ERLI PINTO DOS SANTOS

ORGANIC CARBON MODELING IN TROPICAL SOILS OF BRAZIL THROUGH PROXIMAL AND REMOTE SENSING

Thesis submitted to the Agricultural Engineering Graduate Program of the Universidade Federal de Viçosa in partial fulfillment of the requirements for the degree of *Doctor Scientiae*.

Adviser: Michel Castro Moreira

Co-advisers: Elpídio Inácio Fernandes Filho
Demetrius David da Silva
José Alexandre Melo Demattê

VIÇOSA - MINAS GERAIS
2024

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T

S237o 2024

Santos, Erli Pinto dos, 1995-

Organic carbon modeling in tropical soils of Brazil through proximal and remote sensing / Erli Pinto dos Santos. – Viçosa, MG, 2024.

1 tese eletrônica (159 f.): il. (algumas color.).

Texto em inglês.

Inclui apêndices.

Orientador: Michel Castro Moreira.

Tese (doutorado) - Universidade Federal de Viçosa, Departamento de Engenharia Agrícola, 2024.

Inclui bibliografia.

DOI: https://doi.org/10.47328/ufvbbt.2024.474

Modo de acesso: World Wide Web.

1. Humus. 2. Aprendizado do computador. 3. Sensoriamento remoto. 4. Quimiometria. 5. Mudanças climáticas. I. Moreira, Michel Castro, 1980-. II. Universidade Federal de Viçosa. Departamento de Engenharia Agrícola. Programa de Pós-Graduação em Engenharia Agrícola. III. Título.

CDD 22. ed. 631.417

Bibliotecário(a) responsável: Alice Regina Pinto Pires CRB-6/2523

ERLI PINTO DOS SANTOS

ORGANIC CARBON MODELING IN TROPICAL SOILS OF BRAZIL THROUGH PROXIMAL AND REMOTE SENSING

Thesis submitted to the Agricultural Engineering Graduate Program of the Universidade Federal de Viçosa in partial fulfillment of the requirements for the degree of *Doctor Scientiae*.

APPROVED: July 31, 2024.

Assent:

Erli Pinto dos Santos
Author

Michel Castro Moreira
Adviser

To my parents, Roberval and Marli.

ACKNOWLEDGMENTS

I believe this is the part of the thesis where we discuss the journey leading up to the defense and express our gratitude to those who have supported us along the way.

First and foremost, I would like to thank myself. Without my dedication and perseverance, none of this would have been possible. Embarking on this journey has been a profound personal endeavor, and only I can fully appreciate the sweat and effort poured into it. Each challenge faced and overcome is a testament to my commitment. I do not congratulate myself only for obtaining a Doctorate degree, but rather for enduring and thriving through this demanding stage of my life.

Many people supported and helped me during this process. To avoid being unfair to anyone, besides my advisory board, I'm going to focus on six very important individuals.

First, my parents, Roberval Lopes dos Santos and Marli Pinto dos Santos, to whom I dedicate this work. I also acknowledge my brother, Ieslli. Next, my life partner, Renata, and our little one, Katie (AKA Katiezinha in sweet moments, or "Piranha" when she was being mischievous).

My family has been at the forefront of my life. Renata emerged during this challenging process, and we both know how much she has helped me professionally, personally, and in many other aspects of life. Paraphrasing Carl Sagan, it is a pleasure to share space and time with Renata. I now understand the true meaning of being a companion. Little K gave me the title of "pai de pet", even me denying loving her (a failed try to compensate for her spoiling habits). She stayed there to offer me her friend paw (and her butt to me clean it up).

Finally, I would like to thank my therapist, Fagner, who helped me gain the insights expressed in this acknowledgment section.

I would like to express my gratitude to my advisory committee for sharing their knowledge, support, and life teachings. Thank you to Professor Michel Castro Moreira, my adviser, and my co-advisers Professors Elpídio Inácio Fernandes Filho, Demetrius David da Silva, and José Alexandre Melo Demattê. In their name, I also thank others who partnered with me during this stage, endeavoring to push the boundaries of human knowledge.

I thank the Universidade Federal de Viçosa and its staff. This institution, with one of the most beautiful university campuses in Brazil, is ahead of its time in scientific development and human capacity building.

I would like to thank the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001 for the scholarship.

This study was also financed by the Conselho Nacional de Desenvolvimento Científico e Tecnológico – Brasil (CNPq).

I also would like to thank the Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) for financially supporting this study through the grant APQ-01562-23.

Thank you to everyone who helped me reach this point!

Thank God and Nature for allowing me to live this day.

BIOGRAPHY

Erli Pinto dos Santos was born in the city of Itiúba, in the interior of the state of Bahia, on July 28, 1995. He is son of Roberval Lopes dos Santos and Marli Pinto dos Santos, and brother of Ieslli.

In February 2010 he joined the Instituto Federal de Educação, Ciência e Tecnologia Baiano, at the campus of Senhor do Bonfim – Bahia. There he attended high school along with the technical course in agriculture.

In January 2013 he joined the Agronomy course at the Universidade Estadual de Feira de Santana, in Feira de Santana – Bahia. In July 2018 he completed his Bachelor's degree, obtaining the title of Agronomist.

In August 2018 he started his master's degree at the Universidade Federal de Viçosa, in the graduate Program in Agricultural Engineering. In July 2020 he submitted to the defense of his dissertation, achieving the title of *Magister Scientiae*.

In the same year, he began his PhD in the same Graduate Program. Undergoing the thesis defense in July 2024.

“Imagination is more important than knowledge. Knowledge is limited. Imagination encircles the world.”
(Albert Einstein)

“É impossível existir sem sonho.”
(Paulo Freire)

"Eu não sou um realista. Sou um realista esperançoso, porque acho que a esperança é o sonho do homem acordado."
(Ariano Suassuna)

RESUMO

SANTOS, Erli Pinto dos, D.Sc., Universidade Federal de Viçosa, julho de 2024. **Modelagem do carbono orgânico de solos tropicais por sensoriamento próximo e remoto**. Orientador: Michel Castro Moreira. Coorientadores: Elpídio Inácio Fernandes Filho, Demetrius David da Silva e José Alexandre Melo Demattê.

O carbono orgânico do solo (SOC) é essencial tanto para a segurança alimentar quanto para a regulação do clima, desempenhando um papel crucial no ciclo biogeoquímico do carbono e na manutenção da fertilidade do solo e produção agropecuária. Por isso, nesta tese buscou-se aprimorar as estimativas e compreensão de teores de SOC utilizando métodos de sensoriamento próximo e remoto. A hipótese central é que, em solos tropicais, a acurácia da predição de SOC utilizando bibliotecas espectrais de solo (SSL) nas faixas do visível (Vis: 350 – 700 nm), infravermelho próximo (NIR: 700 – 1000 nm) e infravermelho de ondas curtas (SWIR: 1000 – 2500 nm) pode ser aumentada com a inclusão de variáveis ambientais e índices de vegetação de radar na modelagem, focando na transparência e generalização dos modelos de *machine learning* (ML). Assim, o estudo foi estruturado em três artigos científicos. No primeiro artigo, foi investigada a melhoria do erro de generalização e da transparência de modelos de regressão usando reflectância do solo como covariáveis. Testou-se se o método LASSO (*least absolute shrinkage and selection operator*) poderia produzir modelos mais transparentes que o PLS (*partial least squares*) e a importância de considerar a estrutura espacial vertical do SOC em perfis de solo para evitar *overfitting*. Para isso, SSLs contendo 701 amostras de camadas de solos (pertencentes a 127 perfis) dos biomas brasileiros Mata Atlântica e Caatinga, nas faixas Vis-NIR-SWIR e infravermelho médio, foram utilizadas. Modelos ajustados com validação cruzada orientada aos perfis do solo mostraram-se mais acurados nos testes, sugerindo o uso dessa metodologia para evitar *overfitting*. Além disso, os modelos LASSO foram tão acurados quanto os PLS, mas mais transparentes, possibilitando associar diretamente as bandas espectrais selecionadas com bandas de transições eletrônicas e ligações vibracionais causadas por compostos orgânicos. No segundo artigo foi explorada a aplicação de imagens de radar do satélite Sentinel-1 e seus índices de vegetação para predizer SOC. O objetivo foi testar se os índices de vegetação de radar, como proxies para biomassa da vegetação, poderiam ser usados na predição de SOC em

camadas do solo. Para isso, foram utilizados 123 perfis de solos do bioma Cerrado amostrados em 7 camadas (0 – 5, 5 – 10, 10 – 15, 15 – 20, 20 – 40, 40 – 60 e 60 – 100 cm). Os índices de vegetação por radar foram eficientes em capturar a variabilidade espacial do SOC nas camadas superficiais do solo (até 10 cm de profundidade), mas não foram eficazes ao predizer o SOC em camadas mais profundas (de 10 a 100 cm). No terceiro artigo, assinaturas de reflectância do solo no Vis-NIR-SWIR foram integradas aos índices de vegetação de radar e descritores ambientais para predizer o SOC. Essa abordagem combinou diversas fontes de dados (SSL, classes de uso e cobertura da terra, tipos de solo, clima, elevação e índices de vegetação de radar) e técnicas de modelagem, melhorando significativamente a acurácia e confiabilidade das estimativas de SOC. Em solos da Caatinga e Mata Atlântica, a estimativa de SOC apenas com SSL resultou em raiz do erro quadrático médio (RMSE) de 4,52 g kg^{-1} e coeficiente de determinação (R^2) de 0,62, enquanto a inclusão de todas as covariáveis reduziu o RMSE para 3,93 g kg^{-1} e aumentou o R^2 para 0,72. Na tese foi possível demonstrar a eficácia da combinação de sensoriamento remoto e próximo com métodos avançados de ML para melhorar as estimativas de SOC, contribuindo para a ciência do solo e monitoramento de estratégias de manejo do solo e de sequestro de carbono.

Palavras-chave: matéria orgânica do solo; saúde do solo; reflectância difusa; radar de abertura sintética; quimiometria; aprendizado de máquina.

ABSTRACT

SANTOS, Erli Pinto dos, D.Sc., Universidade Federal de Viçosa, July, 2024. **Organic carbon modeling in tropical soils through proximal and remote sensing**. Adviser: Michel Castro Moreira. Co-advisers: Elpídio Inácio Fernandes Filho, Demetrius David da Silva, and José Alexandre Melo Demattê.

Soil organic carbon (SOC) is vital for both food security and climate regulation, playing a crucial role in the carbon biogeochemical cycle and maintaining soil fertility and agricultural production. Therefore, this thesis aims to improve the estimates and understanding of SOC levels using proximal and remote sensing methods. The central hypothesis posits that in tropical soils, the accuracy of SOC prediction using soil spectral libraries (SSL) in the visible (Vis: 350 – 700 nm), near-infrared (NIR: 700 – 1000 nm), and short-wave infrared (SWIR: 1000 – 2500 nm) ranges can be enhanced by incorporating environmental variables and radar vegetation indices into the modeling, focusing on the transparency and generalization of machine learning (ML) models. The study is structured into three scientific articles. The first article investigates improving the generalization error and transparency of regression models using soil reflectance as covariables. It tests whether the LASSO (least absolute shrinkage and selection operator) method can produce more transparent models than PLS (partial least squares) and highlights the importance of considering the vertical spatial structure of SOC in soil profiles to avoid overfitting. For this purpose, SSLs containing 701 soil layer samples (from 127 profiles) from the Brazilian biomes Atlantic Forest and Caatinga, in the Vis-NIR-SWIR and mid-infrared ranges, are used. Models adjusted with profile-oriented cross-validation show higher accuracy in tests, suggesting this methodology to avoid overfitting. Additionally, LASSO models are as accurate as PLS but more transparent, allowing direct association of the selected spectral bands with bands of electronic transitions and vibrational bonds caused by organic compounds. The second article explores the application of radar images from the Sentinel-1 satellite and their vegetation indices to predict SOC. The goal is to test whether radar vegetation indices, as proxies for vegetation biomass, can predict SOC in soil layers. For this purpose, 123 soil profiles from the Cerrado and Caatinga biomes sampled in 7 soil layers (0 – 5, 5 – 10, 10 – 15, 15 – 20, 20 – 40, 40 – 60, and 60 – 100 cm) are used. Radar vegetation indices efficiently capture the spatial variability of SOC in the topsoil layers (up to 10 cm depth) but are

ineffective in predicting SOC in deeper layers (10 to 100 cm). The third article integrates soil reflectance signatures in the Vis-NIR-SWIR with radar vegetation indices and environmental descriptors to predict SOC. This approach combines various data sources (SSL, land use and cover classes, soil types, climate, elevation, and radar vegetation indices) and modeling techniques, significantly improving the accuracy and reliability of SOC estimates. In soils of the Caatinga and Atlantic Forest, SOC estimation using only SSL results in a root mean square error (RMSE) of 4.52 g kg^{-1} and a coefficient of determination (R^2) of 0.62, while the inclusion of all covariables reduces the RMSE to 3.93 g kg^{-1} and increases the R^2 to 0.72. This thesis demonstrates the effectiveness of combining remote and proximal sensing with advanced ML methods to improve SOC estimates, contributing to soil science and the monitoring of soil management and carbon sequestration strategies.

Keywords: soil organic matter; soil healthy; diffuse reflectance; synthetic aperture radar; chemometrics; machine learning.

LIST OF FIGURES

Article 1 – Improving the generalization error and transparency of regression models to estimate soil organic carbon using soil reflectance data

Article 2 – Sentinel-1 imagery used for estimation of soil organic carbon by dual-polarization SAR vegetation indices

Article 3 – Integrating satellite radar vegetation indices and environmental descriptors with visible-infrared soil spectroscopy improved organic carbon prediction in soils of semi-arid Brazil

LIST OF TABLES

Article 1 – Improving the generalization error and transparency of regression models to estimate soil organic carbon using soil reflectance data

Article 2 – Sentinel-1 imagery used for estimation of soil organic carbon by dual-polarization SAR vegetation indices

Article 3 – Integrating satellite radar vegetation indices and environmental descriptors with visible-infrared soil spectroscopy improved organic carbon prediction in soils of semi-arid Brazil

SUMMARY

1. GENERAL INTRODUCTION

Soil organic carbon (SOC) is a paramount soil property both for food security and climate. This is because the pedosphere is an integrative component of the biogeochemical carbon (C) cycle among four other spheres: the biosphere, the lithosphere, the hydrosphere, and the atmosphere. Additionally, SOC is a key factor for the physical, chemical, and biological quality of the soil, thus influencing agricultural productivity (BALDOCK; BROOS, 2012; JACKSON et al., 2017; SILVA; MENDONÇA, 2007).

Over the past decades, the C cycle has increasingly gained the attention of scientists and society at large (CANADELL et al., 2021). The reason for this heightened interest is the impact of climate change on the quality of ecosystems. This includes alterations of patterns of atmospheric circulation, temperature, the hydrological cycle, and an increased frequency of extreme weather events, all of which threaten biodiversity, soil quality and health, and food security. Ultimately, these changes affect human life quality (CANADELL et al., 2021; LAL, 2004a, 2004b, 2018; MANABE, 2019).

