. R.; NOVAIS, R. F.; FONTES, M. P. F. Chemical and Mineralogical Properties of Kaolinite-Rich Brazilian Soils. **Soil Science Society of America Journal**, v. 65, p. 1324–1333, 2001.

MENDES, W. DE S.; MEDEIROS NETO, L. G.; DEMATTÊ, J. A. M.; GALLO, B. C.; RIZZO, R.; SAFANELLI, J. L.; FONGARO, C. T. Is it possible to map subsurface soil attributes by satellite spectral transfer models? **Geoderma**, v. 343, p. 269–279, 2019.

MILLER, B. A.; KOSZINSKI, S.; WEHRHAN, M.; SOMMER, M. Impact of multi-scale predictor selection for modeling soil properties. **Geoderma**, v. 239–240, p. 97–106, 2015.

MINASNY, B.; MCBRATNEY, A. B. Digital soil mapping: A brief history and some lessons. **Geoderma**, v. 264, p. 301–311, 2016.

MULDER, V. L.; BRUIN, S.; WEYERMANN, J.; KOKALY, R. F.; SCHAEPMAN, M. E. Characterizing regional soil mineral composition using spectroscopy and geostatistics. **Remote Sensing of Environment**, v. 139, p. 415–429, 2013.

MURRAY, N. J.; PHINN, S. R.; DEWITT, M.; FERRARI, R.; JOHNSTON, R.; LYONS, M. B.; CLINTON, N.; THAU, D.; FULLER, R. A. The global distribution and trajectory of tidal flats. **Nature**, v. 565, n. 7738, p. 222–225, 2019.

NAWAR, S.; CORSTANJE, R.; HALCRO, G.; MULLA, D.; MOUAZEN, A. M. Delineation of Soil Management Zones for Variable-Rate Fertilization: A Review. *In*: SPARKS, D. L. B. T.-A. IN A. (Ed.). . **Advances in Agronomy**. [s.l.] Academic Press, 2017. v. 143p. 175–245.

NUSSBAUM, M.; SPIESS, K.; BALTENSWEILER, A.; GROB, U.; KELLER, A.; GREINER, L.; SCHAEPMAN, M. E.; PAPRITZ, A. Evaluation of digital soil mapping approaches with large sets of environmental covariates. **SOIL**, v. 4, n. 1, p. 1–22, 10 Jan. 2018.

OLDEMAN, L. R.; ENGELEN, V. W. P. VAN. A world soils and terrain digital database (SOTER) — An improved assessment of land resources. **Geoderma**, v. 60, n. 1, p. 309–325, 1993.

OMUTO, C.; NACHTERGAELE, F.; ROJAS, R. V. **State of the Art Report on Global and regional Soil Information: Where are we? Where to go?** Rome: FAO, 2013.

PADARIAN, J.; MINASNY, B.; MCBRATNEY, A. B. Machine learning and soil sciences: A review aided by machine learning tools. **SOIL Discuss.**, v. 2019, p. 1–29, 3 Sep. 2019.

PADARIAN, J.; MINASNY, B.; MCBRATNEY, A. B. Using Google's cloud-based platform for digital soil mapping. **Computers & Geosciences**, v. 83, p. 80–88, 1 Oct. 2015.

PARENTE, L.; MESQUITA, V.; MIZIARA, F.; BAUMANN, L.; FERREIRA, L. Assessing the pasturelands and livestock dynamics in Brazil, from 1985 to 2017: A novel approach based on high spatial resolution imagery and Google Earth Engine cloud computing. **Remote Sensing of Environment**, v. 232, p. 111301, 2019.

PEKEL, J.-F.; COTTAM, A.; GORELICK, N.; BELWARD, A. S. High-resolution mapping of global surface water and its long-term changes. **Nature**, v. 540, n. 7633, p. 418–422, 2016.

POPPIEL, R. R.; LACERDA, M. P. C.; DEMATTÊ, J. A. M.; OLIVEIRA JR, M. P.; GALLO, B. C.; SAFANELLI, J. L. Pedology and soil class mapping from proximal and remote sensed data. **Geoderma**, 2019.

POPPIEL, R. R.; LACERDA, M. P. C.; OLIVEIRA JR, M. P.; DEMATTÊ, J. A. M.; ROMERO, D. J.; SATO, M. V.; ALMEIDA JR, L. R.; CASSOL, L. F. M. Surface Spectroscopy of Oxisols, Entisols and Inceptisol and Relationships with Selected Soil Properties. **Revista Brasileira de Ciência do Solo**, v. 42, p. e0160519, 2018.

POULADI, N.; MØLLER, A. B.; TABATABAI, S.; GREVE, M. H. Mapping soil organic matter contents at field level with Cubist, Random Forest and kriging. **Geoderma**, v. 342, p. 85–92, 2019.

RAMCHARAN, A.; HENGL, T.; NAUMAN, T.; BRUNGARD, C.; WALTMAN, S.; WILLS, S.;

THOMPSON, J. Soil Property and Class Maps of the Conterminous United States at 100-Meter Spatial Resolution. **Soil Science Society of America Journal**, v. 82, p. 186–201, 2018.

RIZZO, R.; DEMATTÊ, J. A. M.; LEPSCH, I. F.; GALLO, B. C.; FONGARO, C. T. Digital soil mapping at local scale using a multi-depth Vis–NIR spectral library and terrain attributes. **Geoderma**, v. 274, p. 18–27, 2016.

ROBERTS, D.; WILFORD, J.; GHATTAS, O. Exposed soil and mineral map of the Australian continent revealing the land at its barest. **Nature Communications**, v. 10, n. 1, p. 5297, 2019.

ROGGE, D.; BAUER, A.; ZEIDLER, J.; MUELLER, A.; ESCH, T.; HEIDEN, U. Building an exposed soil composite processor (SCMaP) for mapping spatial and temporal characteristics of soils with Landsat imagery (1984–2014). **Remote Sensing of Environment**, v. 205, p. 1–17, 2018.

ROSSEL, R. A. V.; CHEN, C. Digitally mapping the information content of visible–near infrared spectra of surficial Australian soils. **Remote Sensing of Environment**, v. 115, n. 6, p. 1443–1455, 15 Jun. 2011.

ROSSITER, D. G. Past, present & future of information technology in pedometrics. **Geoderma**, v. 324, n. November 2017, p. 131–137, 2018.

RUTTENBERG, K. C.; SULAK, D. J. Sorption and desorption of dissolved organic phosphorus onto iron (oxyhydr)oxides in seawater. **Geochimica et Cosmochimica Acta**, v. 75, n. 15, p. 4095–4112, Aug. 2011.

SAMUEL-ROSA, A.; DALMOLIN, R. S. D.; MOURA-BUENO, J. M.; TEIXEIRA, W. G.; ALBA, J. M. F. Open legacy soil survey data in Brazil: geospatial data quality and how to improve it. **Scientia Agricola**, v. 77, n. 1, p. e20170430, 2020.

SCHAEFER, C. E. G. R.; FABRIS, J. D.; KER, J. C. Minerals in the clay fraction of Brazilian Latosols (Oxisols): a review. **Clay Minerals**, v. 43, n. 1, p. 137–154, 9 Mar. 2008.

SCHEINOST, A. C.; CHAVERNAS, A.; BARRÓN, V.; TORRENT, J. 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**, v. 46, n. 5, p. 528–536, 1998.

SCHWERTMANN, U.; TAYLOR, R. M. Iron Oxides. *In*: **Minerals in Soil Environments**. [s.l: s.n.]. 1989, p. 379–438.

SCULL, P.; FRANKLIN, J.; CHADWICK, O. A.; MCARTHUR, D. Predictive soil mapping: a review. **Progress in Physical Geography**, v. 27, n. 2, p. 171–197, 2003.

SIEWERT, M. B. High-resolution digital mapping of soil organic carbon in permafrost terrain using machine learning: a case study in a sub-Arctic peatland environment. **Biogeosciences**, v. 15, n. 6, p. 1663–1682, 21 Mar. 2018.

SIMON, T.; ZHANG, Y.; HARTEMINK, A. E.; HUANG, J.; WALTER, C.; YOST, J. L. Predicting the color of sandy soils from Wisconsin, USA. **Geoderma**, p. 114039, 2019.

SOIL SURVEY STAFF. **Keys to Soil Taxonomy**. Washington: United States Department of Agriculture, 2014. v. 12

SOONG, J. L.; JANSSENS, I. A.; GRAU, O.; MARGALEF, O.; STAHL, C.; LANGENHOVE, L. VAN; URBINA, I.; CHAVE, J.; DOURDAIN, A.; FERRY, B.; FREYCON, V.; HERAULT, B.; SARDANS, J.; PEÑUELAS, J.; VERBRUGGEN, E. Soil properties explain tree growth and mortality, but not biomass, across phosphorus-depleted tropical forests. **Scientific Reports**, v. 10, n. 1, p. 2302, 2020.

STENBERG, B.; VISCARRA ROSSEL, R. A.; MOUAZEN, A. M.; WETTERLIND, J. Visible and Near Infrared Spectroscopy in Soil Science. **Advances in Agronomy**, v. 107, p. 163–215, 2010.

UNITED NATIONS - DEPARTMENT OF ECONOMIC AND SOCIAL AFFAIRS - POPULATION DIVISION. **World Population Prospects 2019: Highlights**. New York, USA: United Nations, 2019.

VISCARRA ROSSEL, R. A. Fine-resolution multiscale mapping of clay minerals in Australian soils measured with near infrared spectra. **Journal of Geophysical Research: Earth Surface**, v. 116, n. F4, 2011.

VISCARRA ROSSEL, R. A.; BEHRENS, T.; BEN-DOR, E.; BROWN, D. J.; DEMATTÊ, J. A. M.; SHEPHERD, K. D.; SHI, Z.; STENBERG, B.; STEVENS, A.; ADAMCHUK, V.; AÏCHI, H.; BARTHÈS, B. G.; BARTHOLOMEUS, H. M.; BAYER, A. D.; BERNOUX, M.; BÖTTCHER, K.; BRODSKÝ, L.; DU, C. W.; CHAPPELL, A.; FOUAD, Y.; GENOT, V.; GOMEZ, C.; GRUNWALD, S.; GUBLER, A.; GUERRERO, C.; HEDLEY, C. B.; KNADEL, M.; MORRÁS, H. J. M.; NOCITA, M.; RAMIREZ-LOPEZ, L.; ROUDIER, P.; CAMPOS, E. M. R.; SANBORN, P.; SELLITTO, V. M.; SUDDUTH, K. A.; RAWLINS, B. G.; WALTER, C.; WINOWIECKI, L. A.; HONG, S. Y.; JI, W. A global spectral library to characterize the world's soil. **Earth-Science Reviews**, v. 155, p. 198–230, Apr. 2016.

VISCARRA ROSSEL, R. A.; BUI, E. N.; CARITAT, P. DE; MCKENZIE, N. J. Mapping iron oxides and the color of Australian soil using visible–near-infrared reflectance spectra. **Journal of Geophysical Research**, v. 115, n. F4, p. F04031, 15 Dec. 2010.

VISCARRA ROSSEL, R. A.; CHEN, C.; GRUNDY, M. J.; SEARLE, R.; CLIFFORD, D.; CAMPBELL, P. H. The Australian three-dimensional soil grid: Australia's contribution to the GlobalSoilMap project. **Soil Research**, v. 53, n. 8, p. 845–864, 2015.

VISCARRA ROSSEL, R. A.; HICKS, W. S. Soil organic carbon and its fractions estimated by visible–near infrared transfer functions. **European Journal of Soil Science**, v. 66, n. 3, p. 438–450, 1 May 2015.

WADOUX, A.; MINASNY, B.; MCBRATNEY, A. Machine learning for digital soil mapping: applications, challenges and suggested solutions. **EarthArXiv**, 2020.

CHAPTER 1 — MAPPING AT 30 M RESOLUTION OF PHYSICAL AND CHEMICAL SOIL ATTRIBUTES AT MULTIPLE DEPTHS IN MIDWEST BRAZIL[1]

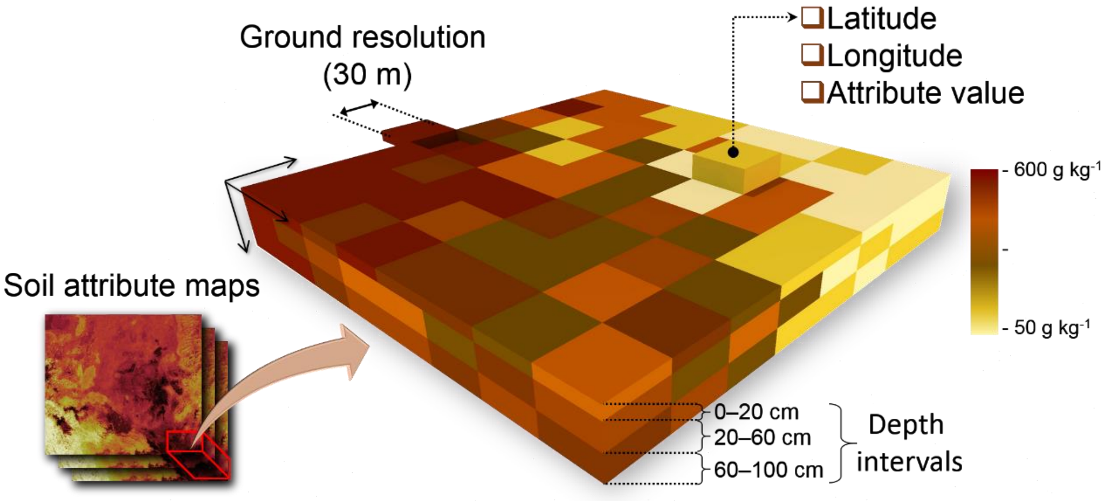
ABSTRACT

The Midwest region in Brazil has the largest and most recent agricultural frontier in the country, where there is no currently detailed soil information to support the agricultural intensification. Producing large-extent digital soil maps demands a huge volume of data and high computing capacity. This paper proposed mapping surface and subsurface key soil attributes with 30 m-resolution in a large extent of Midwest Brazil. These soil maps at multiple depth increments will provide adequate information to guide land use throughout the region. The study area comprises about 851,000 km^2 in the Cerrado biome (savannah) across Brazilian Midwest. We used soil data from 7908 sites of the Brazilian Soil Spectral Library and 231 of the Free Brazilian Repository for Open Soil Data. We selected nine key soil attributes for mapping and aggregated them into three depth intervals: 0–20, 20–60 and 60–100 cm. A total of 33 soil predictors were prepared using Google Earth Engine (GEE), such as climate and geologic features with 1 km-resolution, terrain attributes and two new covariates with 30 m-resolution, based on satellite measurements of the topsoil reflectance and the seasonal variability in vegetation spectra. The *scorpan* model was adopted for mapping of soil variables using random forest regression (RF). We used the model-based optimization by tuning RF hyperparameters and calculated the scaled permutation importance of covariates in R software. Our results were promising, with a satisfactory model performance for physical and chemical attributes at all depth intervals. Elevation, climate and topsoil reflectance were the most important covariates in predicting sand, clay and silt. In general, climatic variables, elevation and vegetation reflectance provided to be the most important covariates for predicting soil chemical attributes, while for organic matter it was a combination of climatic dynamics and reflectance bands from vegetation and topsoil. The multiple depth maps showed that soil attributes largely varied across the study area, from clayey to sandy, suggesting that less than 44% of the studied soils had good natural fertility. We concluded that key soil attributes from multiple depth increments can be mapped using Earth observations data and machine learning methods with good performance.

Keywords: spatial big data; soil attributes; digital soil mapping; random forest; remote sensing; Google Earth Engine; land management

---

[1] Article published in **Remote Sensing**, 5 December 2019, https://doi.org/10.3390/rs11242905



GRAPHICAL ABSTRACT

1. INTRODUCTION

The soil plays essential role for natural and anthropic ecosystems (BÜNEMANN et al., 2018). Reliable spatial soil information can improve natural capital assessment, becoming important for food production, especially in large countries or emerging economies where the major demographic growth is expected (UNITED NATIONS, 2019). Soil mapping is expensive and time-demanding, consequently performing adequate maps in large areas takes several years and require significant economic resources. Such fact is observed in countries like Brazil, which is covered by small scale soil maps, mostly developed by Brazilian government institutions using RADAMBRASIL (Radar on Amazon and Brazil) project data (1:1,000,000 or nominally 2 km) (MCBRATNEY et al., 2003). In this case, such maps are not capable of supporting any decision making in regional or local scales.

Currently, there are no soil attribute maps with complete coverage across the Brazilian Midwest, which could support management and policy decisions. This region has the largest and most recent agricultural frontier in Brazil (PARENTE et al., 2019), which contributes about 34% and more than 10% to the agricultural production and gross domestic product of the country, respectively (IBGE, 2018).

The huge volume of quantitative (pedometric) data required in the production of soil attribute maps, for large geographical extents, limits the feasibility of conventional (traditional) manual (expert-based) soil mapping (HENGL et al., 2017; MCBRATNEY et al., 2003). Several key soil factors are still not fully represented by classical environmental covariates, being necessary to develop new covariates that provide improved proxies for describing soil spatial variations. Advances in Earth observation (satellite images and products), digital elevation models and digital soil mapping (DSM) frameworks, based on machine learning and cloud-based computing, might be a solution to the lack of adequate soil data (HENGL et al., 2018). An Earth observation product that has raised attention in DSM is the satellite image. Such data can retrieve medium- to high-resolution information and are easily acquired. Recently, studies have employed multi-temporal images in soil assessment and mapping. Such data provides measurements of topsoil reflectance, which are directly related to clay content, organic matter, mineralogy, moisture and soil color (STENBERG et al., 2010). The synergy between satellite images and DSM techniques is described by Diek et al. (2016), who performed a multi-temporal composite from the Airborne Prism Experiment (APEX). By overlapping images, the authors doubled the amount of bare soil pixels in the scene and presented an enhanced spatial representation of topsoil. Later, Diek et al. (2017) developed a method for identifying the least-vegetated pixels (e.g. barest pixel) in a dense Landsat time series. Such data was used to estimate soil attributes and evaluate the contribution of remote sensing (RS) to conventional and DSM procedures. Similarly, Rogge et al. (2018) proposed the Soil Composite Mapping Processor (SCMaP), which is an approach able to use per-pixel compositing to address the issue of limited soil exposure. Another bare surface composition technique was proposed by Demattê et al. (2018), called Geospatial Soil System (GEOS3). These authors validated the method by comparing the bare surface data (described as SySI) to laboratory spectral measurements, and found a canonical correlation of 0.93. Later, Fongaro et al. (2018) used such composite images to digitally map soils from southeastern Brazil. These authors described an expressive enhancement in clay content's digital mapping when employing SySI and terrain attributes. The R^2 and root mean squared error (RMSE) improved from 0.64 and 93.44 g kg^{-1} to 0.83 and 65.36 g kg^{-1}, respectively. Finally, Mendes et al. (2019) indicated that besides surface layer mapping, SySI can also aid in the prediction of soil subsurface attributes.

The prediction in DSM is usually based on machine learning techniques, which fit models for the spatial prediction of soil variables (i.e. maps of soil attributes and classes at different resolutions) (HENGL et al., 2017). While machine learning supports the soil spatial predictions (PADARIAN et al., 2019), cloud-based computing provides a superior architecture for the execution of such complex algorithms (GORELICK et al., 2017). These techniques are very attractive, once it result in the automation of processes, reducing overall soil data production costs, combining statistics, data science, soil science, physical geography, RS, geoinformation science and a number of other sciences (PADARIAN et al., 2019; DEMATTÊ et al., 2018; HENGL et al., 2018; MCBRATNEY et al., 2003).

A brief search in literature regarding the terms "soil" and "machine learning" resulted in more than 72,000 publications, from which 7,200 items were published in the first half of 2019 and 4,000 discussed random forest (RF) algorithms. The RF algorithm was first introduced by Breiman (2001) and became a standard nonparametric classification and regression tool. The method establishes prediction rules based on various types of predictor variables, without making any prior assumption on the form of their association with the response variable (PROBST et al., 2019). The RF is one of most popular algorithms in DSM, being employed in several soil mapping studies (AMIRIAN-CHAKAN et al., 2019; GOMES et al., 2019; HENGL et al., 2018; LOISEAU et al., 2019; MA et al., 2017). Many inter-comparisons between machine learning algorithms are described in literature, and in most cases, authors indicated RF as the most adequate algorithm for DSM. Keskin et al. (2019) compared many models to quantify stochastic and/or deterministic components of soil carbon (C) pools. The prediction performance indicated the RF as the best algorithm. The covariables that best described variations in C pools were the biotic and hydro-pedological ones. Lithological and climatic factors had a reasonable influence in C predictions, while topographic factors did not contribute to soil C modeling. Similarly, Nussbaum et al. (2018) evaluated six approaches for digitally mapping 14 soil attributes at four depths. They found small differences in predictive performances, but RF was often the best among all methods. Hengl et al. (2015) mapped 14 soil attributes from African soils, combining quality-controlled point data and a large number of covariates. The RF was the best method, outperforming linear regression with an average decrease of 15%–75% in RMSE across soil properties and depths.

Based on results from studies reviewed, we expect that Earth observation data and machine learning, coupled with Brazilian available legacy soil datasets, to promote a favorable framework for producing accurate soil predictions across this important agricultural region. We assume that it is possible to map physical and chemical soil attributes for three fixed depth intervals (0–20, 20–60 and 80–100 cm) with 30 m-resolution across the Midwest region in Brazil.

Thus, we intend to produce up-to-date pedometric maps of surface and subsurface key soil attributes in a large extension of the Midwest of Brazil. These maps at three fixed depth intervals might provide adequate information to account for the multi-functionality of soil in the region. Therefore, we aimed to (a) produce composite images (described hereafter as SySI and SyVI) using Landsat data, which describes the reflectance variability of bare surfaces and natural vegetation; (b) employ SySI and SyVI coupled with terrain attributes, geologic and climate variables as predictors in the digital mapping of key soil attributes in the Midwest Brazil; (c) fit and assess the performance of the random forest models for the spatial patterns of each soil attribute at three depth intervals; (d) identify the covariates that were most relevant to describe the soil variability in Midwest Brazil; (e) to map the spatial distribution of soil attributes at 30 m resolution for the 0−20, 20−60 and 60−100 cm depth using the optimized models and cloud-based computing for the study area.

2. MATERIAL AND METHODS

2.1. Study Area and Soil Data

The study area comprises about 851,000 km^2 in the Cerrado biome (savanna) of Midwest Brazil (Figure 1), covered by Cerrado vegetation and gallery forest over extensive plateaus. The climate is tropical humid, which has two well-defined seasons, wet in summer and dry in winter, with annual precipitation ranging from 1200 to 1800 mm. According to the 1:1,000,000-scale pedological map (IBGE, 2017), the region was dominated by Ferralsols, Lixisols, Plinthosols, Arenosols and Regosols (IUSS WORKING GROUP WRB, 2015). These soils developed from highly diversified lithologies, consisting of volcanic, metamorphic, and sedimentary rocks, who reworked surface materials (CPRM, 2004).

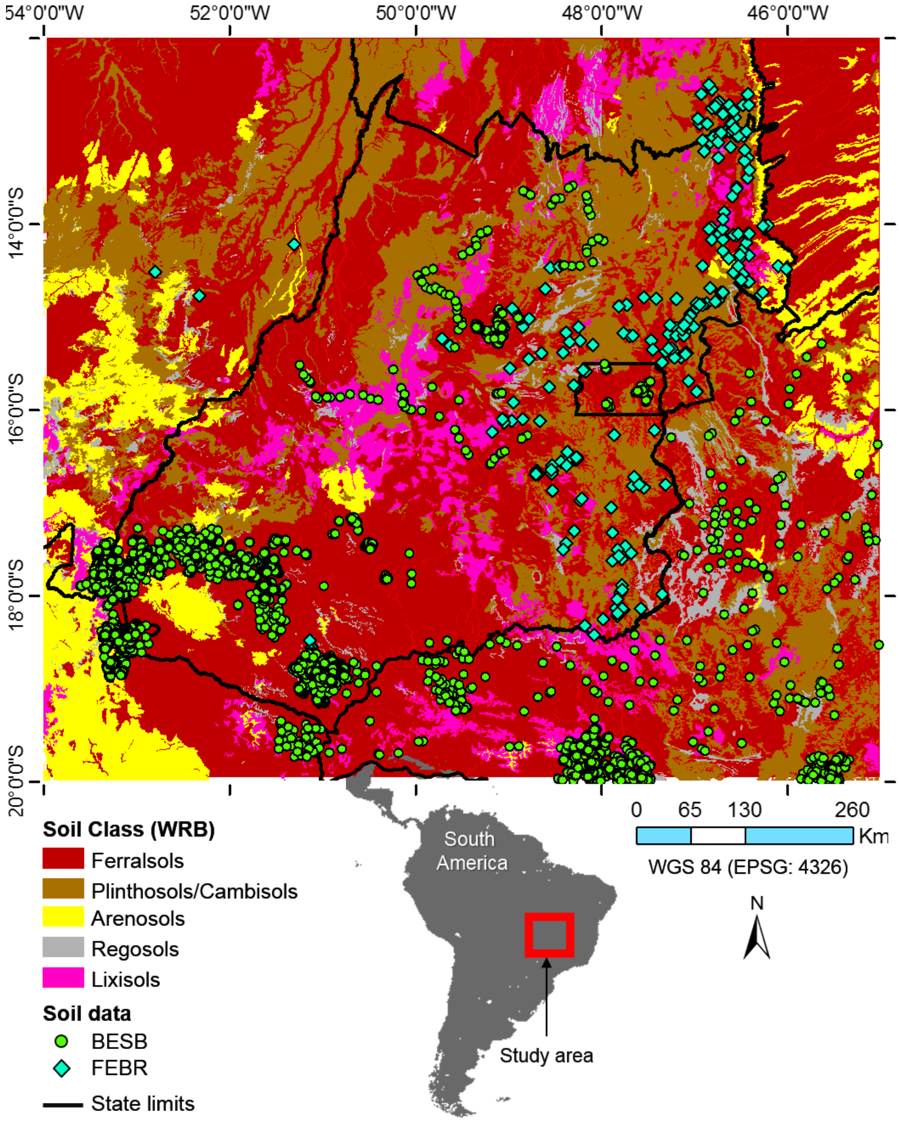


Figure 1. Spatial distribution of soil observations displayed over a 1:1,000,000-scale map of the main soil classes of the study area (IBGE, 2017). Soil classes were defined according to World Reference Base (IUSS WORKING GROUP WRB, 2015).

We obtained physical and chemical soil attribute data from 7908 sites of the Brazilian Soil Spectral Library (BSSL) (DEMATTÊ et al., 2019) and 231 of the Free Brazilian Repository for Open Soil Data (FEBR) (SAMUEL-ROSA et al., 2020). The BSSL started in 1995 as a collaborative network formed by several institutions across Brazil. The FEBR contains legacy soil observations data collected by Brazilian government agencies since the 1960s.

We selected nine soil attributes for mapping in the study area (Figure 1) and aggregated into three depth intervals (0–20, 20–60 and 60–100 cm): sand, silt and clay contents, organic matter ($OM = organic\ carbon\ \times 1.72$), pH measured in water (pH H_2O) and in potassium chloride (pH KCl), cation exchange capacity ($CEC = Ca^{2+} + Mg^{2+} + K^{+} + H^{+} + Al^{3+}$), and base saturation ($V\% = \left((Ca^{2+} + Mg^{2+} + K^{+}) \times 100\right) \div CEC$) and aluminum saturation ($m\% = (Al^{3+} \times 100) \div (Ca^{2+} + Mg^{2+} + K^{+} + Al^{3+})$). These soil attributes are commonly used (as key criteria) to guide agricultural recommendations, to evaluate the locations most suitable for farming and delineation of soil management zones (NAWAR et al., 2017). They are also used for soil classification (IUSS WORKING GROUP WRB, 2015; SOIL SURVEY STAFF, 2014). According to Canadell (1996), maximum rooting depth of crops by far can exceed 100 cm soil depths. Thus, soil attributes from 0 to 100 cm depth can affect plant growth and yield. When exploring the complete dataset, we checked for possible duplicated data and typos. To remove outliers from the dataset before modelling, we used more than one condition by nesting IF functions in Microsoft Excel. For example, to remove sand, silt and clay contents smaller or greater than 1000 g kg^{-1} [=IF(SUM(Sand;Silt;Clay)=1000);"OK";"REMOVE")], or testing IF the relationships (V% vs. pH vs. m%, OM vs. CEC) were coherent. A large proportion of the data had information based on the laboratory method of Embrapa (2017), while the remaining were transformed to the same standard units.

Finally, we performed a chord diagram based on Pearson correlation to check weighted relationships between soil attributes using the circlize package version 0.4.8 (GU, 2019) in the R software (R CORE TEAM, 2018). In that diagram, each soil attribute is represented by a fragment on the outer part of the circular layout, where the size of the connections is proportional to the value of the correlation.

The framework used for digital mapping of soil attributes (Figure 2), from covariates preparing to spatial predictions, was fully implemented via the cloud-based platform of Google Earth Engine (GEE) (GORELICK et al., 2017) and the R environment for statistical computing (R CORE TEAM, 2018).

2.2.Preparing Environmental Covariates

Environmental covariate layers can be used as predictors (“independent variables”) in prediction models. Their preparation is time and resources consuming, involving a huge image processing to transform large environmental databases into relevant predictors for machine learning of soil attributes. Therefore, efforts to produce appropriate predictors to explain the spatial distribution of soil attributes (at detail and generalization) increases the accuracy of the models. Various covariates (e.g. climate, terrain attributes and RS data) representing soil state factors have been widely used in statistical models to predict soil texture, bulk density, organic carbon, nutrients (Ca, Mg, K, Na, N, P), available water capacity, pH and CEC (BALLABIO et al., 2019; GOMES et al., 2019; HENGL et al., 2014, 2015, 2017; LIANG et al., 2019; VISCARRA ROSSEL et al., 2015).

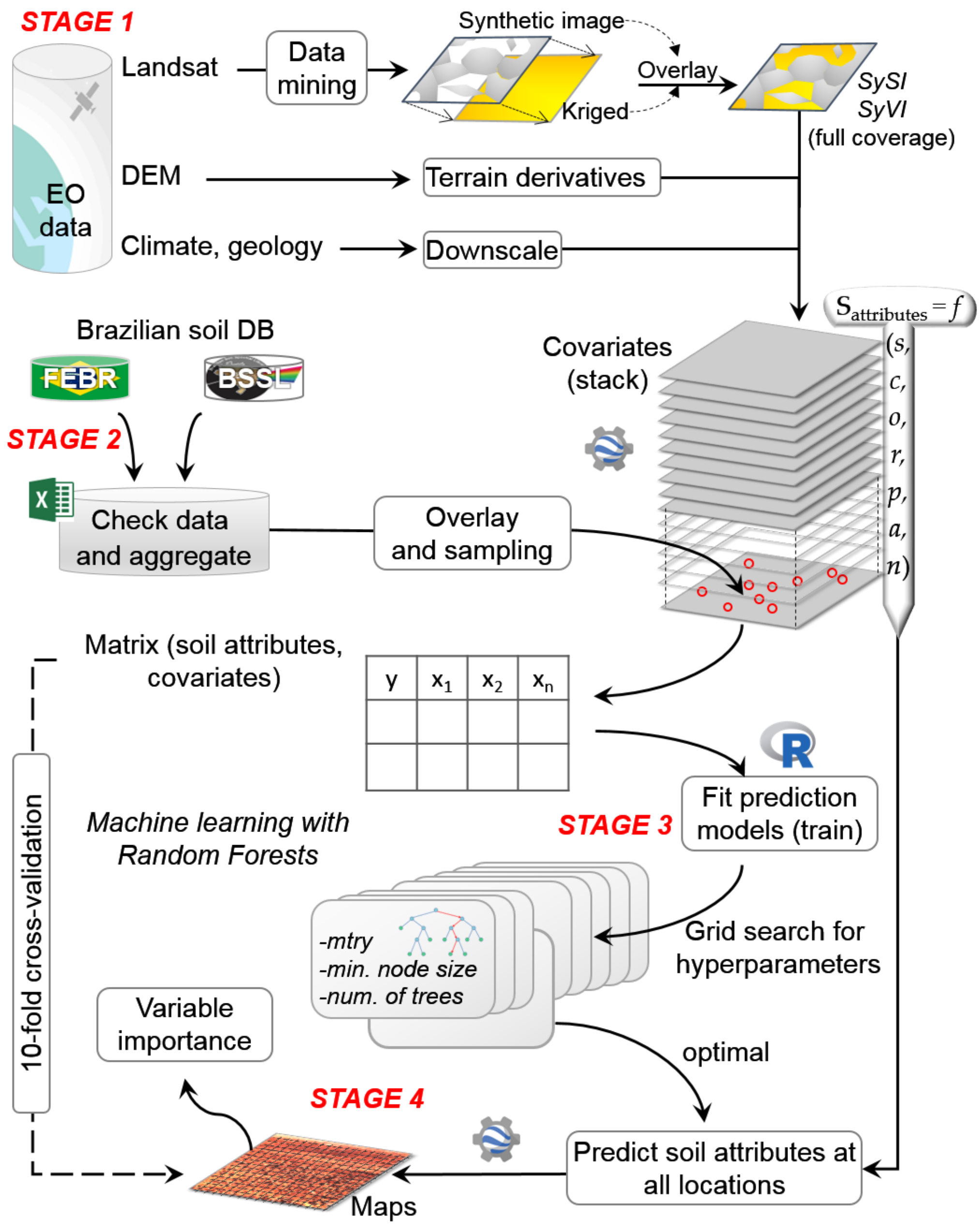


Figure 2. Digital soil mapping framework used for generating soil attribute maps.

For mapping the selected soil attributes, a group of covariates were obtained (Table 1) and used as proxies of the factors of soil formation, according the *scorpan* model (MCBRATNEY et al., 2003), that accounts for soil (s), climate (c), vegetation (o), relief (r), parent material (p), age of surface (a) and spatial position (n). This model assumes that the soils were formed in response to different processes operating over different distances or scales (BAILEY, 1987; MCBRATNEY et al., 2003). These soil spatial patterns can be captured by the use of multi-resolution covariates and used in predictions models (MILLER et al., 2015). For that, we adjusted the coarser-resolution covariates to a target resolution of 30 m, that was in accordance with the majority of covariates used. We employed inverse distance weighting (IDW) interpolator to downscale the 1 km covariates (climate and geology) to 30 m resolution. IDW attenuates the influence of distant points, according to the inverse distance weight, and gives an assumption of positive spatial autocorrelation (AKINYEMI et al., 2008). Furthermore, this interpolator is easy to be implemented and available in Google Earth Engine (GEE) platform (GORELICK et al., 2017). Reducing pixel size did not produce information gain on downscaled covariates, but it enables the simultaneous use of predictors with different spatial resolutions that account for soil spatial patterns at different scales.

2.2.1. Climate Data

Annual temperature average, range and seasonality, and annual precipitation and seasonality values were obtained from the WorldClim dataset (HIJMANS et al., 2005) at a spatial resolution of 1 km, and then were downscaled to 30 m pixel size by IDW. These data layers derived from numerous weather stations data interpolated by thin-plate smoothing spline, using latitude, longitude, and elevation as independent variables (HIJMANS et al., 2005). The WorldClim is a spatially continuous climate dataset with the highest resolution available for the study region.

2.2.2. Relief and Geology Data

We derived local terrain attributes, including elevation, slope, aspect, horizontal and vertical curvature and topographic position index (TPI) from the 30 m ALOS digital elevation model (TADONO et al., 2014) within GEE. Slope, aspect and curvatures, were calculated from the partial derivatives of terrain using a 3 × 3 moving window (FLORINSKY, 2016). The TPI was calculated by subtracting the elevation in meters at a given location (or cell) to the mean elevation of all cells within a neighborhood specified by a radius of 3 km, which best described our region after a radius search. Highly positive values are associated with peaks and ridges, while highly negative values are associated with valley bottoms and sinks. We obtained the density of geological lineaments by counting the meters of structural lines obtained from a 1:1,000,000-scale map (CPRM, 2004) in 1 km grids, and then transformed to raster and downscaled to 30 m pixel size by the IDW method.