The C in the terrestrial system is distributed among five major reservoirs: the oceans, the lithosphere, the pedosphere, the atmosphere, and the biosphere. After the two largest reservoirs – the oceans and geological components (including fossil fuels) – soil is the largest terrestrial reservoir of C (CANADELL et al., 2021). Current estimates report that the pedosphere, including permafrost, holds approximately 2,900 Pg of C (CANADELL et al., 2021). This amount is 3.3 times greater than the C stock in the atmosphere (estimated at 870 Pg-C) and 6.4 times greater than the C stock in the biosphere (vegetation), estimated at 450 Pg-C (CANADELL et al., 2021). C in each reservoir is of significant importance. However, the increase in atmospheric C content, in the form of greenhouse gases (GHG) – notably CO_2, which is the primary C greenhouse gas after CH_4 and N_2O – is the cause of climate change (MANABE, 2019).

Soil is the primary terrestrial reservoir of C, where it is present in both organic (SOC) and inorganic forms. Although there are stocks in inorganic forms (predominantly $CaMg(CO_3)$ and $CaCO_3$), which are very important for soils in arid and semi-arid regions (LAL, 2004a), the largest stock of soil C is in the form of soil organic matter (SOM). Recent studies, based on digital soil mapping, estimate the global stock of SOC down to a depth of 100 cm to be between 1,408 Pg-C (BATJES, 2016) and 1,486 Pg-C (JACKSON et al., 2017).

The SOC fraction is the largest elemental component of SOM, comprising approximately 58% of it. SOM, in turn, originates from the input of aerial and root biomass residues of plants, the release of root exudates, the leaching of soluble plant constituents by rainwater, and the transformation of these carbonaceous materials by soil macro- and microorganisms (COTRUFO et al., 2015; JACKSON et al., 2017).

SOC levels depend on the balance between inputs and the decomposition/mineralization of SOM, combined with stabilization processes. The primary process of organic material input into the soil is through photosynthesis products from autotrophic organisms (JANZEN, 2006; LAL, 2018; SILVA; MENDONÇA, 2007). This input of organic residues, via plant tissues, can be measured in terms of net primary production (NPP). NPP is the net production of organic material by plants, which serves as the primary energy source for other heterotrophic organisms until it becomes part of SOM (SILVA; MENDONÇA, 2007).

The processes of decomposition and mineralization lead to the consumption of SOM and the depletion of SOC, in addition to the release of other products (BALDOCK; BROOS, 2012). The main end products are: CO_2 and CH_4, which are GHGs produced by the action of aerobic and anaerobic heterotrophic organisms, respectively; NH_3, via ammonification, where NH_4^+ is formed in the soil and utilized by plants; NO_3^-, produced by nitrification and used by plants; N_2O, also a GHG, produced during nitrification; SO_4^{2+}, which is utilized by plants and microorganisms; and H_2S, which is used as a substrate by microorganisms.

NPP rates vary among ecosystems (GUO; GIFFORD, 2002; LAL, 2004b) and are influenced by the availability of macro- and micronutrients, as well as water for plants (LAL, 2018; WIESMEIER et al., 2019). In areas managed for agricultural production, the type of management adopted also influences NPP rates (BOLINDER et al., 2007) and affects other soil properties associated with the stabilization of SOM compartments.

The stabilization of SOM occurs through colloidal and physical protection, as well as biochemical stabilization (BALDOCK; BROOS, 2012; BALDOCK; SKJEMSTAD, 2000; SILVA; MENDONÇA, 2007; WIESMEIER et al., 2019). Colloidal protection occurs through the interaction of SOM with soil mineral particles, forming clay-organic complexes. Physical protection occurs because soil aggregates shield non-selective organic compounds from microbial decomposition. Finally, biochemical stabilization occurs due to the structural resistance of some SOM compounds to enzymatic attack.

SOM is important not only for the release of essential nutrients to plants and its participation in the nitrogen, potassium, and sulfur cycles, but also for its effects on other soil properties. Among these, the direct positive effect of SOM on cation exchange capacity and water storage capacity can be highlighted (BALDOCK; SKJEMSTAD, 2000; JACKSON et al., 2017; SILVA; MENDONÇA, 2007). Therefore, SOM is a key component for the chemical, physical, and biological quality of soils. Consequently, SOC levels and stocks reflect soil quality and health.

The importance of SOC for soil quality, agricultural production, and ecosystem health has led to the establishment of strategies aimed at atmospheric C sequestration into soils (JANZEN, 2006; LAL, 2004b). These strategies primarily target agriculture and forest areas, where it is estimated that SOC depletion has been more pronounced due to changes in land use and land cover (DIONIZIO et al., 2020; GUO; GIFFORD, 2002) and inadequate soil management (JANZEN, 2006; LAL, 2004a, 2018). In agricultural areas, C sequestration is important both for climate mitigation and for the sustainability of agricultural production (JANZEN, 2006; SILVA; MENDONÇA, 2007). Therefore, atmospheric C sequestration is considered a win-to-win process, where when adopting conservative soil management practices, we have gains in both removing C from atmosphere and storing into SOM, and gains in soil fertility after decomposition and mineralization of SOM (JANZEN, 2006).

The 21st Conference of the Parties of the United Nations Framework Convention on Climate Change (COP21), held in Paris, France, provides a notable example of discussions on this topic. During COP21, the Paris Agreement was established, setting an international agenda for C sequestration known as "4 per – Soils for Food Security and Climate". This initiative aims to offset anthropogenic CO_2 emissions by increasing SOC by 4‰ (4/1000 or 0.4%) per year from sequestered atmospheric CO_2, given that fossil C emissions into the atmosphere were estimated at 8.9 Pg-C, while the global SOC stock in soils down to 2 meters depth was estimated at 2400 Pg-C (LAL, 2016; MINASNY et al., 2017). Minasny et al. (2017) studied the feasibility of this agenda and concluded that even sequestration rates greater than 0.4% per year can be achieved depending on efforts, technologies, and soil management practices.

A Brazilian national example on the way of soil conservation management practices is the Low Carbon Agriculture Plan (ABC Plan). The ABC Plan is a public policy consisting of actions aimed at increasing the use of sustainable agriculture

technologies with high potential in favoring C storage both in the biota on the agroecosystems and in SOM (MAPA; MDA, 2012). It was presented in 2009 as an international commitment for reducing GHG emissions, particularly in agriculture. The plan, officially named “Mitigation and Adaptation to Climate Change for the Consolidation of a Low-Carbon Economy in Agriculture”, was initially planned to incentivize the adoption of conservative agricultural management practices for the 2010 to 2020 but was later extended to the period of 2020-2030 (BRASIL, 2021). The program focuses on contributing to the sustainable development of agricultural production systems, particularly in recovering degraded pastures and integration crop-livestock-forest systems, as well as other soil conservation management practices.

Among soil conservation management practices are no-till farming and agroforestry systems. The adoption of these practices, aimed at increasing biomass input and maintaining soil cover with living vegetation, as well as reducing soil tillage, has the medium and long-term effect of increasing SOC stocks, depending on initial conditions (J. SILVA et al., 2024; OLIVEIRA et al., 2023, 2024). Such regenerative agricultural practices, focused on environmental and production gains, rely on continuous soil monitoring (ANGELOPOULOU et al., 2020; O’DONOGHUE; MINASNY; MCBRATNEY, 2024).

Soil monitoring is one of the main challenges associated with managing its health and effective capacity for C sequestration and/or storage (PAUSTIAN et al., 2019; SMITH et al., 2020). However, SOC is not an easy property to measure, as traditional laboratory methods (via "wet chemistry" or dry combustion) are often expensive, labor-intensive, and time-consuming (FAO, 2020). With the increasing demand for soil analysis, especially considering the site-specific management required when adopting precision agriculture practices, there is also a growing demand for chemical reagents (such as potassium dichromate: $K_2Cr_2O_7$; sulfuric acid: H_2SO_4; ferrous sulfate: $FeSO_4$; among others) (DEMATTÊ et al., 2019a).

One alternative to traditional chemical analysis of SOC is proximal sensing through diffuse reflectance spectroscopy. In this method, the reflectance of soil samples is measured – most frequently in the visible (Vis), near-infrared (NIR), and short-wave infrared (SWIR) spectra –, soil spectral libraries (SSL) are constructed, and soil properties are estimated using chemometric models (DEMATTÊ et al., 2019b; NOCITA et al., 2015; SORIANO-DISLA et al., 2014; VISCARRA ROSSEL et al., 2009).

Reflectance spectroscopy is a proximal sensing method applied to soils, recognized by the FAO (the Food and Agriculture Organization of the United Nations) for soil C Measurement, Reporting, and Verification systems (FAO, 2020, 2022). Although there is a long history of efficient use of this methodology in predicting soil attributes, particularly SOC (BEN-DOR, 1997; JANIK; SKJEMSTAD, 1995; NOCITA et al., 2015; SILVERO et al., 2020; VISCARRA ROSSEL et al., 2016; VISCARRA ROSSEL; BEHRENS, 2010), it is not yet a fully established methodology for routine soil analysis.

Recent studies indicate possibilities for improving prediction accuracy by expanding SSLs to include other data sources (LI et al., 2023; MOURA-BUENO et al., 2021; SABETIZADE et al., 2021; ZAYANI et al., 2023). These additional sources may include variables obtained from orbital remote sensing and variables associated with soil forming factors. Other studies critique and propose solutions to the transparency issue of chemometric models, which commonly employes machine learning methods, as well as their ability to generalize well in predicting attributes in new soil samples (BRICKLEMYER et al., 2018; BROWN; BRICKLEMYER; MILLER, 2005; MALMIR et al., 2019; MCBRIDE, 2022; POGGIO; BROWN; BRICKLEMYER, 2017; VISCARRA ROSSEL et al., 2022; WADOUX, 2023).

Therefore, considering the importance of SOC for food and climate security, scalable, cost-effective, reliable, and operational methods for assessing SOC in different ecosystems are increasingly necessary. Thus, this thesis aimed to develop and test methodologies for estimating SOC contents in profiles of tropical soils (surveyed in Brazilian biomes: Atlantic Forest, Caatinga, and Cerrado), adopting interpretability and generalization ability as *sine qua non* requisites for the machine learning methods. The main hypothesis of the thesis is that in tropical soils, the accuracy of SOC content prediction using soil spectral libraries in Vis-NIR-SWIR spectra can be increased when environmental variables and radar vegetation indices are added as covariates to the SOC prediction models. Accordingly, the thesis was divided into three scientific articles that sought to address the following key questions:

i) Can the LASSO (least absolute shrinkage and selection operator) regularization method produce accurate estimates of SOC content but improve the transparency of the models, compared to PLS (partial least squares)?

ii) Does the adoption of cross-validation oriented to soil profiles, when SOC samples from the same soil profile are present either only in model calibration, validation, or testing, avoid overfitting of the ML models?

iii) With which accuracy can the remote sensing radar vegetation indices, as an indirect measurement for aboveground biomass, be used in predicting SOC contents along soil layers?

iv) Will expanding SSL to include radar vegetation indices and environmental variables improve the prediction of SOC contents in soil profiles?

1.1. References

ANGELOPOULOU, T. et al. From Laboratory to Proximal Sensing Spectroscopy for Soil Organic Carbon Estimation—A Review. **Sustainability**, v. 12, n. 2, p. 443, jan. 2020.

BALDOCK, J. A.; BROOS, K. Soil organic matter. Em: HUANG, P. M.; LI, Y.; SUMMER, M. E. (Eds.). **Handbook of Soil Sciences: properties and processes**. 2. ed. Boca Raton: CRC Press, 2012.

BALDOCK, J. A.; SKJEMSTAD, J. O. Role of the soil matrix and minerals in protecting natural organic materials against biological attack. **Organic Geochemistry**, v. 31, n. 7, p. 697–710, 1 jul. 2000.

BATJES, N. H. Harmonized soil property values for broad-scale modelling (WISE30sec) with estimates of global soil carbon stocks. **Geoderma**, v. 269, p. 61–68, 1 maio 2016.

BEN-DOR, E. The reflectance spectra of organic matter in the visible near-infrared and short wave infrared region (400–2500 nm) during a controlled decomposition process. **Remote Sensing of Environment**, v. 61, n. 1, p. 1–15, 1 jul. 1997.

BOLINDER, M. A. et al. An approach for estimating net primary productivity and annual carbon inputs to soil for common agricultural crops in Canada. **Agriculture, Ecosystems & Environment**, v. 118, n. 1, p. 29–42, 1 jan. 2007.

BRASIL. **Decreto nº 10.606, de 22 de Janeiro de 2021**. Brasília, 2021. Disponível em: < https://www.planalto.gov.br/ccivil_03/_ato2019-2022/2021/decreto/d10606.htm >

BRICKLEMYER, R. S. et al. Comparing vis–NIRS, LIBS, and Combined vis–NIRS-LIBS for Intact Soil Core Soil Carbon Measurement. **Soil Science Society of America Journal**, v. 82, n. 6, p. 1482–1496, 2018.

BROWN, D. J.; BRICKLEMYER, R. S.; MILLER, P. R. Validation requirements for diffuse reflectance soil characterization models with a case study of VNIR soil C prediction in Montana. **Geoderma**, v. 129, n. 3, p. 251–267, 1 dez. 2005.

CANADELL, J. G. et al. Global Carbon and other Biogeochemical Cycles and Feedbacks. Em: **Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change**. 1. ed. Cambridge & New York: Cambridge University Press, 2021.

COTRUFO, M. F. et al. Formation of soil organic matter via biochemical and physical pathways of litter mass loss. **Nature Geoscience**, v. 8, n. 10, p. 776–779, out. 2015.

DEMATTÊ, J. A. M. et al. Soil analytical quality control by traditional and spectroscopy techniques: Constructing the future of a hybrid laboratory for low environmental impact. **Geoderma**, v. 337, p. 111–121, 1 mar. 2019a.

DEMATTÊ, J. A. M. et al. The Brazilian Soil Spectral Library (BSSL): A general view, application and challenges. **Geoderma**, v. 354, p. 113793, 15 nov. 2019b.

DIONIZIO, E. A. et al. Carbon stocks and dynamics of different land uses on the Cerrado agricultural frontier. **PLOS ONE**, v. 15, n. 11, p. e0241637, 6 nov. 2020.

FAO. **A protocol for measurement, monitoring, reporting and verification of soil organic carbon in agricultural landscapes: GSOC-MRV Protocol**. Rome, Italy: FAO, 2020.

FAO. **A primer on soil analysis using visible and near-infrared (vis-NIR) and mid-infrared (MIR) spectroscopy**. [s.l.] FAO, 2022.

GUO, L. B.; GIFFORD, R. M. Soil carbon stocks and land use change: a meta analysis. **Global Change Biology**, v. 8, n. 4, p. 345–360, 1 abr. 2002.

J. SILVA, L. et al. Soil carbon dynamics in integrated agricultural systems in Minas Gerais state, Brazil: A meta-analysis. **Geoderma Regional**, v. 36, p. e00761, 1 mar. 2024.

JACKSON, R. B. et al. The Ecology of Soil Carbon: Pools, Vulnerabilities, and Biotic and Abiotic Controls. **Annual Review of Ecology, Evolution, and Systematics**, v. 48, n. Volume 48, 2017, p. 419–445, 2 nov. 2017.

JANIK, L.; SKJEMSTAD, J. Characterization and analysis of soils using mid-infrared partial least-squares .2. Correlations with some laboratory data. **Soil Research**, v. 33, n. 4, p. 637, 1995.

JANZEN, H. H. The soil carbon dilemma: Shall we hoard it or use it? **Soil Biology and Biochemistry**, v. 38, n. 3, p. 419–424, 1 mar. 2006.

LAL, R. Soil Carbon Sequestration Impacts on Global Climate Change and Food Security. **Science**, v. 304, n. 5677, p. 1623–1627, 11 jun. 2004a.

LAL, R. Soil carbon sequestration to mitigate climate change. **Geoderma**, v. 123, n. 1–2, p. 1–22, nov. 2004b.

LAL, R. Beyond COP 21: Potential and challenges of the “4 per Thousand” initiative. **Journal of Soil and Water Conservation**, v. 71, n. 1, p. 20A-25A, 1 jan. 2016.

LAL, R. Digging deeper: A holistic perspective of factors affecting soil organic carbon sequestration in agroecosystems. **Global Change Biology**, v. 24, n. 8, p. 3285–3301, 2018.

LI, T. et al. Preliminary Results in Innovative Solutions for Soil Carbon Estimation: Integrating Remote Sensing, Machine Learning, and Proximal Sensing Spectroscopy. **Remote Sensing**, v. 15, n. 23, p. 5571, jan. 2023.

MALMIR, M. et al. Prediction of soil macro- and micro-elements in sieved and ground air-dried soils using laboratory-based hyperspectral imaging technique. **Geoderma**, v. 340, p. 70–80, 15 abr. 2019.

MANABE, S. Role of greenhouse gas in climate change**. **Tellus A: Dynamic Meteorology and Oceanography**, v. 71, n. 1, p. 1620078, 1 jan. 2019.

MAPA, Ministério da Agricultura, Pecuária e Abastecimento; MDA, Ministério do Desenvolvimento Agrário. **Plano setorial de mitigação e de adaptação às mudanças climáticas para a consolidação de uma economia de baixa emissão de carbono na agricultura**. Brasília: Ministério da Agricultura, Pecuária e Abastecimento, 2012.

MCBRIDE, M. B. Estimating soil chemical properties by diffuse reflectance spectroscopy: Promise versus reality. **European Journal of Soil Science**, v. 73, n. 1, p. e13192, 2022.

MINASNY, B. et al. Soil carbon 4 per mille. **Geoderma**, v. 292, p. 59–86, 15 abr. 2017.

MOURA-BUENO, J. M. et al. Environmental covariates improve the spectral predictions of organic carbon in subtropical soils in southern Brazil. **Geoderma**, v. 393, p. 114981, 1 jul. 2021.

NOCITA, M. et al. Soil Spectroscopy: An Alternative to Wet Chemistry for Soil Monitoring. Em: **Advances in Agronomy**. [s.l: s.n.]. v. 132p. 139–159.

O'DONOGHUE, T.; MINASNY, B.; MCBRATNEY, A. Digital Regenerative Agriculture. **npj Sustainable Agriculture**, v. 2, n. 1, p. 1–5, 26 mar. 2024.

OLIVEIRA, D. M. DA S. et al. Climate-smart agriculture and soil C sequestration in Brazilian Cerrado: a systematic review. **Rev. Bras. Ciênc. Solo**, v. 47, n. spe, 6 mar. 2023.

OLIVEIRA, D. M. S. et al. Crop, livestock, and forestry integration to reconcile soil health, food production, and climate change mitigation in the Brazilian Cerrado: A review. **Geoderma Regional**, v. 37, p. e00796, jun. 2024.

PAUSTIAN, K. et al. Quantifying carbon for agricultural soil management: from the current status toward a global soil information system. **Carbon Management**, v. 10, n. 6, p. 567–587, 2 nov. 2019.

POGGIO, M.; BROWN, D. J.; BRICKLEMYER, R. S. Comparison of Vis–NIR on in situ, intact core and dried, sieved soil to estimate clay content at field to regional scales. **European Journal of Soil Science**, v. 68, n. 4, p. 434–448, 2017.

SABETIZADE, M. et al. Combination of MIR spectroscopy and environmental covariates to predict soil organic carbon in a semi-arid region. **CATENA**, v. 196, p. 104844, 1 jan. 2021.

SILVA, I. R. DA; MENDONÇA, E. DE S. Matéria Orgânica do Solo. Em: NOVAIS, R. F. et al. (Eds.). **Fertilidade do Solo**. 1. ed. Viçosa: Sociedade Brasileira de Ciência do Solo, 2007. p. 275–374.

SILVERO, N. E. Q. et al. Effects of water, organic matter, and iron forms in mid-IR spectra of soils: Assessments from laboratory to satellite-simulated data. **Geoderma**, v. 375, p. 114480, 1 out. 2020.

SMITH, P. et al. How to measure, report and verify soil carbon change to realize the potential of soil carbon sequestration for atmospheric greenhouse gas removal. **Global Change Biology**, v. 26, n. 1, p. 219–241, 6 jan. 2020.

SORIANO-DISLA, J. M. et al. The Performance of Visible, Near-, and Mid-Infrared Reflectance Spectroscopy for Prediction of Soil Physical, Chemical, and Biological Properties. **Applied Spectroscopy Reviews**, v. 49, n. 2, p. 139–186, 17 fev. 2014.

VISCARRA ROSSEL, R. A. et al. 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 maio 2009.

VISCARRA ROSSEL, R. A. et al. A global spectral library to characterize the world's soil. **Earth-Science Reviews**, v. 155, p. 198–230, 1 abr. 2016.

VISCARRA ROSSEL, R. A. et al. Diffuse reflectance spectroscopy for estimating soil properties: A technology for the 21st century. **European Journal of Soil Science**, v. 73, n. 4, p. e13271, 2022.

VISCARRA ROSSEL, R. A.; BEHRENS, T. Using data mining to model and interpret soil diffuse reflectance spectra. **Geoderma**, Diffuse reflectance spectroscopy in soil science and land resource assessment. v. 158, n. 1, p. 46–54, 15 ago. 2010.

WADOUX, A. M. J.-C. Interpretable spectroscopic modelling of soil with machine learning. **European Journal of Soil Science**, v. 74, n. 3, p. e13370, 2023.

WIESMEIER, M. et al. Soil organic carbon storage as a key function of soils - A review of drivers and indicators at various scales. **Geoderma**, v. 333, p. 149–162, 1 jan. 2019.

ZAYANI, H. et al. Using Machine-Learning Algorithms to Predict Soil Organic Carbon Content from Combined Remote Sensing Imagery and Laboratory Vis-NIR Spectral Datasets. **Remote Sensing**, v. 15, n. 17, p. 4264, jan. 2023.

2. SCIENTIFIC ARTICLES

2.1. Article 1 – Improving the generalization error and transparency of regression models to estimate soil organic carbon using soil reflectance data [1]

Abstract: Despite the success of using soil spectroscopy in studies to predict soil attributes, like soil organic carbon (SOC), recent work has revealed several limitations to this approach: a tendency for model overfitting and a lack of transparency of machine learning (ML) methods. Thus, we aimed to both test the ability to improve the generalizability of the models to predict SOC using a cross-validation (CV) strategy oriented to soil profiles and to test the gain in model interpretability by using the least absolute shrinkage and selection operator (LASSO) regression method instead of the commonly used partial least squares (PLS) method. We used one soil spectral library composed of 127 soil profiles (n = 701), from Northeast Brazil, containing reflectance data from the visible, near, and short-wave infrared (VNIR) and the mid-infrared (MIR) spectral regions. We tuned the ML models to predict SOC via two CV strategies: the standard k-fold CV and the leave-soil-profile-out (LSPO) CV. We found that LSPO CV can produce models with better generalizability, as they lose less accuracy than the ones trained with k-fold CV. We conclude that disregarding the autocorrelation of SOC within the soil profile can produce models that are prone to overfitting. In addition, LASSO used 105 covariables from VNIR and 190 from MIR for a total of 8,604 and 13,336 covariables, respectively. Moreover, a few LASSO covariables correlated with SOC and are associated with both electronic transitions and vibrational bonds in organic compounds, so the possibility and ease of identifying spectral bands and their correlation with organic carbon indicate that the LASSO models presented more transparent models than the PLS models.

Keywords: diffuse reflectance; machine learning; soil spectroscopy; soil horizons; partial least squares.

---

[1] Article published in **Ecological Informatics**, 2023, volume 77, p. 102240, DOI: https://doi.org/10.1016/j.ecoinf.2023.102240

2.1.1. Introduction

Soil spectroscopy is a technology employed to explore the spectral properties of different soil types and to predict their physical, mineralogical, and chemical attributes (Demattê and Terra, 2014; Nocita et al., 2015). Among the most studied soil attributes using the diffuse reflectance of samples are soil organic matter and/or, more specifically, soil organic carbon (SOC; Hutengs et al., 2019; Roudier et al., 2017; Seidel et al., 2019; Soriano-Disla et al., 2014). Some studies point to soil spectroscopy as an important method to improve regional and local SOC estimates, aiming to enhance the understanding of biogeochemical carbon processes (Bricklemyer et al., 2018; Gehl and Rice, 2007).

Due to the success of these studies, the use of spectroscopy associated with regression methods (chemometrics) has been emphasized as an important complementary tool to estimate soil chemical and physical properties in addition to traditional laboratory analyses with wet chemical reagents, in which the analytical determinations are expensive and time-consuming, two problems that can be tackled by using soil spectroscopy (Demattê et al., 2019b; Paiva et al., 2022). However, there are problems with the operational applicability of spectroscopy, including uncertainties regarding the generalizability of the calibrated models, how modeling methods deal with the data dimensionality (a large number of predicting variables, or spectral bands, and then samples) of spectral libraries, and the lack of interpretability of machine learning (ML) algorithms (the “black-box” problem; McBride, 2022; Viscarra Rossel et al., 2022).

Janik and Skjemstad (1995) were the first researchers to estimate SOC with ML; they fed the models with diffuse soil reflectance using partial least squares (PLS) regression. Since then, PLS has been widely used (Moura-Bueno et al., 2019; Nocita et al., 2015; Santos et al., 2020; Soriano-Disla et al., 2014) because it is capable of dealing with multicollinearity among the predictors, a common phenomenon in spectral curves. PLS combines the predictors (by their covariance) into components, then associates the components with the dependent variable (Geladi and Kowalski, 1986; Yun et al., 2019). Subsequently, the components are used to predict the dependent variable (Boehmke and Greenwell, 2019; Geladi and Kowalski, 1986). Based on these characteristics, PLS has also been reported to be a dimensionality reduction method. However, it does not exactly reduce the number of predictors. Rather, it just makes linear combinations and keeps using the entire spectrum of covariables to predict a soil attribute (Yun et al., 2019).

Although PLS, a common regression method in chemometrics, is not necessarily adopted in spectral library studies as an ML method, as can be observed by the absence of different cross-validation (CV) and model testing strategies (Moura-Bueno et al., 2019; Santos et al., 2020), it is considered a supervised ML method (Boehmke and Greenwell, 2019; Kuhn and Johnson, 2013; Morellos et al., 2016). PLS transforms the reflectance data to predict SOC and continues using the full spectrum of covariates to make the estimates, including spectral bands not associated with SOC. Therefore, considering the increasing demand for the interpretability of ML models (Adadi and Berrada, 2018; Barredo Arrieta et al., 2020; Gilpin et al., 2018; Savage, 2022), other ML methods can be

evaluated and employed to produce regression models between soil spectroscopy and SOC that are even more interpretable (Viscarra Rossel et al., 2022).

To understand the importance of spectral bands in a PLS-based SOC prediction model, it is necessary to use at least variable importance techniques, which are known as model-agnostic methods (Barredo Arrieta et al., 2020). Such techniques measure the loss of accuracy of the model when a covariable is withheld. Santos et al. (2020) used PLS to predict SOC and demonstrated that PLS generates a spectrum of covariable importance. However, selection and penalty regression methods such as least absolute shrinkage and selection operator (LASSO) can be used to improve the interpretability of SOC prediction models with soil spectral library (SSL) data.

LASSO is not a newly developed ML method: It was proposed by Tibshirani (1996), but it has been loosely used to predict soil attributes with reflectance data (Bricklemyer et al., 2018). The closest use was by Dyar et al. (2012), who applied LASSO to rock mineralogy with spectral curves, studying the relationship between the spectral bands selected by LASSO and the minerals in the samples. Hence, there is a gap in the literature regarding the potential transparency and accuracy of LASSO to predict soil attributes.

LASSO can generate more transparent models than PLS because it works on the principle of parsimony, which states that for the model to be useful, it needs to generalize beyond the observations for which it was calibrated, and to achieve this goal, the model should remain as simple as possible (El Naqa et al., 2018). Furthermore, because this approach is based on linear regressions, LASSO models can be considered transparent (Barredo Arrieta et al., 2020). Instead of measuring the importance of a given covariable as a function of the loss in model accuracy, when it is retained, it is possible to identify mathematically how a selected variable relates to the dependent variable, in this case, SOC.

McBride (2022) also pointed out a problem with extrapolation beyond the limits of the observed data for soil attributes that have been used to calibrate spectroscopy-based models. When modeling SOC, which is a soil attribute that exhibits a spatially dependent structure—that is, in each soil profile—SOC samples from closer layers (in depth) tend to have more similar values than those from more distant layers, not considering that the autocorrelation condition of SOC along the profile is an error that can lead to model overfitting.

ML systems are sensitive to correlation between predictors and within the dependent variable. Autocorrelation in the dependent variable is the tendency that values from nearby observations are more similar than more distant observations. More simply, it is the tendency to have an internally dependent structure in time and space, among others (Roberts et al., 2017). This phenomenon has already been demonstrated by Dias et al. (2021) and Meyer et al. (2018), and when there is an internally dependent structure in the data, an appropriate CV strategy and holdout test should be defined to better evaluate the models (Roberts et al., 2017). Although CV and holdout test methods have been proposed to model soil attributes using whole soil profile samples (Brown et al., 2005), there has been limited work applying such methods (Malmir et al., 2019; Poggio et al.,

2017). The fact that SOC contents tend to decrease as the soil depth increases (Hao et al., 2023) indicates the presence of spatial dependence of samples collected from the same soil profile. However, the most recent studies concerning SOC prediction along soil profiles have not considered autocorrelation of SOC across soil horizons. Researchers have employed CV strategies such as leave-one-out (Moura-Bueno et al., 2019; Nocita et al., 2015; Santos et al., 2020; Soriano-Disla et al., 2014).