2.2.3 Landsat-Derived Data

Data. The Landsat program has been observing the Earth continuously from 1972 through the present day. We used Landsat surface reflectance data (Tier 1, Collection 1) of different sensors covering the study area from 1982 to 2019, including the Thematic Mapper (TM, Landsat 4–5), the Enhanced Thematic Mapper Plus (ETM+, Landsat 7), and the Operational Land Imager and Thermal Infrared Sensor (OLI/TIRS, Landsat 8) with 16 days revisiting time and 30 m resolution (USGS, 2019a, b). Considering these products are gridded into common characteristics (resolution, projection, spatial extent, scale values and spectral ranges), we performed an inter-sensor harmonization to combine their separated collections into a single dataset. The bands of each sensor, positioned in equivalent spectral regions, were matched into a common name (e.g. Blue, Green, Red, NIR, $SWIR_1$, $SWIR_2$ and LST) using the specific band number (Table A1). The quality assessment bands were used to remove cloudy and cloud shadow pixels. We calculated the land surface temperature (LST, in degrees Celsius scaled from 0 to 10,000) for each image in three steps: 1) we calculated the normalized difference vegetation index (NDVI, Equation 1); 2) we estimated the land surface emissivity (LSE, Equation 2) using the NDVI-based method (VANDEGRIEND et al., 1992); 3) and we converted the brightness temperature (BT) data to LST using the Stefan–Boltzmann law expressed in Equation 3 (AL-GAADI et al., 2018). This approach enabled to obtain LST from the available Landsat data within GEE.

$$NDVI = \frac{NIR - Red}{\mathrm{NIR} + \mathrm{Red}} \tag{1}$$

$$LSE = 1.009 + 0.047 \times \ln(\mathrm{NDVI}) \tag{2}$$

$$\mathrm{LST} = \left(\left(\frac{1}{LSE^{1/4}}\right) \times BT\right) \tag{3}$$

SySI. We implemented the Geospatial Soil Sensing System (GEOS3) (DEMATTÊ et al., 2018) into GEE to generate a 30 m synthetic soil image (SySI) using the harmonized Landsat data. The GEOS3 is a data mining algorithm that uses classifications rules to identify soil at pixel level on denser satellite time series. The rules are a set of spectral indices and thresholds that mask out non-soil pixels by flagging soil pixels as a valid value and the remaining pixels as unavailable information (NA). We used NDVI (Equation 1), normalized burn ratio 2 (NBR2, Equation 4) and Visible to Shortwave Infrared tendency index (VNSIR, Equation 5). The thresholds for the spectral indices were defined as $-0.15 < \mathrm{NDVI} < 0.25$, $-0.15 < \mathrm{NBR2} < 0.15$ and $\mathrm{VNSIR} < 9{,}000$. These rules selected soil pixels that were aggregated into a single composite (SySI) by computing band-to-band the median of the reflectance values, over the time series. For our study, SySI represents the soil surface of agriculture areas and other natural surfaces with low vegetation cover and rock outcrops, when the vegetation was absent or almost absent, typical for savannas. The GEOS3 has also been implemented across different regions in Brazil for mapping soil variables (FONGARO et al., 2018; GALLO et al., 2018; MENDES et al., 2019). Similar approaches were developed to produce bare soil composites based on Landsat data and accurately employed for soil mapping and management in Germany (ROGGE et al., 2018) and both the Swiss Plateau and Europe (DIEK et al., 2017).

$$NBR2 = \frac{SWIR_1 - SWIR_2}{SWIR_1 + SWIR_2} \tag{4}$$

$$VNSIR = \left(10000 - \left((2 \times RED - GREEN - BLUE) + ((SWIR_2 - NIR) \times 3)\right)\right) \tag{5}$$

SyVI. To take advantage of the spatio-temporal variation of vegetation that might be linked to soil distribution, the GEOS3 (DEMATTÊ et al., 2018) was adapted into GEE to produce a 30 m synthetic vegetation image in the wet ($\mathrm{SyVI_w}$) and dry ($\mathrm{SyVI_d}$) seasons by constraining the harmonized Landsat data. We constrained the wet and dry seasons from November to March and from May to September, respectively, between 1982 and

1994 when natural vegetation predominated over the landscape. In this work, SyVI represents potential natural vegetation (PNV), without or with minimal human intervention, that can be used as a proxy of the factor "organisms" in the *scorpan* model for estimating soil variables (MCBRATNEY et al., 2003). The PNV classification rules were constructed by combining the NDVI (Equation 1), NBR2 (Equation 4), the vegetation spectral shape index (VSI, Equation 6) and soil index (SI, Equation 7). The VSI and SI were elaborated by visual interpretation of the spectral shape of different types of vegetation collected from Landsat images using the MapBiomas dataset as a reference (PARENTE et al., 2019). To retrieve PNV reflectance in the study area, the thresholds were adjusted to NDVI ≥ 0.20, NBR2 ≥ 0.18, VSI < 11,000 and SI > 2. Therefore, selected PNV pixels were aggregated into two composites, for wet (SyVI$_w$) and dry (SyVI$_d$) seasons, by computing band-to-band the median of the reflectance values over the time series for both seasons.

$$VSI = Blue + Green + SWIR_1 + SWIR_2 + 2(Red + NIR) + \left(\frac{SWIR_1}{Green}\right) \times 100 \quad (6)$$

$$SI = \left(\left(\frac{\frac{SWIR_1}{NIR + SWIR_2}}{\frac{NIR + SWIR_2 - SWIR_1}{(NIR + SWIR_2 + SWIR_1)}}\right)^2\right)^{\frac{1}{2}} \quad (7)$$

Kriging. To obtain spatially continuous products (100% coverage) over the study area, we interpolated the gaps using ordinary kriging within GEE. Thus, the spectral values were randomly sampled from the composites (SySI, SyVI$_w$ and SyVI$_d$) using two observations per km^2. For each band, we fitted the spherical model to the empirical semivariogram to obtain the parameters (range, sill, nugget and maximum distance) and make the spatial interpolation of the values (MCBRATNEY et al., 1986). Finally, we overlaid the (original) composites on top of kriged images and merged them to obtain spatially continuous images (original + kriged). Merged images preserved the original values and incorporated the kriged where a gap occurred. Thus, the spatially continuous images (original + kriged), named as SySI, SyVI$_w$ and SyVI$_d$, were used as covariates for mapping soil attributes (Table 1), according the *scorpan* model.

Quality. We assessed the kriging results by sampling band-to-band 1 value per km^{-2} on overlapped areas between the synthetic image (original) and kriged (non-merged to synthetic image) and calculating the Pearson's correlation for the seven spectral bands. We checked the quality of the spatially continuous composites (original + kriged) by assessing 1) the reflectance values on the spectral profile, 2) the soil line method (BARET et al., 1993), and 3) the spatial consistency with land cover classifications (PARENTE et al., 2019). The soil line uses a scatterplot to display the reflectance between Red and NIR spectral regions. Both methodologies can be used to analyze the spectral patterns of the composites and to determine if they are consistent with the classical patterns of soils and vegetation.

Table 1. List of environmental covariates used as proxies of factors of soil formation for digital mapping of soil attributes in the study area.

Covariate	Description	Scale	Source
Soil, Parent Material and Age			
Synthetic Soil Image (SySI)	Bare soil reflectance covering VNIR-SWIR-TIR range (7 bands)	30 m	Landsat 4, 5, 7 and 8
Geological Lineaments Density	Meters of structural features per km^2	1:1M[a]	CPRM
Organisms			
Synthetic Vegetation Image of wet season ($SyVI_w$)	Potential natural vegetation reflectance from November to March covering VNIR-SWIR-TIR range (7 bands)	30 m	Landsat 4 and 5
Synthetic Vegetation Image of dry season ($SyVI_d$)	Potential natural vegetation reflectance from May to September covering VNIR-SWIR-TIR range (7 bands)	30 m	Landsat 4 and 5
Climate			
Annual Precipitation (mm)	Bioclimatic variables obtained from the monthly temperature and rainfall in order to generate more biologically meaningful values.	1 km[b]	WorldClim
Precipitation Seasonality (CV)		1 km[b]	WorldClim
Annual Mean Temperature (ºC)		1 km[b]	WorldClim
Temperature Annual Range(ºC)		1 km[b]	WorldClim
Temperature Seasonality (ºC)		1 km[b]	WorldClim
Relief and Age			
Elevation (m)	Height of terrain above sea level	30 m	ALOS
Slope (degree)	Slope gradient	30 m	ALOS
Aspect (degree)	Compass direction	30 m	ALOS
Topographic Position Index (m)	Distinguishes ridge from valley forms	30 m	ALOS
Horizontal Curvature (m)	Curvature tangent to the contour line	30 m	ALOS
Vertical Curvature (m)	Curvature tangent to the slope line	30 m	ALOS

VNIR: Visible and Near infrared spectral range (~450–900 nm); SWIR: Shortwave infrared spectral range (~1550–2350 nm); TIR: Thermal infrared spectral range (~10,400–12,500 nm). [a] Lines (1,000,000-scale) counted in grids of 1 km^2, transformed to raster and interpolated to 30 m pixel resolution by IDW method. [b] Interpolated to 30 m pixel resolution by Inverse Distance Weighted method; CV: coefficient of variation;.

2.3. Random Forest (RF) Regression

For DSM, we have implemented the *scorpan* model (MCBRATNEY et al., 2003) to predict the spatial patterns of soil attributes (Table 2). We used all soil forming factor proxies (Table 1) at the same time and let the decision tree algorithms reveal the soil patterns. We select RF regression for soil predictions. RF is a tree-based machine learning algorithm which consists of many decision or regression trees where each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the data (BREIMAN, 2001). The output of the model is an average of all the regression trees. The RF method is popular in DSM because it has proven to be efficient for mapping soil attributes across a wide range of data scenarios and scales of soil variability (FAO, 2018; GOMES et al., 2019; HENGL et al., 2014, 2015, 2017; LOISEAU et al., 2019; NUSSBAUM et al., 2018).

2.3.1 Calibration and Model Tuning

To generalize local patterns and to minimize possible artifacts in the final maps, the covariates (Table 1) were smoothed, prior to sampling, by computing the median values within a 4 × 4 moving window. At each soil observation (Table 2), the values were extracted and used as input data for calibrating RF regressions (BREIMAN, 2001) using the ranger package version 0.11.1 (WRIGHT et al., 2017) in the R software (R CORE TEAM, 2018). Usually, most modeling studies employed default hyperparameters, which can prompt models to under or over fit and misinterpretations of results. Thus, to improve the performance of RF models, we performed a grid search for (optimal) tuning hyperparameters (PROBST et al., 2019) investigating a range of values, where $mTry$ was 6, 24, 33, *minimum node size* was 5, 20, 50. The $mTry$ regulates the number of variables that can be randomly sampled in each split of the trees. The *minimum node size* controls the tree depth by setting the minimal number of samples for the terminal nodes. We used 500 trees for stable variable estimates (PROBST et al., 2019).

2.3.2 Validation and Variable Importance

In order to evaluate the models' performance for the prediction of each soil attribute at each of the three depths a 10-fold cross-validation was used. The observations were split into 10-folds by using the caret package version 6.0-84 (KUHN, 2019). According to Padarian et al. (2019), the k-cross-validation is a stable method, where the dataset is partitioned into k groups or folds, where $k - 1$ groups are used for training and 1 group for validation, repeating the training k times, each with a different validation group. For each predictive model, we derived the RMSE, coefficient of determination (R^2) and ratio of the performance to inter-quartile distance ($RPIQ = (Q3 - Q1)/\text{RMSE}$), where Q1 and Q3 are the 1st (25%) and 3rd (75%) quartiles. The RPIQ is based on prediction error and quartiles, which better represents the spread of the population and easier comparable across model validation studies. Generally, smaller values of RMSE and larger R^2 and RPIQ indicate better model performance (BELLON-MAUREL et al., 2010). We selected the optimized model by the minimum RMSE of the 10-fold cross-validation (FAO, 2018; PROBST et al., 2019).

To quantify the most influential covariates used in the models, the scaled permutation importance was calculated for each soil attribute prediction at each depth interval (BREIMAN, 2001), which were graphically displayed using the folds estimates.

2.3.3 Prediction of Continuous Soil Attributes

The optimized models were used to predict the spatial patterns of soil variables (Table 2) using RF optimized hyperparameters within GEE (GORELICK et al., 2017). In this study, the error or inaccuracy were not spatially examined as maps, because the GEE does not support this technique at the current stage of development. Furthermore, the GEE′ RF algorithm in probability mode only works for binary (presence/absence) datasets (GORELICK et al., 2017).

In addition, to verify the correspondence of our spatial predictions to their possible parent materials, we grouped lithologies using the soil attribute maps. Lithological data was obtained from a legacy geological 1:1,000,000-scale map (CPRM, 2004) and used (each geometry) for sampling the mean value from 0 to 100 cm depth interval of each soil attribute map. Afterward, we clustered the lithologies (geometries) into geological domains using the averaged soil value. For each domain (cluster), we identified the main lithotypes according to metadata (table data) of the geological map. Thus, we obtained a new outcome containing geological domains from a pedological viewpoint.

3. RESULTS

3.1. Summary and Relationships between Soil Attributes

The soil dataset (showed in Figure 1 and summarized in Table 2), covered the main peological classes of the study area. Overall, the mean clay content ranged from around 271 g kg^{-1} at the surface to 313 g kg^{-1} in the 60–100 cm depth interval. At the surface, clay content ranged from 10 to 920 g kg^{-1}, while at deeper layers, the maximum values were 930 and 950 g kg^{-1} (Table 2). There is relatively little silt in the studied soils, and the mean values does not vary much with depth. Silt content ranged from around 77 g kg^{-1} (0–20 cm) to 67 g kg^{-1} (60–100 cm). The silt data at the three depths was positively skewed. The average sand content ranged from 652 g kg^{-1} at the surface to 619 g kg^{-1} in the 60–100 cm depth. At all depth intervals, values ranged from 1 to 975 g kg^{-1} (Table 2).

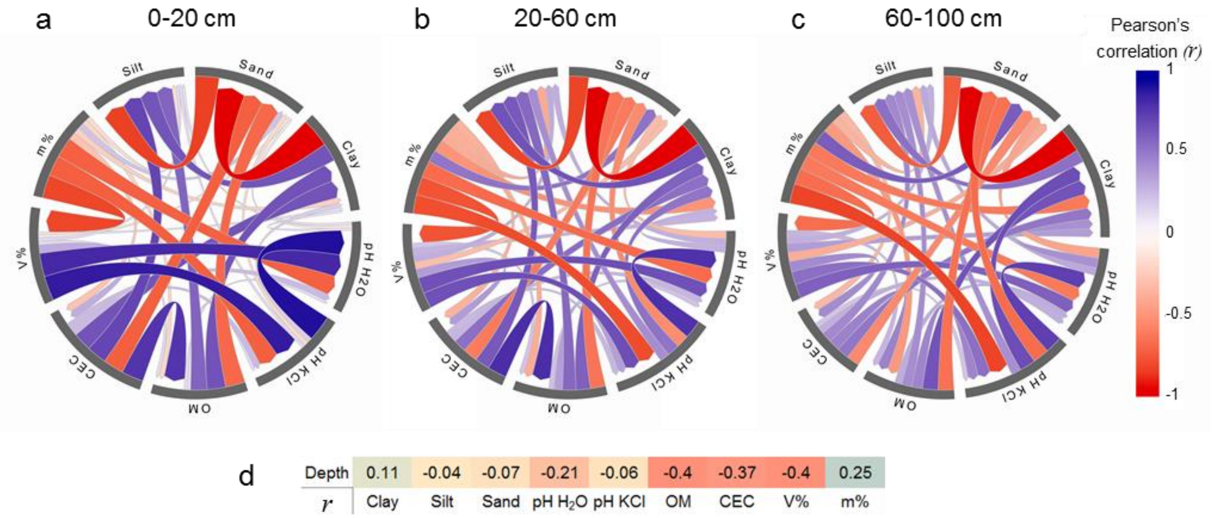
The average OM content ranged from around 21 g kg^{-1} at the surface to 9 g kg^{-1} in the 60–100 cm depth. At all depth intervals, the mean values of pH H_2O were greater than pH KCl, ranging from 5.6 (0–20 cm) to 5.3 (60–100 cm). The average pH KCl ranged from 4.9 at the surface to 4.8 in the 60-100 cm depth. At the surface, CEC ranged from 2 to 641 mmol$_c$ kg^{-1}, while at deepest, the maximum values were between 1 and 582 mmol$_c$ kg^{-1} (Table 2). At all depth intervals, V and m% values ranged from 0 to 100 %, with averages varying inversely with depth (Table 2).

Correlation between soil attributes had similar patterns for each depth interval (Figure 3a–c). Sand and clay presented the highest negative correlation at all depth intervals, while for silt was slightly lower. The pH H_2O, pH KCl and V% positively correlated in the three depths, and negatively with m%. OM positively correlated with CEC at each of the three depth intervals. Chemical attributes, weakly correlated among each other at topsoil, became stronger with increasing depth intervals (Figure 3d), whereas the strongest, slightly decreased at deeper depths.

Table 2. Statistical Summary of Soil Data Aggregated into Different Depth Intervals for Spatial Modelling.

Soil attribute	Depth*	n	Min.	Q1	Mean	Med.	Q3	Max.	Sd	IQR	Skew.
Clay	0-20	7930	10	87	271	176	450	920	221	363	0.8
(g kg^{-1})	20-60	6908	10	100	287	176	482	930	233	382	0.8
	60-100	7520	12	125	314	225	500	950	231	375	0.8
Silt	0-20	7930	1	24	77	38	94	816	89	70	2.3
(g kg^{-1})	20-60	6907	1	24	71	37	82	760	79	58	2.3
	60-100	7520	1	24	67	37	80	794	75	56	2.6
Sand	0-20	7930	1	409	652	783	883	975	280	474	-0.8
(g kg^{-1})	20-60	6907	1	393	643	783	873	973	284	480	-0.8
	60-100	7520	1	377	619	741	848	967	276	471	-0.7
Organic Matter	0-20	7242	0	11	21	17	28	393	14	17	4.8
(g kg^{-1})	20-60	6021	0	7	13	11	17	412	9	10	15.2
	60-100	6808	0	4	9	8	12	98	6	7	2.3
pH H_2O	0-20	6200	3.7	5.2	5.6	5.6	6.0	8.2	0.6	0.8	0.1
(log)	20-60	5149	3.8	4.9	5.3	5.2	5.6	9.0	0.6	0.7	0.7
	60-100	7511	3.8	4.9	5.3	5.2	5.6	9.1	0.5	0.7	0.7
pH KCl	0-20	5596	3.1	4.6	4.9	4.8	5.3	7.7	0.6	1.0	0.6
(log)	20-60	4707	0.4	4.2	4.6	4.4	4.9	7.7	0.5	0.7	1.1
	60-100	7384	3.5	4.3	4.8	4.5	5.2	7.5	0.6	0.9	0.9
CEC	0-20	8010	2	32	53	45	68	641	33	37	3.0
(mmol$_c$ kg^{-1})	20-60	6852	2	22	36	32	45	696	23	23	5.7
	60-100	7655	1	16	26	22	32	582	18	16	6.2
Base Saturation	0-20	8018	0	24	42	42	58	100	22	34	0.2
(V%)	20-60	6860	0	12	25	21	34	100	18	23	1.2
	60-100	7655	0	10	23	18	31	100	17	20	1.5
Aluminum	0-20	7964	0	0	16	4	24	100	23	24	1.6
Saturation	20-60	6841	0	5	33	28	57	100	29	52	0.4
(m%)	60-100	7635	0	3	36	34	62	100	30	59	0.3

* in centimeters; n: number of soil observations; Min.: minimum value; Q1/Q3: 1st (25%) and 3rd (75%) quartiles; Max.: maximum value; Med: median; Sd: standard deviation; IQR: interquartile range; Skew: skewness; Organic Matter $= organic\ carbon \times 1.72$; CEC: cation exchange capacity $= (Ca^{2+} + Mg^{2+} + K^{+} + H^{+} + Al^{3+})$; $V\% = \left((Ca^{2+} + Mg^{2+} + K^{+}) \times 100\right) \div CEC)$; $m\% = \left((Al^{3+} \times 100) \div (Ca^{2+} + Mg^{2+} + K^{+} + Al^{3+})\right)$.



Depth	0.11	-0.04	-0.07	-0.21	-0.06	-0.4	-0.37	-0.4	0.25
r	Clay	Silt	Sand	pH H_2O	pH KCl	OM	CEC	V%	m%

Figure 3. Chord diagram based on Pearson correlation (r) among all measured soil attributes at (**a**) 0–20 cm, (**b**) 20–60 cm and (**c**) 60–100 cm depth intervals; and (**d**) overall correlation with depth intervals. Blue and red colors symbolize positive and negative correlations, respectively. OM: organic matter; V%: base saturation; m%: aluminum saturation.

3.2. Synthetic Soil Image (SySI) and Synthetic Vegetation Image (SyVI)

The harmonized Landsat data and the data mining algorithms implemented in GEE enabled to obtain a SySI with 443,000 km^2 (52%) coverage, which was later kriged to close the gaps. The bare soil frequency (data not presented) at each locations ranged from 1 to 1303 pixels and average of 12 pixels, between the 1982 and 2019 years. For potential natural vegetation from 1982 to 1994, we obtained a SyVI_w with 814,175 km^2 (95.2 %) and a SyVI_d with 847,426 km^2 (99.1 %) coverage during the wet and dry seasons, respectively. The PNV frequency (at every locations) in the wet season ranged from 1 to 85 pixels, and mean values of 6 pixels. During the dry season, the PNV frequency had average values of 19 pixels, ranging from 1 to 185 pixels. The kriged images had satisfactory correlation (Pearson) with the originals, which presented average values of 0.66 (SySI), 0.78 (SyVI_w) and 0.81 (SyVI_d) for the seven spectral bands (Figure 4a–c). The full coverage SySI, SyVI_w and SyVI_d with the NA gaps interpolated by kriging (original + krikeg) were displayed in Figure 4a-c.

The soil line for bare soil (Figure 4d) had an adjustment near to the 1:1 trend line, with highly correlated values (R^2 of 0.95), while raw (unprocessed) pixels extracted from a median composite (between 2017 and 2019) had a scatter distribution. For PNV, the soil line had clustered values with lower reflectance intensities (Figure 4e–f) compared to the raw pixels sampled from the median composite (between 2017 and 2019) over croplands mapped by MapBiomas (PARENTE et al., 2019). The mean NIR reflectance was higher for SyVI_w (2,471) than for SyVI_d (2,123), while the mean Red reflectance was higher for SyVI_d (611) than for SyVI_w (536).

The spectral signature for bare soil (Figure 4g) had a constant ascendant pattern from Blue to SWIR_1 regions, while the PNV (Figure 4h-i) had an opposite overall pattern, ascending from Blue to NIR and then descending to SWIR$_2$. The SySI averaged a LST of 38.7 °C (Figure 4g), higher than for the SyVI$_w$ and SyVI$_d$, with mean values of 22.58 and 23.04 °C (Figure 4h–i), respectively. Remaining soil covariates (Table 1) were placed in the Appendix A as Figure A1.

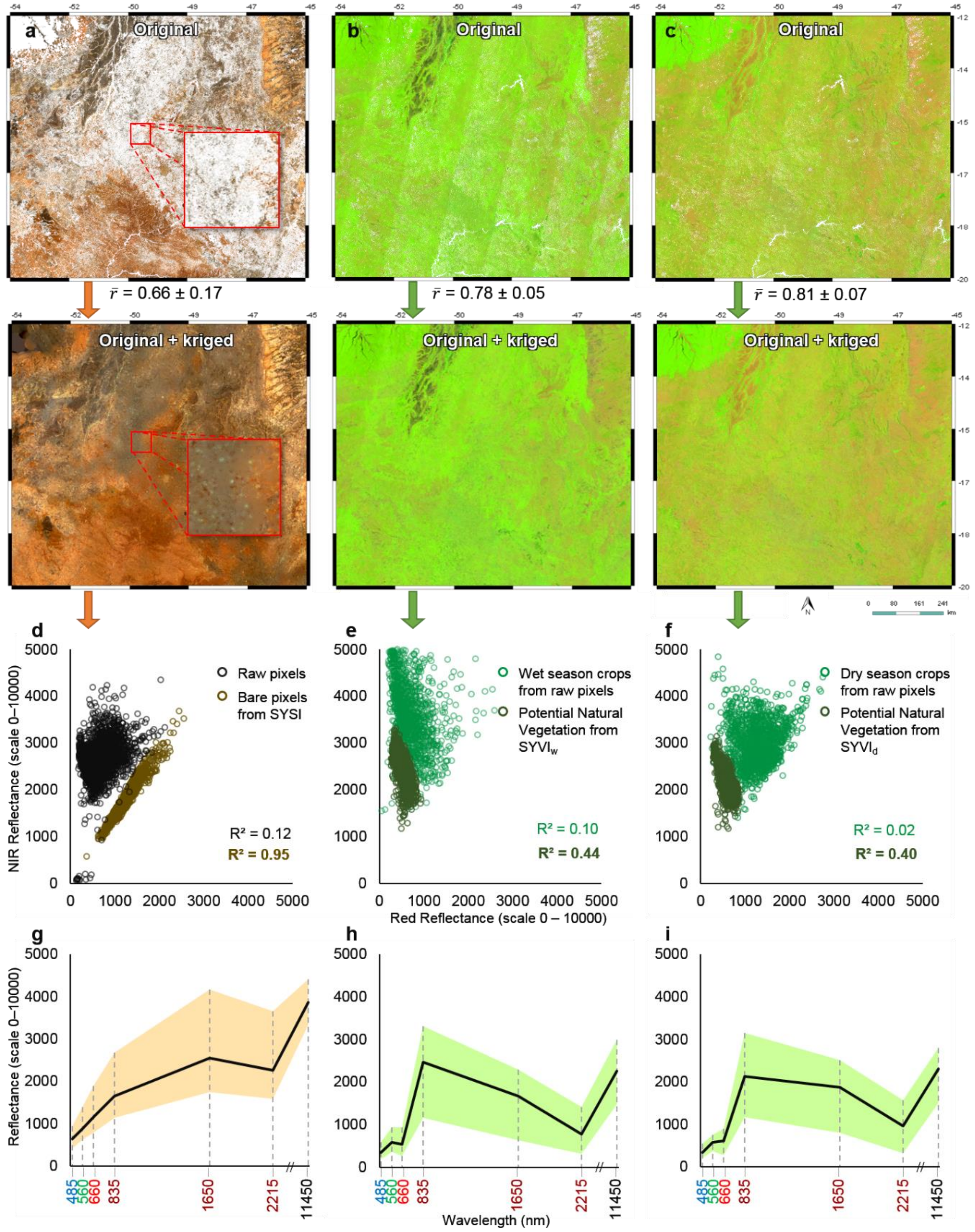


Figure 4. Soil covariates obtained by data mining and statistics of Landsat data. a) SySI (RGB: Red, Green, Blue), b) SyVI$_w$ and c) SyVI$_d$ (RGB: SWIR$_1$, NIR, Red) subsets. Soil line charts for d) SySI vs raw pixels, e) SyVI$_w$ vs wet season crops obtained from raw pixels and f) SyVI$_d$ vs dry season crops obtained from raw pixels. Minimum, average and maximum spectra collected from g) SySI, h)

$SyVI_w$ and i) $SyVI_d$. The visualization of the images was adjusted by stretching the range of pixel values between 2% and 98%. Optical bands are positioned in the mean spectral range from 485 to 2215 nm, and the thermal band at 11,450 nm. $\bar{r}$: average of Pearson's correlation from the seven spectral bands, performed by sampling 1 value km^{-2} on overlapped areas between the synthetic image (original) and kriged (non-merged to synthetic image).

3.3. Model Assessments

Table 3 shows the performance of optimized RF regression models on calibration ($_{cal}$) and validation ($_{10cv}$) sets. Predicted vs observed scatterplots from 10-fold cross-validation derived from the models of sand, silt and clay were placed in the Appendix A as Figure A2, while the remaining soil attributes are displayed in Figure A3.

We obtained decreasing RMSE and increasing RPIQ with increasing depth interval, both in calibration and validation data. The relatively low values of $RMSE_{10cv}$ suggested that the soil variables were slightly overestimated for all the models. On average, $RPIQ_{10cv}$ and R^2_{10cv} increased slightly from 0–20 to 60–100 cm depth, while decreased for silt and CEC. Sand and clay presented the best model' predictive capacity with the highest $RPIQ_{10cv}$ (from 3.8 to 4.3), followed by m% > pH KCl > OM > V% > pH H_2O > CEC > silt, ranging from 1.2 to 3.0 (Table 3).

Overall, the amount of variation explained by the spatial prediction models in validation were reasonable at all depths, with higher values for sand and clay (R^2_{10cv} from 0.81 to 0.85) followed by silt (R^2_{10cv} from 0.64 to 0.66). Chemical attributes were best explained for pH KCl (R^2_{10cv} from 0.19 to 0.64), m% (R^2_{10cv} from 0.26 to 0.56), OM (R^2_{10cv} from 0.30 to 0.53) and CEC (R^2_{10cv} from 0.40 to 0.48). The poorest performances were for V% (R^2_{10cv} from 0.18 to 0.36) and pH H_2O (R^2_{10cv} from 0.21 to 0.35) (Table 3). We observed that R^2_{10cv} and $RPIQ_{10cv}$ had a positive relationship in most models (Figure 5), where higher values indicate greater robustness in predictive capability. Models with poor performance exhibited a scatterplot (predicted vs. observed) with higher dispersion and weaker trend, while good models showed more distributed values following a stronger linear trend (Figures A2 and A3).

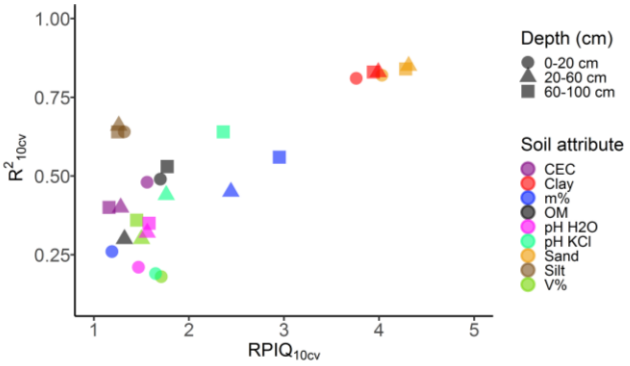


Figure 5. Performance indicators of optimized models used in the soil predictions by depth interval.

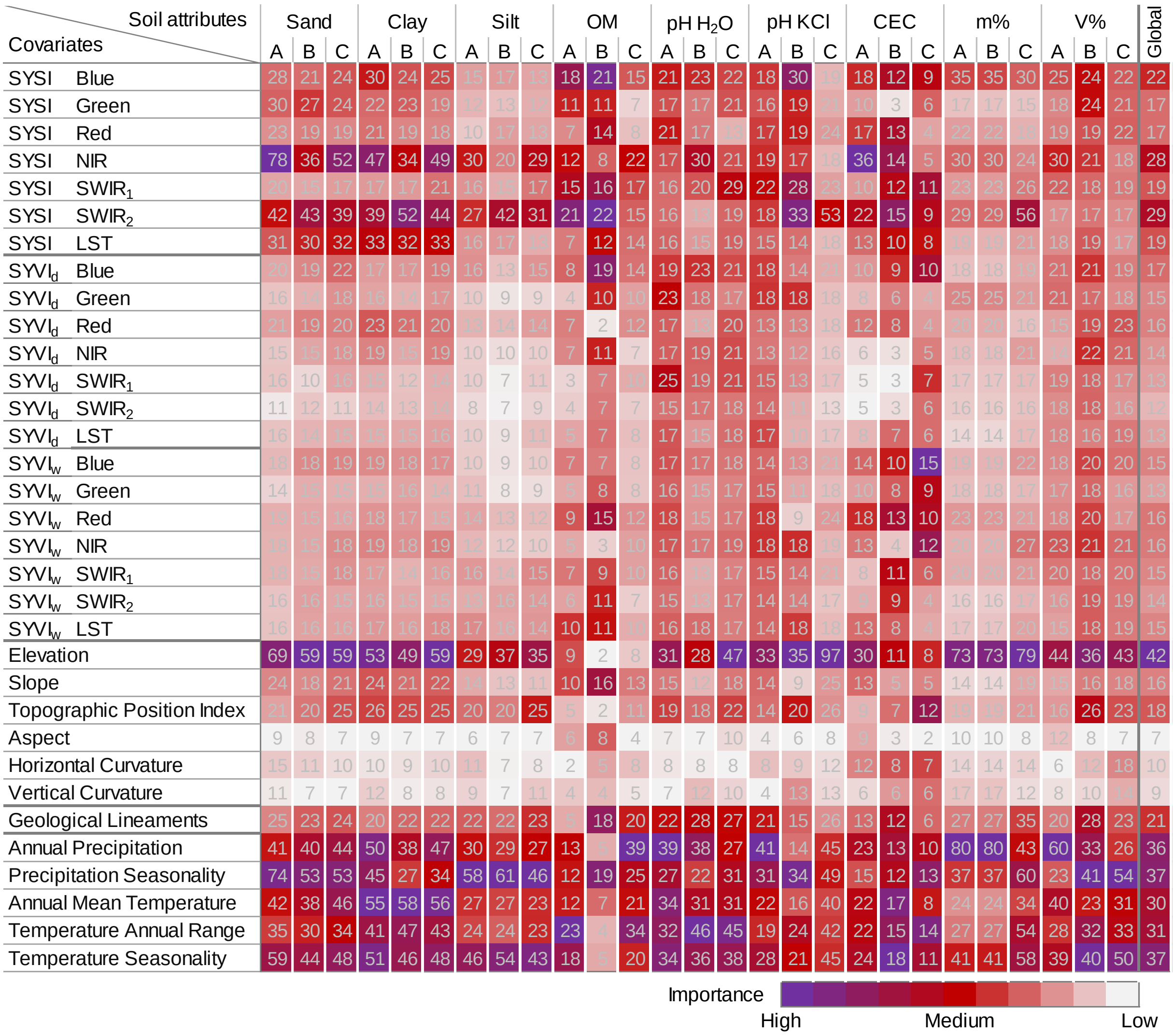
Table 3. Hyperparameters and Performance of the Optimized Models used for Spatial Predictions of Continuous Soil Attributes at Distinct Depth Intervals.

Soil attribute	Depth (cm)	mTry	minNS	$RMSE_{cal}$	$RPIQ_{cal}$	R^2_{cal}	$RMSE_{10cv}$	$RPIQ_{10cv}$	R^2_{10cv}
Clay (g kg^{-1})	0-20	24	5	39	9.4	0.97	96	3.8	0.81
	20-60	24	5	38	10.0	0.97	96	4.0	0.83
	60-100	24	5	38	9.9	0.97	95	4.0	0.83
Silt (g kg^{-1})	0-20	24	5	21	3.3	0.94	53	1.3	0.64
	20-60	33	5	18	3.2	0.95	46	1.3	0.66
	60-100	24	5	18	3.1	0.94	45	1.3	0.64
Sand (g kg^{-1})	0-20	33	5	47	10.1	0.97	118	4.0	0.82
	20-60	24	5	45	10.7	0.98	111	4.3	0.85
	60-100	24	5	44	10.6	0.97	110	4.3	0.84
Organic Matter (g kg^{-1})	0-20	33	5	4	4.1	0.91	10	1.7	0.49
	20-60	33	5	3	3.4	0.90	8	1.3	0.30
	60-100	24	5	2	4.3	0.92	4	1.8	0.53
pH H_2O (log)	0-20	33	5	0.21	3.7	0.88	0.54	1.5	0.21
	20-60	33	5	0.19	3.9	0.89	0.47	1.6	0.32
	60-100	33	5	0.18	3.9	0.90	0.44	1.6	0.35
pH KCl (log)	0-20	33	5	0.23	4.2	0.87	0.57	1.7	0.19
	20-60	33	5	0.16	4.3	0.91	0.40	1.8	0.44
	60-100	24	5	0.15	5.9	0.94	0.38	2.4	0.64
CEC ($mmol_c$ kg-1)	0-20	33	5	10	3.7	0.91	23	1.6	0.48
	20-60	24	5	8	3.0	0.89	18	1.3	0.40
	60-100	24	5	6	2.7	0.89	14	1.2	0.40
Base Saturation (V%)	0-20	33	5	8	4.4	0.87	20	1.7	0.18
	20-60	33	5	6	3.7	0.89	15	1.5	0.30
	60-100	33	5	6	3.6	0.89	14	1.5	0.36
Aluminum Saturation (m%)	0-20	33	5	8	2.9	0.88	20	1.2	0.26
	20-60	33	5	9	6.1	0.91	21	2.4	0.45
	60-100	24	5	8	7.4	0.93	20	3.0	0.56

CEC: Cation exchange capacity; mTry: hyperparameter that regulates the number of variables that can be randomly sampled in each split of the trees; minNS: minimum node size, a hyperparameter that controls the tree depth by setting the minimal number of samples for the terminal nodes. $RMSE_{cal}$: Root Mean Square Error of calibration; $RMSE_{10cv}$: Root Mean Square Error of 10-fold cross-validation; RPIQ*cal*: Ratio of the Performance to Inter-Quartile distance of calibration; RPIQ*10cv*: Ratio of the Performance to Inter-Quartile distance of 10-fold cross-validation; R^2_{cal}: Coefficient of determination of calibration; R^2_{10cv}: Coefficient of determination of 10-fold cross-validation.

3.4. Best Predictors

Figure 6 shows the permutation importance (%) of all the 33 covariates in RF models for the spatial prediction of 9 soil attributes at three depth intervals. From a general view (global values, Figure 6), the results indicated that the most important covariates were elevation, the five climate layers, SWIR$_2$–NIR–Blue reflectance bands derived from SySI, ranging their estimates from 22% to 42%. The importance values did not vary much with depth, except for OM and CEC, which had slight differences. That is because the regional patterns from the coarser covariates could help the RF models in stratifying the region at the coarser level, while the more detailed information from the finer resolution covariates can represent the variability within the regional patterns.