Soil spectroscopy is an important methodology that has advanced the development of large SSL, computational resources, and statistical modeling methods (Wadoux, 2023) to predict soil attributes. However, recent studies have highlighted that many users are skeptical of the predictions made by ML algorithms due to the tendency of model overfitting and the lack of interpretability of "black-box"-type methods (McBride, 2022; Viscarra Rossel et al., 2022; Wadoux, 2023). Both problems can be considered obstacles to the operational use of SSLs and ML methods to predict soil attributes, especially the SOC content. SOC is an important soil attribute because of its relationship with other soil properties and functions, from regulating runoff and water retention, fertility, and soil quality, to combating greenhouse gas emissions by sequestrating atmospheric carbon in organic form (Davis et al., 2018; Hao et al., 2023; Hutengs et al., 2019; Pechanec et al., 2018).

Misiuk and Brown (2023) proposed and highlighted the importance of using independent samples to calibrate and test ML models. However, this approach has not been used in practice for SOC modeling with SSLs from whole soil profiles. Furthermore, Wadoux (2023) proposed ways to improve the interpretability of ML models for soil reflectance, although they are also model-agnostic methods that use Shapley values. Hence, the present study is novel because we consider the need to use independent soil samples to model SOC using SSLs and we propose using transparent models, in which the relationships between spectral bands and SOC can be easily understood, to produce reliable SOC estimates.

Considering the search for transparency and generalizability of SOC prediction models using soil spectroscopy, we explored both the explainability-accuracy and bias-variance trade-off of SOC prediction models using a regional SSL. We aimed (a) to test the improvement in the generalizability of models trained using a CV strategy and holdout test oriented to the soil profile, and (b) to test the gain in interpretability of models built using LASSO compared with models based on PLS.

2.1.2. Material and methods

2.1.2.1. Description of the study area and the analytical data obtained

The database used was obtained from Santos et al. (2020), which was deposited in the Brazilian Soil Spectroscopy Library (Demattê et al., 2019b). These data were collected through field survey sampling of soil profiles in the State of Pernambuco, Northeastern Brazil, between 34°48′ and 38°48′ W and 7°18′ and 9°18′ S (Figure 2.1.1). Data were collected in three climatic regions according to the Köppen climate classification. The first region is in a coastal area and has a tropical monsoon climate

(class Am); the second, entering the continent, also has a tropical zone, but with a dry summer (class As); and the third is a dry and semi-arid zone of low latitude and altitude (class Bsh). In the Bsh and As domains, the predominant biome is the Caatinga, while in the Am domain the predominant biome is the Atlantic Forest. The exact location of the soil profiles with the Köppen climate classification can be found in the Supplementary Material (APPENDIX A: supplementary material for the article 1).

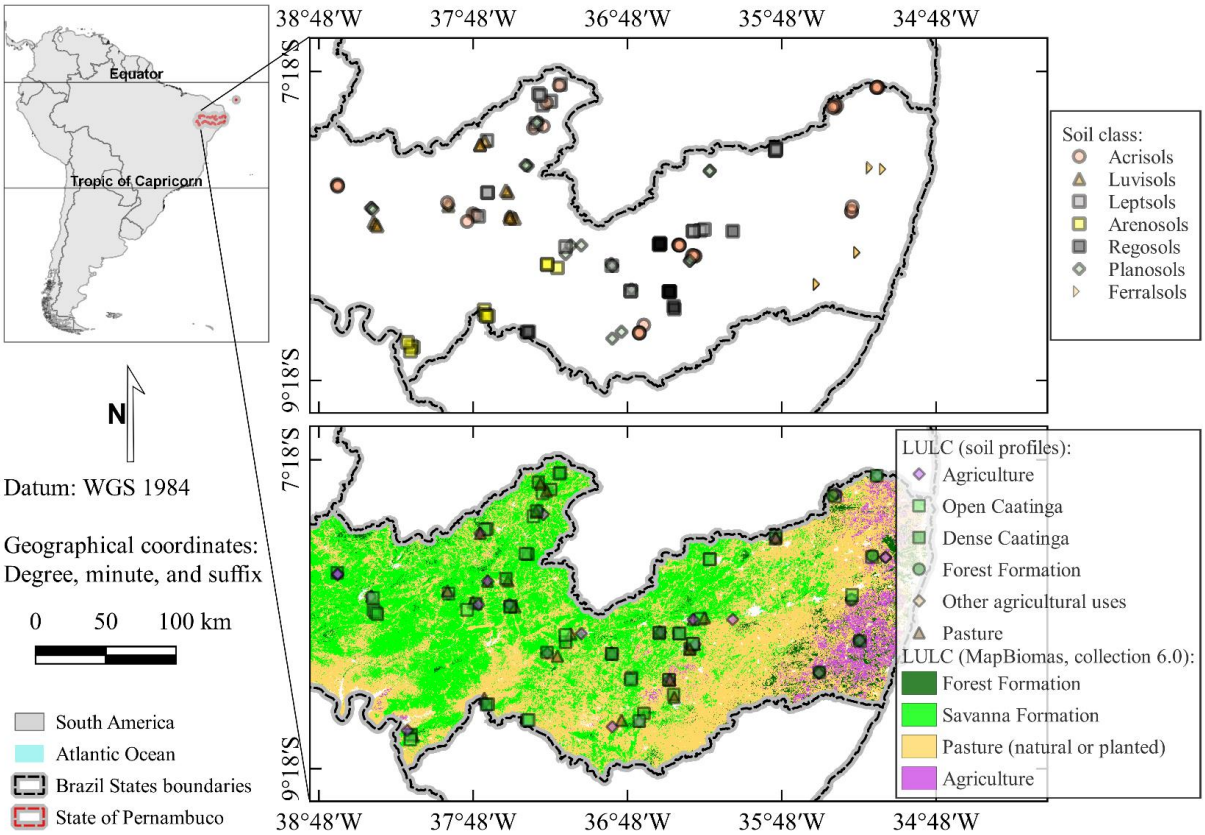


Figure 2.1.1. The figure presents the location and (a) the sampled soil profile classes and (b) the land use and land cover (LULC) class from both field observation and MapBiomas maps (Collection 6.0).

Soil samples were collected for 127 soil profiles in seven different soil classes: Acrisols (35 soil profiles and 210 samples), Ferralsols (11 soil profiles and 77 samples), Leptsols (23 soil profiles and 84 samples), Planosols (23 soil profiles and 117 samples), Arenosols (11 soil profiles and 74 samples), Luvisols (12 soil profiles and 68 samples), and Regosols (12 soil profiles and 71 samples). In addition, samples were collected in the following classes of land use and land cover (LULC): Atlantic Forest, Dense Caatinga, Open Caatinga, Agriculture, and Pastures.

Soil samples were collected from trenches (0.7 × 0.7 m), which were opened just after removing litter on the surface, obeying the horizons at standard depths of 0–10, 10–20, 20–30, 30–40, 40–60, 60–80, and 80–100 cm. A total of 701 samples were collected

and sent for physical and chemical analysis in the laboratory to determine the SOC, clay, silt, and total sand contents (in g kg^{-1}).

The soil samples were air dried and sieved through a 2 mm mesh. The SOC content was determined with subsamples (approximately 10 mg) via dry determination, using the TruSpec CHN-analyzer and a modulated solid sample operated at 900 ºC (LECO® 2006, St. Joseph, MI, USA). While this method can measure total carbon, carbonates were not present in the studied soils, so only organic carbon was determined. Then, the samples were forwarded to obtain spectroscopic measurements, namely the spectral signature of reflectance (ρ) in the visible, near, and shortwave infrared (hereinafter referred to as VNIR) and mid-infrared (MIR) spectral regions.

The FieldSpec 3 sensor was used for readings in the VNIR spectral region, with wavelengths ranging from 350 to 2500 nm (Analytical Spectral Devices Inc., Boulder, CO, USA). Twenty grams of each soil sample was placed in a Petri dish and homogeneously distributed over the flat surface. Two 50-W halogen lights served as a source of electromagnetic radiation. They were positioned 35 cm from the measured sample (both lamps had non-collimated rays with a zenith angle of 30º) with a 90º angle between the lights. A fiber-optic cable, located 8 cm from the center of the sample surface, captured the reflected energy from an approximately 2 cm^2 area. For each sample, the final reflectance was the average of three repetitions of the readings in different positions. Each repetition consisted of 100 sensor readings, aiming to maximize the signal-to-noise ratio. The spectroradiometer was calibrated before sample readings and then every 20 minutes using a white Spectralon plate.

In the MIR region (from 4000 to 600 cm^{-1}), measurements were made using an Alpha Sample Compartment RT-DLaTGS ZnSe (Bruker Optics Corporation, Billerica, MA 01821, USA) equipped with an accessory to detect diffuse reflectance (Drift). One cubic centimeter of the soil subsample received the same treatment used to determine reflectance in the VNIR spectral region. All spectra were recorded with a spectral resolution of 2 cm^{-1}, and for each measurement, 32 scans were performed. Before each reading, the equipment was calibrated with a gold plate, and its reflectance was also measured with 32 scans. Finally, for SOC modeling, the position of the MIR spectral readings was converted from wavenumber (cm^{-1}) to wavelength (nm) (nm = 10,000,000/wave number), making it compatible with VNIR spectra measurements (where $4000\ cm^{-1} = 2500$ nm and $600\ cm^{-1} \cong 16{,}666\ nm$).

2.1.2.2. Preprocessing and feature engineering of spectral curves

Before using the spectral reflectance curves in the modeling process, in which each spectral band represents a predictor, they underwent preprocessing to smooth the curve of each sample using a four-band running average filter (4 nm). In sequence, the smoothed curves underwent feature engineering techniques to extract information from them.

Three techniques were applied: continuum removal (CR), spectral derivative analysis with a first-order derivative (STD), and spectral derivative analysis with a second-order derivative (SCD). The CR technique (Clark and Roush, 1984) aims to easily

identify absorption bands in reflectance spectra by normalizing the reflectance curve as a function of a continuous spectrum. The spectral derivative analysis was applied to identify the variation rate between the bands (Tsai and Philpot, 1998); it was calculated with the Savitzky–Golay method.

After feature engineering processing, the number of spectral predictors obtained was four times larger than the initial number of predictors because in addition to the reflectance spectral bands (ρ), with 2,151 bands in the VNIR spectral region and 3,334 bands in the MIR spectral region, there were three new variables for each band (CR, STD, and SCD). These were joined to ρ in the same datasets, VNIR and MIR separately, and the number of predictors increased from 2,151 and 3,334 to 8,604 and 13,336 spectral covariates, respectively, passing the spectral curves beyond the ρ domain, and creating the CR, STD, and SCD domains.

All data processing, from the preprocessing of spectral curves to the evaluation of the models (detailed hereafter), was performed using the R programming language (R Core Team, 2023). Hence, all the computer libraries mentioned here were implemented in R. Preprocessing and feature engineering were performed using the hsdar library (Lehnert et al., 2019).

2.1.2.3. ML modeling of the SOC content

2.1.2.3.1. Regression methods

Two regression algorithms were used to predict the SOC content (in g kg^{-1}): LASSO and PLS. Both are linear-based methods designed to deal with collinearity between predictors, but in different ways. PLS combines predictors into components (q), associating them by their covariance (Geladi and Kowalski, 1986; Yun et al., 2019). LASSO is based on the principle of parsimony because it decreases the number of predictors by adjusting the penalty parameter (λ) (Zhao and Yu, 2006).

For each covariable, LASSO shrinks the slope coefficient estimates toward zero by adjusting λ via model tuning during CV. As LASSO is a linear regression method, to fit a model it uses the residual sum of squares (RSS), a common loss function. The RSS, which is displayed in Equation 1, is used to minimize the prediction error of a fitted linear regression.

$$RSS = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j \cdot x_{ij}\right)^2 \equiv \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (1)$$

with y_i and $\hat{y}_i$ being the observed sample and the predicted dependent variable, respectively; x_{ij} the observed sample of the j-th covariable; and β_j the slope coefficient (or the weight) of the j-th covariable.

LASSO uses the RSS and adds a penalization term, which uses the absolute value of β_j and contains the penalization parameter, as presented in Equation 2. When finding the slope coefficients that minimize the RSS and increase λ, the penalized residual sum

of squares (PRSS) increases. This means that a LASSO-fitted model presents gains in bias but reductions in variance.

$$PRSS = minimize(RSS) + \lambda \sum_{j=1}^{p} |\beta_j| \quad (2)$$

As λ increases, the value of β_j tends toward zero, and some values of β_j are constrained exactly to zero. When this occurs, the variable is not used for prediction, and then LASSO performs a variable selection task when also dealing with multicollinearity. More details about the LASSO method can be found in the literature (James et al., 2013; Kuhn and Johnson, 2013; Tibshirani, 1997).

2.1.2.3.2. Model fitting

Both PLS and LASSO were applied to predict the SOC content using spectral curve data in the VNIR and MIR spectral regions separately, using the covariates ρ, CR, STD, and SCD. Out of the 127 soil profiles, 80% were allocated for training the models, while the remaining 20% were set aside for testing (holdout test). In other words, samples from 101 soil profiles were used to train the model, and samples from 26 soil profiles were used to test the model. The division of the dataset into training and testing sets was done randomly by assigning soil profile samples to either dataset. This ensured that soil samples from nearby soil layers of the same soil profile were included in only one of the modeling steps.

Because of the nature of LASSO's constriction, by adjusting the penalization hyperparameter (λ), the coefficients are forced to tend toward zero, producing some coefficients that are exactly equal to zero (Tibshirani, 1997). The predictors with coefficients equal to zero are automatically eliminated, so LASSO reduces the dimension of the predictors and then performs the prediction. λ was determined from the algorithm by using a search grid, whose value ranged from 0 to 2 at intervals of 0.05. The PLS models were also adjusted via a search grid, which was designed to make q range from 1 to 10.

The glmnet package (Friedman et al., 2010) was used to tune the LASSO models. In the glmnet implementation, besides λ, the hyperparameter α is required for model training. α is the elastic net mixing parameter, and to tune a LASSO model, α was held constant at 1 for the LASSO penalty (Friedman et al., 2010; Tay et al., 2023).

2.1.2.3.3. Adopted CV strategies

To study the bias-variance trade-off of the models considering the autocorrelation of SOC within the soil profiles, two CV methods were chosen: the leave-soil-profile-out (LSPO CV) and the commonly used k-fold (k-fold CV). LSPO CV is a special case of k-fold CV in which instead of random data samples being sorted for each folder (for calibration and CV), soil profiles are sorted, and their samples are allocated to the folders for calibration and CV. The LSPO CV strategy consisted of creating an identification

variable to assign a unique ID to all soil samples from the same soil profile. This ID variable was used to guide the subsampling for CV of k-folds during training. In other words, the ID variable was used to determine which soil profile and its samples were allocated to one of the k-folds to tune the models, making LSPO CV a target-oriented k-fold CV where the "target" is the soil profile. LSPO CV was applied using the CreateSpacetimeFolds function from the CAST package (Meyer et al., 2018): 10 folds were defined, and the soil profiles were randomly sorted by the algorithm. For the standard k-fold CV, there were 10 folds and five repetitions.