Soil attributes / Covariates	Sand			Clay			Silt			OM			pH H_2O			pH KCl			CEC			m%			V%			Global
	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	
SYSI Blue	28	21	24	30	24	25	15	17	13	18	21	15	21	23	22	18	30	19	18	12	9	35	35	30	25	24	22	22
SYSI Green	30	27	24	22	23	19	12	13	12	11	11	7	17	17	21	16	19	21	10	3	6	17	17	15	18	24	21	17
SYSI Red	23	19	19	21	19	18	10	17	13	7	14	8	21	17	13	17	19	24	17	13	4	22	22	18	19	19	22	17
SYSI NIR	78	36	52	47	34	49	30	20	29	12	8	22	17	30	21	19	17	18	36	14	5	30	30	24	30	21	18	28
SYSI $SWIR_1$	20	15	17	17	17	21	16	15	17	15	16	17	16	20	29	22	28	23	10	12	11	23	23	26	22	18	19	19
SYSI $SWIR_2$	42	43	39	39	52	44	27	42	31	21	22	15	16	13	19	18	33	53	22	15	9	29	29	56	17	17	17	29
SYSI LST	31	30	32	33	32	33	16	17	13	7	12	14	16	15	19	15	14	18	13	10	8	19	19	21	18	19	17	19
$SYVI_d$ Blue	20	19	22	17	17	19	16	13	15	8	19	14	19	23	21	18	14	21	10	9	10	18	18	19	21	21	19	17
$SYVI_d$ Green	16	14	18	16	14	17	10	9	9	4	10	10	23	18	17	18	18	18	8	6	4	25	25	21	21	17	18	15
$SYVI_d$ Red	21	19	20	23	21	20	13	14	14	7	2	12	17	13	20	13	13	18	12	8	4	20	20	16	15	19	23	16
$SYVI_d$ NIR	15	15	18	19	15	19	10	10	10	7	11	7	17	19	21	13	12	16	6	3	5	18	18	21	14	22	21	14
$SYVI_d$ $SWIR_1$	16	10	16	15	12	14	10	7	11	3	7	10	25	19	21	15	13	17	5	3	7	17	17	17	19	18	17	13
$SYVI_d$ $SWIR_2$	11	12	11	14	13	14	8	7	9	4	7	7	15	17	18	14	11	13	5	3	6	16	16	16	18	18	16	12
$SYVI_d$ LST	16	14	15	15	15	16	10	9	11	5	7	8	17	15	18	17	10	17	8	7	6	14	14	17	18	16	19	13
$SYVI_w$ Blue	18	18	19	19	18	17	10	9	10	7	7	8	17	17	18	14	13	21	14	10	15	19	19	22	18	20	20	15
$SYVI_w$ Green	14	15	15	15	16	14	11	8	9	5	8	8	16	15	17	15	11	18	10	8	9	18	18	17	17	18	16	13
$SYVI_w$ Red	19	15	16	18	17	15	14	13	12	9	15	12	18	15	17	18	9	24	18	13	10	23	23	21	18	20	17	16
$SYVI_w$ NIR	18	15	18	19	18	19	12	12	10	5	3	10	17	17	19	18	18	19	13	4	12	20	20	27	23	21	21	16
$SYVI_w$ $SWIR_1$	18	15	18	17	14	16	16	14	15	7	9	10	16	13	17	15	14	21	8	11	6	20	20	21	20	18	20	15
$SYVI_w$ $SWIR_2$	16	16	15	16	15	15	13	16	14	6	11	7	15	13	17	14	14	17	9	9	4	16	16	17	16	19	19	14
$SYVI_w$ LST	16	16	16	17	16	18	17	16	14	10	11	10	16	18	17	14	18	18	13	8	4	17	17	20	15	18	19	15
Elevation	69	59	59	53	49	59	29	37	35	9	2	8	31	28	47	33	35	97	30	11	8	73	73	79	44	36	43	42
Slope	24	18	21	24	21	22	14	13	11	10	16	13	15	12	18	14	9	25	13	5	5	14	14	19	15	16	18	16
Topographic Position Index	21	20	25	26	25	25	20	20	25	5	2	11	19	18	22	14	20	26	9	7	12	19	19	21	16	26	23	18
Aspect	9	8	7	9	7	7	6	7	7	6	8	4	7	7	10	4	6	8	9	3	2	10	10	8	12	8	7	7
Horizontal Curvature	15	11	10	10	9	10	11	7	8	2	5	8	8	8	8	8	9	12	12	8	7	14	14	14	6	12	18	10
Vertical Curvature	11	7	7	12	8	8	9	7	11	4	4	5	7	12	10	4	13	13	6	6	6	17	17	12	8	10	14	9
Geological Lineaments	25	23	24	20	22	22	22	22	23	5	18	20	22	28	27	21	15	26	13	12	6	27	27	35	20	28	23	21
Annual Precipitation	41	40	44	50	38	47	30	29	27	13	5	39	39	38	27	41	14	45	23	13	10	80	80	43	60	33	26	36
Precipitation Seasonality	74	53	53	45	27	34	58	61	46	12	19	25	27	22	31	31	34	49	15	12	13	37	37	60	23	41	54	37
Annual Mean Temperature	42	38	46	55	58	56	27	27	23	12	7	21	34	31	31	22	16	40	22	17	8	24	24	34	40	23	31	30
Temperature Annual Range	35	30	34	41	47	43	24	24	23	23	4	34	32	46	45	19	24	42	22	15	14	27	27	54	28	32	33	31
Temperature Seasonality	59	44	48	51	46	48	46	54	43	18	5	20	34	36	38	28	21	45	24	18	11	41	41	58	39	40	50	37

Figure 6. Permutation importance (%) of covariates for prediction soil attributes at 0–20 cm (**A**), 20–60 cm (**B**) and 60–100 cm (**C**) depth intervals. The mean values were calculated from the importance obtained by the 10-fold used in cross-validation. CEC: cation exchange capacity; m%: Aluminum saturation; V%: Base saturation. Global represents averaged importance values for all soil attributes.

Elevation, climate and soil reflectance derived from SYSI (NIR, $SWIR_2$ and LST) were the most important covariates in predicting sand, clay and silt (Figure 6). In general, for predicting soil chemical attributes, climatic variables, elevation and SYSI (NIR and $SWIR_2$ bands) seemed to be the most important, while for OM it was a combination of climatic dynamics and reflectance bands derived from SYSI, i.e. $SWIR_2$, Blue and $SWIR_1$.

Furthermore, the results indicated that PNV reflectance and temperature derived from $SyVI_w$ and $SyVI_d$, geological lineaments density, topographic position index and slope were mid important (from 12% to 21%) for whole soil attributes at all depths, with slightly higher values for the chemicals such as OM, pH, CEC and V% (Figure 6). In all cases, the least important were aspect, horizontal and vertical curvatures, which had an average importance of less than 10%.

3.5. Soil Maps at Multiple Depths

Figure 7a–c shows the maps of sand, silt and clay contents (g kg^{-1}) in each of the three depth intervals. These maps were made publicly available for download as integer GeoTIFF format at 250 m-resolution (POPPIEL et al., 2019a). The soils of the study region were dominated by high to moderate amounts of sand, moderate clay and little silt. Sand and clay maps were inversely distributed in the region (Figure 7a, c), due to their negative correlation (Figure 3). The largest sand contents were located southwest of the study area, decreasing gradually to the north and severely to the east. The silt and clay maps followed a very similar spatial distribution between them (Figure 7b, c), due to their positive correlation (Figure 3). There was more clay and silt in the east highlands of the study area, while a decreasing value was observed on the west lowlands.

In general, mean sand content decreased with depth from around 522 g kg^{-1} in the surface to 467 g kg^{-1} in the deepest layer, while at the same depths, mean clay content increased from 336 to 400 g kg^{-1}. Average silt content remained relatively uniform with increasing depth (Figure 7).

For each depth, a map representing the sum of the sand, silt and clay contents (Figure 7d) were used to show where the estimates of soil texture diverged from 1000 g kg^{-1}. On average, 87% of the predicted summed for the three depths ranged from 800 to 1200 g kg^{-1}. Under and overestimates in soil texture were visually more related to the spatial patterns of the silt map (Figure 7b). Maps of the soil chemical attributes (Table 2) for all depths were placed in the Appendix A, as Figure A4.

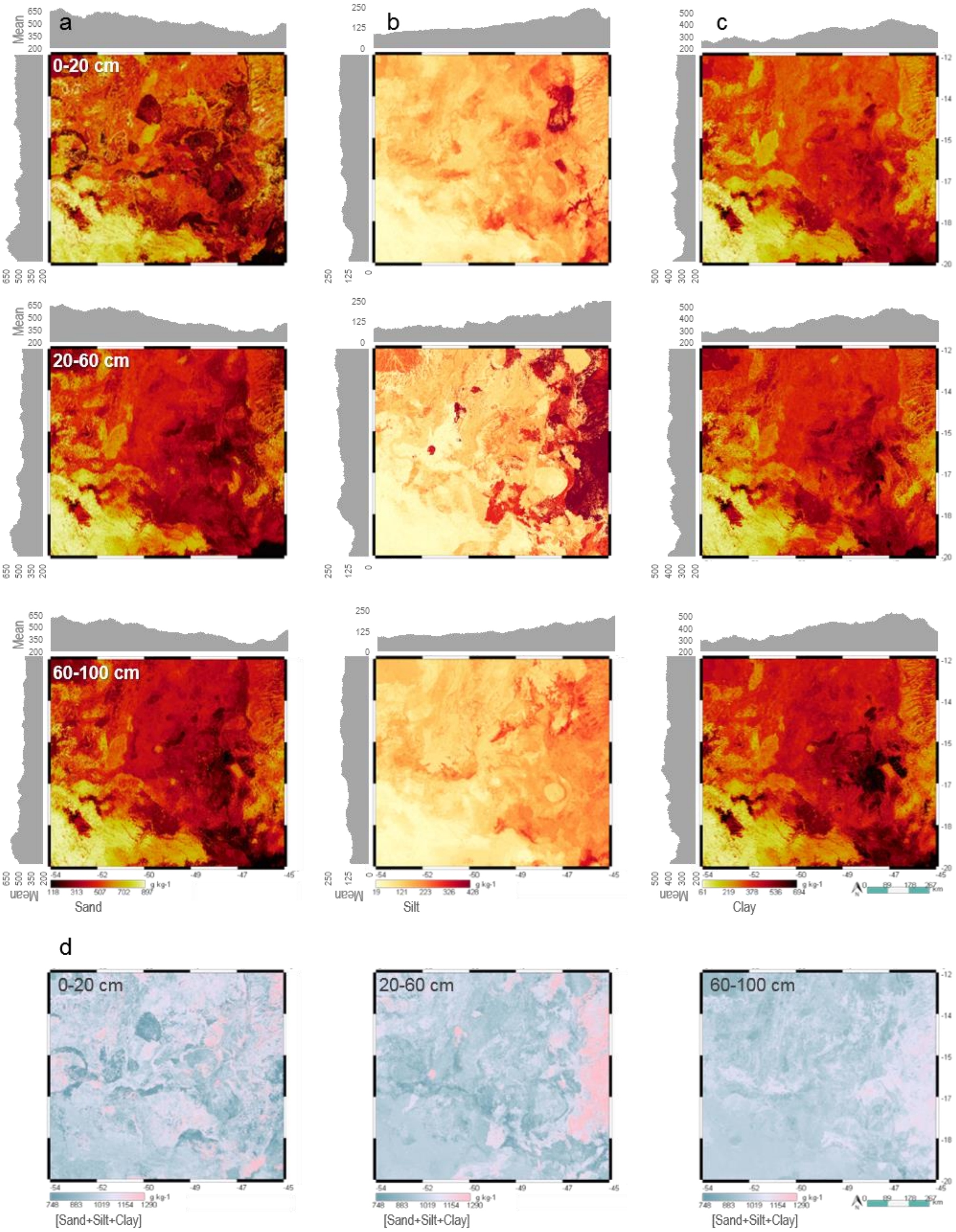


Figure 7. Maps and mean lat/long distribution chart of (**a**) sand, (**b**) silt and (**c**) clay content (g kg^{-1}) for different depth intervals (0–20 cm, 20–60 cm and 60–100 cm). The sum of the sand, silt and clay contents [Sand+Silt+Clay] for each depth interval appears in (**d**). The visualization of the images was adjusted by stretching the range of pixel values between 2% and 98%.

4. DISCUSSION

4.1. Soil Data

The soil data (BESB and FEBR) aggregated into depth intervals yielded consistent information for mapping selected key soil attributes (Table 2). Data showed that the amount of soil attributes and the intensity of their relationships were soil depth-dependent. The influence of soil forming factor (MCBRATNEY et al., 2003) causes pedogenic processes with different intensities across space and soil profile, resulting in gradients of soil attributes with different correlations among them and between different depth intervals (MA et al., 2019). These gradients can provide differences in nutrient release to soil solution for vegetation, to which they may reply back with different root systems. Similar patterns were described by Goebes et al. (2019), who distinguished between stable (e.g. sand, silt and clay) and dynamic (e.g. soil pH, nutrient contents, base saturation) soil attributes throughout the whole soil profile.

Despite soil observations were spatially dense and accounted for the main soil classes that developed in the region (Figure 1), there are still some gaps in terms of spatial coverage. Natural vegetation and pasturelands remained under-represented. This is because many soil observations, from different studies and existing repositories for the study area, are still publicly unavailable for users, which could become open access to increase spatial coverage and improve predictions (SAMUEL-ROSA et al., 2020). Considering the large extent of our study area, we performed relatively low-cost mapping by using legacy soil observations and comparatively fewer new soil observations, coupled with free RS-based covariates and open-source algorithms for data processing and visualization, available within R software and Google Earth Engine cloud-based platform.

4.2. Machine Learning

RF was satisfactory for DSM even with soil observations that partially covered the large extent of the study area (~851,000 km^2), where relationship between soil attributes and covariates is usually complex and non-linear (BREIMAN, 2001; HENGL et al., 2017). Therefore, the regression models used covariates that captured spatial patterns from broader to more local levels across different landscapes (HENGL et al., 2018). Our validation results were similar or even higher to those obtained in other DSMs using machine learning and cross-validation (GOMES et al., 2019; HENGL et al., 2015; LOISEAU et al., 2019; MA et al., 2017; NUSSBAUM et al., 2018; VAUDOUR et al., 2019; VISCARRA ROSSEL et al., 2015). Most of recently DSM studies also found tree-based models as the best option for soil spatial predictions (HENGL et al., 2015; LOISEAU et al., 2019; NUSSBAUM et al., 2018). Performance in these cases usually vary between 0.3 to 0.5 (R^2), with clay content being the best predicted attribute. Variable importance indicated satellite images (BUI et al., 2006; LOISEAU et al., 2019), elevation and climate data (BUI et al., 2009; GOMES et al., 2019; LOISEAU et al., 2019) as relevant covariates.

The good performance of the models showed that RF optimization was able to generate robust and accurate spatial predictions. This approach agreed with Probst et al. (PROBST et al., 2019), who provided different optimization strategies and reported that tuning the RF hyperparameters improved the performance of regression models. Sand, silt and clay had the best performances because they are stable soil attributes, and the chemicals are dynamic along the soil profile (HENGL et al., 2017). The pH and nutrient contents may change relatively quickly (within years) related to biological processes, vegetation cover and management practices (GOEBES et al., 2019). Gomes et al. (2019) mapped soil organic carbon at five standard depths (from 0 to 100 cm) for Brazilian territory, where RF showed the best performance for all depths, with the highest performance at 30–60 cm for validation ($R^2 = 0.33$). Bui et al. (2009) reported similar performances for topsoil (R^2=0.49) and subsoil (R^2 = 0.36) when used analogous covariates and data mining for mapping soil organic carbon in Australia.

The better model performance in lower layers are related to soil conditions at such depths. A possible factor impacting surface-subsurface predictions are the agricultural practices, where soil management could be increasing the system's complexity (MENDES et al., 2019). While the chemical and physical weathering are more intense and active in surface, alterations in depth tend to be less intense (BUI et al., 2006). This suggests that the models for topsoil were more influenced by climatic variables, i.e. precipitation and temperature, which lowered the performances. Therefore, subsurface soils usually have conditions closer to the ones observed in pristine areas, and could have a better relationship with soil forming factors and covariates considered in our study.

4.3. Interpretation of Covariate Layers

We did not perform covariates selection (elimination) because this approach could generated additional load of interpretation to the project, and because RF can be used to fit models with large number of covariates (HENGL et al., 2018). For instance, Nussbaum et al. (2018) evaluated six approaches for DSM of several soil variables (totaling 48 responses) using from 300 to 500 environmental covariates, where RF models had the highest overall performances. Miller et al. (2015) demonstrated that the best performing model was produced when using multi-resolution covariates, compared to a single resolution, for modeling the distribution of soil attributes at surface and subsurface. Relevant covariates for soil prediction had large importance values, whereas covariates not associated with the soil attributes showed values close to zero (Figure 6).

Our results showed that direct measurement (where soil areas were exposed) of topsoil reflectance patterns by RS was a strong contributor to soil mapping. The topsoil reflectance from SySI (Figure 4a) was important for the spatial prediction of soil attributes at the rooting depth of crops (CANADELL et al., 1996) in Midwest Brazil (Figure 6). That was possible because the spectral patterns of SySI can provide valuable information on pedogenic processes, which are useful for understanding and predicting soil variation (MA et al., 2019). The SySI also can indicate the soil weathering products, which cause spatial variations in the soil color and temporally stable soil attributes, such as mineralogy and texture (DEMATTÊ et al., 2018; POPPIEL et al., 2019b). Thus, complementary RS data can improve prediction models, as reported by Loiseau et al. (2019) where adding RS covariates increased the R^2 and decreased the bias of the clay content estimation on bare topsoil layers (e.g., 0–30 cm).

It is recognized within the soil science community that vegetation plays significant roles on soil formation (MCBRATNEY et al., 2003). However, Savin et al. (2019) stated that the use of vegetation patterns from RS for soil interpretation is insufficiently studied. Some previous works found that the spectral response of vegetation in natural conditions can be confidently used as an indirect indicator of soil attributes (HENGL, WALSH, et al., 2018; MAYNARD et al., 2017; SERTESER et al., 2008). Our results pointed out that PNV (from $SyVI_w$ and $SyVI_d$) was influential for modelling soil attributes at all depths, especially for chemical variables (Figure 6). The soil–vegetation connection can be due to the spatial and seasonal differences in reflectance intensities between wet and dry conditions (Figure 4b–c), that showed relations between the spectral patterns of natural vegetation and soil attributes from 0 to 100 cm (rooting) depth (Figure 6). Since vegetation is temporally dynamic, the relationships are largely controlled by available soil moisture and, to a lesser extent, chemical soil properties such as pH and fertility (MAYNARD et al., 2017). Thus, average seasonal spectral patterns of vegetation provide a better indication of soil variables than only a single snapshot of surface reflectance, and it is probably the best way to effectively represent the cumulative influence of living organisms on soil formation (HENGL et al., 2017).

Climate, relief and geology played significant roles in model prediction (Figure 6) because they can significantly influence the soil-vegetation feedback, as described by McBratney et al. (2003). The climate and geologic heterogeneity of the study region affected soil patterns at the macroscale (regional), followed by relief (especially elevation), which moderates many of the macroclimatic regimes, and landforms, at the meso and microscales (BAILEY, 1987; FLORINSKY, 2016). Das Sumit (2019) demonstrated that geological lineaments density was strongly related to drainage density, soil texture and soil depth, controlling the movement of groundwater through soil.

Landforms affect surface water dynamic and exposure to radiant solar energy, which directly influence soil-forming processes (FLORINSKY, 2016). Within a landform, there exists slight differences in local edaphic conditions, such as soil texture and mineralogy, and soil moisture and temperature regimes (BAILEY, 1987). These local conditions provides the most significant alterations of the soil reflectance patterns (DEMATTÊ et al., 2018; POPPIEL et al., 2019b) and segregation of the plant communities (BAILEY, 1987), which could be captured and measured by the SySI, $SyVI_w$ and $SyVI_d$ at the finest (local) resolution. Generally, the Keys to Soil Taxonomy (SOIL SURVEY STAFF, 2014) uses the same differing criteria to define families of soils.

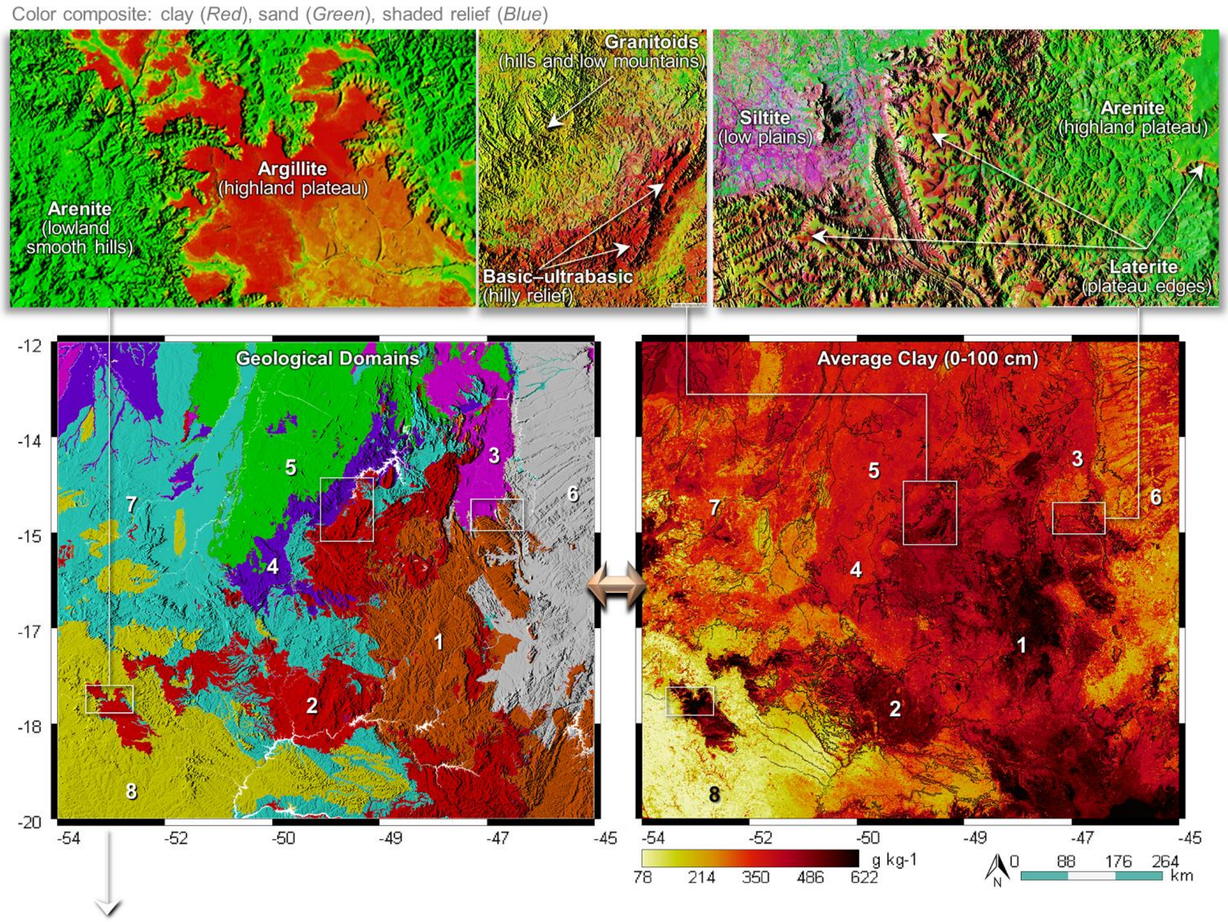
Individual relationships between soil variables and environmental covariates can also be interpreted and understood in terms of pedological knowledge. For instance, higher SWIR reflectance may be associated with high amounts of sand in soil and hence lower CEC; higher precipitation and cooler temperatures frequently increase the OM content, due to the speed of accumulation is higher than the speed of decomposition. For a large number of soil attributes, however, relationships are not clearly linear and often many soil covariates are equally important (HENGL et al., 2017).

4.4. Reliability and Interpretation of Soil Maps

The spatial patterns of soil on our predicted maps were consistent with pedological expert knowledge of the region and with their parent materials (Figure 8). Soil attributes largely varied across the area (Figure 7 and A4). This can be due to the tropical climate that exposed the parent materials of the studied soils (with different resistances) to intense weathering (VIEIRA et al., 2015).

We identified clayey and nutrient-poor soils throughout the southeast, covering 13% of the studied area (Figure 8). The region developed upon metasedimentary rocks (mostly argillite, siltite, arenite) which formed smooth hills, and over ferruginous laterite crusts supporting residual lowered plateaus in continuous dissection process (MORAES, 2014). The soils from these rocks (geological domain 1) had the highest clay contents with lowest chemical fertility and, in some cases, can be very acidic and contain ferruginous concretions (typically reddish color) that hinder the farming. Nevertheless, it is possible to observe several cropland areas distributed on soils of this domain (PARENTE et al., 2019), probably after undergoing soil chemical correction.

Clayey and medium textured soils with the best chemical conditions covered 38% of the area, widely distributed along the central portion over domains 2 to 5 (Figure 8). These domains were represented by basic-ultrabasic volcanic rocks such as basalt, diabase and gabbro, and sedimentary rocks such as argillite, siltite and calcarenite (CPRM, 2004). According to the geodiversity of the region (MORAES, 2014), the areas also were constituted by an association of metamorphosed volcanic and sedimentary (metavolcanosedimentary) rocks frequently containing amphibolite, serpentine, dunites and peridotites, metacarbonates, phyllite and paragneiss. All these lithologies reworked nutrient-richer surface materials, which released nutrients into the soil and provided better fertility conditions (see domains 2 and 3 in Figure 8). Furthermore, granitoids occurred sparsely mixed with calcareous and schist (see domain 4 in Figure 8), providing more dissected relief than the neighboring lands, such as hills and low mountains that hinder the agricultural mechanization (CPRM, 2004). In those areas, higher elevations with denser vegetation had larger soil OM contents (Figures A1 and A4), mainly due to cooler and wetter climate regimes and lesser human disturbance, which promoted accumulation processes (VIEIRA et al., 2015). In the floodplain areas over domain 5 (Figure 8), fertility conditions may be linked to the good fertility of the areas that surround it, from where it receives a high volume of water, sediments and wastes (MORAES, 2014).



Geological Domains \ Soil attributes	Area	Sand	Silt	Clay	OM	pH H2O	pH KCl	CEC	V	m
	%	g kg^{-1}				log		mmol$_c$ kg^{-1}	%	
1 Metasedimentary and lateritic	13	300 ± 74	203 ± 34	525 ± 70	29 ± 6	5.3 ± 0.2	4.6 ± 0.2	70 ± 13	24 ± 9	36 ± 10
2 Volcanic and sedimentary	15	360 ± 77	161 ± 30	500 ± 50	31 ± 6	5.6 ± 0.3	4.8 ± 0.2	65 ± 11	40 ± 10	20 ± 9
3 Sedimentary and Metavolcanosedimentary	4	367 ± 82	275 ± 38	389 ± 38	33 ± 8	5.9 ± 0.4	4.8 ± 0.2	91 ± 22	54 ± 11	19 ± 9
4 Sedimentary and granitoids	7	456 ± 54	143 ± 30	402 ± 31	49 ± 7	5.7 ± 0.3	4.8 ± 0.2	64 ± 21	42 ± 12	27 ± 10
5 Metasedimentary	12	465 ± 55	152 ± 34	385 ± 30	38 ± 6	6.3 ± 0.4	4.9 ± 0.2	73 ± 18	53 ± 11	30 ± 10
6 Sedimentary	13	496 ± 109	261 ± 32	319 ± 60	21 ± 6	5.3 ± 0.2	4.5 ± 0.2	70 ± 13	24 ± 8	32 ± 9
7 Metavolcanosedimentary	23	504 ± 80	132 ± 27	357 ± 51	27 ± 6	5.4 ± 0.3	4.7 ± 0.2	56 ± 14	33 ± 10	33 ± 11
8 Sedimentary and acid-subacid volcanic	14	731 ± 93	59 ± 18	211 ± 68	15 ± 5	5.3 ± 0.3	4.5 ± 0.2	36 ± 10	28 ± 8	37 ± 10

Soil attributes with averaged values from 0 to 100 cm depth.
Predominant lithotypes: ^{1}argillite, siltite, arenite, laterite; ^{2}basalt, diabase, gabbro, amphibolite, serpentine, dunites and peridotites, ferromagnetic minerals, argillite, siltite; ^{3}argillite, siltite, arenite, calcarenite, paragneiss and orthogneiss; ^{4}arenite, granite, argilite, siltite, calcareus, schist; ^{5}arenite, argillite, phyllite, paragneiss; ^{6}arenite, conglomerate, siltite, calcareous; ^{7}metaconglomerate, quartzite, phyllite, orthogneiss; ^{8}arenite, andesite, rhyolite.

Figure 8. Geological domains and summary values of soil attributes averaged from the three depth intervals. The geological domains were obtained by clustering lithologies using the averaged soil data. The main lithotypes were identified within each domain according the geological map (CPRM, 2004).

Sandy soils were expressive in the region, comprising 32% of the studied area (Figure 8). The lowest occurrence was in the northeast with 13% of sandy soils developed from sedimentary rocks (domain 6). They widely occurred in the southwest and midwest, developed over metavolcanosedimentary rocks (23%) and sedimentary and acid-subacid volcanic rocks (14%), domains 7 and 8 respectively. Such geological domains were mainly formed by arenite, conglomerate, siltite, calcareous, metaconglomerate, quartzite, phyllite, orthogneiss, andesite and rhyolite, which naturally tend to generate flattened reliefs such as smooth hills and plateaus (MORAES, 2014). These lithologies generally develop sandy soils with low chemical fertility (Figure 8). However, its high permeability and smooth reliefs facilitate agricultural mechanization after soil acidity correction and fertilization.

4.5. Possible Applications of the Maps

It is important to note that there are currently no detailed soil attribute maps with complete coverage over Midwest Brazil and that their production costs money. Our soil attribute maps can be used for different purposes, at different spatial scales from farm, local to regional. They provide a first complete assessment of key soil attributes across the Midwest region, and can be used to, for example, as input data in biological-chemical-physical modelling and in assessments of dynamic environmental processes. Together with other information, the maps can be used to obtain basic information for the implementation of colonization projects, rural subdivisions, integrated studies of micro-basins, local planning for the use and conservation of soils in areas projected for the development of agricultural, livestock and forestry projects, as well as civil engineering. The maps can also guide future soil sampling for inventory at different scales.

5. CONCLUSIONS AND FINAL CONSIDERATIONS

We have demonstrated that key soil attributes from multiple depth increments can be mapped using Earth observation data and machine learning with good performances. These maps had a satisfactory performance for physical ($0.64 > R^2_{10cv} > 0.85$) and chemical ($0.18 > R^2_{10cv} > 0.64$) attributes at all depth intervals (0–20, 20–60 and 60–100 cm), being spatially consistent with the main lithologies from which they originated.

The methodological approach was able to capture the spatial distribution of nine soil variables. The predicted soil maps suggest that less than 38% of the studied soils had good natural fertility. Nevertheless, the dominant smooth reliefs of the region facilitate agricultural mechanization, which allow the soil pH correction and fertilization.

Although we had representative soil observations, chemical attributes were particularly more challenging to map because to their high dynamic, with their concentration changing in a short space of time due to many natural and human-induced factors.

Our results support the use of multi-resolution covariates for DSM, where topsoil and natural vegetation reflectance are important predictors of soil variability together with relief and climate data.

Since covariates widely used in digital soil mapping are globally available, such as elevation and climate data, this approach may be useful for other initiatives where obtaining the soil (SySI) and vegetation (SyVI) covariates is feasible, that is, locations around the world with bare soil and natural vegetation occurring with enough coverage within the considered satellite time series.

6. LIST OF ABBREVIATIONS

ALOS	Advanced Land Observing Satellite
DSM	Digital Soil Mapping
GEE	Google Earth Engine
OM	Organic Matter
CEC	Cation Exchange Capacity
V%	Base Saturation
m%	Aluminum Saturation
SySI	Synthetic Soil Image
$SyVI_w$	Synthetic Vegetation Image of wet season
$SyVI_d$	Synthetic Vegetation Image of dry season
PNV	Potential Natural Vegetation
IDW	Inverse Distance Weighted
DEM	Digital Elevation Model
NIR	Near infrared spectral band
$SWIR_1$	First shortwave infrared spectral band
$SWIR_2$	Second shortwave infrared spectral band
LST	Land surface temperature
RMSE	Root Mean Square Error
RPIQ	Ratio of the Performance to Inter-Quartile distance

7. APPENDIX A

Table A1 shows the specific band number of each Landsat sensor, positioned in equivalent spectral regions, which were matched into a common name (e.g. Blue, Green, Red, NIR, $SWIR_1$, $SWIR_2$ and LST) for an inter-sensor harmonization.

Figure A1 displays 12 of the 33 covariates used to support the spatial predictions of soil variables. These covariates were obtained using the Google Earth Engine (GEE) cloud-based platform (GORELICK et al., 2017), according to their possible representativeness of the soil forming factors (MCBRATNEY et al., 2003). The density of geological lineaments was obtained by counting the meters of structural lines obtained from a 1:1,000,000-scale map (CPRM, 2004) per 1 km^2.

Figures A2 and A3 exhibits the predicted vs observed scatterplots of 10-fold cross-validation derived from optimized models for sand, silt and clay and the chemical attributes. The 30 m resolution maps of predicted soil chemical attributes at three distinct depth intervals are shown in the Figure A4.

Table A1. Harmonized Landsat Surface Reflectance Data Set.

Band	L4 TM	L5 TM	L7 ETM+	L8 OLI/TIRS
Blue	1 (450–520 nm)	1 (450–520 nm)	1 (450–520 nm)	2 (452–512 nm)
Green	2 (520–600 nm)	2 (520–600 nm)	2 (520–600 nm)	3 (533–590 nm)
Red	3 (630–690 nm)	3 (630–690 nm)	3 (630–690 nm)	4 (636–673 nm)
NIR	4 (770–900 nm)	4 (770–900 nm)	4 (770–900 nm)	5 (851–879 nm)
$SWIR_1$	5 (1550–1750 nm)	5 (1550–1750 nm)	5 (1550–1750 nm)	6 (1566–1651 nm)
$SWIR_2$	7 (2080–2350 nm)	7 (2080–2350 nm)	7 (2080–2350 nm)	7 (2107–2294 nm)
LST	6 (10,400–12,500 nm)	6 (10,400–12,500 nm)	6 (10,400–12,500 nm)	10 (10,600–11,190 nm)
Period	1982–1993	1984–2012	1999–present	2013–present

L4 TM: Landsat 4 Thematic Mapper; L5 TM: Landsat 5 Thematic Mapper; L7 ETM+: Landsat 7 Enhanced Thematic Mapper Plus; L8 OLI/TIRS: Landsat 8 Operational Land Imager/ Thermal Infrared Sensor; NIR: Near infrared band; $SWIR_1$: First shortwave infrared band; $SWIR_2$: Second shortwave infrared band; LST: Land surface temperature; Native spectral ranges are in parenthesis.