LSPO CV is a target-oriented CV method that during optimization of the model hyperparameter does not allow similar samples—that is, samples that are nearby in space—in the training and CV datasets. When using common k-fold CV, there is the possibility of having soil samples from the same soil profile, including nearby samples (in soil depth), in the training, CV, and testing datasets, a phenomenon that can lead to overfitting. On the other hand, withholding an entire soil profile (and its samples) for each modeling step can avoid overfitting and produce a better bias-variance trade-off (Brown et al., 2005; Malmir et al., 2019; Poggio et al., 2017).

CV and testing of the models that utilize soil spectroscopy data, oriented to the soil profile, were first proposed by Brown et al. (2005). Subsequent papers in other environmental modeling studies have described the importance of using independent samples in each modeling step to produce realistic accuracy metrics. They also proposed methods such as leave-location-out (LLO) CV (Meyer et al., 2016, 2018). The difference between LLO and LSPO CV is that in LLO CV, both individual locations and groups of locations, such as soil profile groups, can be configured as LLO CV. Therefore, to distinguish it from LLO CV, the soil profile-oriented CV method was named LSPO CV for convenience. Figure 2.1.2 provides a visualization of how the k-fold and LSPO CV methods work.

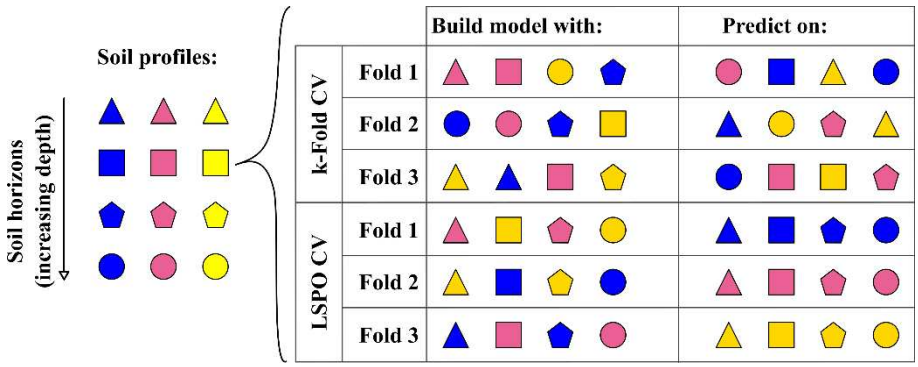


Figure 2.1.2. An illustrative scheme of how both the employed cross-validation (CV) strategies—k-fold CV and leave-soil-profile-out CV (LSPO CV)—work. Each geometric figure represents a soil horizon and each color indicates a different soil profile.

Each algorithm (LASSO and PLS) adjusted with each spectral database (VNIR and MIR) using two CV strategies (LSPO CV and k-fold CV) was trained 100 times with different randomization seeds, giving rise to 800 models. This process is important to evaluate the variability of predictions because different subsets of data (in k-fold

resampling) can generate different results and, consequently, different performances for each model (Gomes et al., 2019; Siqueira et al., 2022). For reproducibility, a fixed vector containing 100 randomization seeds with truly random numbers was used.

After fitting models through the k-fold and LSPO CV strategies, the models were tested for their ability to predict the SOC content in unknown soil profiles (26 soil profiles). Figure 2.1.3 displays, step-by-step, the entire training, CV, and model testing procedures. The training and CV steps were performed using features of the caret package (Kuhn, 2020) with kernels: for LASSO, the glmnet package, and for PLS, the pls package (Liland et al., 2021). The source code for all modeling procedures is open and available from the Github repository (https://github.com/eupassarinho/soil-carbon-and-spectroscopy-with-PLS-and-LASSO.git).

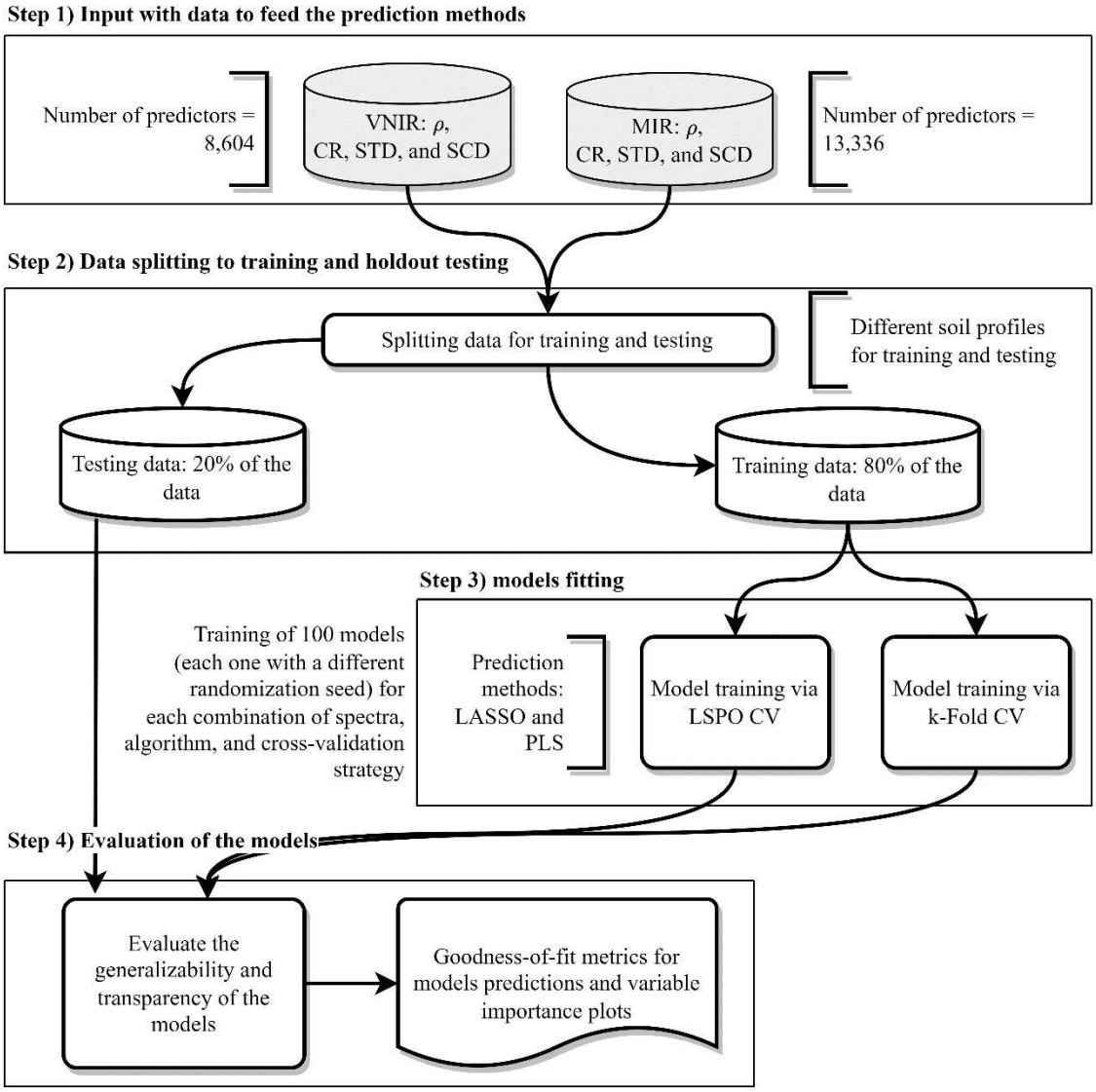


Figure 2.1.3. The model fitting scheme for each spectral region, which includes subsampling of soil profiles to evaluate model training, testing, and performance. LASSO, least absolute shrinkage and selection operator; PLS, partial least squares; LSPO, leave-soil-profile-out cross-validation (CV); MIR, mid-infrared; VNIR, visible,

near- and shortwave-infrared; ρ, original spectral reflectance; CR, the normalized reflectance; STD, the first derivative of ρ; and SCD, and the second derivative of ρ.

2.1.2.4. Evaluation of models

The model performance was evaluated in the training step based on the metrics coefficient of determination (R^2), root mean squared error (RMSE), and mean absolute error (MAE). With the fitted models, the variable importance ranking was obtained using the vip package (Variable Importance Plots; Greenwell et al., 2020). For the PLS models, the importance of each feature is computed based the impact it had in reducing the sums of squares, while the feature importance for the LASSO models is just the slope coefficient of each selected feature rescaled from zero to 100.

The nonparametric Kruskal–Wallis test, with Dunn's *post hoc* test, was used to evaluate the RMSE, MAE, and R^2 values to find statistical differences in the models in both modeling steps: CV (training) and holdout testing. The models and their 100 repetitions were considered as parameters, and the significance level was set at $P < 0.05$.

Other metrics were considered to evaluate the performance of the models in the holdout test: Pearson's correlation test (r), Willmott's agreement index (d), and the mean bias error (MBE) (Willmott, 1982; Willmott and Matsuura, 2005; Willmott et al., 2012). The d index and the MBE were computed with features from the hydroGOF package (Zambrano-Bigiarini, 2017).

The aforementioned set of metrics was adopted to involve metrics with different characteristics in the model evaluation. First, to evaluate the performance of model fitting for both k-fold and LSPO CV, standard metrics of caret were used: RMSE, MAE, and R^2. The RMSE squares the difference between simulated and observed values, making it a good indicator of outliers in simulated data (Chai and Draxler, 2014). On the other hand, the MAE does not have great sensitivity to outliers (Willmott and Matsuura, 2005). R^2 is based on how much the model explains the variance in the observations (Onyutha, 2020). A limitation of R^2 is that it is inadequate to evaluate model performance because it does not provide information about the model's residues (Onyutha, 2020). Although R^2 is used in CV, both the r and d indices were adopted to evaluate the model's goodness-of-fit in the holdout test. The d index was chosen because it is normalized (from 0 to 1) and is a non-dimensional measure (Ji and Gallo, 2006), features that are proposed to circumvent the limitations associated with R^2 (Willmott, 1981, 1982). Similarly to the RMSE, the d index squares the difference between predictions and observations; consequently, it is sensitive to outliers (Willmott et al., 2012). The MBE was included to identify the bias of super- and underestimation because it is not normalized and varies from negative to positive values.

2.1.3. Results

2.1.3.1. CV results

Figure 2.1.4 shows the distribution of the adjustment metrics for model fitting, obtained from hyperparameter tuning. Each boxplot displays 100 values of each metric (RMSE, MAE, and R^2) for each set of models. In the MIR and VNIR spectral regions, the PLS method produced the most accurate models (lower RMSE and MAE values and higher R^2 values) trained via k-fold CV. Moreover, the models adjusted via LSPO CV presented a greater amplitude of adjustment metrics, while in the models adjusted via k-fold CV, the RMSE, MAE, and R^2 were very close (Figure 2.1.4). Thus, by changing the randomization seed—and therefore the choice of soil profiles for training—the regression parameters can vary greatly, causing the models to present very different adjustment metric values.

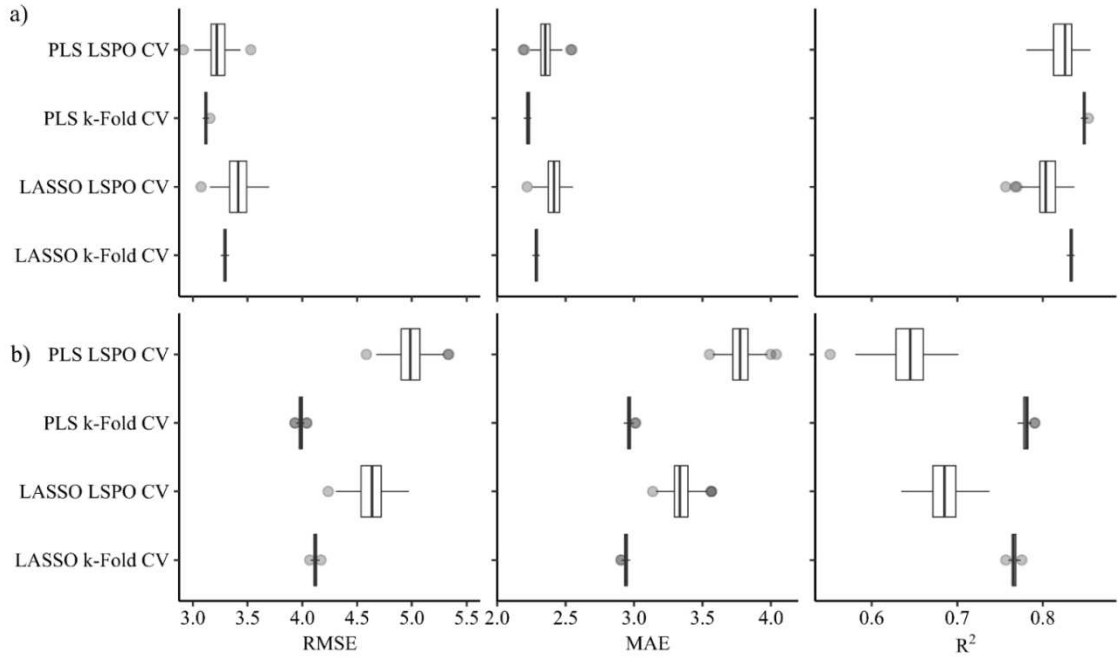


Figure 2.1.4. Characterization of the adjustment metrics of the models in the cross-validation (CV) step (leave-soil-profile-out [LSPO] and k-fold CV) to predict the soil organic carbon (SOC) content (in g kg^{-1}). The panels present the root mean squared error (RMSE), the mean absolute error (MAE), and the coefficient of determination (R^2) for the (a) mid-infrared and (b) visible, near, and short-wave infrared spectral models.

2.1.3.2. Bias-variance trade-off in the models

We explored the CV performance of the models by analyzing the mean and standard deviation of RMSE, MAE, and R^2 (Table 2.1.1) alongside the holdout testing metrics. First, the models fed the MIR spectral covariables were more accurate than the models fed the VNIR spectral covariables. Next, in the training step, the PLS-based models generally produced the best adjustment metrics. However, in the testing step, the LASSO-based models had the best metrics, indicating that they have better generalizability.

Regarding CV strategies, the models trained via k-fold CV in the VNIR spectral region were better than those trained via LSPO CV when assessed only in the training stage. In the testing stage, however, there was almost no difference in the metrics between the k-fold and LSPO CV models for both the MIR and VNIR spectral regions.

Table 2.1.1. Statistical metrics of the models in the cross-validation (training) and holdout testing steps.

Spectrum	Algorithm	Cross-validation strategy	RMSE		MAE		R^2	
			$\bar{x}$	sd	$\bar{x}$	sd	$\bar{x}$	sd
Cross-validation (training) performance								
MIR	LASSO	k-fold	3.293	0.016	2.286	0.011	0.833	0.002
		LSPO	3.411	0.110	2.411	0.066	0.805	0.015
	PLS	k-fold	3.119	0.014	2.226	0.013	0.849	0.002
		LSPO	3.227	0.099	2.352	0.063	0.823	0.016
VNIR	LASSO	k-fold	4.117	0.018	2.939	0.013	0.766	0.003
		LSPO	4.634	0.138	3.346	0.085	0.684	0.021
	PLS	k-fold	3.985	0.020	2.963	0.018	0.781	0.004
		LSPO	4.989	0.138	3.777	0.092	0.646	0.025
Holdout testing performance								
MIR	LASSO	k-fold	3.632	0.000	2.331	0.000	0.722	0.000
		LSPO	3.630	0.015	2.329	0.013	0.722	0.002
	PLS	k-fold	3.926	0.037	2.515	0.028	0.688	0.006
		LSPO	3.875	0.083	2.531	0.050	0.693	0.009
VNIR	LASSO	k-fold	4.147	0.013	3.359	0.004	0.579	0.004
		LSPO	4.184	0.015	3.292	0.034	0.559	0.003
	PLS	k-fold	4.401	0.381	3.423	0.261	0.571	0.063
		LSPO	4.695	0.287	3.638	0.177	0.519	0.040

$\bar{x}$ and sd refer to the mean and standard deviation, respectively, of the statistical metric values. LASSO, least absolute shrinkage and selection operator; LSPO, leave-soil-profile-out; MAE, mean absolute error; MIR, mid-infrared; PLS, partial least squares; R^2, coefficient of determination; RMSE, root mean squared error; VNIR, visible, near, and short-wave infrared.

The models tuned with the k-fold CV strategy produced over-optimistic statistical metrics in training and had a greater accuracy drop in testing compared with the LSPO CV–tuned models. Because SOC in the dataset showed a trend to decrease as the soil depth increased (Figure 2.1.5), it can be assumed that samples from closer soil layers have more similar SOC contents than samples from more distant soil layers. For example, the SOC content of the 0–10 cm layer is more similar to the content of the 10–20 cm layer of the same profile than that of the 80–100 cm layer. The same is valid for the reflectance data. This overfitting behavior of the k-fold CV models is because in k-fold CV, very similar samples (neighboring samples) could end up in both calibration and CV folds, leading to overestimated accuracy results.

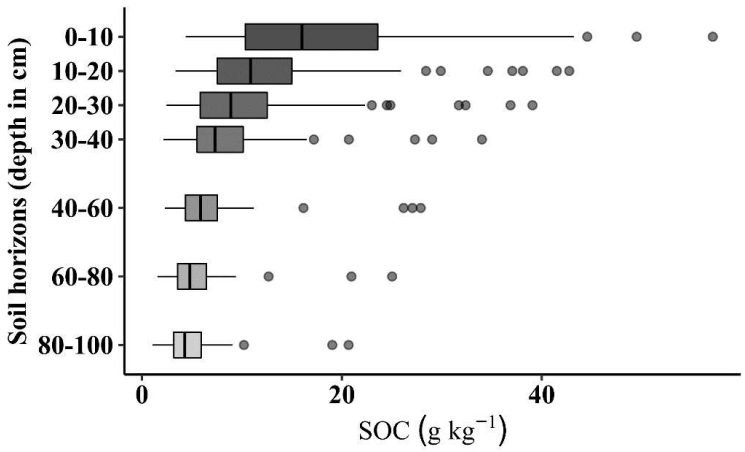


Figure 2.1.5. Boxplots exhibiting the distribution of soil organic carbon (SOC, in g kg^{-1}) in each soil horizon.

We noted significant differences in performance between all models for both CV (training) and holdout testing. The Kruskal–Wallis test (and Dunn's *post hoc* test) results are presented in the Supplementary Material (APPENDIX A: supplementary material for the article 1).

Because each model performed differently, we prepared Figure 2.1.6 to display the loss of accuracy between the training and testing of each model. This figure presents the slope of R^2 obtained for the training and testing of each model set. It is evident that the LASSO- and MIR-based model, tuned with LSPO CV, presented the lowest loss of accuracy. Considering the PLS-based models, the MIR PLS LSPO CV models showed better generalizability.

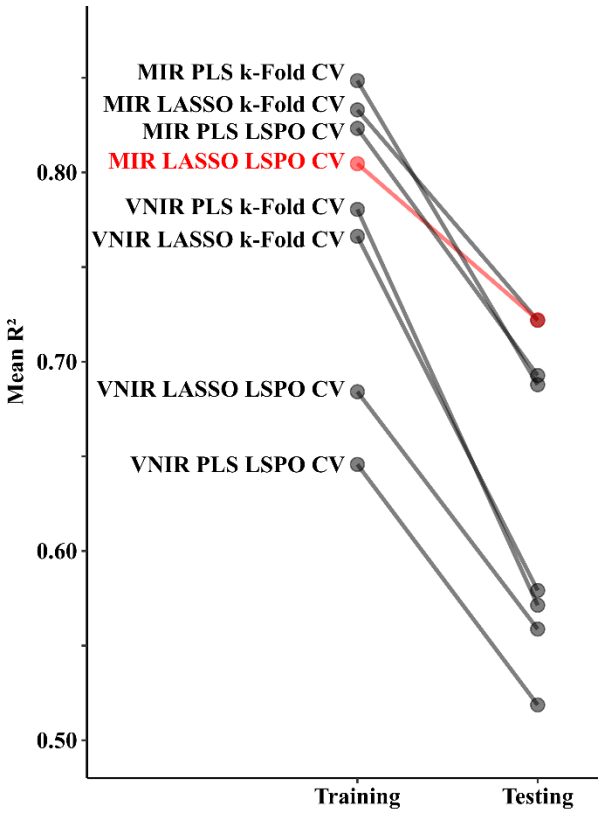


Figure 2.1.6. Slope graph of the mean R^2 values for all models obtained for training and testing. The MIR LASSO LSPO CV model (red) presented a better bias-variance trade-off, or less of a difference in R^2 between the training and testing steps. CV, cross-validation; LASSO, least absolute shrinkage and selection operator; LSPO, leave-soil-profile-out; MIR, mid-infrared; PLS, partial least squares; R^2, coefficient of determination; VNIR, visible, near, and short-wave infrared.

We assumed that the MIR LASSO LSPO CV model had the best bias-variance trade-off and used it, along with the MIR PLS LSPO CV model, to predict the SOC content (Figure 2.1.7). The scatterplots show an average of 100 predictions for each SOC value observed from the testing dataset.

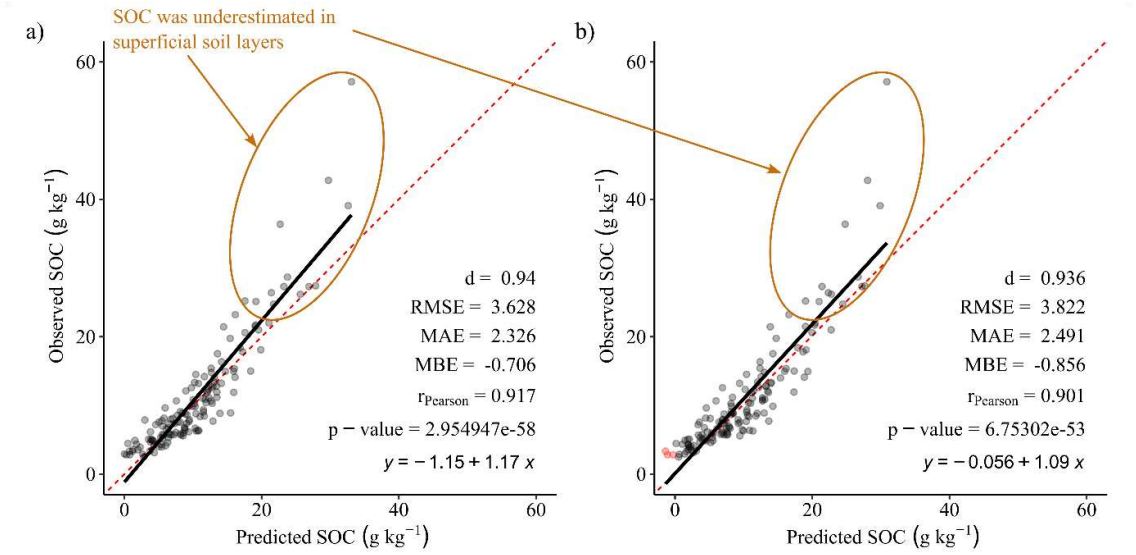


Figure 2.1.7. The scatterplots show the predicted versus observed SOC content (g kg^{-1}) from the holdout testing dataset, with predictions made by the (a) MIR LASSO LSPO CV and (b) MIR PLS LSPO CV models. Predicted SOC values that are negative are shown as red dots. LASSO, least absolute shrinkage and selection operator; LSPO, leave-soil-profile-out; MAE, mean absolute error; MBE, mean bias error; MIR, mid-infrared; PLS, partial least squares; R^2, coefficient of determination; RMSE, root mean squared error; SOC, soil organic carbon.

When analyzing the metrics in Figure 2.1.7, we observed finetuning between the predictions and observations, because the d index and the r value are very close to the unit. However, when we analyzed the MBE, we noticed that both models tended to underestimate the SOC content, especially when it was > 20 g kg^{-1}. By analyzing these underestimated samples, we determined that they are mainly in the surface soil horizons, where the organic matter content usually tends to be higher than in the deeper soil horizons. In Figure 2.1.7, all statistical metrics displayed are coherent, which means that we detected better adjustment between predictions using the LASSO method compared with the PLS method. In addition, they both have an anomaly because they estimated negative SOC values. The minimum estimated value in the PLS model was -2.16 g kg^{-1}, while for LASSO, it was -0.04 g kg^{-1}.

In general, from the data in Table 2.1.1 and Figure 2.1.6, the LASSO models have better generalizability for new data, regardless of the CV strategy (LSPO CV or k-fold CV). The models for each spectral region (VNIR and MIR) are based on the same set of soil observations and classes, including the same control object—that is, the same configuration of different soil profiles for training and validation. So, that fact that LASSO-based models lost less accuracy in testing than PLS-based models indicates the better generalizability of the former.

2.1.3.3. Accuracy-transparency trade-off of regression methods

As the MIR-based models trained via LSPO CV were the most accurate and provided better generalizability to predict the SOC content, we explored the accuracy-

transparency trade-off of the MIR LASSO LSPO CV and MIR PLS LSPO CV models, as well as the transparency of the LASSO-based models. We also extended this analysis to the VNIR LASSO LSPO CV models.

Figure 2.1.8 shows the importance of the MIR covariables used to predict the SOC content by both the PLS and LASSO methods. For the PLS method, Figure 2.1.8a displays a spectrum of covariable importance in the domains of ρ, CR, STD, and SCD, because all covariables are used in modeling. For the LASSO method, where when selecting the covariables, the method eliminates others, Figure 2.1.8b displays the importance of 190 covariables from the total of 13,336 that entered the modeling. Of the 13,336 spectral covariables in the MIR spectral region, only 190 were used to predict the SOC content (g kg^{-1}), which represents approximately 1.5% of all features the LASSO algorithm received. In the VNIR spectral region, 105 were selected, representing approximately 1.2% of the total covariables. On the other hand, the PLS method used all predictors and performed worse than the LASSO method in all scenarios in the testing sets (see Table 2.1.1 and Figure 2.1.6).

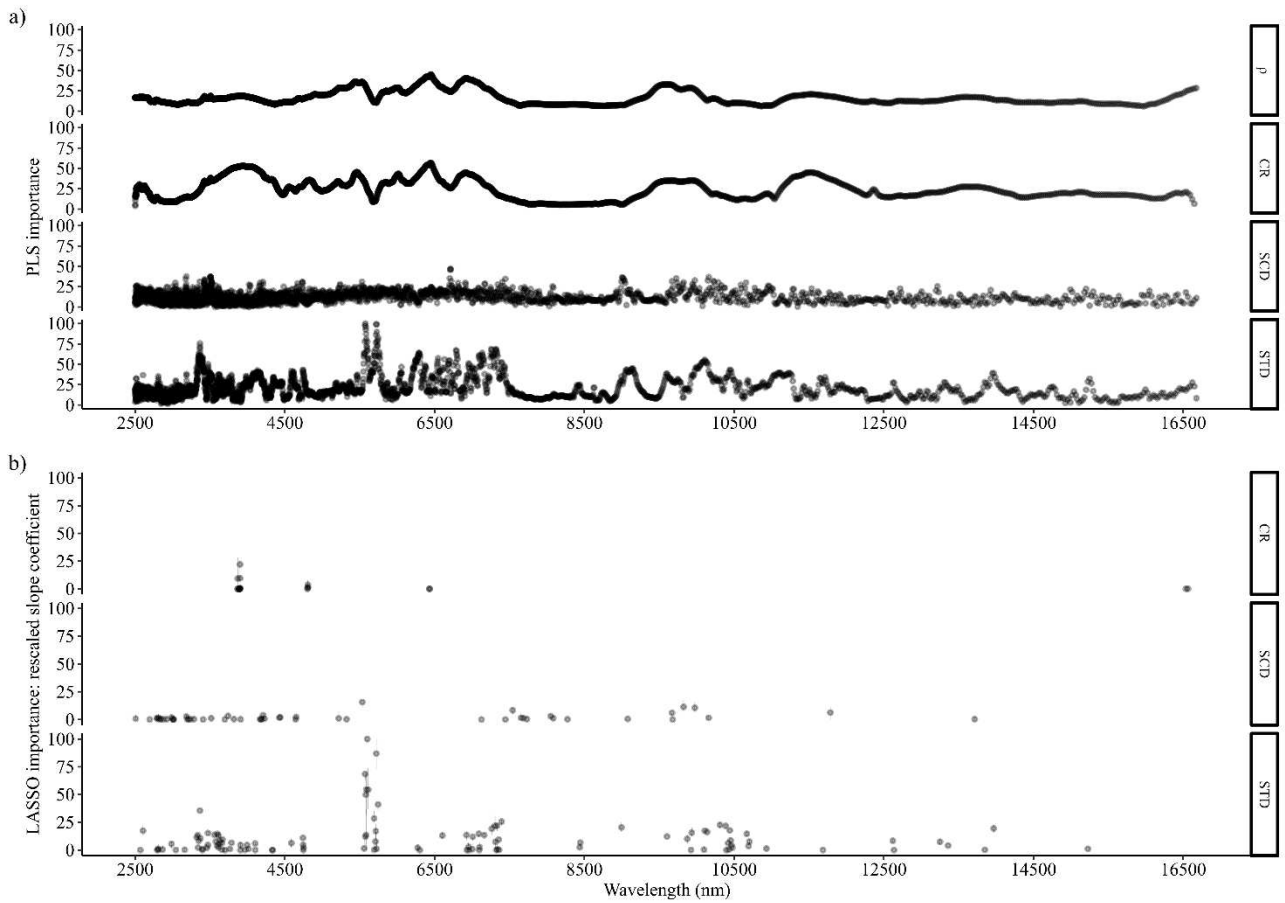


Figure 2.1.8. Variable importance plots of the mid-infrared (MIR) spectral covariates used to predict the soil organic carbon (SOC) content with (a) the partial least squares (PLS) method and (b) the variables selected and used by the least absolute shrinkage and selection operator (LASSO) method. The facets of the plots represent the spectral covariates in the domain of the original spectral reflectance (ρ), the normalized reflectance (CR), and the first and second derivative of ρ (STD and SCD, respectively).