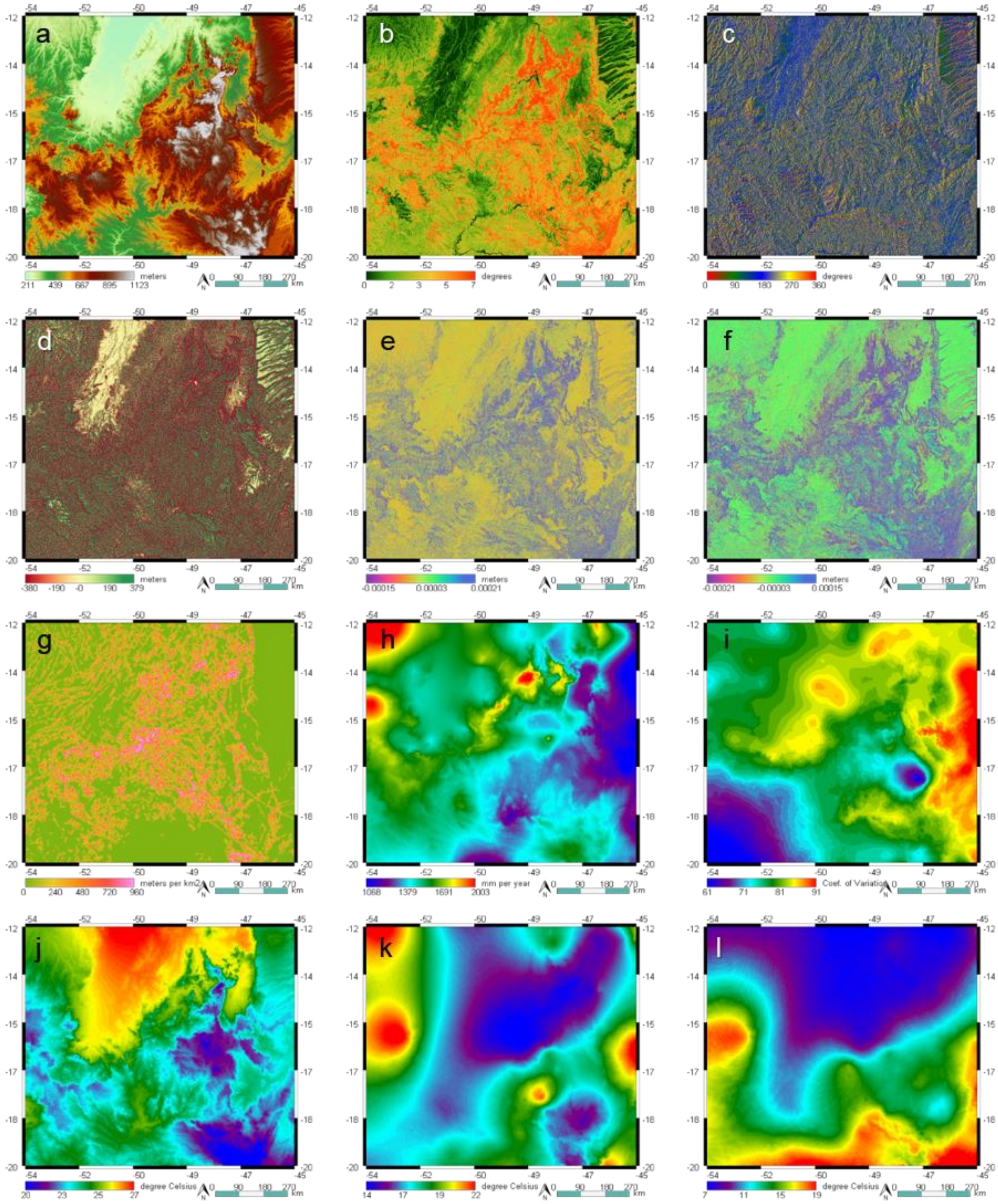


Figure A1. Environmental covariates used in the Random Forest modelling of soil attributes data. Terrain features derived from ALOS digital elevation model: (**a**) Elevation in meters, (**b**) Slope in degrees, (**c**) Aspect in degree, (**d**) Topographic Position Index, (**e**) Horizontal Curvature and (**f**) Vertical Curvature in meters. (**g**) Geological Lineaments Density representing meters of structural features per km², derived from legacy maps of the Geological Survey of Brazil (CPRM). Climate data obtained from WorldClim: (**h**) Annual Precipitation in mm, (**i**) Coefficient of variation of the Precipitation Seasonality, (**j**) Annual Mean Temperature, (**k**) Temperature Annual Range and (**l**) Temperature Seasonality in ºC.

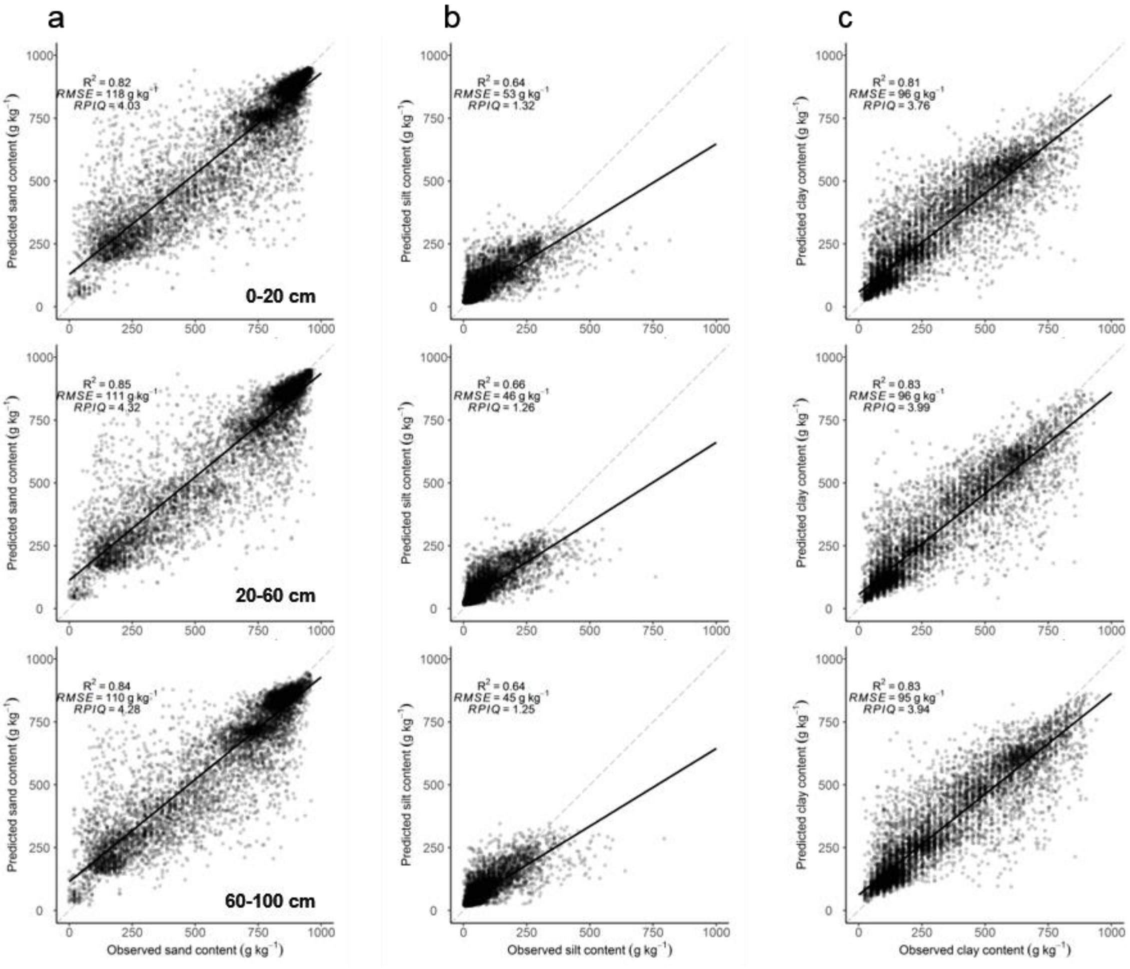


Figure A2. Predicted vs. observed (**a**) sand, (**b**) silt and (**c**) clay contents by depth intervals of 10-fold cross-validation derived from optimized random forest regression.

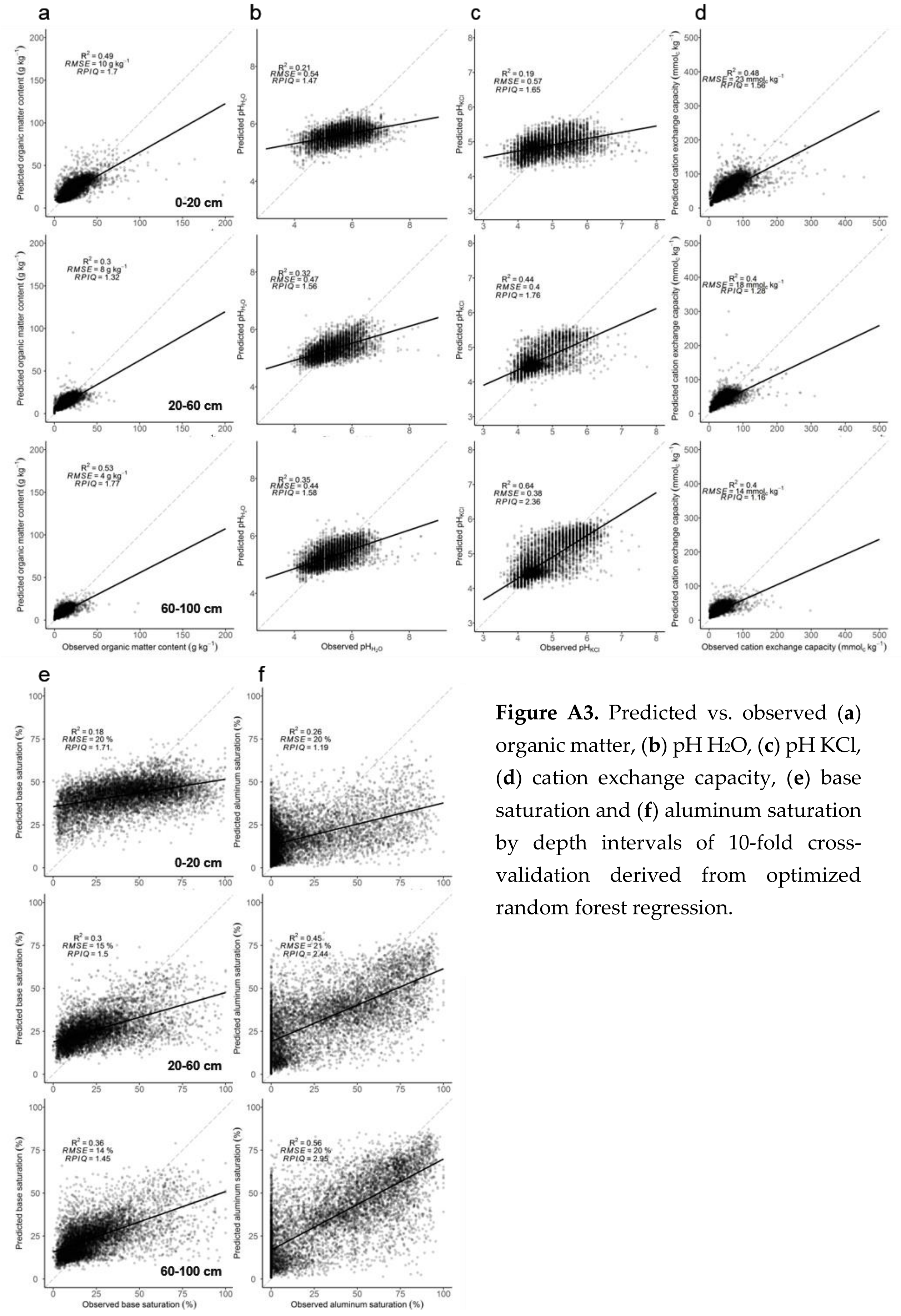
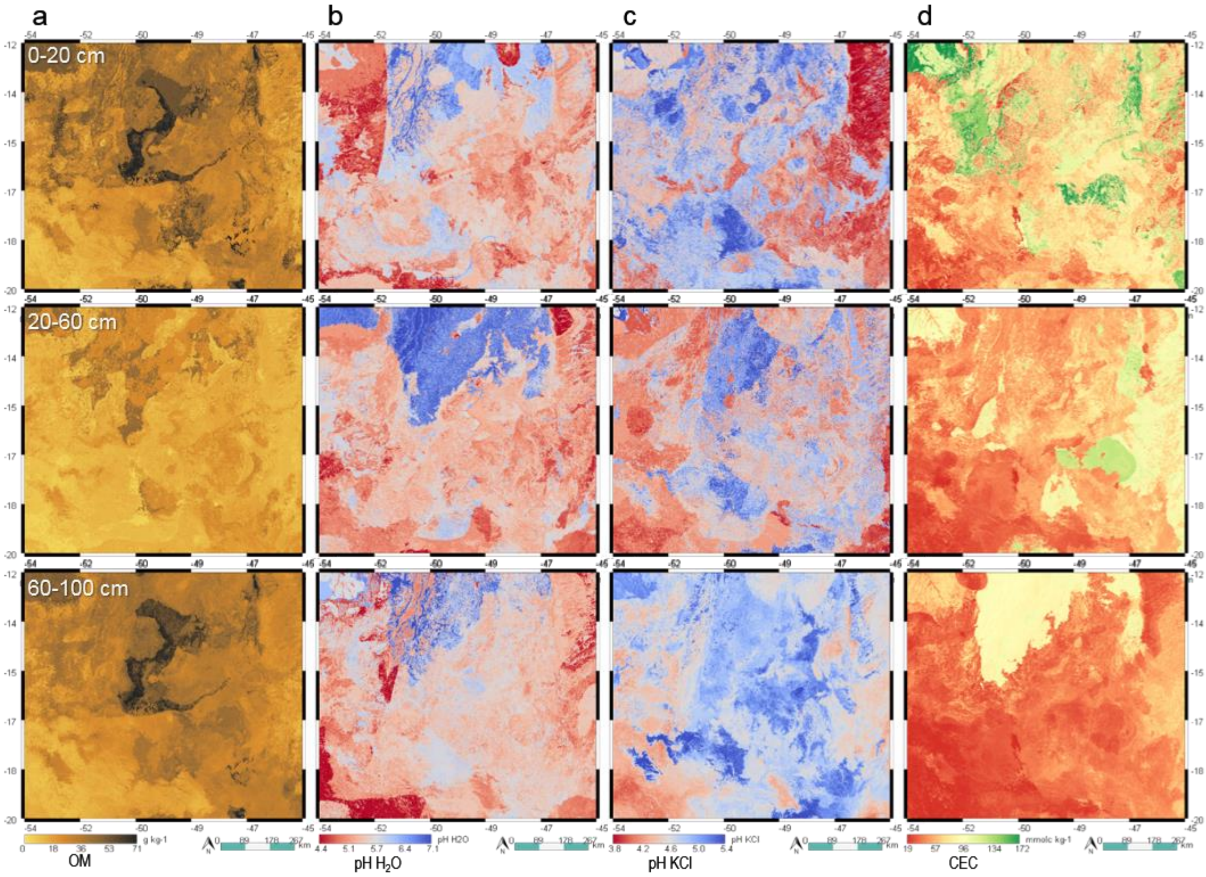


Figure A3. Predicted vs. observed (**a**) organic matter, (**b**) pH H_2O, (**c**) pH KCl, (**d**) cation exchange capacity, (**e**) base saturation and (**f**) aluminum saturation by depth intervals of 10-fold cross-validation derived from optimized random forest regression.



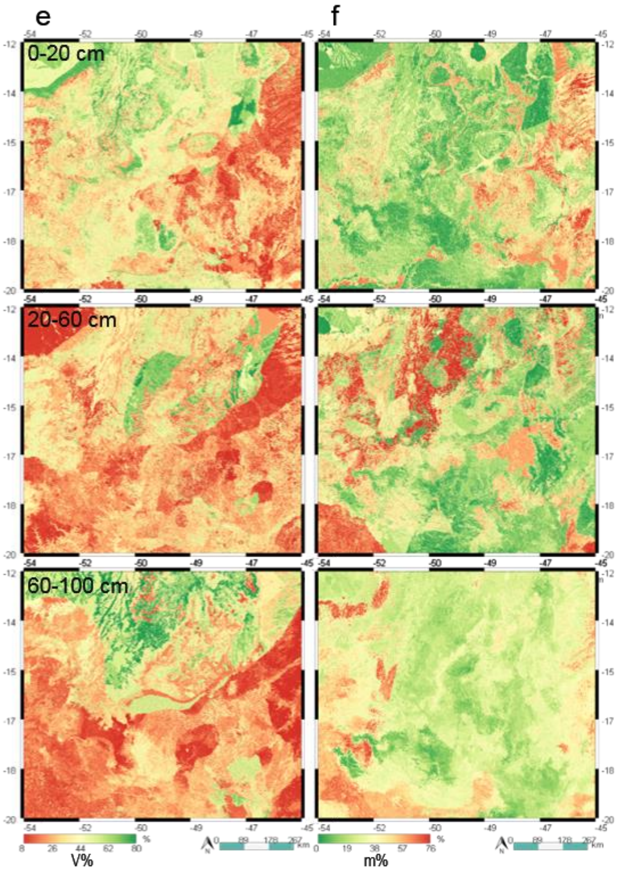


Figure A4. Maps of (**a**) organic matter, (**b**) pH H2O, (**c**) pH KCl, (**d**) cation exchange capacity, (**e**) base saturation and (**f**) aluminum saturation predicted at three depth intervals (0–20 cm, 20–60 cm and 60–100 cm). The visualization of the images was adjusted by stretching the range of pixel values between 2% and 98%.

8. REFERENCES

AKINYEMI, F.; ADEJUWON, J. A GIS-Based Procedure for Downscaling Climate Data for West Africa. **Transactions in GIS**, v. 12, n. 5, p. 613–631, 1 Oct. 2008.

AL-GAADI, K. A.; HASSABALLA, A. A.; TOLA, E.; KAYAD, A. G.; MADUGUNDU, R.; ASSIRI, F.; ALBLEWI, B. Characterization of the spatial variability of surface topography and moisture content and its influence on potato crop yield. **International Journal of Remote Sensing**, v. 39, n. 23, p. 8572–8590, 2 Dec. 2018.

AMIRIAN-CHAKAN, A.; MINASNY, B.; TAGHIZADEH-MEHRJARDI, R.; AKBARIFAZLI, R.; DARVISHPASAND, Z.; KHORDEHBIN, S. Some practical aspects of predicting texture data in digital soil mapping. **Soil and Tillage Research**, v. 194, p. 104289, 2019.

BAILEY, R. G. Suggested hierarchy of criteria for multi-scale ecosystem mapping. **Landscape and Urban Planning**, v. 14, n. C, p. 313–319, 1987.

BALLABIO, C.; LUGATO, E.; FERNÁNDEZ-UGALDE, O.; ORGIAZZI, A.; JONES, A.; BORRELLI, P.; MONTANARELLA, L.; PANAGOS, P. Mapping LUCAS topsoil chemical properties at European scale using Gaussian process regression. **Geoderma**, v. 355, p. 113912, 2019.

BARET, F.; JACQUEMOUD, S.; HANOCQ, J. F. About the soil line concept in remote sensing. **Advances in Space Research**, v. 13, n. 5, p. 281–284, May 1993.

BELLON-MAUREL, V.; FERNANDEZ-AHUMADA, E.; PALAGOS, B.; ROGER, J.-M.; MCBRATNEY, A. Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy. **TrAC Trends in Analytical Chemistry**, v. 29, n. 9, p. 1073–1081, Oct. 2010.

BREIMAN, L. **Random forests**. v. 45, n. 1, p. 5–32, 2001.

BUI, E.; HENDERSON, B.; VIERGEVER, K. Using knowledge discovery with data mining from the Australian Soil Resource Information System database to inform soil carbon mapping in Australia. **Global Biogeochemical Cycles**, v. 23, n. 4, 1 Dec. 2009.

BUI, E. N.; HENDERSON, B. L.; VIERGEVER, K. Knowledge discovery from models of soil properties developed through data mining. **Ecological Modelling**, v. 191, n. 3, p. 431–446, 2006.

BÜNEMANN, E. K.; BONGIORNO, G.; BAI, Z.; CREAMER, R. E.; DEYN, G. DE; GOEDE, R. DE; FLESKENS, L.; GEISSEN, V.; KUYPER, T. W.; MÄDER, P.; PULLEMAN, M.; SUKKEL, W.; GROENIGEN, J. W. VAN; BRUSSAARD, L. Soil quality – A critical review. **Soil Biology and Biochemistry**, v. 120, p. 105–125, 2018.

CANADELL, J.; JACKSON, R. B.; EHLERINGER, J. B.; MOONEY, H. A.; SALA, O. E.; SCHULZE, E.-D. Maximum rooting depth of vegetation types at the global scale. **Oecologia**, v. 108, n. 4, p. 583–595, 1996.

CPRM - COMPANHIA DE PESQUISA DE RECURSOS MINERAIS. **Carta Geológica do Brasil ao Milionésimo: sistema de informações geográficas-SIG [Geological Map of Brazil 1:1.000.000 scale: geographic information system-GIS]**. Brasília: CPRM, 2004.

DAS, S. Comparison among influencing factor, frequency ratio, and analytical hierarchy process techniques for groundwater potential zonation in Vaitarna basin, Maharashtra, India. **Groundwater for Sustainable Development**, v. 8, p. 617–629, 2019.

DEMATTÊ, J. A. M.; DOTTO, A. C.; PAIVA, A. F. S.; SATO, M. V; DALMOLIN, R. S. D.; ARAÚJO, M. DO S. B.; SILVA, E. B.; NANNI, M. R.; CATEN, A. TEN; NORONHA, N. C.; LACERDA, M. P. C.; ARAÚJO FILHO, J. C.; RIZZO, R.; BELLINASO, H.; FRANCELINO, M. R.; SCHAEFER, C. E. G. R.; VICENTE, L. E.; SANTOS, U. J.; SÁ BARRETTO SAMPAIO, E. V DE; MENEZES, R. S. C.; 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.; MENEZES, M. D.; CURI, N.;

COUTO, E. G.; ANJOS, L. H. C.; CEDDIA, M. B.; PINHEIRO, É. F. M.; GRUNWALD, S.; VASQUES, G. M.; MARQUES JÚNIOR, J.; SILVA, A. J.; BARRETO, M. C. DE V.; NÓBREGA, G. N.; SILVA, M. Z.; SOUZA, S. F.; VALLADARES, G. S.; VIANA, J. H. M.; SILVA TERRA, F. DA; HORÁK-TERRA, I.; FIORIO, P. R.; SILVA, R. C.; FRADE JÚNIOR, E. F.; LIMA, R. H. C.; ALBA, J. M. F.; 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.; SOUSA MENDES, W.; SAFANELLI, J. L.; GUIMARÃES, C. C. B.; POPPIEL, R. R.; SOUZA, A. B.; QUESADA, C. A.; COUTO, H. T. Z. The Brazilian Soil Spectral Library (BSSL): A general view, application and challenges. **Geoderma**, v. 354, p. 113793, 2019.

DEMATTÊ, J. A. M.; FONGARO, C. T.; RIZZO, R.; SAFANELLI, J. L. Geospatial Soil Sensing System (GEOS3): A powerful data mining procedure to retrieve soil spectral reflectance from satellite images. **Remote Sensing of Environment**, v. 212, p. 161–175, Jun. 2018.

DIEK, S.; FORNALLAZ, F.; SCHAEPMAN, M.; JONG, R. DE. Barest Pixel Composite for Agricultural Areas Using Landsat Time Series. **Remote Sensing**, v. 9, n. 12, p. 1245, 2017.

DIEK, S.; SCHAEPMAN, M. E.; JONG, R. DE. Creating multi-temporal composites of airborne imaging spectroscopy data in support of digital soil mapping. **Remote Sensing**, 2016.

EMBRAPA - BRAZILIAN AGRICULTURAL RESEARCH CORPORATION - NATIONAL SOILS RESEARCH CENTER. **Manual of soil analysis methods**. 3. ed. Brasilia, DF: Embrapa Solos, 2017.

FAO. **Soil Organic Carbon Mapping Cookbook**. 2. ed. Rome, Italy: FAO, 2018.

FLORINSKY, I. V. **Digital Terrain Analysis in Soil Science and Geology**. [s.l.] Academic press, 2016.

FONGARO, C.; DEMATTÊ, J.; RIZZO, R.; LUCAS SAFANELLI, J.; MENDES, W.; DOTTO, A.; VICENTE, L.; FRANCESCHINI, M.; USTIN, S. Improvement of Clay and Sand Quantification Based on a Novel Approach with a Focus on Multispectral Satellite Images. **Remote Sensing**, v. 10, n. 10, p. 1555, 27 Sep. 2018.

GALLO, B.; DEMATTÊ, J.; RIZZO, R.; SAFANELLI, J.; MENDES, W.; LEPSCH, I.; SATO, M.; ROMERO, D.; LACERDA, M. Multi-Temporal Satellite Images on Topsoil Attribute Quantification and the Relationship with Soil Classes and Geology. **Remote Sensing**, v. 10, n. 10, p. 1571, 1 Oct. 2018.

GOEBES, P.; SCHMIDT, K.; SEITZ, S.; BOTH, S.; BRUELHEIDE, H.; ERFMEIER, A.; SCHOLTEN, T.; KÜHN, P. The strength of soil-plant interactions under forest is related to a Critical Soil Depth. **Scientific Reports**, v. 9, n. 1, p. 8635, 2019.

GOMES, L. C.; FARIA, R. M.; SOUZA, E. DE; VELOSO, G. V.; SCHAEFER, C. E. G. R.; FILHO, E. I. F. Modelling and mapping soil organic carbon stocks in Brazil. **Geoderma**, v. 340, p. 337–350, 2019.

GORELICK, N.; HANCHER, M.; DIXON, M.; ILYUSHCHENKO, S.; THAU, D.; MOORE, R. Google Earth Engine: Planetary-scale geospatial analysis for everyone. **Remote Sensing of Environment**, v. 202, p. 18–27, 2017.

GU, Z. **circlize: Circular Visualization**. URL: https://cran.r-project.org/web/packages/circlize/index.html

HENGL, T.; HEUVELINK, G. B. M.; KEMPEN, B.; LEENAARS, J. G. B.; WALSH, M. G.; SHEPHERD, K. D.; SILA, A.; MACMILLAN, R. A.; MENDES DE JESUS, J.; TAMENE, L.; TONDOH, J. E. Mapping Soil Properties of Africa at 250 m Resolution: Random Forests Significantly Improve Current Predictions. **PLOS ONE**, v. 10, n. 6, p. e0125814, 25 Jun. 2015.

HENGL, T.; JESUS, J. M. DE; MACMILLAN, R. A.; BATJES, N. H.; HEUVELINK, G. B. M.; RIBEIRO, E.; SAMUEL-ROSA, A.; KEMPEN, B.; LEENAARS, J. G. B.; WALSH, M. G.; GONZALEZ, M. R. SoilGrids1km — Global Soil Information Based on Automated Mapping.

PLoS ONE, v. 9, n. 8, p. e105992, 29 Aug. 2014.

HENGL, T.; MENDES DE JESUS, J.; HEUVELINK, G. B. M.; RUIPEREZ GONZALEZ, M.; KILIBARDA, M.; BLAGOTIĆ, A.; SHANGGUAN, W.; WRIGHT, M. N.; GENG, X.; BAUER-MARSCHALLINGER, B.; GUEVARA, M. A.; VARGAS, R.; MACMILLAN, R. A.; BATJES, N. H.; LEENAARS, J. G. B.; RIBEIRO, E.; WHEELER, I.; MANTEL, S.; KEMPEN, B. SoilGrids250m: Global gridded soil information based on machine learning. **PLoS ONE**, v. 12, n. 2, p. e0169748, 16 Feb. 2017.

HENGL, T.; NUSSBAUM, M.; WRIGHT, M. N.; HEUVELINK, G. B. M.; GRÄLER, B. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. **PeerJ**, v. 6, p. e5518, 2018.

HENGL, T.; WALSH, M. G.; SANDERMAN, J.; WHEELER, I.; HARRISON, S. P.; PRENTICE, I. C. Global mapping of potential natural vegetation: an assessment of machine learning algorithms for estimating land potential. **PeerJ**, v. 6, p. e5457, 2018.

HIJMANS, R. J.; CAMERON, S. E.; PARRA, J. L.; JONES, P. G.; JARVIS, A. Very high resolution interpolated climate surfaces for global land areas. **International Journal of Climatology**, v. 25, n. 15, p. 1965–1978, 1 Dec. 2005.

IBGE - INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Produção Agrícola Municipal [Municipal Agricultural Production]**. URL: https://sidra.ibge.gov.br/pesquisa/pam/tabelas. Acesso em: 29 sep. 2019.

IBGE. **Pedologia [Pedological maps of Brazil]**. URL: https://www.ibge.gov.br/geociencias/informacoes-ambientais/pedologia/10871-pedologia.html?=&t=downloads. Acesso em: 30 sep. 2019.

IUSS WORKING GROUP WRB. **World reference base for soil resources 2014: International soil classification system for naming soils and creating legends for soil maps**. Rome: Food and Agriculture Organization, 2015.

KESKIN, H.; GRUNWALD, S.; HARRIS, W. G. Digital mapping of soil carbon fractions with machine learning. **Geoderma**, v. 339, p. 40–58, 2019.

KUHN, M. **caret: Classification and Regression Training**. URL: https://cran.r-project.org/web/packages/caret/index.html

LIANG, Z.; CHEN, S.; YANG, Y.; ZHOU, Y.; SHI, Z. High-resolution three-dimensional mapping of soil organic carbon in China: Effects of SoilGrids products on national modeling. **Science of The Total Environment**, v. 685, p. 480–489, 2019.

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. Satellite data integration for soil clay content modelling at a national scale. **International Journal of Applied Earth Observation and Geoinformation**, v. 82, p. 101905, Oct. 2019.

MA, Y.; MINASNY, B.; MALONE, B. P.; MCBRATNEY, A. B. Pedology and digital soil mapping (DSM). **European Journal of Soil Science**, v. 70, n. 2, p. 216–235, 1 Mar. 2019.

MA, Y.; MINASNY, B.; WU, C. Mapping key soil properties to support agricultural production in Eastern China. **Geoderma Regional**, v. 10, p. 144–153, 2017.

MAYNARD, J. J.; LEVI, M. R. Hyper-temporal remote sensing for digital soil mapping: Characterizing soil-vegetation response to climatic variability. **Geoderma**, v. 285, p. 94–109, 2017.

MCBRATNEY, A. B.; MENDONÇA SANTOS, M. L.; MINASNY, B. On digital soil mapping. **Geoderma**, v. 117, n. 1–2, p. 3–52, Nov. 2003.

MCBRATNEY, A. B.; WEBSTER, R. Choosing functions for semi-variograms of soil properties and fitting them to sampling estimates. **Journal of Soil Science**, v. 37, n. 4, p. 617–639, 1 Dec. 1986.

MENDES, W. DE S.; MEDEIROS NETO, L. G.; DEMATTÊ, J. A. M.; GALLO, B. C.; RIZZO, R.; SAFANELLI, J. L.; FONGARO, C. T. Is it possible to map subsurface soil attributes by satellite spectral transfer models? **Geoderma**, v. 343, p. 269–279, 2019.

MILLER, B. A.; KOSZINSKI, S.; WEHRHAN, M.; SOMMER, M. Impact of multi-scale predictor selection for modeling soil properties. **Geoderma**, v. 239–240, p. 97–106, 2015.

MORAES, J. M. **Geodiversidade do estado de Goiás e do Distrito Federal [Geodiversity of Goiás State and the Federal District, Brazil]**. Goiânia, GO: CPRM, 2014.

NAWAR, S.; CORSTANJE, R.; HALCRO, G.; MULLA, D.; MOUAZEN, A. M. Delineation of Soil Management Zones for Variable-Rate Fertilization: A Review. *In*: SPARKS, D. L. B. T.-A. IN A. (Ed.). **Advances in Agronomy**. Academic Press, 2017. v. 143p. 175–245.

NUSSBAUM, M.; SPIESS, K.; BALTENSWEILER, A.; GROB, U.; KELLER, A.; GREINER, L.; SCHAEPMAN, M. E.; PAPRITZ, A. Evaluation of digital soil mapping approaches with large sets of environmental covariates. **SOIL**, v. 4, n. 1, p. 1–22, 10 Jan. 2018.

PADARIAN, J.; MINASNY, B.; MCBRATNEY, A. B. Machine learning and soil sciences: A review aided by machine learning tools. **SOIL Discuss.**, v. 2019, p. 1–29, 3 Sep. 2019.

PARENTE, L.; MESQUITA, V.; MIZIARA, F.; BAUMANN, L.; FERREIRA, L. Assessing the pasturelands and livestock dynamics in Brazil, from 1985 to 2017: A novel approach based on high spatial resolution imagery and Google Earth Engine cloud computing. **Remote Sensing of Environment**, v. 232, p. 111301, 2019.

POPPIEL, R. R.; LACERDA, M. P. C.; SAFANELLI, J. L.; RIZZO, R.; OLIVEIRA JR, M. P.; NOVAIS, J. J.; DEMATTÊ, J. A. M. 250 m-gridded soil texture at multiple depths of Midwest Brazil. **Data Mendeley**, 2019a.

POPPIEL, R. R.; LACERDA, M. P. C.; DEMATTÊ, J. A. M.; OLIVEIRA, M. P.; GALLO, B. C.; SAFANELLI, J. L. Pedology and soil class mapping from proximal and remote sensed data. **Geoderma**, v. 348, p. 189–206, Aug. 2019b.

PROBST, P.; WRIGHT, M. N.; BOULESTEIX, A.-L. Hyperparameters and tuning strategies for random forest. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, v. 9, n. 3, p. e1301, 1 May 2019.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. Vienna, Austria: R Foundation for Statistical Computing, 2018. URL: https://www.r-project.org/

ROGGE, D.; BAUER, A.; ZEIDLER, J.; MUELLER, A.; ESCH, T.; HEIDEN, U. Building an exposed soil composite processor (SCMaP) for mapping spatial and temporal characteristics of soils with Landsat imagery (1984–2014). **Remote Sensing of Environment**, v. 205, p. 1–17, 2018.

SAMUEL-ROSA, A.; DALMOLIN, R. S. D.; MOURA-BUENO, J. M.; TEIXEIRA, W. G.; ALBA, J. M. F. Open legacy soil survey data in Brazil: geospatial data quality and how to improve it. **Scientia Agricola**, v. 77, n. 1, p. e20170430, 2020.

SAVIN, I. Y.; ZHOGOLEV, A. V; PRUDNIKOVA, E. Y. Modern Trends and Problems of Soil Mapping. **Eurasian Soil Science**, v. 52, n. 5, p. 471–480, 2019.

SERTESER, A.; KARGIOĞLU, M.; IÇAĞA, Y.; KONUK, M. Vegetation as an Indicator of Soil Properties and Water Quality in the Akarçay Stream (Turkey). **Environmental Management**, v. 42, n. 5, p. 764, 2008.

SOIL SURVEY STAFF. **Keys to Soil Taxonomy**. Washington: United States Department of Agriculture, 2014. v. 12 URL: http://www.nrcs.usda.gov/Internet/FSE_DOCUMENTS/nrcs142p2_051546.pdf

STENBERG, B.; VISCARRA ROSSEL, R. A.; MOUAZEN, A. M.; WETTERLIND, J. Visible and

Near Infrared Spectroscopy in Soil Science. **Advances in Agronomy**, v. 107, p. 163–215, 2010.

TADONO, T.; ISHIDA, H.; ODA, F.; NAITO, S.; MINAKAWA, K.; IWAMOTO, H. Precise global DEM generation by ALOS PRISM. **ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences**, v. II–4, p. 71–76, 31 Mar. 2014.

UNITED NATIONS - DEPARTMENT OF ECONOMIC AND SOCIAL AFFAIRS - POPULATION DIVISION. **World Population Prospects 2019: Highlights**. New York, USA: United Nations, 2019. URL: https://population.un.org/wpp/Publications/Files/WPP2019_Highlights.pdf

USGS - UNITED STATES GEOLOGICAL SURVEY. **Landsat 4-7 Surface Reflectance Code LEDAPS Product Guide**. Department of the Interior, USGS, 2019a. URL: https://www.usgs.gov/media/files/landsat-4-7-surface-reflectance-code-ledaps-product-guide

USGS. **Landsat 8 Surface Reflectance Code LaSRC Product Guide**. Department of the Interior, USGS, 2019b. URL: https://www.usgs.gov/media/files/land-surface-reflectance-code-lasrc-product-guide

VANDEGRIEND, A.; OWE, M.; VUGTS, H.; RAMOTHWA, G. **Botswana water and surface energy balance research program. Part 1: Integrated approach and field campaign results**. Greenbelt, MD, USA: NASA Goddard Space Flight Center, 1992. URL: https://ntrs.nasa.gov/search.jsp?R=19930011702

VAUDOUR, E.; GOMEZ, C.; FOUAD, Y.; LAGACHERIE, P. Sentinel-2 image capacities to predict common topsoil properties of temperate and Mediterranean agroecosystems. **Remote Sensing of Environment**, v. 223, p. 21–33, 2019.

VIEIRA, B. C.; SALGADO, A. A. R.; SANTOS, L. J. C. **Landscapes and Landforms of Brazil**. [s.l.] Springer, 2015.

VISCARRA ROSSEL, R. A.; CHEN, C.; GRUNDY, M. J.; SEARLE, R.; CLIFFORD, D.; CAMPBELL, P. H. The Australian three-dimensional soil grid: Australia's contribution to the GlobalSoilMap project. **Soil Research**, v. 53, n. 8, p. 845–864, 2015.

WRIGHT, M. N.; ZIEGLER, A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. **Journal of Statistical Software**, v. 1, n. 1, 31 Mar. 2017.