Despite the inclusion of 190 or 105 covariables, we acknowledge that the models cannot be considered parsimonious (according to some criteria). Furthermore, it is worth noting that not all of the selected covariables exhibit a significant correlation with the SOC content. We explored this relationship and prepared Figure 2.1.9 to present the correlation between the selected spectral covariables (for both the VNIR and MIR spectral regions) and the SOC content. In general, the highest r values are close to 0.6 or -0.6, indicating a moderate correlation, and the spectral variables (ρ, CR, STD, and SCD) and SOC do not necessarily follow a linear relationship.

We only selected the ρ bands in the VNIR spectral regions; they are at wavelengths of approximately 603, 2260, and 2490 nm. Similarly to the ρ covariables, we selected CR covariables in the same band regions, plus at approximately 1180, 1240, and 1900 nm. We selected the first derivative of reflectance near 1900 and 2260 nm. The second derivative had more selected covariables, and many of them did not have a significant correlation with SOC, which can also be observed in the MIR spectral region.

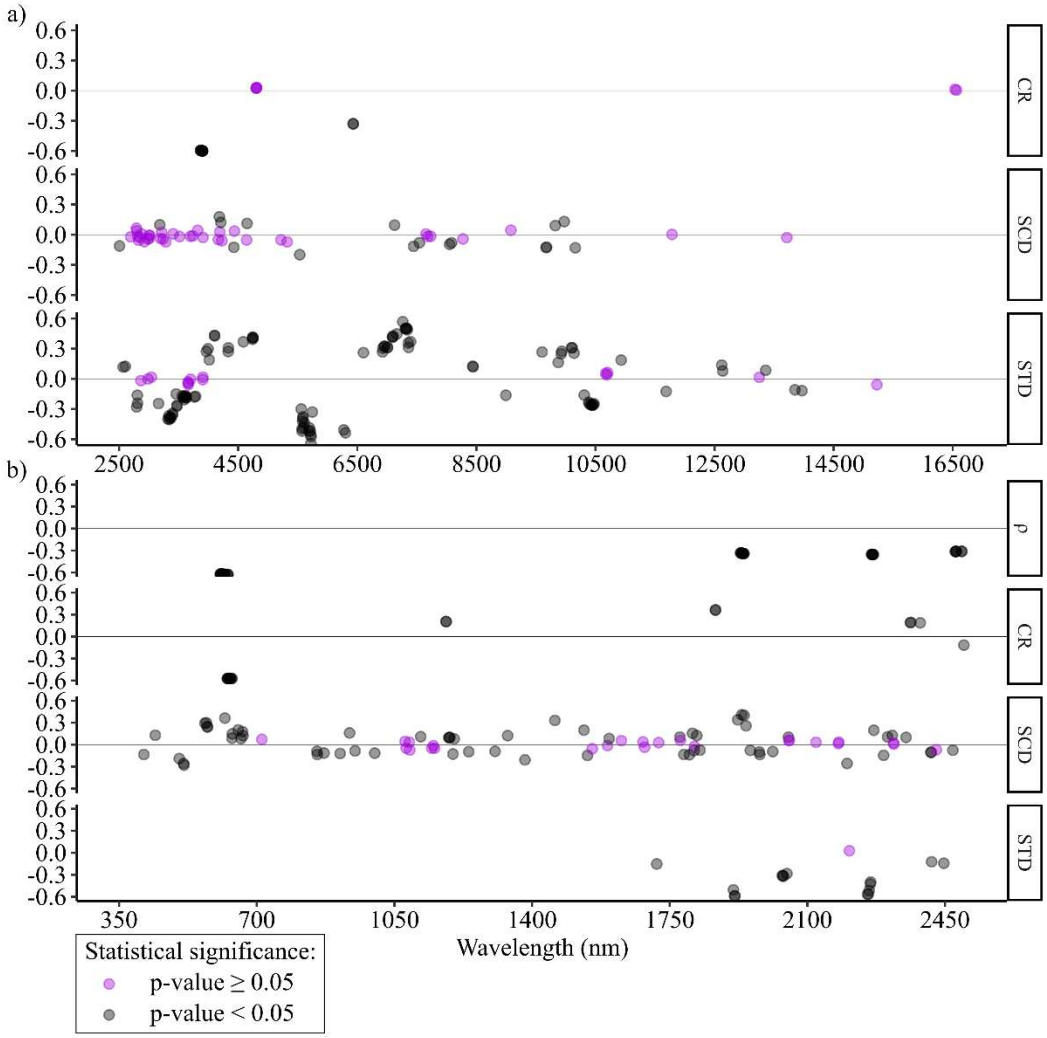


Figure 2.1.9. Pearson correlation analysis between the SOC content (g kg^{-1}) and the selected covariables in the (a) MIR LASSO LSPO CV and (b) VNIR LASSO LSPO CV models. Each point indicates the level of correlation, while the color of the point indicates

its statistical significance. ρ, original spectral reflectance; CR, normalized reflectance; LASSO, least absolute shrinkage and selection operator; LSPO, leave-soil-profile-out; MIR, mid-infrared; SOC, soil organic carbon; STD and SCD, the first and second derivative of ρ, respectively.

2.1.4. Discussion

The models trained using LSPO CV showed the lowest loss of accuracy to predict SOC. This is because in k-fold CV, samples from the same soil profile can be present in both the training and CV subsets (or folds). When LSPO CV-trained models are tested with data from new soil profiles, they tend to maintain higher accuracy because similar neighboring samples are not present in both the training and testing sets, reducing the risk of overfitting. Brown et al. (2005) stated that ensuring the independence of validation samples is essential to obtain realistic accuracy in models.

Brungard et al. (2021) evaluated prediction models of different soil attributes at different depths, but without adopting a CV strategy like LSPO CV. The authors proposed a regional scale model covering six states of the United States of America: Arizona, Colorado, Nevada, New Mexico, Utah, and Wyoming. Although they did not develop the models by using object-oriented strategies, they argued that model validation by region has implications for digital soil mapping. Among the implications is the possibility of identifying regions where the model has low accuracy and where additional resources can be allocated (intensify sampling, add covariates, and improve modeling efforts). So, Brungard et al. (2021) described the advantages of adopting evaluations of the models based on the location where soils were sampled. In the present study, as highlighted by Figure 2.1.7, the models tended to underestimate the SOC content in the superficial layers.

When comparing the results between the spectral regions, we found greater accuracy to predict the soil SOC content when using the MIR rather than the VNIR spectral region. In this respect, our study corroborates data from other studies (see Table 2.1.1, Figure 2.1.4, and Figure 2.1.6), such as those by Soriano-Disla et al. (2014), Santos et al. (2020), and Terra et al. (2019). However, we observed that all the obtained models tended to underestimate the SOC content at higher concentrations (> 20 g kg^{-1}; Figure 2.1.7).

Vasques et al. (2010) studied organic soils and reported similar results, and this same behavior can be observed in the scatterplots presented in the studies by Stevens et al. (2013), Liu et al. (2016), and Liu et al. (2019), and even in the study by Francos et al. (2021), who evaluated organic matter. There could be saturation in the spectral response with higher SOC contents, which are commonly found in topsoil layers. One reason could be that topsoil samples are generally underrepresented in the dataset. However, considering the use of entire soil core samples or soil layers from each soil profile and that LSPO CV ensures that each horizon is properly presented in calibrations and validation sets, SOC underestimation in samples with a higher SOC content could be a limitation of linear regression methods such as PLS and LASSO.

Based on Table 2.1.1 and Figure 2.1.6, there is an increase in generalizability when adopting LASSO as the prediction method. SOC and spectroscopy-based models have reached values of about 1.5–6.5 g kg^{-1}, as can be observed in the studies by de Soriano-Disla et al. (2014) and Santos et al. (2020), which means that LASSO can replace PLS without loss of accuracy, as Dyar et al. (2012) concluded, mainly due to the feasible transparency of LASSO-based models.

The spectral covariables selected by the LASSO models to predict SOC were correlated with the SOC content in the VNIR and MIR spectral regions (Figure 2.1.9). In the near and shortwave infrared spectral region, key components of the organic matter peaked at approximately 1930 nm (Knadel et al., 2015), behavior that is associated with what was detected in the first derivative (which presented a moderate and significant correlation). Vasques et al. (2010) reported that wavelengths around 1400 nm and from 1800 to 2400 nm are important for SOC estimates; however, they made these observations by creating a correlation spectrum. These bands are related to molecular vibrations of functional groups present in organic compounds (Coates, 2006).

In the visible spectrum, organic matter influences the red region, specifically around 665 nm (Ben-Dor, 1997; Nocita et al., 2015; Stevens et al., 2013). In this region, both reflectance and CR bands present covariables with a moderate and negative correlation, which means an indirect but proportional association with SOC. This indicates what is already well known in the literature regarding the effects of organic matter on soil darkening. These visible absorption features are related to the electronic transition (Coates, 2006; Soriano-Disla et al., 2014).

Among the LASSO-selected covariables in the MIR spectral region, the ones with the highest correlation with SOC were in the CR and STD domains. In the CR domain, the most important covariables were in three regions: approximately 3870, 3890, and 5450 nm. In the STD domain, bands between 4760 and 5880 nm and between 7140 and 7690 nm were selected and presented a moderate and positive correlation with SOC. As Barra et al. (2021) showed when synthesizing the work of Du and Zhou (2009), the spectral curves of soil samples in the MIR spectral region can be divided into four parts: from 2500 to 4000 nm, representing fundamental vibrations caused by stretching of O-H, C-H, and N-H; from 4000 to 5000 nm, representing vibrations of triple-bonds; from 5000 to 6666 nm, representing vibrations of double-bonds; and from 6666 to 25000 nm, representing the fingerprint region.

Many covariables in the MIR spectral region between ~2940 and 6660 nm can be attributed to bonds in organic molecules, such as aliphatic hydrocarbons, aromatic compounds, carboxylic acid, amines, amides, and other sulfur- and phosphorus-containing compounds (McDowell et al., 2012). Between ~2860 and 5000 nm, Silvero et al. (2020) identified, after chemically removing organic matter from soil samples, that organic matter masks mineral diagnostic features, as it reduces the intensity of reflectance in this region. These authors identified organic matter absorption features between ~3419 and 3517 nm and observed that, after chemical removal, the absorption features disappeared. In the present study, no absorption band (CR) was selected in this region;

however, in the STD domain, spectral covariables were selected and presented a moderate correlation.

In addition, bands between 7142 and 7692 nm were selected in the STD domain, with a high correlation with SOC. Among these bands, Calderón et al. (2011) identified the presence of esters, carboxylic acids, and aliphatic and phenolic compounds. Although bands between 5000 and 6666 nm were selected in the CR and STD domains, they could be ambiguous because both organic compounds, such as amides, carboxylic acids, and others containing single bonds ($C-H$), double bonds ($C=O$), and triple bonds ($C \equiv O$), as well as common silicate minerals, such as quartz and kaolinite, have features that overlap in this spectral region (Nguyen et al., 1991; Stuart, 2004).

We noted something interesting in the VNIR and MIR spectral regions: The largest number of covariables were selected in the SCD domain (Figure 2.1.8). However, as shown in Figure 2.1.9, there is a weak correlation with SOC; furthermore, many of the covariables are not significantly correlated. We applied a correlation filter to detect multicollinearity in the ρ domain. We detected 3,324 highly correlated bands (with $|r| \geq 0.95$), while in the SCD domain, there were only 135 in the MIR spectral region. LASSO was able to detect very important covariables for the SOC prediction problem, but it is necessary to consider that the method was also designed to deal with multicollinearity, which explains the number of covariables selected in the SCD domain.

In this study, we explored the potential for LASSO to generate more interpretable models through the selection of individual spectral covariates. However, in addition to LASSO, other penalty and parsimony methods can be investigated for soil attribute prediction using diffuse reflectance spectroscopy data. One such method is group LASSO, which is a generalization of LASSO specifically designed for applications where the selection and utilization of multiple interrelated factors (covariables) are important for prediction (Hastie et al., 2015; Yuan and Lin, 2006).

Soil is the largest terrestrial pool of organic carbon in the biosphere: It stores more carbon than plants and the atmosphere combined (Jobbágy and Jackson, 2000; Schlesinger, 1977; Yost and Hartemink, 2019). In the soil, the abundance of organic carbon affects and is affected by plant production, the main input of organic matter into the soil (via deposition of plant debris, previously formed via photosynthesis and atmospheric CO_2 assimilation; Wiesmeier et al., 2019). In addition, SOC plays a key role in controlling soil fertility and agricultural production, a phenomenon that has been recognized for more than a century (Jobbágy and Jackson, 2000), and SOC has an important role in the biogeochemical cycle of CO_2 (one of the main greenhouse gases; Hao et al., 2023). The importance of SOC has led to the development and adoption of sustainable management practices in agricultural areas as a strategy to remove greenhouse gases from the atmosphere (FAO, 2020; Lal et al., 2018; Paustian et al., 2019; Smith et al., 2020). Therefore, SOC monitoring plays a key role in surveying soil fertility and monitoring soil as a mediator of climate change.

National monitoring and reporting of SOC is becoming increasingly important to fulfill global conventions and mechanisms (FAO, 2020). Rapid, low-cost estimates of soil properties are considered imperative to monitor and report agricultural soil conditions and

to implement site-specific management practices (Angelopoulou et al., 2020) because traditional methods are expensive and consume a lot of chemical reagents. Indeed, Demattê et al. (2019a) argued that the high global demand for soil analyses would lead to high consumption of chemical reagents, such as dichromate, ammonium ferrous sulfate, and sulfuric acid, which are used to measure the SOC content. In this sense, visible and infrared soil spectroscopy is an alternative method that is relatively fast, low-cost, and non-destructive, to conventional SOC testing (Bellon-Maurel and McBratney, 2011; FAO, 2020; Viscarra Rossel et al., 2016).

Although soil spectroscopy has been recognized as an important method that can be combined with conventional laboratory analyses and produces faster and cheaper soil attribute estimates (Demattê et al., 2019a), the method faces skepticism from users (Wadoux, 2023). Among the problems pointed out in the literature are the underestimated generalization errors of ML models and the "black-box" nature of these models, in which analysts do not have direct access to the complex associations that regression methods make between the measured diffuse reflectance and the modeled soil attribute (McBride, 2022; Viscarra Rossel et al., 2022; Wadoux, 2023).

ML modeling is essential for the use of spectroscopy as a method to estimate soil attributes, but the trend of overfitting ML models has not been addressed in recent soil spectroscopy studies. The tendency to overfit is a problem present in several ML applications to model environmental variables, which lack an understanding of the variables of interest to be modeled (Aldossari et al., 2022; de Bruin et al., 2022; Misiuk and Brown, 2023). In the case of SOC, a variable whose structure is dependent on nearby layers in the same soil profile, many studies have not considered knowledge of the distribution of SOC throughout the profile when performing ML modeling. Therefore, they can present overfitting, as we detected in the trained models without LSPO CV. This is because the use of "pseudo-independent" validation samples overestimates the relative accuracy of the models (Brown et al., 2005).