CHAPTER 2 — SOIL COLOR AND MINERALOGY MAPPING USING PROXIMAL AND REMOTE SENSING IN MIDWEST BRAZIL[2]

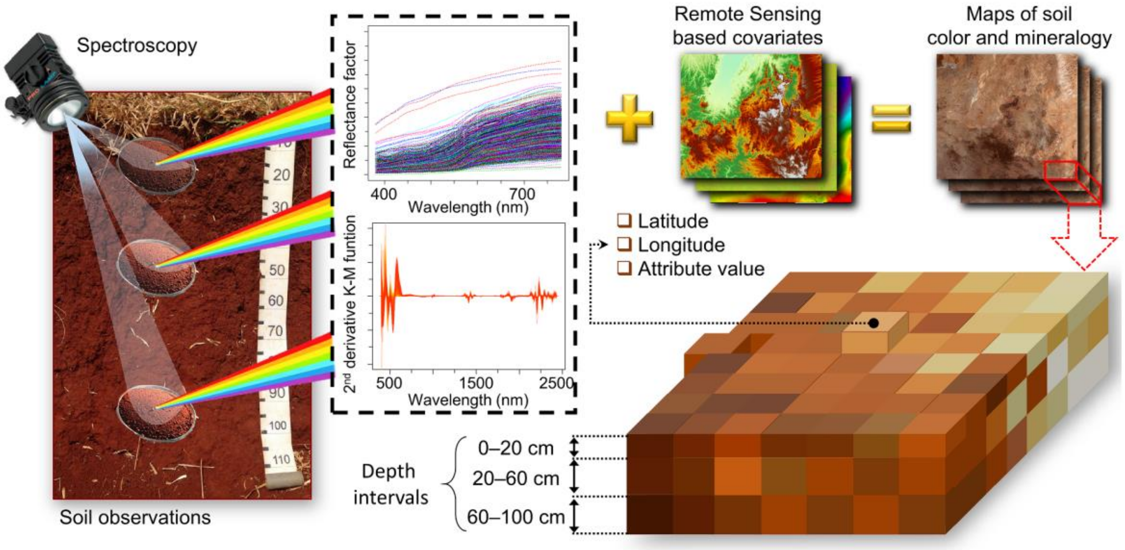
ABSTRACT

Soil color and mineralogy are used as diagnostic criteria to distinguish different soil types. In the literature, spectra (350–2500 nm) was successfully used to predict soil color and mineralogy, but these attributes currently were not mapped for most Brazilian soils. In this paper, we provided the first large-extent maps with 30 m resolution of soil color and mineralogy at three depth intervals for 851,000 km^2 of Midwest Brazil. We obtained soil spectra (350–2500 nm) from 1397 sites of the Brazilian Soil Spectral Library at 0–20 cm, 20–60 and 60–100 cm depths. Spectra was used to derive Munsell hue, value and chroma, and also second derivative spectra of the Kubelka–Munk function, where key spectral bands were identified and their amplitude measured for mineral quantification. Landsat composites of topsoil and vegetation reflectance, together with relief and climate data, were used as covariates to predict Munsell color and Fe–Al oxides, 1:1 and 2:1 clay minerals of topsoil and subsoil. We used Random Forest for soil modeling and 10-fold cross-validation. Soil spectra and remote sensing data accurately mapped color and mineralogy at topsoil and subsoil in Midwest Brazil. Hematite showed high prediction accuracy (R^2>0.71), followed by Munsell value and hue. Satellite topsoil reflectance at blue spectral region was the most relevant predictor (25% global importance) for soil color and mineralogy. Our maps were consistent with pedological expert knowledge, legacy soil observations and legacy soil class map of the study region.

Keywords: Reflectance spectroscopy; Munsell color system; derivative spectra; remote sensing; Google Earth Engine; data mining; Random Forest; digital soil mapping; soil attributes.

---

[2] Article submitted to **Remote Sensing**, in the Special Issue "*Bridging the Proximal and Remote Sensing Spectroscopy for Soil Properties Estimation and Monitoring*".



GRAPHICAL ABSTRACT

1. INTRODUCTION

The color is the most noticeable feature of soil that can be easily determined in field or laboratory (SCHWERTMANN et al., 1989). The main factors that influence soil color are the organic matter (AITKENHEAD et al., 2013) and mineralogy, especially iron oxides (SCHWERTMANN, 1993). Soil organic matter causes the darkness of soil by decreasing the Munsell value and chroma (SCHULZE et al., 1993). The most frequent pedogenic oxides in soil are hematite (usually associated to goethite) with hues between 10R and 5YR, and goethite (without hematite), that have hues between 7.5YR and 2.5Y (SCHWERTMANN, 1993). Goethite is common in diverse climates and parent materials, while hematite is abundant in well-drained tropical soils with strong pigmenting effect, and is absent in young soils from temperate humid climates (CURI et al., 1984; SCHWERTMANN et al., 1989).

In tropical Midwest Brazil, the surface materials are rich in Al, Si, and Fe-bearing minerals, where most of the soils developed from these rocks tend to be low in exchangeable bases and highly weathered (MORAES, 2014). The secondary minerals of these soils are frequently dominated by iron oxides, kaolinite and gibbsite, and 2:1 clay minerals that may resist in the silt and sand fractions (SCHAEFER et al., 2008). When iron oxides are completely removed (after reduction processes) from soil particles under anaerobic conditions, and if organic matter is negligible, the soil matrix achieves the background color of the minerals resulting in shades of gray (SCHWERTMANN, 1993). Consequently, color can be used to indicate the presence of minerals and the redoximorphic conditions of the soil (TORRENT et al., 1993). However, in tropical soils, mineralogy cannot be inferred from color alone because hematite has stronger pigmenting effect than the other soil minerals, including goethite (BARRÓN et al., 1986).

The iron oxides and soil color are effective pedoenvironmental indicators (SCHWERTMANN, 1993; SILVA et al., 2020). Various soil classification criteria used by non-scientists, like ethnic groups and farmers across the world, frequently are based on color (BARRERA-BASSOLS et al., 2006). Soil management zones, with different productivity potential, can be successfully delineated using bare soil color and topography (NAWAR et al., 2017). Some national (Embrapa, 2018) and international (IUSS WORKING GROUP WRB, 2015) soil classification systems used the color patterns and mineral contents of soil as diagnostic criteria. Patterns of color were also used to characterize soil parent materials (HURST, 1977). The standard method to describe color in soil science is the Munsell system (MUNSELL, 1907), which allows for direct comparison of soils worldwide based on three measurable components: hue, value, and chroma. Soil color and mineralogy play an important role in soil cartography, since they carry important information for pedological classification or soil attributes prediction (ZHANG et al., 2019).

Reflectance spectroscopy data (350–2500 nm) was successfully used in pedometry as predictor of the soil color and mineralogy (ESCADAFAL et al., 1988; FERNANDEZ et al., 1987; MARQUES et al., 2019; MATTIKALLI, 1997; RIZZO et al., 2016; SCHEINOST et al., 1998; SIMON et al., 2019), nevertheless, only a small set of researches mapped their spatial patterns. At the moment of this work, Viscarra Rossel et at. (2010) performed one of the few studies on soil color mapping, where the authors also mapped iron oxides of Australian soil using reflectance spectra (350–2500 nm) and geostatistics.

Studies on soil mineralogy mapping, such as Viscarra Rossel and Chen (2011), summarized the information content of spectra (350–2500 nm) by principal components to construct linear models using 31 predictors, and map the mineral of Australian topsoils robustly ($0.69 < R^2 < 0.85$). Likewise, Viscarra Rossel (2011) measured the relative abundances of kaolinite, illite and smectite at 0–20 and 60–80 cm soil depths, using continuum-removed reflectance (350–2500 nm) to derive statistical models and map the minerals with good cross-validation ($0.40 < R^2 < 0.61$). Malone et al. (2014) also used continuum-removed spectra (350–2500 nm) for the detection of iron oxides, kaolinite and smectite prior to mapping their spatial distribution (as class or categorical) in Australia, with overall accuracy ranging from 44 to 80%. Mulder et al. (2013) used reflectance spectroscopy (350–2500 nm) to derive soil minerals, and multinomial logistic regression, for mapping the likelihood of “absence” or “presence” of kaolinite, mica and smectite with high overall accuracy (>0.74). Other studies (DUCART et al., 2016; MADEIRA NETTO et al., 1997; ROBERTS et al., 2019) used enhanced mineral mapping techniques to produce a thematic mineral map of soil using the spectral response of Landsat imagery.

Nevertheless, the spatial patterns of soil color and mineralogy under current conditions remains not mapped for most Brazilian soils, both qualitatively and quantitatively. The main reason for that might be that in Brazilian repositories (SAMUEL-ROSA et al., 2020) there is a lack of mineralogical data, possible due to traditional methods, such as x-ray diffraction (XRD) are time consuming and resource intensive (MULDER et al., 2013). Besides that, XRD measurements are qualitative and not conducive to numerical analyses (VISCARRA ROSSEL et al., 2009). Soil color in national datasets (SAMUEL-ROSA et al., 2020) also lacks or do not contain spatial referencing or was visually approximated, that has been proved to be less consistent than modern colorimeter measurements (MARQUES et al., 2019).

Soils of tropical Midwest Brazil usually present high weathering degree and tend to have relatively homogenous profiles (CURI et al., 1984; SCHAEFER et al., 2008). Some studies has been shown that topsoil spectral patterns are related to the subsoil pattern variations and dynamic processes which occurs within the soil profile (MENDES et al., 2019; POPPIEL et al., 2018; POPPIEL et al., 2019a). In addition, bare topsoil reflectance composites produced from Landsat time series (DEMATTÊ et al., 2018; POPPIEL et al., 2019b; ROGGE et al., 2018) were considered as reliable proxies of topsoil spatial patterns, that can be integrated with other datasets by machine learning to better capture information from deeper layer of the Earth (ROBERTS et al., 2019).

Machine learning emerged in the 1990 as a tool for digital soil mapping (LAGACHERIE, 2008). Between the algorithms, Breiman (2001) proposed "random forests" (RF) that is currently the most popular for regression. RF combines several randomized decision trees and aggregates their predictions by their average. RF is often used by researchers for regressing the response Y to covariates X. Scornet et al. (2015) remarked that RF´s tree aggregation models are able to estimate linear and non-linear patterns and seeks to minimize the chance of overfitting during the splitting of trees, by selecting a reduced subset of covariates at each split.

Revealing the spatial patterns of the color and mineralogy in soils of Midwest Brazil may support our understanding of soil function to improve land use and management, as well as to operate as predictor for geological mapping, mineral exploration and digital soil mapping.

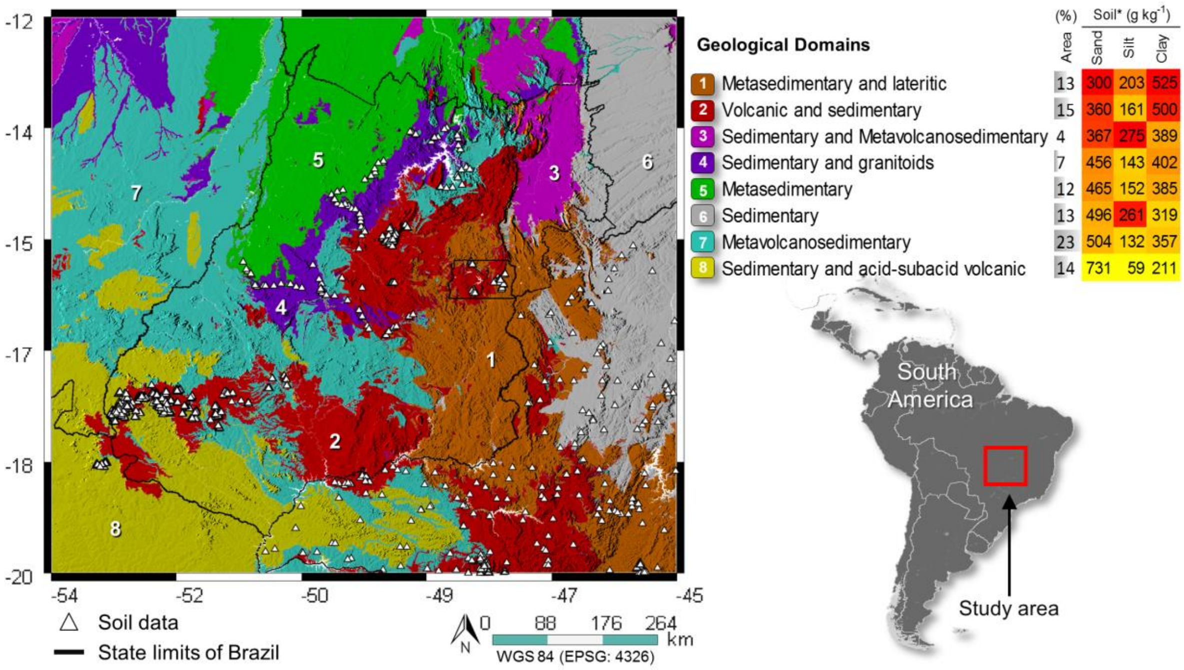
We expect that proximal soil sensing data have potential to provide accurate information on soil color and mineralogy, and that the use of predictors based on remote sensing data can provide accurate representations of the topsoil and subsoil spatial patterns over a large geographical extent. Then, proximal and remote sensing data can be coupled to accurately produce digital soil maps.

In this paper, we assessed the efficiency of proximal and remote sensing for mapping the soil color and mineralogy with 30 m resolution at three fixed depth intervals over 851,000 km^2 of Midwest Brazil. For that, we aimed: 1) to predict the soil color in Munsell notation from laboratorial spectra (350–2500 nm), 2) to measure and report the relative abundance of minerals in soil (hematite, goethite, kaolinite, gibbsite and 2:1 clay minerals) from their spectra (350–2500 nm), 3) to map their spatial distribution at 30 m resolution for the 0–20, 20–60 and 60–100 cm depth and verify the spatial patterns of the predicted maps with legacy soil information.

2. MATERIALS AND METHODS

2.1. *Study Area and Soil Data*

The study area is located in the Midwest of Brazil (Figure 1) comprising near 851,000 km^2. The landscape consists of extensive plateaus covered by Cerrado vegetation and gallery forest, within Cerrado biome (savanna). The humid tropical climate of the region exposed the highly diversified lithologies to intense weathering (VIEIRA et al., 2015), who reworked surface materials (Figure 1), and resulting in soils with attributes largely varying across the area (POPPIEL et al., 2019b). Thus, rocks from domains 1 (metasedimentary) and 2 (volcanic) developed clayey soils, typically redder than domains 7 and 8, which generated sandier soils with higher hue values. Such conditions allowed the genesis of Ferralsols, Lixisols, Plinthosols, Arenosols and Regosols across the region (IBGE, 2017).



-12
-14
-15
-17
-18
-20
-54
-52
-50
-49
-47
-45
1
2
3
4
5
6
7
8
△ Soil data
State limits of Brazil
N
0 88 176 264 km
WGS 84 (EPSG: 4326)
Geological Domains
(%) Area
Soil* (g kg^{-1}) Sand Silt Clay
1 Metasedimentary and lateritic 13 300 203 525
2 Volcanic and sedimentary 15 360 161 500
3 Sedimentary and Metavolcanosedimentary 4 367 275 389
4 Sedimentary and granitoids 7 456 143 402
5 Metasedimentary 12 465 152 385
6 Sedimentary 13 496 261 319
7 Metavolcanosedimentary 23 504 132 357
8 Sedimentary and acid-subacid volcanic 14 731 59 211
South America
Study area

Figure 1. Soil observations and limits of Brazil's states over shaded geological domains of the study area (POPPIEL, et al., 2019b). * Soil attributes averaged from 0 to 100 cm depths, where red represent clayey soils and yellow indicate sandy soils.

We obtained soil data from 1397 sites (Figure 1) of the Brazilian Soil Spectral Library (BSSL) (DEMATTÊ et al., 2019), at 0–20 cm, 20–60 and 60–100 cm depth intervals. Those layers represent the rooting depths where soil attributes can affect the growth of plants (CANADELL et al., 1996). The location of soil observations was recorded using GNSS (Global Navigation Satellite System) receivers with positioning accuracy greater than 10 m, that matched the spatial resolution of covariates. The data was acquired from soil samples dried at 45 °C, ground and sieved with 2 mm mesh and then homogeneously distributed in Petri dishes prior the measurement of the spectra in laboratory, between 350 and 2500 nm, using the Fieldspec 3 spectroradiometer (Analytical Spectral Devices, ASD, Boulder, CO). Splices positioned at 1000 and 1800 nm were corrected by linear interpolation of 10 bands using the prospectr package version 0.1.3 (STEVENS et al., 2013) in the R software (R CORE TEAM, 2018). The flow diagram of the proposed method is shown at the end of this section in Figure 2.

2.2. Reflectance to Soil Color

Soil scientists usually use the Munsell system to represent the color of soil, resembling the natural way that humans perceive the color (TORRENT et al., 1993). The Munsell notation is a cylindrical system based on three components, hue (the color: red, yellow, etc), value (lightness) and chroma (purity, similar to saturation), which can be estimated from spectra using mathematical formulas. We used spectral reflectance data to calculate the Munsell soil color at three depth intervals (0–20 cm, 20–60 and 60–100 cm), according to Marques et al. (2019) and Rizzo et al. (2016). The method used as input only the reflectance values between 380 and 780 nm (visible spectral range), and followed the steps: 1) spectra were converted to the XYZ color system for illuminant D_{65} (daylight) and 2nd standard observer (WYSZECKI et al., 1982); 2) XYZ tristimulus values were converted to the CIELAB color system (L*a*b*); 3) coordinates a* and b* were used to calculate hue angles and chroma, while value was estimated by L*; 4) hue angle was converted to Munsell notation using a color conversion table (CENTORE, 2014). All steps were implemented within the R software (R CORE TEAM, 2018), using the pracma (BORCHERS, 2019) and CircStats (AGOSTINELLI, 2018) packages.

For mapping purposes, Munsell hue was converted into a numerical scale of continuous values following the arrangement of the Munsell Soil Color Book, as suggested by Hurst (1977). In this system the hue charts of interest for our soil dataset were numbered as follow: 7.5 R was 7.5; 10 R was 10; 2.5 YR was 12.5; 5 YR was 15; 7.5 YR was 17.5; 10 YR was 20; 2.5 Y was 22.5, at 0.1 increments. The Munsell notation for selected hues (letter-number combination) used R (red), YR (yellow-red) and Y (yellow) preceded by a number from 1 to 10 to indicate position around the hue circle.

2.3. Reflectance to Soil Mineralogy

2.3.1. Spectral processing

Soils are mixtures of mineral and organic particles which partly absorb and partly scatter the incident light. When the dimensions of the mixed particles are comparable with the wavelength of the incident light, the absorption and scattering processes can be described by the Kubelka–Munk function [$KM = (1-R)^2/2R$; where R is reflectance] (BARRÓN et al., 1986). KM curves (likewise original spectra) show broad, strongly overlapping bands at different wavelengths. Therefore, to determine the positions of these bands, the resolution may be mathematically enhanced by calculating the derivatives of the spectra. The second derivative (SD) of the KM function is a promising method for spectral quantitative analysis (TORRENT et al., 2002), with sensibility for soil minerals detection slightly smaller than x-ray diffraction (SCHEINOST et al., 1998; SILVA et al., 2020). Thus, we transformed the reflectance data of soils into the KM and then calculated the SD using Savitzky–Golay method (fitting 2nd polynomial order to 40-smoothing points), within The Unscrambler software (CAMO SOFTWARE INC, 2007). This combination provided well-resolved spectral features, low background noise with little loss of spectral information for data collected at 1-nm intervals.

2.3.2. Key spectral bands for mineral quantification

The SD of the KM curve has spectral features originated from electronic transitions and non-fundamental vibrations of minerals (SCHEINOST et al., 1998), where minimum and maximum values match with the positions of the absorption bands in the original spectrum. The difference between derivative values at maxima and minima, determines the intensity of the "band amplitude", that is equivalent to the amount of mineral in the soil sample (KOSMAS et al., 1984). Therefore, to assess the soil mineralogy we: 1) selected the main minerals by checking its occurrence with previous works on soil mineralogy in the study area (GOMES et al., 2004; MACEDO et al., 1987; TERRA et al., 2018; ZINN et al., 2007); 2) defined the position of key spectral bands, at specific wavelengths (λ), for the main minerals of soil, summarized in Table 1; and, 3) calculated the band amplitudes for mineralogical quantification [$A = Max_{\lambda} - Min_{\lambda}$]. The intensity values of these band amplitudes were used as proxies of the soil minerals in the study area. Ternary diagrams were obtained by calculating the proportion of band amplitude between minerals for each plot using ggtern package (HAMILTON, 2018) in R.

Table 1. Position of the spectral bands in the SD KM curve, used to calculate the amplitude for the main minerals of soils.

Soil mineral	Minima band position (nm)	Maxima band position (nm)	Band Amplitude	Reference for band positions
Goethite	~415*	~455*	A_{Gt}	(KOSMAS et al., 1984; SCHEINOST et al., 1998)
Hematite	535*	580*	A_{Ht}	(SCHEINOST et al., 1998)
2:1 clay minerals[1]	1900–1925	1870–1895	$A_{2:1}$	(CLARK et al., 1990)
Kaolinite	2205	2225	A_{Kt}	(CLARK et al., 1990)
Gibbsite	2265	2295	A_{Gb}	(CLARK et al., 1990)

[1] Illite, Chlorite, Vermiculite, Montmorillonite. * Band positions relatively stable (lowest shift to neighboring wavelengths) for Al-substitution in both goethite and hematite.

2.4. Environmental Covariates

We used environmental predictors as proxies of the soil formation factors described by the *scorpan* model (MCBRATNEY et al., 2003) for the purpose of digital soil mapping (DSM). The DSM approach assumes that a soil attribute is a function of a spatial representation of soil forming factors: soil (s), climate (c), vegetation (o), relief (r), parent material (p), age of surface (a) and spatial position (n). Thus, we acquired a set of covariates (33 layers) from Poppiel et al. (2019b) to act as proxies of each factor of soil formation (Table 2). These covariates were prepared using big databases of remote sensing at multiple spatial resolution within Google Earth Engine (GEE) (GORELICK et al., 2017). Then, coarser-resolution predictors were downscaled into a target grid resolution of 30 m. For further details on how the covariates were prepared and quality assessed, see Poppiel et al. (2019b).

Table 2. Soil forming factor proxies from Poppiel et al. (2019b) used for mapping the soil color and main minerals.

Factor	**Covariate**	**Description**
Soil, Parent Material and Age	SySI	Synthetic Soil Image based Landsat 4, 5, 7 and 8 (7 bands), representing bare soil reflectance at 30 m resolution.
	Geological Lineaments	Meters of structural features per km^2 from CPRM data at 1:1,000,000 scale (CPRM, 2004).
Organisms	SyVI$_w$ and SyVI$_d$	Synthetic Vegetation Image of wet (Nov-Mar) and dry (May-Sep) seasons based Landsat 4 and 5 (7 bands), representing potential natural vegetation reflectance at 30 m resolution.
Climate	Annual Precipitation (mm) Precipitation Seasonality (CV) Annual Mean Temperature (°C) Temperature Annual Range(°C) Temperature Seasonality (°C)	Bioclimatic variables obtained from the WorldClim dataset at 1 km resolution (HIJMANS et al., 2005).
Relief	Elevation (m) Slope (degree) Aspect (degree) Topographic Position Index (m) Horizontal Curvature (m) Vertical Curvature (m)	Terrain attributes obtained from the 30 m ALOS digital elevation model (TADONO et al., 2014)

CV: coefficient of variation.

2.5. Soil modelling by Random Forest (RF)

In DSM studies (GOMES et al., 2019; HENGL et al., 2014, 2015, 2017; LEENAARS et al., 2019; LOISEAU et al., 2019; SILVA et al., 2019; WADOUX et al., 2019), Random Forests (BREIMAN, 2001) is increasingly being used to infer relationships between diverse soil attributes (at single and multiple depths), and several covariates (from multiple sources and resolutions) across landscapes. One of the reasons relies on that RF can handle both linear and nonlinear relationships of the data. Thus, we used RF regression for DSM of the soil color (Munsell hue, value and chroma) and the main soil minerals (Table 1) at 0–20, 20–60 and 60–100 cm depth intervals. For that, we used the full set of covariates (Table 2), on factors of soil formation–*scorpan* model (MCBRATNEY et al., 2003), and let the decision tree algorithm reveal the patterns. Therefore, a different model was adjusted to each soil attribute, at each one of our depths, counting 24 models. RF can fit models with large number of predictors (HENGL et al., 2018).

2.5.1. Model tuning

We filtered possible artifacts in the covariates (Table 2) by computing the median values within a 4 × 4 moving window. These covariates were sampled at each soil observation and the values were used as input data for calibrating RF regressions (BREIMAN, 2001) using the ranger package version 0.11.1 (WRIGHT et al., 2017) in the R software (R CORE TEAM, 2018). According to Probst et al. (2019), a proper tuning of hyperparameters ensures the RF's consistency. For that, we performed a grid search examining a range of values, where the number of covariates randomly selected at each node ($mTry$) was 6, 24, 33; and the tree depth by minimal number of samples "or leaves" for the terminal nodes ($minimum\ node\ size$) was 5, 20, 50. We fixed 500 trees to obtain stable estimates (PROBST et al., 2019).

2.5.2. Model performance

In order to assess the prediction models, we calculated performance metrics such as the root mean squared error (RMSE), coefficient of determination (R^2) and ratio of the performance to inter-quartile distance ($RPIQ = (Q3 - Q1)/\text{RMSE}$), where Q1 and Q3 are the 1st (25%) and 3rd (75%) quartiles. The RPIQ is based on prediction error and quartiles, which evaluates the spread of the dataset to the models accuracy making easier the comparison among soil attribute models and other studies. We derived these metrics for each one of the 24 models to assess the goodness of fit in the calibration step, and the robustness in the validation step. Validation was performed for each one of the 24 models by 10-fold cross-validation, using the caret package version 6.0-84 (KUHN, 2019). The k-cross-validation maximizes the quantity of points in the training dataset, where the points are divided into k groups or folds, where $k - 1$ groups are used for training and 1 group for validation, repeating the training k times, each with a different validation group (PADARIAN et al., 2019). We selected the optimized model by the minimum RMSE of the 10-fold cross-validation (FAO, 2018; PROBST et al., 2019). Generally, smaller values of RMSE and larger R^2 and RPIQ indicate higher model performance (BELLON-MAUREL et al., 2010).

2.5.3. Covariates importance

RF supply some abilities to interpret the model by providing measures for variable importance (GOMES et al., 2019; HENGL et al., 2014, 2015, 2017; LEENAARS et al., 2019; LOISEAU et al., 2019; SILVA et al., 2019; WADOUX et al., 2019), based on the increase in mean square error when a covariate is randomly permuted. Thus, we used the folds estimates to calculate the mean frequency of use for the covariates in the models and reported as a measure of the scaled permutation importance for each soil attribute prediction (BREIMAN, 2001), using the ranger package version 0.11.1 (WRIGHT et al., 2017) in R (R CORE TEAM, 2018). Interpreting this output is quite straightforward: the more importance, the more relevant the variable is, according to the model.

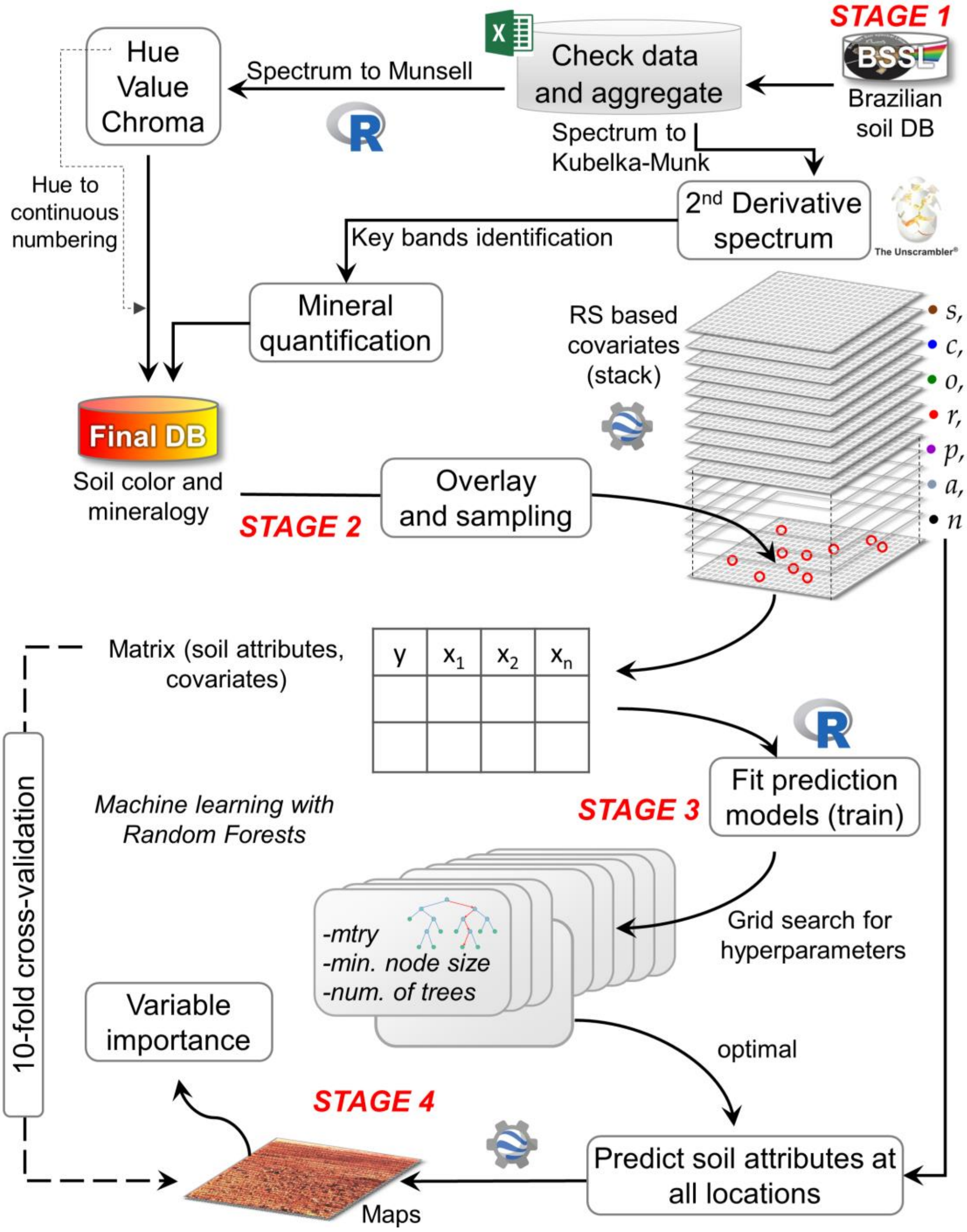


Figure 2. Flow diagram of the proposed methodology for soil color and mineralogy mapping using proximal and remote sensing data.

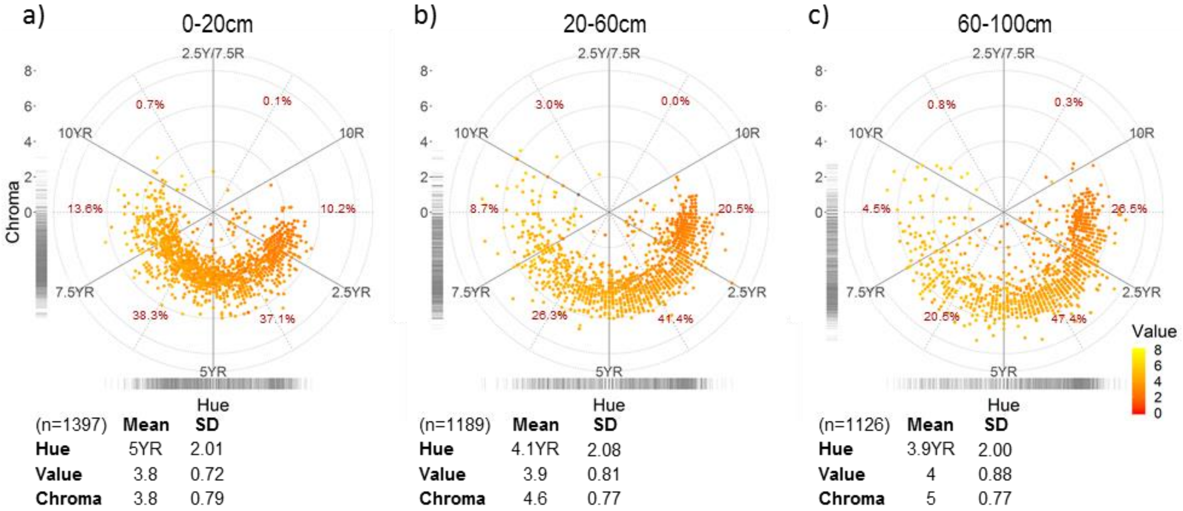
2.5.1. Soil Mapping

The optimized models (tuned hyperparameters in R) of soil attributes were implemented into cloud-based platform of GEE (GORELICK et al., 2017) to predict their spatial distribution in the study area using RF algorithm. In this study, the uncertainty was not examined as maps because this technique was not implemented at the current development stage of GEE (GORELICK et al., 2017). Therefore, to verify the correspondence of the spatial patterns of our predictions, we performed Pearson's correlation between our maps (at the three depth intervals) with legacy soil observations acquired from a national dataset (SAMUEL-ROSA et al., 2020), and also with weathering degree and hue, both inferred from a 1:1,000,000-scale legacy soil class map that covered the study area (IBGE, 2017).

3. RESULTS

3.1. *Soil attributes derived from spectra*

Spectra (350-2500 nm) contain information on important attributes of the soil: minerals, color, organic material, texture and water. The reflectance in the visible spectral interval revealed that in our dataset the soil color ranged from 8.9R (red) to 2.5Y (yellow), and reached more than 50% of samples up to 5YR (yellow-red) (Figure 3). Value and chroma ranged from 1.7 to 8 and from 0.7 to 8 with mean values of 3.9 and 4.5, respectively. Overall, the hue decreased and the value and chroma increased as the soil depth interval increased, that is, the soil color was redder, lighter and purer (or saturated) at deeper layers. The amplitude between key spectral bands in the SD KM curve (Table 1) indicated that the soils were dominated by hematite, goethite and kaolinite, with relative amounts between them of about 38%, 36% and 25%, respectively. These minerals were mixed in soils with smaller amounts of gibbsite and 2:1 clay minerals, where its proportions in relation to kaolinite were near 19%, 15% and 66%, respectively (Figure 4).

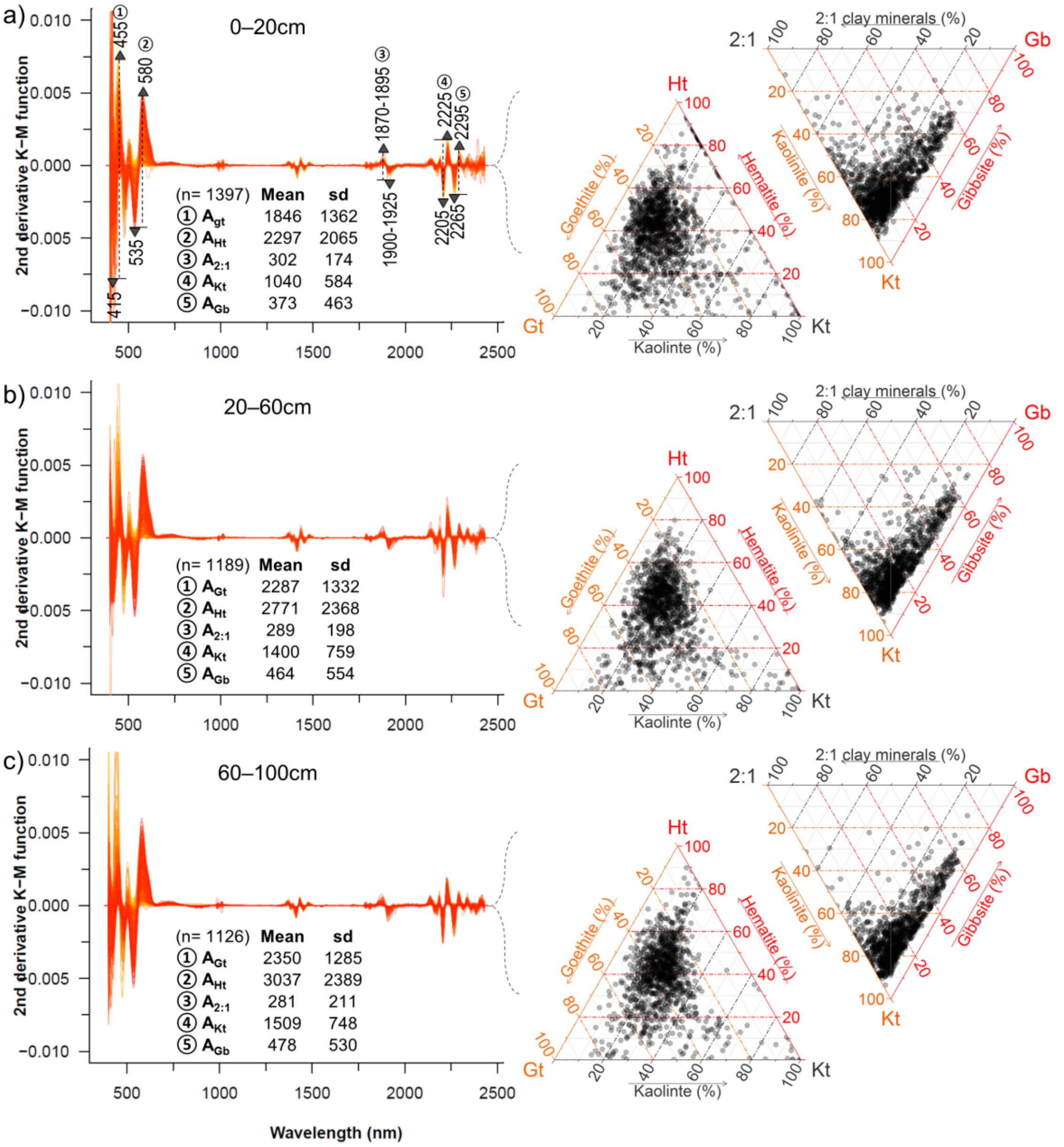


(n=1397)	Mean	SD
Hue	5YR	2.01
Value	3.8	0.72
Chroma	3.8	0.79

(n=1189)	Mean	SD
Hue	4.1YR	2.08
Value	3.9	0.81
Chroma	4.6	0.77

(n=1126)	Mean	SD
Hue	3.9YR	2.00
Value	4	0.88
Chroma	5	0.77

Figure 3. Polar plot of soil color in the Munsell system (hue, value, chroma) determined from spectra at (**a**) 0–20, (**b**) 20–60 and (**c**) 60–100 cm depth intervals. Hue values were displayed on circular grid beginning at 7.5YR, increasing values clockwise up to 2.5Y. Chroma values were presented in Y axis, increasing from the center outwards. Value was shown as a color scale, increasing from red to yellow. The number of soil samples (n) used to calculate soil color and their mean values and standard deviation (SD) were summarized at the bottom of each panel.



a)

(n= 1397)	Mean	sd
① A_{gt}	1846	1362
② A_{Ht}	2297	2065
③ $A_{2:1}$	302	174
④ A_{Kt}	1040	584
⑤ A_{Gb}	373	463

b)

(n= 1189)	Mean	sd
① A_{Gt}	2287	1332
② A_{Ht}	2771	2368
③ $A_{2:1}$	289	198
④ A_{Kt}	1400	759
⑤ A_{Gb}	464	554

c)

(n= 1126)	Mean	sd
① A_{Gt}	2350	1285
② A_{Ht}	3037	2389
③ $A_{2:1}$	281	211
④ A_{Kt}	1509	748
⑤ A_{Gb}	478	530

Figure 4. Second derivative of the KM spectra (left) and ternary diagrams of soil minerals (right) at three depth intervals: (**a**) 0–20, (**b**) 20–60 and (**c**) 60–100 cm. The amount of mineral was quantified by the measurements of the amplitude between values at minima and maxima specifics bands, graphically exemplified in the left panel **a**: A_{Gt} (goethite), A_{Ht} (hematite), $A_{2:1}$ (2:1 clay minerals), A_{Kt} (kaolinite), A_{Gb} (gibbsite). The ternary diagrams were constructed by assessing the proportion between band amplitudes of the minerals. The number of soil samples (n) used to derive soil minerals and their mean values and standard deviation (sd) were summarized (scale factor 1 x 10^{-6}) at the bottom of the spectral curves.

The significant correlations ($p < 0.01$) for goethite ($-0.3 < r < -0.66$) and hematite ($-0.79 < r < -0.88$) with hue and value, suggested that iron oxides decreased these color attributes, and promoted the redness and darkness of the soils at the three depth intervals (Figure 5). Iron oxides also were correlated with chroma (average r of 0.29) which caused the saturation of soil color.

All minerals were positively correlated with each other (Figures 5a–c), where gibbsite and 2:1 clay minerals showed the smallest values between them ($r < 0.21$). Likewise, the proportion of minerals slightly increased with depth (Figure 5d), since they are relatively dominant at finest fractions, and there is more clay in the subsurface of the studied soils. Gibbsite was relatively constant across depth (Figure 5d), while the 2:1 clay minerals were a little more abundant in the topsoil. The hue decreased with depth while chroma increased, both due to small amounts of iron oxides which pigmented the soil at deeper layers. Value increased with depth (Figure 5d), since it depends on reflectance, which increases with less amounts (masking effects) of organic matter on mineral particles.

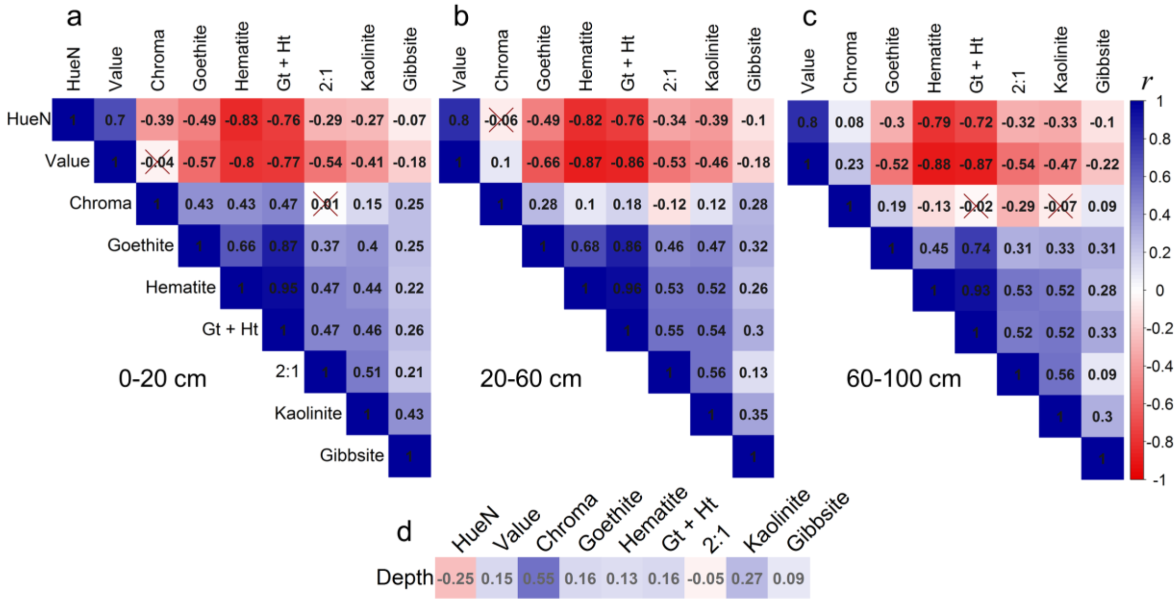


Figure 5. Correlogram based on Pearson's correlation (*r*) between soil color components and minerals derived from spectra at (**a**) 0–20 cm, (**b**) 20–60 cm and (**c**) 60–100 cm depth intervals; and (**d**) overall correlation with depth intervals analyzed. Blue and red colors symbolize positive and negative correlations, respectively. Insignificant correlation coefficient values (p-value > 0.01) were marked by crosses (X). HueN: Hue number; 2:1: 2:1 clay minerals. The sum of Gt+Ht (Goethite + Hematite) was added to the plot only for comparisons.

3.2. Performance of spatial models

The RF models proved to be robust for mapping soil color and mineralogy at three depth intervals in Midwest Brazil (Table 2), with high prediction accuracy for hematite ($R^2_{10cv} > 0.71$). The prediction of Munsell value and hue, gibbsite, kaolinite, 2:1 minerals and goethite were accurate ($0.43 < R^2_{10cv} < 0.65$). The models for goethite produced lower validation metrics than for hematite (Table 2), probably because their spectral bands used for relative quantification (Table 1), especially at 60–100 cm depth ($R^2_{10cv} = 0.24$). Munsell chroma at all depths had worse prediction accuracy ($0.24 < R^2_{10cv} < 0.38$). Albeit some models had low R^2 for validation, they all showed a god performance ($RPIQ_{10cv} > 1.3$) and scatterplots with values following linear trends (Figure A1).

Table 2. Hyperparameters and performance metrics for calibration (goodness of fit) and validation (robustness) of the models used for mapping soil attributes at surface and subsurface.

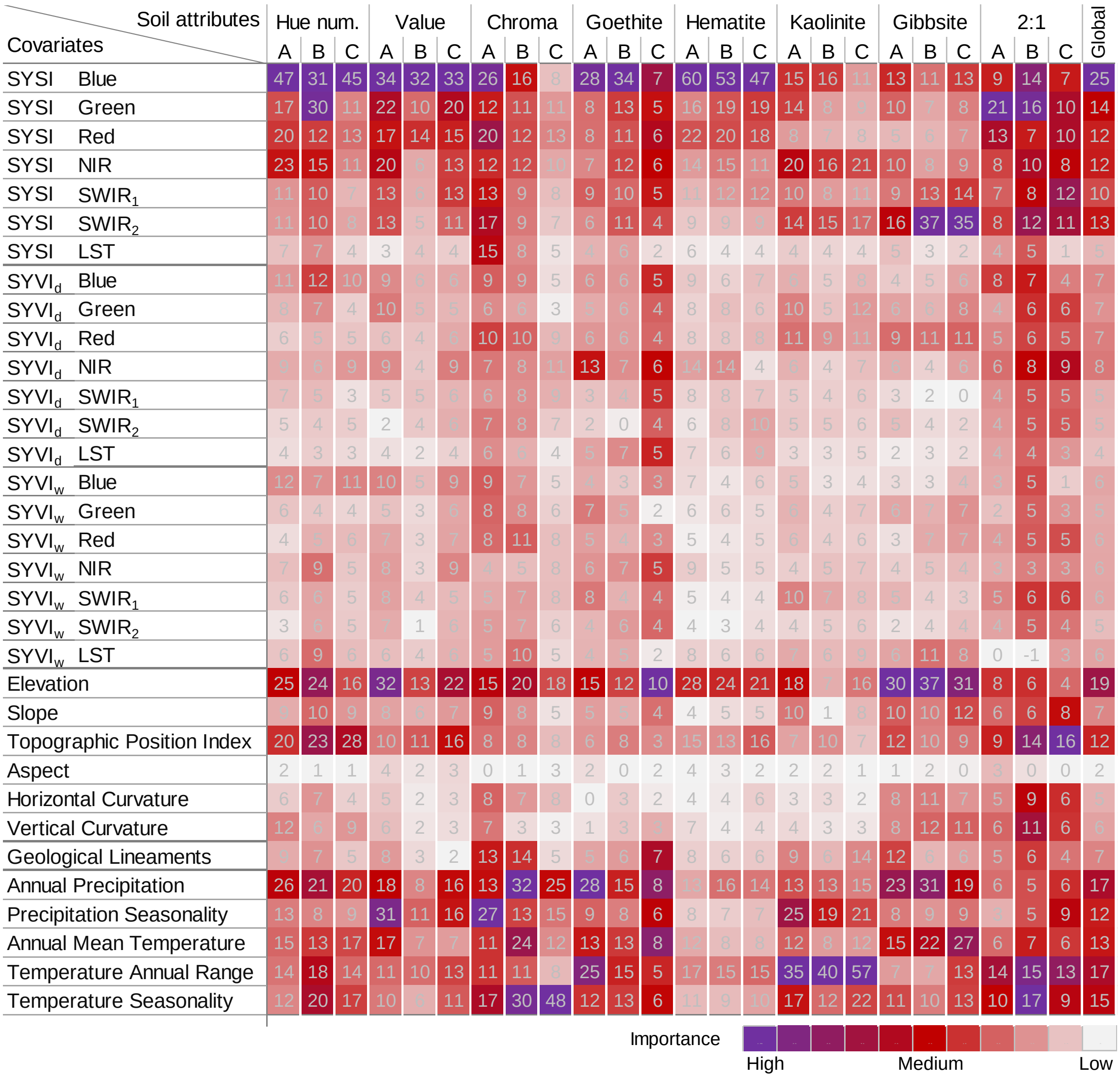
Soil attribute	**Depth**^{2}	**mTry**	**minNS**	**RMSE**$_{cal}$	**RPIQ**$_{cal}$	**R**$^2_{cal}$	**RMSE**$_{10cv}$	**RPIQ**$_{10cv}$	**R**$^2_{10cv}$
Hue number^1	0–20	24	5	0.53	5.89	0.93	1.30	2.35	0.58
	20–60	24	5	0.61	5.45	0.93	1.50	2.17	0.54
	60–100	33	5	0.56	4.80	0.92	1.40	1.91	0.50
Value	0–20	24	5	0.19	4.84	0.93	0.50	1.95	0.59
	20–60	24	5	0.24	5.85	0.92	0.60	2.44	0.55
	60–100	24	5	0.21	5.79	0.94	0.50	2.32	0.64
Chroma	0–20	33	5	0.27	3.67	0.89	0.70	1.45	0.31
	20–60	33	5	0.31	3.56	0.88	0.80	1.41	0.24
	60–100	33	5	0.29	3.41	0.90	0.70	1.36	0.38
Goethite	0–20	24	5	414*	4.07	0.91	1008*	1.67	0.45
	20–60	24	5	396*	4.06	0.91	990*	1.62	0.45
	60–100	6	5	494*	2.97	0.85	1122*	1.31	0.24
Hematite	0–20	24	5	436*	6.49	0.96	1102*	2.57	0.71
	20–60	24	5	496*	6.49	0.96	1254*	2.57	0.72
	60–100	24	5	504*	6.46	0.96	1264*	2.58	0.72
Kaolinite	0–20	33	5	171*	4.41	0.91	424*	1.78	0.47
	20–60	33	5	205*	4.86	0.93	508*	1.96	0.55
	60–100	33	5	190*	5.20	0.94	481*	2.05	0.59
Gibbsite	0–20	24	5	123*	3.23	0.93	309*	1.28	0.55
	20–60	33	5	132*	3.55	0.94	335*	1.40	0.64
	60–100	24	5	124*	4.09	0.95	312*	1.62	0.65
2:1 minerals	0–20	33	5	54*	3.99	0.90	132*	1.63	0.43
	20–60	24	5	56*	3.75	0.92	139*	1.52	0.51
	60–100	24	5	63*	3.13	0.91	151*	1.31	0.49

^{1} see Munsell hue in section 2.2; ^{2} Depth in cm; * Scale factor 1 x 10^{-6}; mTry: hyperparameter of Random Forest regression that controls the number of variables that can be randomly sampled in each split of the trees; minNS: minimum node size, a hyperparameter of Random Forest that controls the tree depth by setting the minimal number of samples for the terminal nodes. RMSE$_{cal}$: Root Mean Square Error of calibration; RMSE$_{10cv}$: Root Mean Square Error of 10-fold cross-validation; R$^2_{cal}$: Coefficient of determination of calibration; R$^2_{10cv}$: Coefficient of determination of 10-fold cross-validation.

3.3. Relevance of covariates

The importance of each covariate on predicting Munsell color and mineralogy of soil is shown in Figure 6. The main predictors (global importance > 10%) for most of the attributes and depths of soil in the study area were (decreasing sequence) SySI$_{Blue}$, elevation, annual precipitation, temperature annual range, temperature seasonality, SySI$_{Green}$, SYSI$_{Swir2}$, annual mean temperature, SYSI$_{NIR}$, precipitation seasonality, SySI$_{Red}$, topographic position index and SySI$_{Swir1}$. These covariates remained unchanged and usually in the same sequence at each depth, with bare topsoil reflectance at blue spectral region (SySI$_{Blue}$) as the most relevant predictor for our conditions. They are proxies of the soil forming factors *s*, *c*, *r*, *p* and *a* (Table 2), which all interact to influence spatial distribution of color and minerals of soil in Midwest Brazil.

The forming factor *o*, represented by the potential natural vegetation reflectance of dry and wet seasons, especially at blue, green, red and near-infrared spectral ranges, had medium to low importance (global < 10%) to predict soil color and mineralogy at all depths (Figure 6). The reason is that vegetation had a more local effect on the spatial distribution of soil attributes, followed by slope and density of geological lineaments. Horizontal and vertical curvatures had low importance, while aspect was frequently no important as predictor for our conditions (Figure 6), possibly because the sparse and uneven distribution of our soil dataset failed to described important short-range patterns of soil variation contained in these covariates.



Covariates \ Soil attributes	Hue num.			Value			Chroma			Goethite			Hematite			Kaolinite			Gibbsite			2:1			Global
	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	
SYSI Blue	47	31	45	34	32	33	26	16	8	28	34	7	60	53	47	15	16	11	13	11	13	9	14	7	25
SYSI Green	17	30	11	22	10	20	12	11	11	8	13	5	16	19	19	14	8	9	10	7	8	21	16	10	14
SYSI Red	20	12	13	17	14	15	20	12	13	8	11	6	22	20	18	8	7	8	5	6	7	13	7	10	12
SYSI NIR	23	15	11	20	6	13	12	12	10	7	12	6	14	15	11	20	16	21	10	8	9	8	10	8	12
SYSI SWIR$_1$	11	10	7	13	6	13	13	9	8	9	10	5	11	12	12	10	8	11	9	13	14	7	8	12	10
SYSI SWIR$_2$	11	10	8	13	5	11	17	9	7	6	11	4	9	9	9	14	15	17	16	37	35	8	12	11	13
SYSI LST	7	7	4	3	4	4	15	8	5	4	6	2	6	4	4	4	4	4	5	3	2	4	5	1	5
SYVI$_d$ Blue	11	12	10	9	6	6	9	9	5	6	6	5	9	6	7	6	5	8	4	5	6	8	7	4	7
SYVI$_d$ Green	8	7	4	10	5	5	6	6	3	5	6	4	8	8	6	10	5	12	6	6	8	4	6	6	7
SYVI$_d$ Red	6	5	5	6	4	6	10	10	9	6	6	4	8	8	8	11	9	11	9	11	11	5	6	5	7
SYVI$_d$ NIR	9	6	9	9	4	9	7	8	11	13	7	6	14	14	4	6	4	7	6	4	6	6	8	9	8
SYVI$_d$ SWIR$_1$	7	5	3	5	5	6	6	8	9	3	4	5	8	8	7	5	4	6	3	2	0	4	5	5	5
SYVI$_d$ SWIR$_2$	5	4	5	2	4	6	7	8	7	2	0	4	6	8	10	5	5	6	5	4	2	4	5	5	5
SYVI$_d$ LST	4	3	3	4	2	4	6	6	4	5	7	5	7	6	9	3	3	5	2	3	2	4	4	3	4
SYVI$_w$ Blue	12	7	11	10	5	9	9	7	5	4	3	3	7	4	6	5	3	4	3	3	4	3	5	1	6
SYVI$_w$ Green	6	4	4	5	3	6	8	8	6	7	5	2	6	6	5	6	4	7	6	7	7	2	5	3	5
SYVI$_w$ Red	4	5	6	7	3	7	8	11	8	5	4	3	5	4	5	6	4	6	3	7	7	4	5	5	6
SYVI$_w$ NIR	7	9	5	8	3	9	4	5	8	6	7	5	9	5	5	4	5	7	4	5	4	3	3	3	6
SYVI$_w$ SWIR$_1$	6	6	5	8	4	5	5	7	8	8	4	4	5	4	4	10	7	8	5	4	3	5	6	6	6
SYVI$_w$ SWIR$_2$	3	6	5	7	1	6	5	7	6	4	6	4	4	3	4	4	5	6	2	4	4	4	5	4	5
SYVI$_w$ LST	6	9	6	6	4	6	5	10	5	4	5	2	8	6	6	7	6	9	6	11	8	0	-1	3	6
Elevation	25	24	16	32	13	22	15	20	18	15	12	10	28	24	21	18	7	16	30	37	31	8	6	4	19
Slope	9	10	9	8	6	7	9	8	5	5	5	4	4	5	5	10	1	8	10	10	12	6	6	8	7
Topographic Position Index	20	23	28	10	11	16	8	8	8	6	8	3	15	13	16	7	10	7	12	10	9	9	14	16	12
Aspect	2	1	1	4	2	3	0	1	3	2	0	2	4	3	2	2	2	1	1	2	0	3	0	0	2
Horizontal Curvature	6	7	4	5	2	3	8	7	8	0	3	2	4	4	6	3	3	2	8	11	7	5	9	6	5
Vertical Curvature	12	6	9	6	2	3	7	3	3	1	3	3	7	4	4	4	3	3	8	12	11	6	11	6	6
Geological Lineaments	9	7	5	8	3	2	13	14	5	5	6	7	8	6	6	9	6	14	12	6	6	5	6	4	7
Annual Precipitation	26	21	20	18	8	16	13	32	25	28	15	8	13	16	14	13	13	15	23	31	19	6	5	6	17
Precipitation Seasonality	13	8	9	31	11	16	27	13	15	9	8	6	8	7	7	25	19	21	8	9	9	3	5	9	12
Annual Mean Temperature	15	13	17	17	7	7	11	24	12	13	13	8	12	8	8	12	8	12	15	22	27	6	7	6	13
Temperature Annual Range	14	18	14	11	10	13	11	11	8	25	15	5	17	15	15	35	40	57	7	7	13	14	15	13	17
Temperature Seasonality	12	20	17	10	6	11	17	30	48	12	13	6	11	9	10	17	12	22	11	10	13	10	17	9	15

Figure 6. Covariate's permutation importance (%) for soil attributes mapping at 0–20 cm (**A**), 20–60 cm (**B**) and 60–100 cm (**C**) depth intervals. We used the 10-fold importance in cross-validation to calculate mean values. Hue num: Hue number (see Munsell hue in section 2.2); 2:1: 2:1 clay minerals. Global is averaged importance values for all soil attributes (per row). NIR: Near infrared spectral band; SWIR1: First shortwave infrared spectral band; SWIR2: Second shortwave infrared spectral band; LST: Land surface temperature.

3.4. Digital maps of the soil surface and subsurface

3.4.1. Gridded Munsell soil color

We used the maps of Munsell hue, value and chroma to obtain RGB composites of the true Munsell color of soil for the three depth intervals (Figure 7). It allowed us to simultaneously assess and compare the spatial patterns of the three components.

On average, the study area had 49% of soils with hues redder (lower) than 5YR across the three depths (Table 3), which were mainly represented by Rhodic Ferralsols (and some Rhodic Nitisols and Acrisols of lower occurrence), followed by some Dystric Cambisols and Petric Plinthosols with redder hues in the soil matrix (see areas highlighted with red dashed lines in Figure 7). Within this set of soils, 7% were redder than 2.YR, due to the presence of ferralic (also ferritic) horizon of some Ferralsols (and Nitisols or Acrisols) developed from basalt in the study area.

About 51% of soils of the study area had Munsell hues yellower (higher) than 5YR up to 100 cm depth, which were represented by Haplic Ferralsols, Haplic Acrisol, Arenosols, Haplic Plinthosols and Petric Plinthosols with yellower matrix (see areas highlighted with yellow dashed lines in Figure 7). Among these soils, the 7% exhibited hues yellower than 7.5YR due to: 1) lower ratio hematite/(hematite+goethite), where higher contents of goethite pigmented the soil, such as in Xanthic Ferralsols, or 2) reduction and removal (or partial removal) of iron oxides from Gleysols.

The orange dashed lines in Figure 7 highlighted areas with Plinthosols (mainly Petric) in the study area. These soils contain petroplinthite (rich in Fe and Al) within a latosolic matrix, which can range from yellowish (10YR) to reddish (10R), according to the parent material. Their hue also can vary across the same soil profile. This features are important to understand the maps, because in such areas the Munsell color was more changing between depths.

The maps (Figure 7) showed that most of the soils in the study area with hues between 2.5YR and 7.5YR became redder with depth (Table 3). In addition, soils with Munsell hues < 5YR usually presented lower values and higher chromas than yellower soils with hues ≥ 5YR. The predicted true color of soils showed a comprehensible spatial correspondence with taxonomic classes of the legacy soil map. It can be observed in detail on Figure 7e, where the distinction of the spatial pattern of hue and true soil color is clearly evident between a Rhodic Ferralsol (redder and darker) and a Haplic Acrisol (yellower, lighter and brighter).

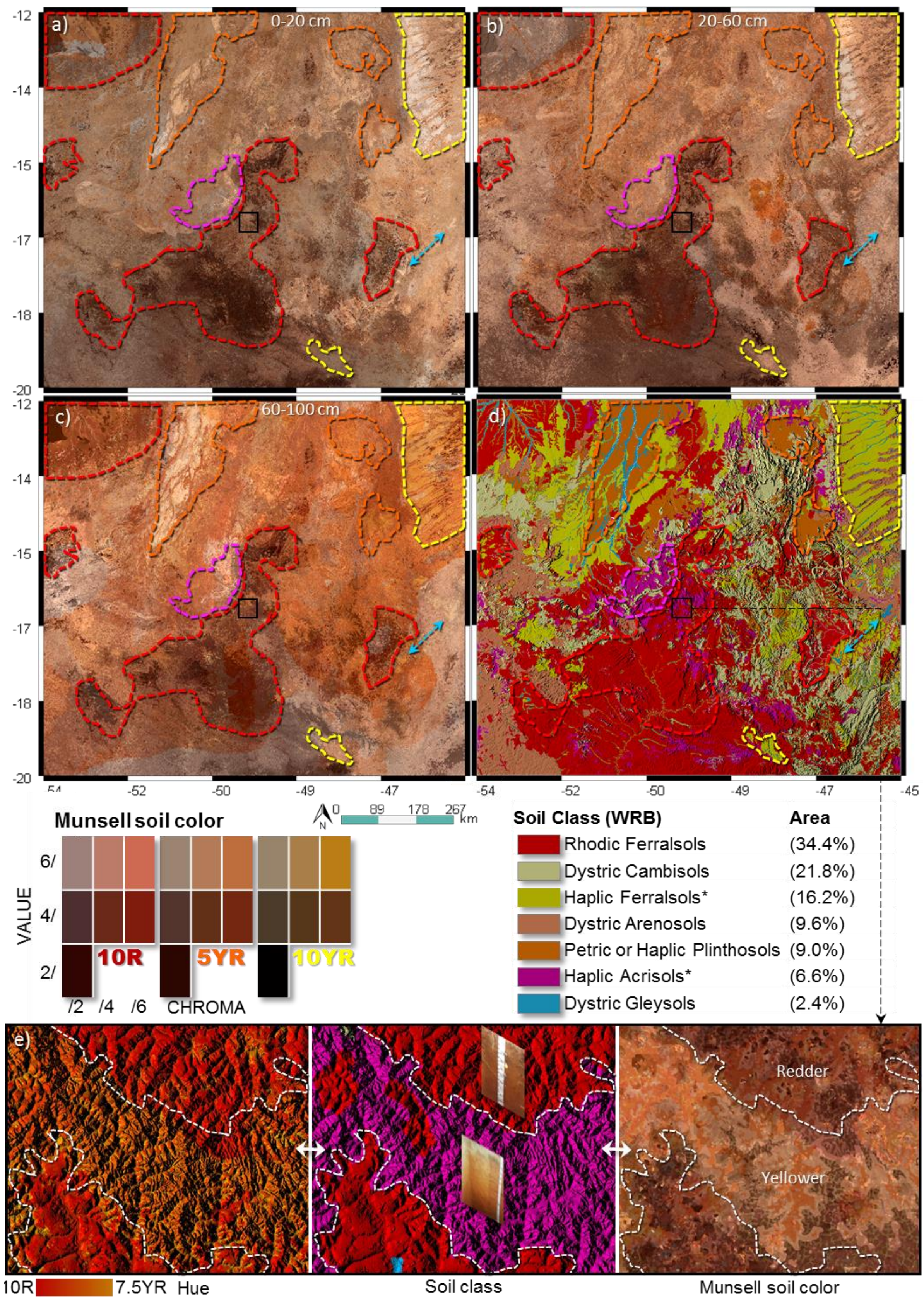


Figure 7. Maps of the true Munsell soil color across the study area and relationship with soil class legacy map. True soil color at (**a**) 0–20, (**b**) 20–60 and (**c**) 60–100 cm depth intervals; (**d**) legacy soil map (1:1,000,000-scale) (IBGE, 2017) with simplified classes according to the WRB (IUSS WORKING GROUP WRB, 2015); (**e**) expanded area showing in detail (from left to right) the spatial

pattern of soil hue, soil classes and true soil color. The dashed lines linked areas between maps with homogeneous soil color to a dominant soil class. *Soils with a hue of 5YR or yellower.

Table 3. Area quantification of the Munsell soil hue maps at three depth intervals.

Depth	**Hue < 2.5YR**	**2.5YR ≤ Hue < 5YR**	**5YR ≤ Hue ≤ 7.5YR**	**7.5YR < Hue**
(cm)		Area (%)		
0–20	1	27	66	6
20–60	4	49	46	1
60–100	16	49	23	13
Average	7	42	45	7

3.4.2. Spatial patterns of the main minerals in studied soils

For simultaneous assessment of the spatial patterns of soil mineralogy at each depth, we separately obtained RGB compositions for hematite, goethite and kaolinite (Figure 8a–c), and for gibbsite, 2:1 clay minerals and kaolinite (Figure 8d–f).

More than 50% of the study area was covered by highly-weathered soils with high relative proportions of hematite, goethite and kaolinite, followed by gibbsite and 2:1 clay minerals (Table 4). The relative proportions of iron oxides in the soil ranged from 6 to 66% as the surface materials were Fe-rich. The highest proportions of hematite (49% < Ht ≤ 66%) were found in 8% of soils, that accounted for nearly 7% of soils with hues redder than 2.5YR (Table 3). About 45% of soils had hematite contents ranging from 31 to 49% (see areas with red dashed lines in Figure 8a–c, and Table 4), that agreed with ~42% of soils with reddish hues between 2.5YR and 5YR (Table 3). This iron oxide also occurred in 47% of soils at lower contents (9% < Ht ≤ 31%), possibly coexisting with most of the 65% of soils with goethite amounts ranging between 24 and 37%, that may account for ~45% of soils with yellowish hues ranging from 5YR to 7.5YR. The lowest amounts of goethite, ranging from 6 to 24%, might be distributed in the redder soil masked by pigmenting effects of hematite. Conversely, 21% of soils presented high amounts of goethite ranging between 37 and 50% (see areas with green dashed lines in Figure 8a–c, and Table 4), which may account for the color of soils with hues yellower than 7.5YR.

The study area had about 56% of soils with high kaolinite contents ranging from 19 to 50% (blue shades in Figure 8d–f), which seemed to coexist in equilibrium with most of the 64% of soils with low amounts of gibbsite (%1 < Gb ≤ 9%). On the other hand, a large proportion of weathered soils (44%) with low kaolinite contents (%4 < Kt ≤ 19%) might coexist with the 36% of soils with highest gibbsite contents, ranging from 9 to 29% (see areas in shades of magenta with red dashed lines in Figure 8d–f, and Table 4). These highly-weathered soils were typical on highland surfaces, where long-term weathering resulted in intensive leaching of silica from soil particles.

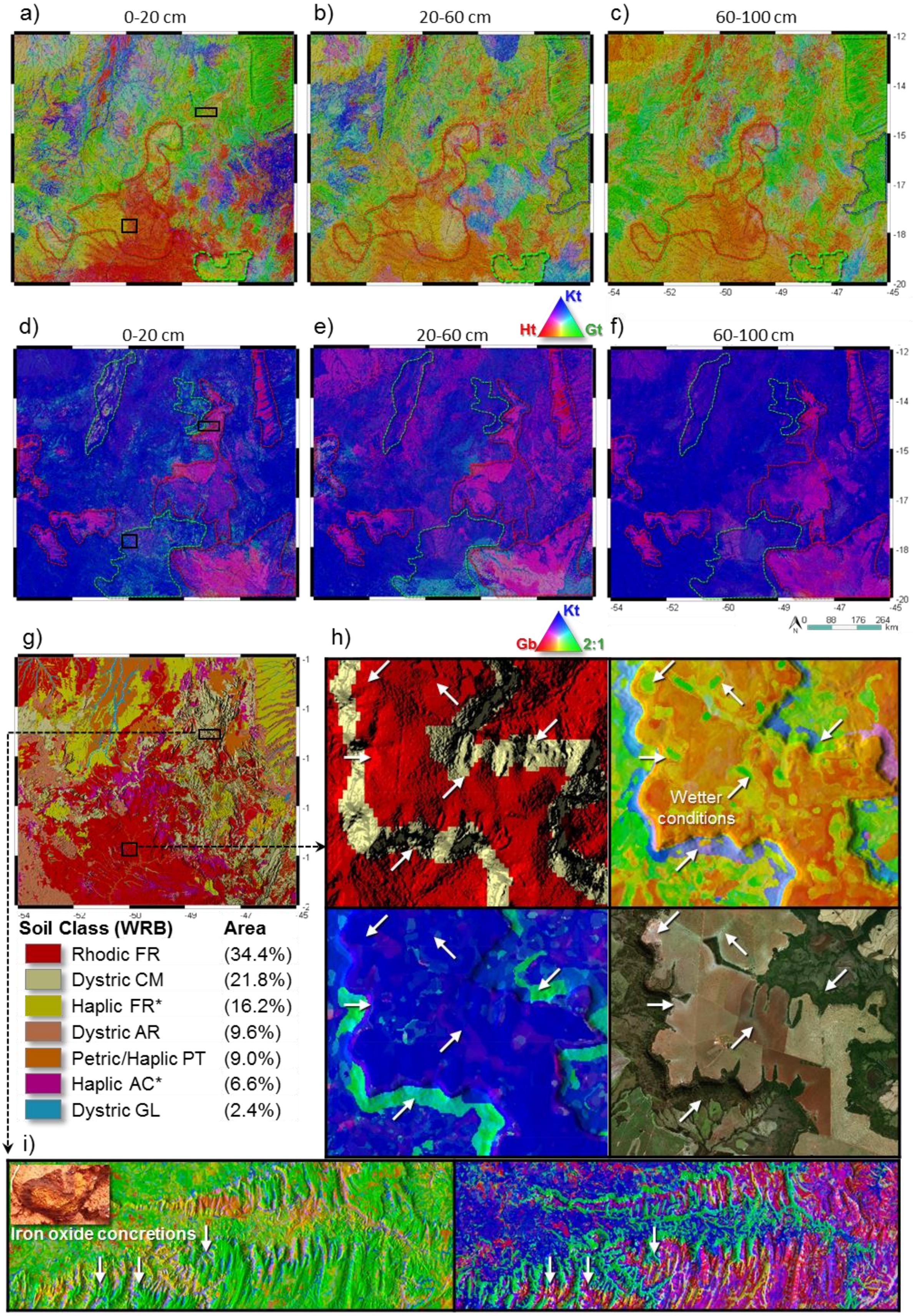


Figure 8. Maps of the soil mineralogy in RGB composition across study area and relationship with a soil class legacy map. RGB: Hematite (Ht), Goethite (Gt) and Kaolinite (Kt) at (**a**) 0–20 cm, (**b**) 20–60 cm and (**c**) 60–100 cm depth; RGB: Gibbsite (Gb), 2:1 clay minerals (2:1) and Kaolinite (Kt) at (**d**) 0–20 cm, (**e**) 20–60

cm and (**f**) 60–100 cm depth; (**g**) legacy soil map (1:1,000,000-scale) (IBGE, 2017) with simplified classes according to the WRB (IUSS WORKING GROUP WRB, 2015); (**h–i**) expanded areas showing in detail the spatial pattern of soil classes and mineralogy. The dashed lines linked areas between maps with homogeneous soil mineralogy. *Soils with a hue of 5YR or yellower. FR: Ferralsol, CM: Cambisol, AR: Arenosol, PT: Plinthosol, AC: Acrisol, GL: Gleysol.

Traces of 2:1 clay minerals (< 7%) were found in most of soils in the study area (76%), while the remaining 24% of soils showed higher contents (7% < 2:1 ≤ 18%), displayed with shades of cyan in Figure 8d–f (especially in areas highlighted with cyan dashed lines). Table 4 shows that iron oxides and kaolinite contents increased with depth, while gibbsite increased with less intensity. Higher proportions of 2:1 clay minerals seemed to be more abundant at topsoil.

We showed in detail the spatial patterns of soil mineralogy in two expanded areas linked to the legacy soil map (Figure 8g). The first area (Figure 8h) is on a highland plateau developed upon arenite and covered by Ferralsols, where flat areas in red shades were predicted with redder color (3YR 3/5) and proportions of Ht = 46%, Gt = 36%, Kt = 10%, Gb = 5% and 2:1 = 3%, indicating that hematite was dominant in well-drained conditions. On concave areas or lower slopes surrounding drainages with green shades in the map (marked by black arrows) showed yellower color (6YR 4/4) and proportions of Ht = 27%, Gt = 47%, Kt = 18%, Gb = 3% and 2:1 = 6%, suggesting that goethite was dominant in wetter drainage conditions. Conversely, on plateau edges covered by Cambisols with cyan shades in the map, presented color 3YR 4/3 and amounts of Ht = 38%, Gt = 36%, Kt = 8%, Gb = 2% and 2:1 = 16%, suggesting less weathered conditions and younger soils.

The second area (Figure 8i) presented two scenarios, one developed upon siltite, where smoother relief (shades of blue) showed color 5.5YR 4/5 and proportions of Ht = 30%, Gt = 40%, Kt = 15%, Gb = 5% and 2:1 = 10%, suggesting less weathered conditions. Areas over ferruginous laterite crusts with hilly relief (shades of magenta), and occurrence of iron oxide concretions, presented color 7YR 4/4 and amounts of Ht = 16%, Gt = 43%, Kt = 8%, Gb = 29% and 2:1 = 4%, indicating highly-weathered conditions on an ancient surface. Thus, the first scenario accounted for younger soils (e.g. Cambisols), while the second corresponded to older soils, such as Petric Plinthosols.

Table 4. Area quantification of the soil mineral maps at 0–20, 20–60 and 60–100 cm depth intervals.

Relative amount*	**0–20 cm**	**20–60 cm**	**60–100 cm**	**Average**
(%)		Area (%)		
$9 < Ht \leq 31$	51	48	42	47
$31 < Ht \leq 49$	36	46	52	45
$49 < Ht \leq 66$	13	5	6	8
$6 < Gt \leq 24$	23	12	6	14
$24 < Gt \leq 37$	55	69	72	65
$37 < Gt \leq 50$	22	19	23	21
$4 < Kt \leq 19$	46	44	43	44
$19 < Kt \leq 31$	34	38	50	41
$31 < Kt \leq 50$	21	17	7	15
$1 < Gb \leq 9$	69	63	60	64
$9 < Gb \leq 17$	26	20	38	28
$17 < Gb \leq 29$	5	17	2	8
$2 < 2{:}1 \leq 7$	58	79	91	76
$7 < 2{:}1 \leq 13$	34	20	9	21
$13 < 2{:}1 \leq 18$	8	0.4	0	3

*Amount of given mineral in the < 2 mm fraction of soils. Gt: goethite; Ht: hematite; Kt: kaolinite; Gb: gibbsite; 2:1: 2:1 clay minerals.

4. DISCUSSION

4.1. Relationships between soil color and mineralogy

We did not verify the accuracy of the color estimations at each site because: 1) lacked of colorimeter records in our dataset; 2) spectral data were acquired under the same conditions that in reference works (MARQUES et al., 2019; RIZZO et al., 2016); 3) the mathematical procedures of reference, implemented in this section, provided similar color estimations to the colorimeter measurements, with R^2 ranging from 0.68 to 0.96 and RMSE between 0.19 and 0.57 (MARQUES et al., 2019; RIZZO et al., 2016).

Aitkenhead et al. (2013) demonstrated that inherent color of soil is mainly controlled by organic compounds and iron oxides. Soil organic matter causes the darkness of soil by decreasing the Munsell value and chroma (SCHULZE et al., 1993). Poppiel et al. (2019b) found organic matter inversely correlated (r = -0.4) with soil depth for the same area of Brazil, where average content ranged from 21 g kg^{-1} at the surface to 9 g kg^{-1} in the 60–100 cm depth. These findings agreed with our results, where value and chroma increased with depth, while organic matter decreased, suggesting a negative correlation between them, as reported by Simon et al. (2019).

The most frequent pedogenic oxides in tropical soils are hematite (usually associated to goethite) with hues between 10R and 5YR, and goethite that has hues between 7.5YR and 2.5Y (SCHWERTMANN, 1993). Munsell color varies with mineral concentration, where higher contents reduce the value and increase the chroma of soil (FERNANDEZ et al., 1992). According to the geodiversity of the region (MORAES, 2014), the most surface materials (Figure 1) contain Al, Si, and Fe-bearing minerals, that released these elements during their weathering (hydrolysis) and it favored the formations of oxide pigments (e.g. hematite and goethite) (SCHWERTMANN et al., 1989), common to the majority of the studied soils (Figure 3) (BARBOSA et al., 2009; RODRIGUES, 1977; ZINN et al., 2007, 2016). Goethite (FeOOH) usually occurs in wetter, colder, and more acidic (pH 4) pedoenvironments, with seasonal anaerobic conditions and slow Fe release (SCHWERTMANN et al., 1989). When the pedoclimate becames drier, warmer and less acidic (pH 7) under higher Fe release, the ferrihydrite (precursor) is formed and then dehydrated to hematite (Fe_2O_3); or also, goethite can dehydrate to hematite (SCHWERTMANN, 1993). Usually, in red soils widely distributed in our study area (e.g. Rhodic Ferralsols), the yellowish hues (10YR) of coexisting goethite are masked by the higher pigmenting effects of hematite with reddish hues (10R) (BARRÓN et al., 1986). Hematite, a less stable mineral, is generally negligible or absent in yellow soils (e.g. Xanthic Ferralsols) from Central Plateau of Brazil (CURI et al., 1984).

When iron oxides are completely removed (after mobilization by microbial reduction) under anaerobic conditions from soil particles, and if organic matter is negligible, the soil achieves the base color of the matrix minerals resulting in shades of gray (gleyic) (SCHWERTMANN, 1993). Reducing conditions can dramatically reduce the chroma and increase the value of gleyed horizons, suggesting saturation by water in concave areas of the landscape, characteristic of Gleysols (IUSS WORKING GROUP WRB, 2015).

The highest kaolinite content in the < 2 mm fraction of soils (see ternary graphs in Figure 4) might result from primary minerals, which weathered directly into kaolinite under intense warm and wet leaching in tropical conditions (MELO et al., 2001). Gibbsite, a pedogenic $Al(OH)_3$, is formed by desilication of kaolinite or primary minerals, at low silica concentration and low pH (5-6), when leaching rates are rather high in well-drained tropical soils (SCHAEFER et al., 2008). Relatively large amounts of this mineral were found in the clay fraction of deeply weathered soils in Central Brazil (RODRIGUES, 1977).

The 2:1 minerals are derived from their parent materials and can be present: 1) in the clay fraction along the profile of younger (less weathered) soils, or 2) strongly interlayered with Al in older soils, which decrease the cation exchange capacity by blocking exchange sites and provide greater stability, as reported by some works for the same region (BARBOSA et al., 2009; ZINN et al., 2007, 2016). In addition, weathered soils (e.g. Ferralsols) can contain up to 5, 17 and 5% of 2:1 minerals in the sand, silt and clay fractions, respectively (RODRIGUES, 1977). The first two fractions gently decreased their concentration with depth in the region (POPPIEL et al., 2019), and therefore, the 2:1 minerals were reduced as well (Figure 4 right panels and Figure 5d).

Soil color allows to infer about the conditions of aeration and drainage of the soil and, consequently, of pedogenetic processes. Thus, red soils (hematite and goethite) are in well-drained interflows; yellow soils (goethite), on moderately drained slopes, and grey soils develop in poorly drained foothills. Mineralogical composition can be used to estimate the degree of weathering of soils, where the next sequence indicates an increasing degree of evolution (from younger to older): 2:1 < kaolinite < hematite < goethite < gibbsite. Thus, the majority of soil presented an advanced weathering degree with well drainage condition, developed in flattened or smoothed reliefs.

4.2. Use of regression models for mapping soil properties

As soil color and mineralogy are important proxies used to distinguish different soil types or to infer related soil attributes (ZHANG et al., 2019), they play an important role in soil cartography. Some studies have used reflectance spectroscopy (350–2500 nm) as input data to estimate the color and/or its mineralogy (ESCADAFAL et al., 1988; FERNANDEZ et al., 1987; MARQUES et al., 2019; MATTIKALLI, 1997; RIZZO et al., 2016; SCHEINOST et al., 1998; SIMON et al., 2019), but only a small number of works mapped their spatial distribution. At the moment, Viscarra Rossel et al. (2010) performed one of the few studies on soil color mapping, where they accurately mapped ($R^2 \cong 0.67$) iron oxides and the color of Australian soil using reflectance spectra (350–2500 nm) and geostatistics.

Studies on mapping the soil mineralogy such as Viscarra Rossel and Chen (2011), summarized the information content of spectra (350–2500 nm) by principal components to construct linear models, and map the mineral (the first three PCA scores) of Australian topsoils robustly ($0.69 < R^2_{10cv} < 0.85$). Likewise, Viscarra Rossel (2011) measured the relative abundances of kaolinite, illite and smectite at 0–20 and 60–80 cm soil depths, using continuum-removed reflectance (350–2500 nm) to derive statistical models and map the minerals with good cross-validation results ($0.40 < R^2_{10cv} < 0.61$). Malone et al. (2014) also used continuum-removed spectra (350–2500 nm) for the detection of iron oxides, kaolinite and smectite prior to mapping their spatial distribution in Australia, such as ordinal classes at fixed mineral abundance intervals, with overall accuracy ranging from 44 to 80%. Mulder et al. (2013) used reflectance spectroscopy (350–2500 nm) to derive soil minerals, and multinomial logistic regression, for mapping the likelihood of "absence" or "presence" of kaolinite, mica and smectite with high overall accuracy (>0.74). Other studies (ROBERTS et al., 2019; DUCART et al., 2016; MADEIRA NETTO et al., 1997) used enhanced mineral mapping techniques to produce a thematic mineral map of soil using the spectral response of Landsat imagery.

Our performance metrics were consistent with studies mentioned above, where most of them used *scorpan* model (MCBRATNEY et al., 2003) for DSM and reported a decline of prediction accuracy from calibration to validation, as summarized in Table 2. The unexplained part of soil variation in our study area can be due to two aspects. The first might be a limited number of sparse soil observations, with one site per ~2 km^2 (denser) to ~800 km^2 (less dense) and ~600 km^2 on average in the study area, as also reported by Liu et al. (2020) when mapped the texture of Chinese soils using RF algorithm. This may be not sufficient to capture and describe short-range patterns of soil variation (HENGL et al., 2019). The second, and the most relevant for soil mapping and its cross validation, can be an uneven spatial distribution of observations (WADOUX et al., 2019). In large extent mapping, more landscapes are usually included with sampling sites not uniformly distributed over space. Figure 1 shows that the northwestern and northeastern portions of the study area had less soil observations than other parts. That is because: 1) the soil data, used in this study, was acquired by survey activities with limited funding, along different periods of time and without a statistical design, but purposive; 2) relatively smaller soil spatial variation in the northwestern and northeastern, developed over more uniform conditions of geology and relief, that were considered by our observations.

Despite our dataset covered the main soil-landscape conditions across the study area, 10-fold cross validation was performed on uneven distribution of observations. This method selects 10% of total sites for validation, leading to relatively smaller number of data for modelling in the areas with sparse observations (LIU et al., 2020). In addition, although the model performances were robust for the whole extent, its prediction may have biased in local areas.

The worse spatial prediction accuracy for chroma can be a consequence of a possible lower performance in their determination from spectra, since this Munsell component is influenced by the organic matter, which decreases in depth, where chroma model had a slightly better performance (R^2_{10cv} = 0.38). In addition, Liles et al. (2013) reported that soils developed over sedimentary rocks, as most of our study area, showed an increasing in the coefficient of variation for Munsell chroma. Silva et al. (2020) found that the spatial variability of goethite was about twice higher than hematite in soils from the Western Paulista Plateau of Brazil, strongly influenced by the parent material. Thus, the lowest model performance for chroma may be related to effect of the density and locations of soil observations used for color predictions, combined with the high occurrence of sedimentary parent materials in the study area.

The substitution of Fe by Al in goethite, that is greater than in hematite, ranging from 7 to 40% for Brazilian soils (SCHAEFER et al., 2008), may be produced their lower performance. This process cause less stability in the absorption feature of goethite (KOSMAS et al., 1984; SCHEINOST et al., 1998), and consequently lower prediction performances, especially at subsoil layers (R^2_{10cv} = 0.24), where we had a relatively smaller number of soil samples.

4.3 Influence of environmental predictors in soil color and mineralogy patterns

Most influential covariates were important predictors of the soil color and mineralogy because they captured the soil spatial patterns at shorter distances or local variations (detail), and also at longer distances or regional variations (generalization) across different landscapes (HENGL et al., 2019). Therefore, SySY (soil), SyVI (vegetation), DEM and derived relief attributes describe at detail the factors of soil formation, while temperature, precipitation and geological lineaments generalized their patterns (VISCARRA ROSSEL, 2011). Then we were able to spatialize our soil predictions from detailed to successively coarser levels of generalization in our study area. The impact of use multi-scale and -source predictors for modelling soil attributes was demonstrated by Miller et al. (2015). They reported that the parallel use of covariates at multiple levels of spatial representation for DSM, improved the model performance, promoting R^2 increases of up to 70%.

Advantages are being taking from the petabyte-scale Landsat datasets widely available within GEE (GORELICK et al., 2017). The covariates SySI and SyVI (Table 1) are examples of that benefits (POPPIEL, et al., 2019b), which provide improved proxies for describing several soil forming factors, e.g. *s*, *o*, *p* and *a* (MCBRATNEY et al., 2003). SySI can provide direct and interpretable information from Earth bare surfaces, from which inferences can be made about the main soil attributes, e.g. the soil color, mineralogy and texture, among others (DEMATTÊ et al., 2018; POPPIEL et al., 2019b). In a recently study, Roberts et al. (ROBERTS et al., 2019) robustly estimated the spectral response of the bare surfaces using the full temporal archive of Landsat images across Australia. The authors highlighted the broad application of the topsoil reflectance mosaic, which can be combined with machine learning for enhanced geological mapping, mineral exploration and digital soil mapping. Likewise, Post et al (POST et al., 1994) reported a very strong correlation ($0.68 < r < 0.85$) between Munsell soil color measured with a colorimeter and Landsat reflectance on semiarid rangelands, where they precisely and accurately determined the color of bare topsoil using remotely sensed spectral data.

When we examined individually the relevance of predictors for each soil attribute, we found that $\text{SySI}_{\text{Blue}}$, $\text{SySI}_{\text{Green}}$ and SySI_{Red}, were the most important spectral bands to predict Munsell hue (from 12 to 47%), value (from 14 to 34%) and chroma (from 11 to 26%), see Figure 6. This is because Munsell color system described different soil components with absorption features (due to electronic transitions) in the visible range between 380 and 780 nm (TORRENT et al., 1993), where the blue, green and red Landsat spectral bands are situated. The $\text{SySI}_{\text{Blue}}$ was by far the most important predictor for geothite (from 7 to 34%) and hematite (from 47 to 60%), followed by $\text{SySI}_{\text{Green}}$ and SySI_{Red} (between 8 and 22%). Goethite and hematite had stronger absorption features situated between the blue and red spectral ranges (Figure 4), with a weaker effect in the near-infrared interval (KOSMAS et al., 1984; SCHEINOST et al., 1998). The $\text{SySI}_{\text{Swir1}}$ and $\text{SySI}_{\text{Swir2}}$ were important (from 9 to 37%) for gibbsite and kaolinite because they both exhibit molecular vibrations (involving stretching and bending) between ~1400 and ~2300 nm (CLARK et al., 1990). Also, $\text{SySI}_{\text{Blue}}$ and SySI_{NIR} were important (from 11 to 21%) for gibbsite and kaolinite, because these minerals are usually associated with iron oxides in tropical soils (SCHAEFER et al., 2008; SCHWERTMANN et al., 1989), which had spectral response between 380 and 1000 nm (KOSMAS et al., 1984; SCHEINOST et al., 1998). $\text{SySI}_{\text{Swir2}}$ was important (from 8 to 12%) for predicting 2:1 clay minerals, due to their typical absorption features (by molecular vibrations) near 1900 nm (CLARK et al., 1990). $SySI_{Blue}$, $SySI_{Green}$ and $SySI_{Red}$ also were highly important (from 7 to 21%) to predict 2:1, since they usually were associated with iron oxides in soils of our study area (BARBOSA et al., 2009; ZINN et al., 2007, 2016).

SyVI provides vegetation feedback dynamic patterns attributed to the differences in topsoil and subsoil conditions, across the rooting depth (POPPIEL et al., 2019b), that can help to distinguish, for example: 1) warmer and more rapidly drying sands, from colder and slower drying wet clays (LIU et al., 2012); 2) different levels of chemical soil attributes, such as pH and fertility (MAYNARD et al., 2017); among others. These soil conditions are all interlinked with other soil attributes (e.g. soil color and mineralogy) as a result of pedogenic processes (MCBRATNEY et al., 2003).

Among the terrain attributes, elevation, topographic position index and slope were the most important covariates for modelling soil color and mineralogy. They control the water dynamic of the relief, which influenced the intensity of erosion, redistribution and sorting processes of soil particles (LIU et al., 2020). In addition to that, the density of geological lineaments strongly influenced the surface drainage density, soil texture and soil depth (DAS, 2019), which controlled the internal drainage through soil and the leaching rates (CURI et al., 1984). Relief attributes such as horizontal and vertical curvatures and aspect had relatively low importance, because they usually controlled local moisture, thermal conditions and short-range mass redistribution over landscapes (MCBRATNEY et al., 2003).

Especially for gibbsite, the elevation of terrain was a very important predictor of their spatial patterns in our study area (Figure 6). According to the study of Reatto et al. (2008), the spatial variation of gibbsite in the Brazilian Central Plateau depended on two aspects. First, the spatial variability of gibbsite at regional levels was mainly related to the age (*a*) of the surface, since the higher the elevation, the greater the time the soils were exposed to weathering and hydrolysis process in tropical climate conditions, resulting in older soils (e.g. Ferralsols, Plinthosols) with a higher gibbsite content. Second, local spatial pattern of gibbsite was related to the local topographic position on landscape (*r*), where conditions that favored the percolation of water through the soil and the hydrolysis processes, presented greater amounts of gibbsite. These conditions on soil water and temperature regimes, also affect the genesis of iron oxides and organic matter oxidation rates, which strongly influence the soil attributes, such as color, aggregation of soil particles, the retention of cations and anions (SCHWERTMANN, 1993).

Climate conditions of relatively high annual temperature (> 20 °C) and precipitation (> 1000 mm) and low temperature changes in the study area lead to strong weathering of surface materials (Al, Si, Fe-rich) (SCHWERTMANN et al., 1989) and intensive silica leaching, that provided conditions for accumulation of specific mineral products (SCHAEFER et al., 2008), such as iron oxides that pigmented the soils (SCHWERTMANN, 1993). Using a similar approach, Ramcharan et al. (2018) found that climate covariates, followed by elevation and satellite data derived from MODIS and Landsat, were the most important predictors for both soil property and taxonomic classes, across the United States.

4.4. Comparison with legacy data and maps

In this section, the most significant issue was but to account for the trend of spatial patterns between the data instead of measure the bias or error between predicted maps and legacy observations. We demonstrated that DSM using proximal and remote sensing data can reach realistic spatial representations of the soil.

The spatial patterns of soil on our predicted maps were consistent with pedological expert knowledge of the region and with legacy data presented in Table 5. Predicted Munsell color was negatively correlated with total elements, especially with the Fe that reduced the hue ($-0.18 \leq r \leq -0.35$) and value ($-0.39 \leq r \leq -0.53$) of soil at the three depth intervals. Higher Fe, Al and Ti concentrations tend to darken the soils, by reducing the brightness and increasing the yellowness, redness, or brownness of soil (SCHWERTMANN, 1993). Chroma was poorly correlated with total elements and not entirely consistent at 60–100 cm depth, where was influenced by Fe ($r = -0.30$) possibly due to their worse spatial prediction accuracy (Table 2). These findings agreed with Simon et al. (2019), who reported negative correlations of hue and value with Fe ($-0.25 \leq r \leq -0.37$) and Al ($r \leq -0.64$), and weak for chroma ($r \leq 0.06$). The Munsell color's spatial patterns from our maps were coherently correlated with the Munsell color from legacy observations ($0.14 \leq r \leq 0.63$), although the latter was determined visually in wet conditions. These relationships reinforce the accuracy and representativeness of our spatial predictions.

Maps of soil minerals mostly were correlated with total elements of soil ($0.20 \leq r \leq 0.56$), determined from clay fraction by sulfuric acid digestion method (Table 5). It was because Ht, Gt, Gb and Kt from Midwest Brazil are Fe and Al-bearing minerals (CURI et al., 1984). The correlations between predicted soil minerals and Ti ($r \leq 0.29$) occurred because titanium probably was absorbed or incorporated into the crystal framework of iron oxides as impurities (SCHWERTMANN et al., 1989). Goethite showed correlation with Al in soils ranging from 0.1 to 0.33 (Table 5), likely because the yellower soils of the region contained more goethite (e.g. Xanthic and Haplic Ferralsols), which was found to have more Al-substituted than hematite (CURI et al., 1984; SCHAEFER et al., 2008). In addition, predicted iron oxides were inversely related with observed hue and value ($-0.24 \leq r \leq -0.46$), suggesting that these two minerals (mainly hematite) reddened and darkened the soil color. Conversely, goethite, gibbsite and kaolinite tended to brighter the soil by increasing the chroma ($0.11 \leq r \leq 0.32$), as suggested in Figure 5 and Table 5. The correlations with Ti may be resulted from the ferralic (also ferritic) horizon of some Ferralsols and Nitisols developed from basalt in the study area (MELO et al., 2001; RODRIGUES, 1977). Ferralic horizons are rich in iron oxides (especially hematite), where the clay fraction can reach 5.3% of Ti-bearing minerals, mainly ilmenite and anatase

(SCHAEFER et al., 2008). Also, some Ti might be substituted in the kaolinite structure or surface-sorbed (MELO et al., 2001).

Table 5. Verification of the spatial correspondence, based on Pearson's correlation (p-value < 0.05), between our predicted maps (at the three depth intervals) with legacy soil observations acquired from a national dataset (SAMUEL-ROSA et al., 2020), and weathering degree and hue, both inferred from a legacy soil class map of the study area (IBGE, 2017).

Depth	**Legacy data**		**Our predicted maps**							
(cm)	***Total elements***[1]	n	HueN	Value	Chroma	Ht	Gt	Gb	Kt	2:1
0–20	Fe_2O_3	225	**-0.35**	**-0.39**	0.10	**0.39**	-0.03	0.06	0.09	0.12
	Al_2O_3	878	-0.01	-0.12	0.05	0.09	0.10	**0.25**	**0.25**	0.01
	TiO_2	782	-0.12	-0.16	0.03	0.11	0.04	0.13	0.06	0.19
20–60	Fe_2O_3	124	**-0.26**	**-0.39**	-0.12	**0.38**	0.11	0.06	0.00	0.09
	Al_2O_3	729	-0.03	0.10	0.04	0.04	**0.23**	**0.20**	**0.23**	0.06
	TiO_2	639	**-0.27**	-0.12	-0.03	0.20	0.28	0.10	0.14	0.07
60–100	Fe_2O_3	174	**-0.18**	**-0.53**	-0.30	**0.56**	**0.30**	0.15	0.00	0.51
	Al_2O_3	532	-0.06	-0.04	-0.09	0.06	**0.33**	**0.37**	0.06	0.05
	TiO_2	479	-0.06	-0.23	-0.14	0.22	0.29	0.14	-0.15	0.20
	Munsell color[2]									
0–20	Hue number	230	**0.53**	0.38	-0.24	**-0.39**	**-0.26**	-0.01	-0.06	-0.10
	Value	230	0.32	**0.37**	0.01	**-0.24**	-0.07	0.06	-0.07	-0.18
	Chroma	230	0.01	0.02	**0.16**	0.01	**0.23**	**0.32**	**0.15**	-0.05
20–60	Hue number	195	**0.63**	0.44	-0.11	**-0.40**	**-0.30**	0.02	-0.11	-0.02
	Value	195	0.48	**0.46**	0.03	**-0.35**	**-0.32**	0.04	-0.10	0.04
	Chroma	195	0.05	0.17	**0.14**	-0.04	**0.11**	**0.24**	**0.19**	0.13
60–100	Hue number	143	**0.35**	0.46	-0.16	**-0.44**	**-0.15**	0.05	0.12	-0.06
	Value	143	0.42	**0.58**	-0.23	**-0.46**	**-0.38**	-0.06	0.09	-0.11
	Chroma	143	-0.01	0.05	**0.19**	-0.06	**0.21**	**0.12**	0.01	-0.12
	Legacy soil map									
0–20	Weather. degree[3]	5k*	**-0.38**	**-0.34**	0.08	**0.42**	**0.23**	**0.19**	0.09	-0.02
	Hue number[4]	5k*	**0.48**	**0.47**	-0.10	**-0.52**	**-0.27**	**-0.17**	-0.05	-0.13
20–60	Weather. degree[3]	5k*	**-0.27**	**-0.35**	0.03	**0.40**	**0.35**	**0.17**	-0.02	0.10
	Hue number[4]	5k*	**0.39**	**0.47**	0.03	**-0.49**	**-0.42**	**-0.10**	-0.03	-0.22
60–100	Weather. degree[3]	5k*	**-0.15**	**-0.38**	-0.09	**0.37**	**0.31**	**0.23**	-0.03	0.02
	Hue number[4]	5k*	**0.29**	**0.48**	0.20	**-0.47**	**-0.38**	**-0.18**	0.00	-0.17

[1]Fe, Al and Ti were determined from clay fraction by sulfuric acid digestion method; [2]Munsell color of soil determined visually in wet conditions. We used the soil classes of the legacy soil map to infer a theoretical number sequence, according to the WRB (IUSS WORKING GROUP WRB, 2015), for: [3]weathering degree from 1 (less weathered) to 10 (more weathered), as follow Leptosols, Arenosols, Gleysols, Cambisols, Plinthosols, Acrisols, Nitisols and Ferralsols (Xanthic, Haplic and Rhodic); and [4]Munsell hue number from 10 (redder, 10R) to 22.5 (yellower, 2.5Y), as follow Rhodic Ferralsols, Rhodic Nitisols, Plinthosols, Leptosols, Cambisols, Haplic Ferralsols, Haplic Acrisols, Arenosols, Xanthic Ferralsols, Gleysols. *5k: 5000 random points. n: numbers of observations or random points used for sampling the maps; HueN: Munsell hue number; Ht: hematite; Gt: goethite; Gb: gibbsite; Kt: kaolinite; 2:1: 2:1 clay minerals.

Low predicted Munsell hues (redder) and values suggested ($-0.15 \leq r \leq -0.38$) higher degrees of soil weathering inferred from a legacy map of soil classes (Table 3). Therefore, nearly 50% of the study area was dominated by weathered soils (Figure 7), such as Rhodic and Haplic Ferralsols (IUSS WORKING GROUP WRB, 2015), which presented high amounts of iron oxides that pigmented the soil color (reddened or yellowed) and absorbed the sunlight (darkened) (SCHWERTMANN, 1993). Higher relative proportions of predicted iron oxides and gibbsite correlated with higher theoretical soil weathering degrees ($0.17 \leq r \leq 0.42$) and lower theoretical Munsell hues ($-0.10 \leq r \leq -0.52$).

Thus, we achieved accurate large extent soil maps because our models dealt with the complex relationships between factors of soil formation across the region, that were well-described by covariates at multiple resolutions. The linkage of our spatial predictions with legacy data provided a good correspondence at both local and regional levels, provided by correlations with soil observations that were relatively uniform spatial distributed and the associations with regionals patterns derived from a legacy soil map (IBGE, 2017). This map of soil classes at coarse 1:1,000,000-scale was performed several years ago by Brazilian government agencies, and is currently the best available pedological information covering the study area.

5. CONCLUSIONS

Reflectance spectra (350–2500 nm) can be used to accurately determine the Munsell color of soil and the relative abundance of hematite, goethite, kaolinite, gibbsite and 2:1 clay minerals in tropical soils. Once the method was defined, only a few minutes were required for application of any of the steps described in 2.2 and 2.3 sections, apart from the time necessary for drying, grinding and sieving the soil samples. Sample mount in Petri dishes and measurement required only a short time and low-cost without chemical solutions, thus making the method suitable for use on a routine basis. We encouraged the soil scientists to implement and improve this clean technology into their research.

The Random Forest models proved to be robust for mapping soil color and mineralogy (derived from spectra) at three depth intervals in Midwest Brazil. Validation showed high prediction accuracy for hematite ($R^2_{10cv} > 0.71$), followed Munsell value and hue, gibbsite, kaolinite, 2:1 minerals and goethite at topsoil and subsoil ($0.43 < R^2_{10cv} < 0.65$). Munsell chroma at all depths had worse prediction accuracy ($0.24 < R^2_{10cv} < 0.38$).

The most relevant predictor of the spatial patterns of soil color and mineralogy at surface and subsurface in Midwest Brazil was the blue spectral region of satellite topsoil reflectance ($SySI_{Blue}$) with 25% of global importance, followed by elevation, precipitation and temperature. These covariates are proxies of the soil forming factors *s*, *c*, *r*, *p* and *a*.

More than 50% of the study area was covered by highly-weathered soils, where 45% of soils had 31 to 49% of hematite accounting for 42% of soils with reddish hues between 2.5YR and 5YR. Nearly 56% of soils had 19 to 50% of kaolinite while 36% of weathered soils presented highest gibbsite contents between 9 and 29%. Traces of 2:1 clay minerals (< 7%) were found resisting in most of soils in the study area (76%).

The soil spatial patterns on our predicted maps were consistent with pedological expert knowledge of the region and with legacy soil observations and legacy soil class map. Therefore, we have proven that large-extent DSM at a fine resolution using proximal and remote sensing data can reach realistic spatial representations of soil color and mineralogy in tropical conditions.

6. APPENDIX A

Figure A1 exhibits the predicted vs observed scatterplots of 10-fold cross-validation derived from optimized models for Munsell hue number, value and chroma, goethite, hematite, kaolinite, gibbsite and 2:1 clay minerals at three depth intervals (0−20 cm, 20−60 cm and 60−100 cm).

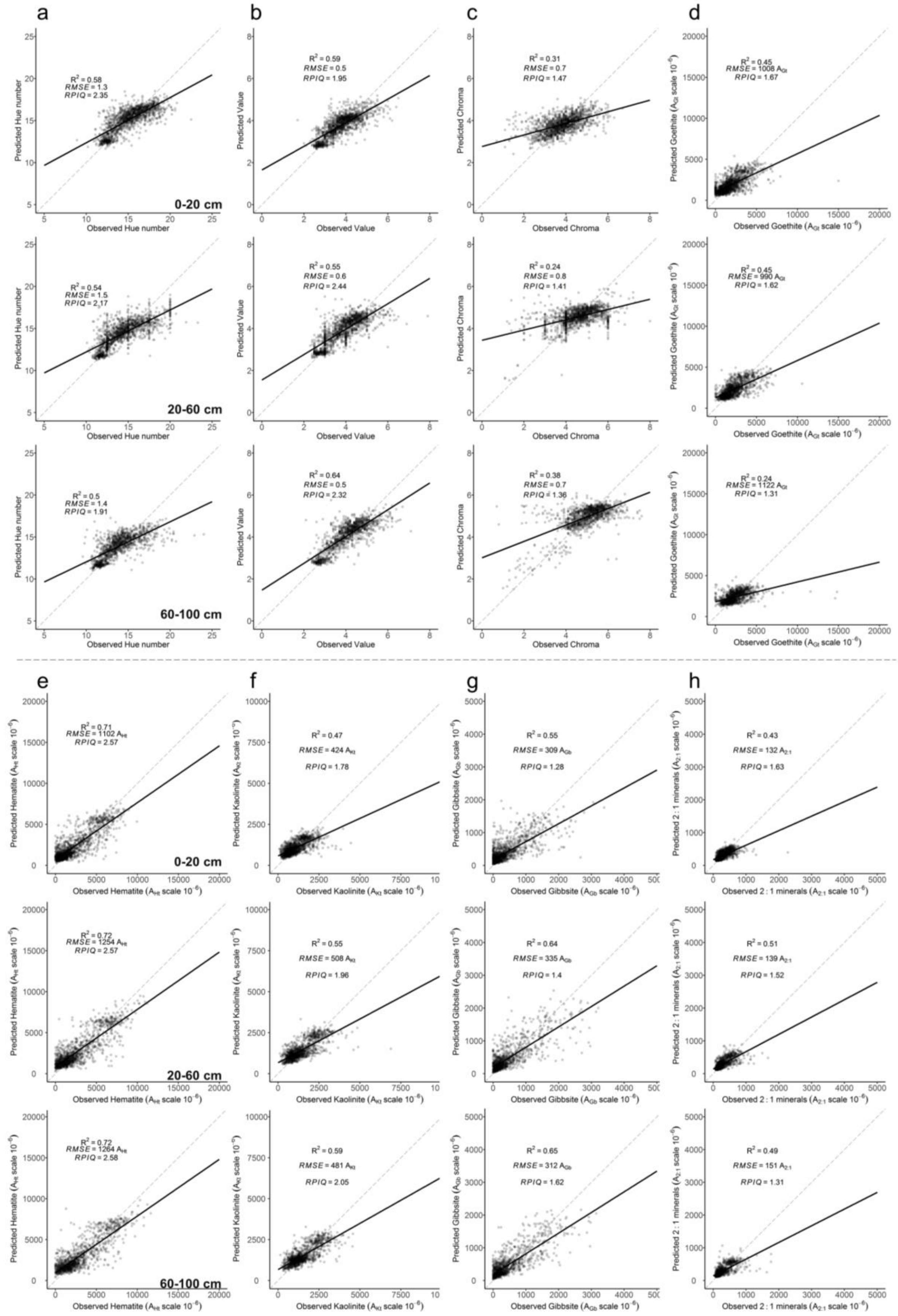


Figure A1. Predicted vs. observed (**a**) hue number, (**b**) value, (**c**) chroma, (**d**) goethite, (**e**) hematite, (**f**) kaolinite, (**g**) gibbsite and (**h**) 2:1 clay minerals by depth interval of 10-fold cross-validation.

7. REFERENCES

AGOSTINELLI, C. **CircStats: Circular Statistics, from "Topics in Circular Statistics"**. URL: https://cran.r-project.org/package=CircStats

AITKENHEAD, M. J.; COULL, M.; TOWERS, W.; HUDSON, G.; BLACK, H. I. J. Prediction of soil characteristics and colour using data from the National Soils Inventory of Scotland. **Geoderma**, v. 200–201, p. 99–107, 2013.

BARBOSA, I. O.; LACERDA, M. P. C.; BILICH, M. R. Pedomorphogeological relations in the chapadas elevadas of the Distrito Federal, Brazil. **Revista Brasileira de Ciência do Solo**, v. 33, n. 5, p. 1373–1383, 2009.

BARRERA-BASSOLS, N.; ALFRED ZINCK, J.; RANST, E. VAN. Symbolism, knowledge and management of soil and land resources in indigenous communities: Ethnopedology at global, regional and local scales. **CATENA**, v. 65, n. 2, p. 118–137, 2006.

BARRÓN, V.; TORRENT, J. Use of the Kubelka—Munk Theory to Study the Influence of Iron Oxides on Soil Colour. **Journal of Soil Science**, v. 37, p. 499–510, 1 Dec. 1986.

BELLON-MAUREL, V.; FERNANDEZ-AHUMADA, E.; PALAGOS, B.; ROGER, J.-M.; MCBRATNEY, A. Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy. **TrAC Trends in Analytical Chemistry**, v. 29, n. 9, p. 1073–1081, Oct. 2010.

BORCHERS, H. W. **pracma: Practical Numerical Math Functions**. URL: https://cran.r-project.org/package=pracma

BREIMAN, L. **Random forests**. v. 45, n. 1, p. 5–32, 2001.

CAMO SOFTWARE INC. **The Unscrambler version 9.7**. New Jersey, USA: CAMO Software AS, 2007.

CANADELL, J.; JACKSON, R. B.; EHLERINGER, J. B.; MOONEY, H. A.; SALA, O. E.; SCHULZE, E.-D. Maximum rooting depth of vegetation types at the global scale. **Oecologia**, v. 108, n. 4, p. 583–595, 1996.

CENTORE, P. **The Munsell and Kubelka-Munk Toolbox**. URL: http://centore.isletech.net/~centore/MunsellAndKubelkaMunkToolbox/MunsellAndKubelkaMunkToolbox.html

CLARK, R. N.; KING, T. V. V; KLEJWA, M.; SWAYZE, G. A.; VERGO, N. High spectral resolution reflectance spectroscopy of minerals. **Journal of Geophysical Research: Solid Earth**, v. 95, n. B8, p. 12653–12680, 1990.

CPRM - COMPANHIA DE PESQUISA DE RECURSOS MINERAIS. **Carta Geológica do Brasil ao Milionésimo: sistema de informações geográficas-SIG [Geological Map of Brazil 1:1.000.000 scale: geographic information system-GIS]**. Brasília: CPRM, 2004.

CURI, N.; FRANZMEIER, D. P. Toposequence of Oxisols from the Central Plateau of Brazil. **Soil Science Society of America Journal**, v. 48, n. 2, p. 341–346, 1984.

DAS, S. Comparison among influencing factor, frequency ratio, and analytical hierarchy process techniques for groundwater potential zonation in Vaitarna basin, Maharashtra, India. **Groundwater for Sustainable Development**, v. 8, p. 617–629, 2019.

DEMATTÊ, J. A. M.; DOTTO, A. C.; PAIVA, A. F. S.; SATO, M. V; DALMOLIN, R. S. D.; ARAÚJO, M. DO S. B.; SILVA, E. B.; NANNI, M. R.; CATEN, A. TEN; NORONHA, N. C.; LACERDA, M. P. C.; ARAÚJO FILHO, J. C.; RIZZO, R.; BELLINASO, H.; FRANCELINO, M. R.; SCHAEFER, C. E. G. R.; VICENTE, L. E.; SANTOS, U. J.; SÁ BARRETTO SAMPAIO, E. V DE; MENEZES, R. S. C.; 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.; MENEZES, M. D.; CURI, N.;

COUTO, E. G.; ANJOS, L. H. C.; CEDDIA, M. B.; PINHEIRO, É. F. M.; GRUNWALD, S.; VASQUES, G. M.; MARQUES JÚNIOR, J.; SILVA, A. J.; BARRETO, M. C. DE V.; NÓBREGA, G. N.; SILVA, M. Z.; SOUZA, S. F.; VALLADARES, G. S.; VIANA, J. H. M.; SILVA TERRA, F. DA; HORÁK-TERRA, I.; FIORIO, P. R.; SILVA, R. C.; FRADE JÚNIOR, E. F.; LIMA, R. H. C.; ALBA, J. M. F.; 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.; SOUSA MENDES, W.; SAFANELLI, J. L.; GUIMARÃES, C. C. B.; POPPIEL, R. R.; SOUZA, A. B.; QUESADA, C. A.; COUTO, H. T. Z. The Brazilian Soil Spectral Library (BSSL): A general view, application and challenges. **Geoderma**, v. 354, p. 113793, 2019.

DEMATTÊ, J. A. M.; FONGARO, C. T.; RIZZO, R.; SAFANELLI, J. L. Geospatial Soil Sensing System (GEOS3): A powerful data mining procedure to retrieve soil spectral reflectance from satellite images. **Remote Sensing of Environment**, v. 212, p. 161–175, Jun. 2018.

EMBRAPA - BRAZILIAN AGRICULTURAL RESEARCH CORPORATION - NATIONAL SOILS RESEARCH CENTER. **Brazilian Soil Classification System**. 5. ed. Brasilia, Brazil: Embrapa-Cnps, 2018. URL: https://www.embrapa.br/busca-de-publicacoes/-/publicacao/1094001/brazilian-soil-classification-system

ESCADAFAL, R.; GIRARD, M. C.; DOMINIQUE, C. Modeling the relationships between Munsell soil color and soil spectral properties. **International Agrophysics**, v. 4, n. 3, p. 249–261, 1 Jan. 1988.

FAO. **Soil Organic Carbon Mapping Cookbook**. 2. ed. Rome, Italy: FAO, 2018.

FERNANDEZ, R. N.; SCHULZE, D. G. Calculation of Soil Color from Reflectance Spectra. **Soil Science Society of America Journal**, v. 51, p. 1277–1282, 1987.

FERNANDEZ, R. N.; SCHULZE, D. G. Munsell Colors of Soils Simulated by Mixtures of Goethite and Hematite with Kaolinite. **Zeitschrift für Pflanzenernährung und Bodenkunde**, v. 155, n. 5, p. 473–478, 1992.

GOMES, J. B. V; CURI, N.; SCHULZE, D. G.; MARQUES, J. J. G. S. M.; KER, J. C.; MOTTA, P. E. F. Mineralogia, morfologia e análise microscópica de solos do bioma cerrado. **Revista Brasileira de Ciência do Solo**, v. 28, n. 4, p. 679–694, Aug. 2004.

GOMES, L. C.; FARIA, R. M.; SOUZA, E. DE; VELOSO, G. V.; SCHAEFER, C. E. G. R.; FILHO, E. I. F. Modelling and mapping soil organic carbon stocks in Brazil. **Geoderma**, v. 340, p. 337–350, 2019.

GORELICK, N.; HANCHER, M.; DIXON, M.; ILYUSHCHENKO, S.; THAU, D.; MOORE, R. Google Earth Engine: Planetary-scale geospatial analysis for everyone. **Remote Sensing of Environment**, v. 202, p. 18–27, 2017.

HAMILTON, N. **ggtern: An Extension to "ggplot2", for the Creation of Ternary Diagrams**. URL: https://cran.r-project.org/package=ggtern

HENGL, T.; HEUVELINK, G. B. M.; KEMPEN, B.; LEENAARS, J. G. B.; WALSH, M. G.; SHEPHERD, K. D.; SILA, A.; MACMILLAN, R. A.; MENDES DE JESUS, J.; TAMENE, L.; TONDOH, J. E. Mapping Soil Properties of Africa at 250 m Resolution: Random Forests Significantly Improve Current Predictions. **PLOS ONE**, v. 10, n. 6, p. e0125814, 25 Jun. 2015.

HENGL, T.; JESUS, J. M. DE; MACMILLAN, R. A.; BATJES, N. H.; HEUVELINK, G. B. M.; RIBEIRO, E.; SAMUEL-ROSA, A.; KEMPEN, B.; LEENAARS, J. G. B.; WALSH, M. G.; GONZALEZ, M. R. SoilGrids1km — Global Soil Information Based on Automated Mapping. **PLoS ONE**, v. 9, n. 8, p. e105992, 29 Aug. 2014.

HENGL, T.; MACMILLAN, R. A. **Predictive Soil Mapping with R**. Wageningen, the Netherlands: OpenGeoHub foundation, 2019.

HENGL, T.; MENDES DE JESUS, J.; HEUVELINK, G. B. M.; RUIPEREZ GONZALEZ, M.;

KILIBARDA, M.; BLAGOTIĆ, A.; SHANGGUAN, W.; WRIGHT, M. N.; GENG, X.; BAUER-MARSCHALLINGER, B.; GUEVARA, M. A.; VARGAS, R.; MACMILLAN, R. A.; BATJES, N. H.; LEENAARS, J. G. B.; RIBEIRO, E.; WHEELER, I.; MANTEL, S.; KEMPEN, B. SoilGrids250m: Global gridded soil information based on machine learning. **PLoS ONE**, v. 12, n. 2, p. e0169748, 16 Feb. 2017.

HENGL, T.; NUSSBAUM, M.; WRIGHT, M. N.; HEUVELINK, G. B. M.; GRÄLER, B. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. **PeerJ**, v. 6, p. e5518, 2018.

HIJMANS, R. J.; CAMERON, S. E.; PARRA, J. L.; JONES, P. G.; JARVIS, A. Very high resolution interpolated climate surfaces for global land areas. **International Journal of Climatology**, v. 25, n. 15, p. 1965–1978, 1 Dec. 2005.

HURST, V. J. Visual estimation of iron in saprolite. **GSA Bulletin**, v. 88, n. 2, p. 174–176, 1 Feb. 1977.

IBGE - INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Pedologia [Pedological maps of Brazil]**. URL: https://www.ibge.gov.br/geociencias/informacoes-ambientais/pedologia/10871-pedologia.html?=&t=downloads. Acesso em: 30 sep. 2019.

IUSS WORKING GROUP WRB. **World reference base for soil resources 2014: International soil classification system for naming soils and creating legends for soil maps**. Rome: Food and Agriculture Organization, 2015.

KOSMAS, C. S.; CURI, N.; BRYANT, R. B.; FRANZMEIER, D. P. Characterization of Iron Oxide Minerals by Second-Derivative Visible Spectroscopy. **Soil Science Society of America Journal**, v. 48, n. 2, p. 401, 1984.

KUHN, M. **caret: Classification and Regression Training**. URL: https://cran.r-project.org/web/packages/caret/index.html

LAGACHERIE, P. Digital soil mapping: A state of the art. *In*: HARTEMINK, A.; MCBRATNEY, A.; MENDONÇA-SANTOS, M. L. (Eds.). . **Digital Soil Mapping with Limited Data**. [s.l.] Springer Netherlands, 2008. p. 3–14.

LEENAARS, J. G. B.; ELIAS, E.; WÖSTEN, J. H. M.; RUIPEREZ-GONZÁLEZ, M.; KEMPEN, B. Mapping the major soil-landscape resources of the Ethiopian Highlands using random forest. **Geoderma**, p. 114067, 2019.

LILES, G. C.; BEAUDETTE, D. E.; O'GEEN, A. T.; HORWATH, W. R. Developing predictive soil C models for soils using quantitative color measurements. **Soil Science Society of America Journal**, v. 77, n. 6, p. 2173–2181, 1 Nov. 2013.

LIU, F.; GENG, X.; ZHU, A.-X.; FRASER, W.; WADDELL, A. Soil texture mapping over low relief areas using land surface feedback dynamic patterns extracted from MODIS. **Geoderma**, v. 171–172, p. 44–52, 2012.

LIU, F.; ZHANG, G.-L.; SONG, X.; LI, D.; ZHAO, Y.; YANG, J.; WU, H.; YANG, F. High-resolution and three-dimensional mapping of soil texture of China. **Geoderma**, v. 361, p. 114061, 2020.

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. Satellite data integration for soil clay content modelling at a national scale. **International Journal of Applied Earth Observation and Geoinformation**, v. 82, p. 101905, Oct. 2019.

MACEDO, J.; BRYANT, R. B. Morphology, Mineralogy, and Genesis of a Hydrosequence of Oxisols in Brazil. **Soil Science Society of America Journal**, v. 51, p. 690–698, 1987.

MADEIRA NETTO, J. S.; BEDIDI, A.; CERVELLE, B.; POUGET, M.; FLAY, N. 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**, v. 18, n. 13, p. 2835–2852, 1997.

MALONE, B. P.; HUGHES, P.; MCBRATNEY, A. B.; MINASNY, B. A model for the identification of terrons in the Lower Hunter Valley, Australia. **Geoderma Regional**, v. 1, p. 31–47, 2014.

MARQUES, K. P.; RIZZO, R.; DOTTO, A. C.; SOUZA, A. B.; MELLO, F. A.; NETO, L. G.; ANJOS, L. H. C.; DEMATTÊ, J. A. How qualitative spectral information can improve soil profile classification? **Journal of Near Infrared Spectroscopy**, p. 096703351882196, 3 Jan. 2019.

MATTIKALLI, N. M. Soil color modeling for the visible and near-infrared bands of Landsat sensors using laboratory spectral measurements. **Remote Sensing of Environment**, v. 59, n. 1, p. 14–28, 1997.

MAYNARD, J. J.; LEVI, M. R. Hyper-temporal remote sensing for digital soil mapping: Characterizing soil-vegetation response to climatic variability. **Geoderma**, v. 285, p. 94–109, 2017.

MCBRATNEY, A. B.; MENDONÇA SANTOS, M. L.; MINASNY, B. On digital soil mapping. **Geoderma**, v. 117, n. 1–2, p. 3–52, Nov. 2003.

MELO, V. F.; SINGH, B.; SCHAEFER, C. E. G. R.; NOVAIS, R. F.; FONTES, M. P. F. Chemical and Mineralogical Properties of Kaolinite-Rich Brazilian Soils. **Soil Science Society of America Journal**, v. 65, p. 1324–1333, 2001.

MENDES, W. DE S.; MEDEIROS NETO, L. G.; DEMATTÊ, J. A. M.; GALLO, B. C.; RIZZO, R.; SAFANELLI, J. L.; FONGARO, C. T. Is it possible to map subsurface soil attributes by satellite spectral transfer models? **Geoderma**, v. 343, p. 269–279, 2019.

MILLER, B. A.; KOSZINSKI, S.; WEHRHAN, M.; SOMMER, M. Impact of multi-scale predictor selection for modeling soil properties. **Geoderma**, v. 239–240, p. 97–106, 2015.

MORAES, J. M. **Geodiversidade do estado de Goiás e do Distrito Federal [Geodiversity of Goiás State and the Federal District, Brazil]**. Goiânia, GO: CPRM, 2014.

MULDER, V. L.; BRUIN, S.; WEYERMANN, J.; KOKALY, R. F.; SCHAEPMAN, M. E. Characterizing regional soil mineral composition using spectroscopy and geostatistics. **Remote Sensing of Environment**, v. 139, p. 415–429, 2013.

MUNSELL, A. H. **A Color Notation**. Boston: G. H. Ellis Company, 1907.

NAWAR, S.; CORSTANJE, R.; HALCRO, G.; MULLA, D.; MOUAZEN, A. M. Delineation of Soil Management Zones for Variable-Rate Fertilization: A Review. *In*: SPARKS, D. L. B. T.-A. IN A. (Ed.). . **Advances in Agronomy**. Academic Press, 2017. v. 143p. 175–245.

PADARIAN, J.; MINASNY, B.; MCBRATNEY, A. B. Machine learning and soil sciences: A review aided by machine learning tools. **SOIL Discuss.**, v. 2019, p. 1–29, 3 Sep. 2019.

POPPIEL, R. R.; LACERDA, M. P. C.; OLIVEIRA JR, M. P.; DEMATTÊ, J. A. M.; ROMERO, D. J.; SATO, M. V.; ALMEIDA JR, L. R.; CASSOL, L. F. M. Surface Spectroscopy of Oxisols, Entisols and Inceptisol and Relationships with Selected Soil Properties. **Revista Brasileira de Ciência do Solo**, v. 42, p. e0160519, 2018.

POPPIEL, RAUL R.; LACERDA, M. P. C.; DEMATTÊ, J. A. M.; OLIVEIRA, M. P.; GALLO, B. C.; SAFANELLI, J. Pedology and soil class mapping from proximal and remote sensed data. **Geoderma**, v. 348, p. 189–206, Aug. 2019a.

POPPIEL, R.R.; LACERDA, M. P. C.; SAFANELLI, J. L.; RIZZO, R.; OLIVEIRA, M. P.; NOVAIS, J. J.; DEMATTÊ, J. A. M. Mapping at 30 m Resolution of Soil Attributes at Multiple Depths in Midwest Brazil. **Remote Sensing**, v. 11, n. 24, p. 2905, 5 Dec. 2019b.

POST, D. F.; LUCAS, W. M.; WHITE, S. A.; EHASZ, M. J.; BATCHILY, A. K.; HORVATH, E. H. Relations between Soil Color and Landsat Reflectance on Semiarid Rangelands. **Soil Science**

Society of America Journal, v. 58, p. 1809–1816, 1994.

PROBST, P.; WRIGHT, M. N.; BOULESTEIX, A.-L. Hyperparameters and tuning strategies for random forest. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, v. 9, n. 3, p. e1301, 1 May 2019.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. Vienna, Austria: R Foundation for Statistical Computing, 2018. URL: https://www.r-project.org/

RAMCHARAN, A.; HENGL, T.; NAUMAN, T.; BRUNGARD, C.; WALTMAN, S.; WILLS, S.; THOMPSON, J. Soil Property and Class Maps of the Conterminous United States at 100-Meter Spatial Resolution. **Soil Science Society of America Journal**, v. 82, p. 186–201, 2018.

REATTO, A.; BRUAND, A.; SOUZA MARTINS, E. DE; MULLER, F.; SILVA, E. M. DA; CARVALHO, O. A. DE; BROSSARD, M. Variation of the kaolinite and gibbsite content at regional and local scale in Latosols of the Brazilian Central Plateau. **Comptes Rendus Geoscience**, v. 340, n. 11, p. 741–748, 2008.

RIZZO, R.; DEMATTÊ, J. A. M.; LEPSCH, I. F.; GALLO, B. C.; FONGARO, C. T. Digital soil mapping at local scale using a multi-depth Vis–NIR spectral library and terrain attributes. **Geoderma**, v. 274, p. 18–27, 2016.

ROBERTS, D.; WILFORD, J.; GHATTAS, O. Exposed soil and mineral map of the Australian continent revealing the land at its barest. **Nature Communications**, v. 10, n. 1, p. 5297, 2019.

RODRIGUES, T. E. **Mineralogy and genesis of a sequence of Cerrados soils in the Federal District**. Dissertation (Master in Agronomy) - Faculty of Agronomy, University of Rio Grande do Sul, Porto Alegre, 1977. URL: https://www.ufrgs.br/agronomia/materiais/19777dt.pdf

ROGGE, D.; BAUER, A.; ZEIDLER, J.; MUELLER, A.; ESCH, T.; HEIDEN, U. Building an exposed soil composite processor (SCMaP) for mapping spatial and temporal characteristics of soils with Landsat imagery (1984–2014). **Remote Sensing of Environment**, v. 205, p. 1–17, 2018.

SAMUEL-ROSA, A.; DALMOLIN, R. S. D.; MOURA-BUENO, J. M.; TEIXEIRA, W. G.; ALBA, J. M. F. Open legacy soil survey data in Brazil: geospatial data quality and how to improve it. **Scientia Agricola**, v. 77, n. 1, p. e20170430, 2020.

SCHAEFER, C. E. G. R.; FABRIS, J. D.; KER, J. C. Minerals in the clay fraction of Brazilian Latosols (Oxisols): a review. **Clay Minerals**, v. 43, n. 1, p. 137–154, 9 Mar. 2008.

SCHEINOST, A. C.; CHAVERNAS, A.; BARRÓN, V.; TORRENT, J. 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**, v. 46, n. 5, p. 528–536, 1998.

SCHULZE, D. G.; NAGEL, J. L.; VAN-SCOYOC, G. E.; HENDERSON, T. L.; BAUMGARDNER, M. F.; STOTT, D. E. Significance of Organic Matter in Determining Soil Colors. *In*: BIGHAM, J. M.; CIOLKOSZ, E. J. (Eds.). . **Soil Color**. 31. ed. Madison: SSSA, 1993. p. 71–90.

SCHWERTMANN, J. M. Relations Between Iron Oxides, Soil Color, and Soil Formation. *In*: CIOLKOSZ, E. J.; BIGHAM, U. (Eds.). . **Soil Color**. Madison: [s.n.]. p. 51–69. URL: https://dl.sciencesocieties.org/publications/books/abstracts/sssaspecialpubl/soilcolor/51

SCHWERTMANN, U.; TAYLOR, R. M. Iron Oxides. *In*: **Minerals in Soil Environments**. [s.l: s.n.]. p. 379–438.

SCORNET, E.; BIAU, G.; VERT, J.-P. Consistency of random forests. **The Annals of Statistics**, v. 43, n. 4, p. 1716–1741, 2015.

SILVA, B. P. C.; SILVA, M. L. N.; AVALOS, F. A. P.; MENEZES, M. D. DE; CURI, N. Digital soil mapping including additional point sampling in Posses ecosystem services pilot watershed, southeastern Brazil. **Scientific Reports**, v. 9, n. 1, p. 13763, 2019.

SILVA, L. S.; MARQUES JÚNIOR, J.; BARRÓN, V.; GOMES, R. P.; TEIXEIRA, D. D. B.;

SIQUEIRA, D. S.; VASCONCELOS, V. Spatial variability of iron oxides in soils from Brazilian sandstone and basalt. **CATENA**, v. 185, p. 104258, 2020.

SIMON, T.; ZHANG, Y.; HARTEMINK, A. E.; HUANG, J.; WALTER, C.; YOST, J. L. Predicting the color of sandy soils from Wisconsin, USA. **Geoderma**, p. 114039, 2019.

STEVENS, A.; RAMIREZ-LOPEZ, L. **prospectr: Processing and sample selection for vis-NIR spectral data**. URL: https://cran.r-project.org/package=prospectr

TADONO, T.; ISHIDA, H.; ODA, F.; NAITO, S.; MINAKAWA, K.; IWAMOTO, H. Precise global DEM generation by ALOS PRISM. **ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences**, v. II–4, p. 71–76, 31 Mar. 2014.

TERRA, F. S.; DEMATTÊ, J. A. M.; VISCARRA ROSSEL, R. A. Proximal spectral sensing in pedological assessments: vis–NIR spectra for soil classification based on weathering and pedogenesis. **Geoderma**, v. 318, p. 123–136, 15 May 2018.

TORRENT, J.; BARRÓN, V. Laboratory Measurement of Soil Color: Theory and Practice. *In*: **Soil Color**. SSSA Special Publication SV - 31. Madison, WI: Soil Science Society of America, 1993. p. 21–33.

TORRENT, J.; BARRÓN, V. Diffuse Reflectance Spectroscopy of Iron Oxides. **Encyclopedia of Surface and Colloid Science**, v. 1, 1 Jan. 2002.

VIEIRA, B. C.; SALGADO, A. A. R.; SANTOS, L. J. C. **Landscapes and Landforms of Brazil**. [s.l.] Springer, 2015.

VISCARRA ROSSEL, R. A. Fine-resolution multiscale mapping of clay minerals in Australian soils measured with near infrared spectra. **Journal of Geophysical Research: Earth Surface**, v. 116, n. F4, p. n/a-n/a, 2011.

VISCARRA ROSSEL, R. A.; BUI, E. N.; CARITAT, P. DE; MCKENZIE, N. J. Mapping iron oxides and the color of Australian soil using visible–near-infrared reflectance spectra. **Journal of Geophysical Research**, v. 115, n. F4, p. F04031, 15 Dec. 2010.

VISCARRA ROSSEL, R. A.; CATTLE, S. R.; ORTEGA, A.; FOUAD, Y. In situ measurements of soil colour, mineral composition and clay content by vis–NIR spectroscopy. **Geoderma**, v. 150, n. 3–4, p. 253–266, 15 May 2009.

VISCARRA ROSSEL, R. A.; CHEN, C. Digitally mapping the information content of visible–near infrared spectra of surficial Australian soils. **Remote Sensing of Environment**, v. 115, n. 6, p. 1443–1455, 15 Jun. 2011.

WADOUX, A. M. .-C.; BRUS, D. J.; HEUVELINK, G. B. M. Sampling design optimization for soil mapping with random forest. **Geoderma**, v. 355, p. 113913, 2019.

WRIGHT, M. N.; ZIEGLER, A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. **Journal of Statistical Software**, v. 1, n. 1, 31 Mar. 2017.

WYSZECKI, G.; STILES, W. S. **Color Science: Concepts and Methods, Quantitative Data and Formulae**. 2. ed. Wiley, New York: John Wiley & Sons, 1982.

ZHANG, Y.; HARTEMINK, A. E. Digital mapping of a soil profile. **European Journal of Soil Science**, v. 70, n. 1, p. 27–41, 2019.

ZINN, Y. L.; BIGHAM, J. M. Pedogenic and lithogenic gravels as indicators of soil polygenesis in the Brazilian Cerrado. **Soil Research**, v. 54, n. 4, p. 440–450, 2016.

ZINN, Y. L.; LAL, R.; BIGHAM, J. M.; RESCK, D. V. S. Edaphic Controls on Soil Organic Carbon Retention in the Brazilian Cerrado: Texture and Mineralogy. **Soil Science Society of America Journal**, v. 71, p. 1204–1214, 2007.

CHAPTER 3 — 250 M-GRIDDED SOIL TEXTURE AT MULTIPLE DEPTHS OF MIDWEST BRAZIL[3]

ABSTRACT

The dataset provide relevant gridded soil information of a large-extent area located in Midwest Brazil. This region is the largest and most recent agricultural frontier in Brazil that lacks of accurate and up-to-dated soil information to support current agricultural and environmental demands. Soil data was delivered as multiband GeoTIFF files at 250 m resolution, which comprise spatially continuous predictions of clay, silt and sand contents in g kg^{-1} at 0–20 cm, 20–60 cm and 60–100 cm (rooting) depth intervals. The spatial predictions were performed using soil observations at more than 7,000 locations and 33 Earth observation based covariates by Random Forest regression models into Google Earth Engine.

Keywords: soil texture; pedometric mapping; multi-depth; soil management; soil services.

---

[3] Data Article published in **Data Mendeley**, 28 January 2020, http://dx.doi.org/10.17632/52cfcm3xr7.4

1. TABLE OF SPECIFICATIONS

Subject	Earth and Soil Sciences
Specific subject area	Pedometry, Pedology, Soil Geography, Geology, Soil Management, Ecology
Related research article	Chapter 1 – Mapping at 30 m Resolution of Soil Attributes at Multiple Depths in Midwest Brazil
Type of data	Raster maps containing soil clay, silt and sand contents.
Data source location	Soil maps covered Midwest Brazil, from 12° S to 20° S and from 45° W to 54° W.
Data accessibility	Repository name: 250 m-gridded soil texture at multiple depths of Midwest Brazil DOI: 10.17632/52cfcm3xr7.4 Link: http://dx.doi.org/10.17632/52cfcm3xr7.4
Description of data collection	We aggregated soil clay, silt and sand data from 7,908 sites of the Brazilian Soil Spectral Library and 231 of the Free Brazilian Repository for Open Soil Data. We predicted values at 0-20 cm, 20-60 and 60-100 depth intervals using 33 soil covariates and Random Forest modelling-based optimization in Google Earth Engine. Native resolution of 30 m was reduced to 250 m by computing the mean values.
Data format	Multiband, integer GeoTIFF
Spatial resolution	250 meters
Unit	g kg^{-1}
Valid range	0–1000
Band:	name:
Band1:	claygkg_0_20cm
Band2:	siltgkg_0_20cm
Band3:	sandgkg_0_20cm
Band4:	claygkg_20_60cm
Band5:	siltgkg_20_60cm
Band6:	sandgkg_20_60cm
Band7:	claygkg_60_100cm
Band8:	siltgkg_60_100cm
Band9:	sandgkg_60_100cm

2. VALUE OF THE DATA

- The gridded soil data is important because, under current conditions, there are no soil texture maps with complete coverage of Midwest Brazil. Spatially continuous soil texture at topsoil and subsoil will subsidize public policies in rural and urban areas, at regional, state and municipal levels, among other applications.
- Data will bring numerous benefits to society as a whole, especially farmers and agricultural companies, to evaluate the locations most suitable for farming and delineation of soil management zones. Also supporting scientific community, governments and creditor banks. This data will provide an invaluable legacy for Brazil.
- Spatial soil data can be used as input in biological-chemical-physical modelling and in assessments of dynamic environmental processes.
- Coupled with other information, the maps can be used to improve decision making, to evaluation the price of land for purchase and sale, increase crop and livestock production and help to reduce investments risk and planning for environmental conservation.
- The gridded soil information can also guide future soil surveys for inventory programs.

3. DATA

The raster maps, available for download from the repository of Mendeley (http://dx.doi.org/10.17632/52cfcm3xr7.4), were divided into 12 tiles, each one with 2 x 3 degree in size and 0.1 overlapping degree, according to Figure 1. Each raster tile at 250 m resolution is delivered as integer GeoTIFF format. Each GeoTIFF file holds nine bands with specific information of soil clay, silt and sand contents in g kg^{-1} at 0-20 cm, 20-60 cm and 60-100 depth intervals, as illustrated in Figure 2.

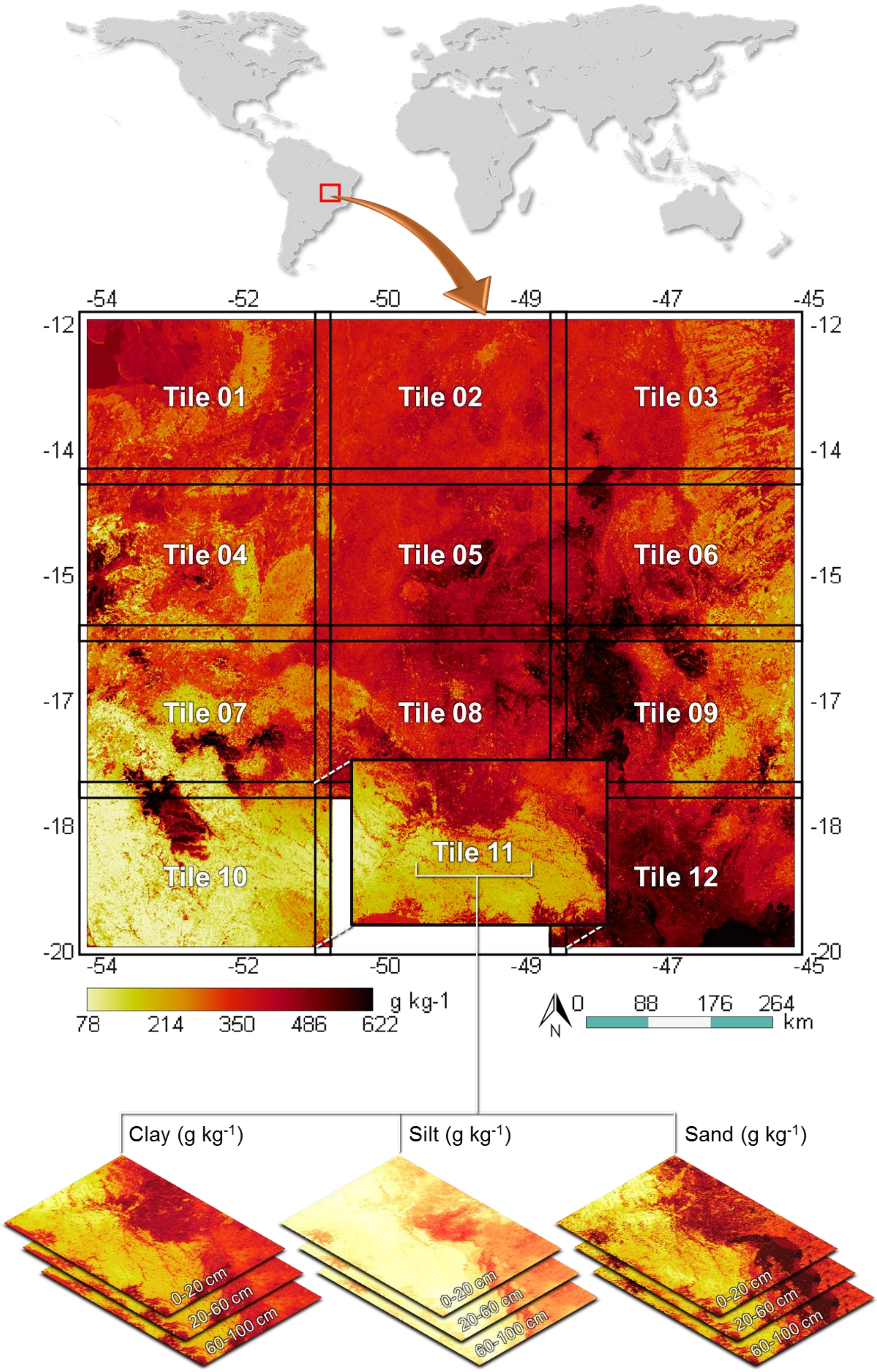
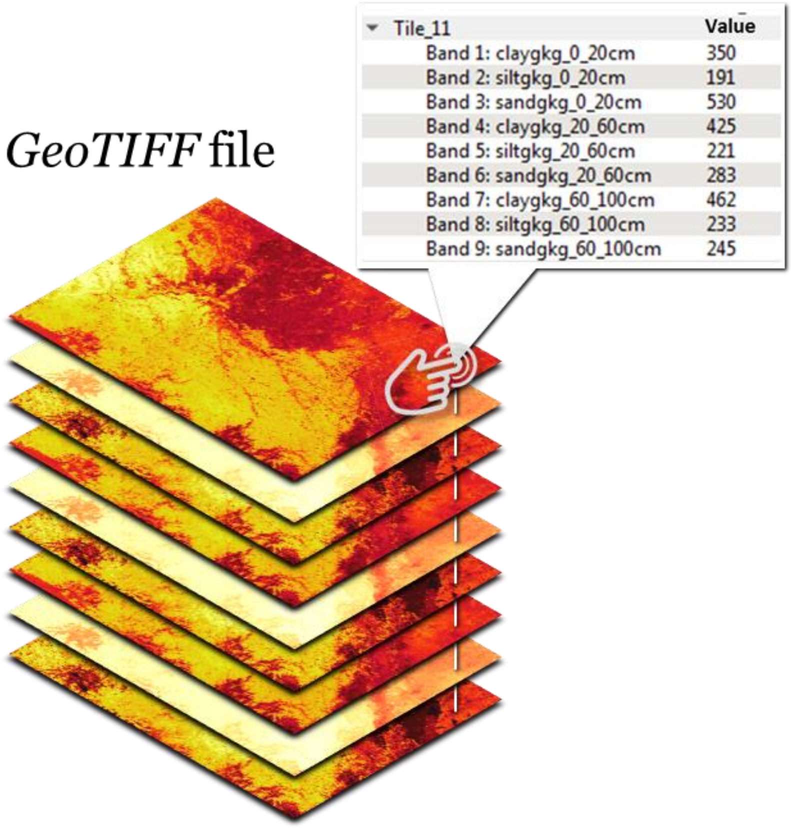


Fig. 1. Location of the soil dataset and reference grid with the identification of the tiles.



Tile_11	Value
Band 1: claygkg_0_20cm	350
Band 2: siltgkg_0_20cm	191
Band 3: sandgkg_0_20cm	530
Band 4: claygkg_20_60cm	425
Band 5: siltgkg_20_60cm	221
Band 6: sandgkg_20_60cm	283
Band 7: claygkg_60_100cm	462
Band 8: siltgkg_60_100cm	233
Band 9: sandgkg_60_100cm	245

Fig. 2. Example of a single raster tile showing the information contained in each pixel at 250 m resolution.

4. MATERIALS AND METHODS

Raster maps of soil texture were produced by Poppiel et al. (2020) using soil observations at more than 7,000 locations from the Brazilian Soil Spectral Library (DEMATTÊ et al., 2019) and the Free Brazilian Repository for Open Soil Data (SAMUEL-ROSA et al., 2020). Soil texture was predicted using 33 Earth observations based covariates and Random Forest (RF) regression (BREIMAN, 2001). Regression models were optimized by tuning RF hyperparameters (PROBST et al., 2019) in R software. Optimal models, with the lowest error in the 10-fold cross-validation step, were used to predict soil clay, silt and sand at three (rooting) depth intervals by RF within Google Earth Engine (GORELICK et al., 2017). The predicted maps had coefficient of determination of 10-fold cross-validation for clay, silt and sand ranging from 0.64 to 0.85. The native resolution (30 m) of the maps was reduced to 250 m by comtupting the mean values. We designed a reference grid to tile the large raster maps into small, manageable areas (Tiles) that were stored in the repository of Mendeley provided in this data article.

5. REFERENCES

BREIMAN, L. **Random forests**. v. 45, n. 1, p. 5–32, 2001.

DEMATTÊ, J. A. M.; DOTTO, A. C.; PAIVA, A. F. S.; SATO, M. V; DALMOLIN, R. S. D.; ARAÚJO, M. DO S. B.; SILVA, E. B.; NANNI, M. R.; CATEN, A. TEN; NORONHA, N. C.; LACERDA, M. P. C.; ARAÚJO FILHO, J. C.; RIZZO, R.; BELLINASO, H.; FRANCELINO, M. R.; SCHAEFER, C. E. G. R.; VICENTE, L. E.; SANTOS, U. J.; SÁ BARRETTO SAMPAIO, E. V DE; MENEZES, R. S. C.; 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.; MENEZES, M. D.; CURI, N.; COUTO, E. G.; ANJOS, L. H. C.; CEDDIA, M. B.; PINHEIRO, É. F. M.; GRUNWALD, S.; VASQUES, G. M.; MARQUES JÚNIOR, J.; SILVA, A. J.; BARRETO, M. C. DE V.; NÓBREGA, G. N.; SILVA, M. Z.; SOUZA, S. F.; VALLADARES, G. S.; VIANA, J. H. M.; SILVA TERRA, F. DA; HORÁK-TERRA, I.; FIORIO, P. R.; SILVA, R. C.; FRADE JÚNIOR, E. F.; LIMA, R. H. C.; ALBA, J. M. F.; 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.; SOUSA MENDES, W.; SAFANELLI, J. L.; GUIMARÃES, C. C. B.; POPPIEL, R. R.; SOUZA, A. B.; QUESADA, C. A.; COUTO, H. T. Z. The Brazilian Soil Spectral Library (BSSL): A general view, application and challenges. **Geoderma**, v. 354, p. 113793, 2019.

GORELICK, N.; HANCHER, M.; DIXON, M.; ILYUSHCHENKO, S.; THAU, D.; MOORE, R. Google Earth Engine: Planetary-scale geospatial analysis for everyone. **Remote Sensing of Environment**, v. 202, p. 18–27, 2017.

POPPIEL, R.R.; LACERDA, M. P. C.; SAFANELLI, J. L.; RIZZO, R.; OLIVEIRA, M. P.; NOVAIS, J. J.; DEMATTÊ, J. A. M. Mapping at 30 m Resolution of Soil Attributes at Multiple Depths in Midwest Brazil. **Remote Sensing**, v. 11, n. 24, p. 2905, 5 Dec. 2019.

PROBST, P.; WRIGHT, M. N.; BOULESTEIX, A.-L. Hyperparameters and tuning strategies for random forest. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, v. 9, n. 3, p. e1301, 1 May 2019.

SAMUEL-ROSA, A.; DALMOLIN, R. S. D.; MOURA-BUENO, J. M.; TEIXEIRA, W. G.; ALBA, J. M. F. Open legacy soil survey data in Brazil: geospatial data quality and how to improve it. **Scientia Agricola**, v. 77, n. 1, p. e20170430, 2020.

GENERAL CONCLUDING REMARKS AND FUTURE OUTLOOK

Physical and chemical soil attributes derived from traditional analysis, and soil color and mineralogy derived from laboratorial reflectance spectra (350–2500 nm), collected at three depth intervals (0−20, 20−60 and 60−100 cm), can be mapped at 30 m resolution using pedometric techniques and multi-resolution covariates with consistent spatial predictions in Midwest Brazil. Texture maps from this research publicly available will be a valuable information for future studies.

The best model performances at topsoil and subsoil were obtained for sand, clay, hematite, Munsell value and hue, and the worst were obtained for silt, cation exchange capacity and Munsell chroma. The most relevant covariates for predicting soil attributes were elevation, bare topsoil reflectance, climate and vegetation reflectance.

Under current demands, it is very important to look for new approaches and systematizations for the mapping of Brazilian soils, considering that the last innovative proposal was performed several decades ago with the RADAMBRASIL (Radar on Amazon and Brazil) project using radar images. This thesis shows that the integration of covariates based on remote sensing data with soil observations by mean of machine learning algorithms is a robust framework for DSM. This innovative framework will contribute greatly to achieve a better level of knowledge at a national scale of important soil attributes, that are essential (key) for soil classification, management and conservation.

Future studies should be performed using recent multispectral and radar sensors, like those onboard the Sentinel satellites, or hyperspectral instruments like Hyperion, that provides detailed spectral absorption features (242 spectral bands) of Earth surface with 30 m resolution. Hyperspectral sensors probably are the future of remote sensing. New covariates for soil predictions may be produced by mining data from a single sensor or from the integration of multiple sensors (at multiple resolutions). Special attention should be paid to the thermal infrared spectral bands.

For DSM purposes, soil reflectance spectra (350–2500 nm) need to be evaluated for further information about suitable spectral absorption bands for practical determination of soil minerals, e.g. by assessing different spectral bands at different Al-substitution percentages for mineralogical determination. The medium infrared spectral range should also be considered for soil evaluation, since this spectral range can provide information about soil geneses and weathering degree, among other valuable pedological information.