Considering the interpretability of the ML models, or the understanding of the relationships between the soil attribute and the measured reflectance in the different bands and spectral regions, we have demonstrated the applicability of LASSO as a transparent method. With LASSO it was possible to directly study the relationship and correlation between the spectral bands selected to predict the SOC content. Wadoux (2023) also proposed ways to improve the interpretability of ML models (more complex than LASSO), such as the use of Shapley values in random forest models, but we have proposed a simple alternative. Thus, the present study contributes to improve knowledge regarding the uncertainties of estimating the SOC content with spectroscopy-based models and more transparent models that generate reliable SOC estimates.

2.1.5. Conclusions

The models trained using soil profile–oriented CV (LSPO CV) showed worse training performance than the models trained with k-fold CV; however, the models trained with LSPO CV had better generalizability. This indicates that disregarding the

autocorrelation of organic carbon within the soil profile can produce models that do not correctly detect patterns in the data and are more prone to overfitting.

When assessing the models in the testing phase, the LASSO models presented better generalizability than the PLS models. We attributed this outcome to the ability to efficiently deal with multicollinearity in predictors. In addition, many spectral bands that were selected by the LASSO method are correlated with SOC. Therefore, the possibility and ease of identifying spectral bands and their correlation with organic carbon indicate that the LASSO models presented a better accuracy-transparency trade-off than PLS models.

2.1.6. References

Adadi, A., & Berrada, M. (2018). Peeking inside the Black-Box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Aldossari, S., Husmeier, D., & Matthiopoulos, J. (2022). Transferable species distribution modelling: Comparative performance of generalised functional response models. *Ecological Informatics*, *71*, 101803. https://doi.org/10.1016/j.ecoinf.2022.101803

Angelopoulou, T., Balafoutis, A., Zalidis, G., & Bochtis, D. (2020). From laboratory to proximal sensing spectroscopy for soil organic carbon estimation—A review. *Sustainability*, *12*(2), Artigo 2. https://doi.org/10.3390/su12020443

Barra, I., Khiari, L., Haefele, S. M., Sakrabani, R., & Kebede, F. (2021). Optimizing setup of scan number in FTIR spectroscopy using the moment distance index and PLS regression: Application to soil spectroscopy. *Scientific Reports*, *11*(1), 13358. https://doi.org/10.1038/s41598-021-92858-w

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Bellon-Maurel, V., & McBratney, A. (2011). Near-infrared (NIR) and mid-infrared (MIR) spectroscopic techniques for assessing the amount of carbon stock in soils – Critical review and research perspectives. *Soil Biology and Biochemistry*, *43*(7), 1398–1410. https://doi.org/10.1016/j.soilbio.2011.02.019

Ben-Dor, E. (1997). The reflectance spectra of organic matter in the visible near-infrared and short wave infrared region (400–2500 nm) during a controlled decomposition process. *Remote Sensing of Environment*, *61*(1), 1–15. https://doi.org/10.1016/S0034-4257(96)00120-4

Boehmke, B., & Greenwell, B. (2019). *Hands-on machine learning with R*. Chapman and Hall/CRC. https://doi.org/10.1201/9780367816377

Bricklemyer, R. S., Brown, D. J., Turk, P. J., & Clegg, S. (2018). Comparing vis–NIRS, LIBS, and combined vis–NIRS-LIBS for intact soil core soil carbon measurement.

Soil Science Society of America Journal, *82*(6), 1482–1496. https://doi.org/10.2136/sssaj2017.09.0332

Brown, D. J., Bricklemyer, R. S., & Miller, P. R. (2005). Validation requirements for diffuse reflectance soil characterization models with a case study of VNIR soil C prediction in Montana. *Geoderma*, *129*(3), 251–267. https://doi.org/10.1016/j.geoderma.2005.01.001

Brungard, C., Nauman, T., Duniway, M., Veblen, K., Nehring, K., White, D., Salley, S., & Anchang, J. (2021). Regional ensemble modeling reduces uncertainty for digital soil mapping. *Geoderma*, *397*, 114998. https://doi.org/10.1016/j.geoderma.2021.114998

Calderón, F. J., Reeves, J. B., Collins, H. P., & Paul, E. A. (2011). Chemical differences in soil organic matter fractions determined by diffuse-reflectance mid-infrared spectroscopy. *Soil Science Society of America Journal*, *75*(2), 568–579. https://doi.org/10.2136/sssaj2009.0375

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*(3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

Clark, R. N., & Roush, T. L. (1984). Reflectance spectroscopy: Quantitative analysis techniques for remote sensing applications. *Journal of Geophysical Research: Solid Earth*, *89*(B7), 6329–6340. https://doi.org/10.1029/JB089iB07p06329

Coates, J. (2006). Interpretation of infrared spectra, a practical approach. In *Encyclopedia of analytical chemistry*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470027318.a5606

Davis, M. R., Alves, B. J. R., Karlen, D. L., Kline, K. L., Galdos, M., & Abulebdeh, D. (2018). Review of soil organic carbon measurement protocols: A US and Brazil comparison and recommendation. *Sustainability*, *10*(1), 1. https://doi.org/10.3390/su10010053

de Bruin, S., Brus, D. J., Heuvelink, G. B. M., van Ebbenhorst Tengbergen, T., & Wadoux, A. M. J.-C. (2022). Dealing with clustered samples for assessing map accuracy by cross-validation. *Ecological Informatics*, *69*, 101665. https://doi.org/10.1016/j.ecoinf.2022.101665

Demattê, J. A. M., Dotto, A. C., Bedin, L. G., Sayão, V. M., & Souza, A. B. e. (2019a). Soil analytical quality control by traditional and spectroscopy techniques: Constructing the future of a hybrid laboratory for low environmental impact. *Geoderma*, *337*, 111–121. https://doi.org/10.1016/j.geoderma.2018.09.010

Demattê, J. A. M., Dotto, A. C., Paiva, A. F. S., Sato, M. V., Dalmolin, R. S. D., de Araújo, M. do S. B., da Silva, E. B., Nanni, M. R., ten Caten, A., Noronha, N. C., Lacerda, M. P. C., de Araújo Filho, J. C., Rizzo, R., Bellinaso, H., Francelino, M. R., Schaefer, C. E. G. R., Vicente, L. E., dos Santos, U. J., de Sá Barretto Sampaio, E. V., … do Couto, H. T. Z. (2019b). The Brazilian Soil Spectral Library (BSSL): A general view, application and challenges. *Geoderma*, *354*, 113793. https://doi.org/10.1016/j.geoderma.2019.05.043

Demattê, J. A. M., & Terra, F. da S. (2014). Spectral pedology: A new perspective on evaluation of soils along pedogenetic alterations. *Geoderma*, *217–218*, 190–200. https://doi.org/10.1016/j.geoderma.2013.11.012

Dias, S. H. B., Filgueiras, R., Fernandes Filho, E. I., Arcanjo, G. S., Silva, G. H. da, Mantovani, E. C., Cunha, F. F. da, Filho, E. I. F., Arcanjo, G. S., Da Silva, G. H., Mantovani, E. C., & Da Cunha, F. F. (2021). Reference evapotranspiration of Brazil modeled with machine learning techniques and remote sensing. *PLoS One*, *16*(2), e0245834. https://doi.org/10.1371/journal.pone.0245834

Du, C., & Zhou, J. (2009). Evaluation of soil fertility using infrared spectroscopy – A review. In E. Litchfouse (Ed.), *Climate Change, intercropping, pest control and beneficial microorganisms* (pp. 453–483). Springer Netherlands. https://doi.org/10.1007/978-90-481-2716-0_16

Dyar, M. D., Carmosino, M. L., Breves, E. A., Ozanne, M. V., Clegg, S. M., & Wiens, R. C. (2012). Comparison of partial least squares and lasso regression techniques as applied to laser-induced breakdown spectroscopy of geological samples. *Spectrochimica Acta Part B: Atomic Spectroscopy*, *70*, 51–67. https://doi.org/10.1016/j.sab.2012.04.011

El Naqa, I., Ruan, D., Valdes, G., Dekker, A., McNutt, T., Ge, Y., Wu, Q. J., Oh, J. H., Thor, M., Smith, W., Rao, A., Fuller, C., Xiao, Y., Manion, F., Schipper, M., Mayo, C., Moran, J. M., & Ten Haken, R. (2018). Machine learning and modeling: Data, validation, communication challenges. *Medical Physics*, *45*(10), e834–e840. https://doi.org/10.1002/mp.12811

FAO. (2020). *A protocol for measurement, monitoring, reporting and verification of soil organic carbon in agricultural landscapes: GSOC-MRV Protocol*. FAO. https://doi.org/10.4060/cb0509en

Francos, N., Ogen, Y., & Ben-Dor, E. (2021). Spectral assessment of organic matter with different composition using reflectance spectroscopy. *Remote Sensing 2021, Vol. 13, Page 1549*, *13*(8), 1549. https://doi.org/10.3390/RS13081549

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22. https://doi.org/10.18637/jss.v033.i01

Gehl, R. J., & Rice, C. W. (2007). Emerging technologies for in situ measurement of soil carbon. *Climatic Change*, *80*(1), 43–54. https://doi.org/10.1007/s10584-006-9150-2

Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: A tutorial. *Analytica Chimica Acta*, *185*(9), 1–17. https://doi.org/10.1016/0003-2670(86)80028-9

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, 80–89. https://doi.org/10.1109/DSAA.2018.00018

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

Greenwell, B., Boehmke, B., & Gray, B. (2020). *vip: Variable Importance Plots*. https://cran.r-project.org/package=vip

Hao, W., Xia, B., Li, J., & Xu, M. (2023). Deep soil CO2 flux with strong temperature dependence contributes considerably to soil-atmosphere carbon flux. *Ecological Informatics*, *74*, 101957. https://doi.org/10.1016/j.ecoinf.2022.101957

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity: The lasso and generalizations*. CRC Press. https://doi.org/10.1201/b18401

Heuvelink, G. B. M., Angelini, M. E., Poggio, L., Bai, Z., Batjes, N. H., Bosch, R., Bossio, D., Estella, S., Lehmann, J., Olmedo, G. F., & Sanderman, J. (2021). Machine learning in space and time for modelling soil organic carbon change. *European Journal of Soil Science*, *72*(4), 1607–1623. https://doi.org/10.1111/ejss.12998

Huang, H., Luo, F., Liu, J., & Yang, Y. (2015). Dimensionality reduction of hyperspectral images based on sparse discriminant manifold embedding. *ISPRS Journal of Photogrammetry and Remote Sensing*, *106*, 42–54. https://doi.org/10.1016/j.isprsjprs.2015.04.015

Hughes, G. F. (1968). On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory*, *14*(1), 55–63. https://doi.org/10.1109/TIT.1968.1054102

Hutengs, C., Seidel, M., Oertel, F., Ludwig, B., & Vohland, M. (2019). In situ and laboratory soil spectroscopy with portable visible-to-near-infrared and mid-infrared instruments for the assessment of organic carbon in soils. *Geoderma*, *355*, 113900. https://doi.org/10.1016/j.geoderma.2019.113900

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. In *Performance evaluation* (Vol. 103, pp. 9–12). Springer. https://doi.org/10.1007/978-1-4614-7138-7

Janik, L., & Skjemstad, J. (1995). Characterization and analysis of soils using mid-infrared partial least-squares .2. Correlations with some laboratory data. *Soil Research*, *33*(4), 637. https://doi.org/10.1071/SR9950637

Ji, L., & Gallo, K. (2006). An agreement coefficient for image comparison. *Photogrammetric Engineering & Remote Sensing*, *72*(7), 823–833. https://doi.org/10.14358/PERS.72.7.823

Jobbágy, E. G., & Jackson, R. B. (2000). The vertical distribution of soil organic carbon and its relation to climate and vegetation. *Ecological Applications*, *10*(2), 423–436. https://doi.org/10.1890/1051-0761(2000)010[0423:TVDOSO]2.0.CO;2

Knadel, M., Thomsen, A., Schelde, K., & Greve, M. H. (2015). Soil organic carbon and particle sizes mapping using vis–NIR, EC and temperature mobile sensor platform. *Computers and Electronics in Agriculture*, *114*, 134–144. https://doi.org/10.1016/j.compag.2015.03.013

Kuhn, M. (2020). *caret: Classification and regression training*. https://cran.r-project.org/package=caret

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer. https://doi.org/10.1007/978-1-4614-6849-3

Lal, R., Smith, P., Jungkunst, H. F., Mitsch, W. J., Lehmann, J., Nair, P. K. R., McBratney, A. B., Sá, J. C. de M., Schneider, J., Zinn, Y. L., Skorupa, A. L. A., Zhang, H.-L., Minasny, B., Srinivasrao, C., & Ravindranath, N. H. (2018). The carbon sequestration potential of terrestrial ecosystems. *Journal of Soil and Water Conservation*, *73*(6), 145A-152A. https://doi.org/10.2489/jswc.73.6.145A

Lehnert, L. W., Meyer, H., Obermeier, W. A., Silva, B., Regeling, B., & Bendix, J. (2019). Hyperspectral data analysis in R: The hsdar package. *Journal of Statistical Software*, *89*(12). https://doi.org/10.18637/jss.v089.i12

Liland, K. H., Mevik, B.-H., & Wehrens, R. (2021). *pls: Partial least squares and principal component regression*. https://cran.r-project.org/package=pls

Liu, L., Ji, M., & Buchroithner, M. (2017). Combining partial least squares and the gradient-boosting method for soil property retrieval using visible near-infrared shortwave infrared spectra. *Remote Sensing*, *9*(12), 1299. https://doi.org/10.3390/rs9121299

Liu, L., Ji, M., Dong, Y., Zhang, R., & Buchroithner, M. (2016). Quantitative retrieval of organic soil properties from visible near-infrared shortwave infrared (Vis-NIR-SWIR) spectroscopy using fractal-based feature extraction. *Remote Sensing*, *8*(12), 1035. https://doi.org/10.3390/rs8121035

Liu, S., Shen, H., Chen, S., Zhao, X., Biswas, A., Jia, X., Shi, Z., & Fang, J. (2019). Estimating forest soil organic carbon content using vis-NIR spectroscopy: Implications for large-scale soil carbon spectroscopic assessment. *Geoderma*, *348*, 37–44. https://doi.org/10.1016/j.geoderma.2019.04.003

Malmir, M., Tahmasbian, I., Xu, Z., Farrar, M. B., & Bai, S. H. (2019). Prediction of soil macro- and micro-elements in sieved and ground air-dried soils using laboratory-based hyperspectral imaging technique. *Geoderma*, *340*, 70–80. https://doi.org/10.1016/j.geoderma.2018.12.049

Mariotto, I., Thenkabail, P. S., Huete, A., Slonecker, E. T., & Platonov, A. (2013). Hyperspectral versus multispectral crop-productivity modeling and type discrimination for the HyspIRI mission. *Remote Sensing of Environment*, *139*, 291–305. https://doi.org/10.1016/J.RSE.2013.08.002

McBride, M. B. (2022). Estimating soil chemical properties by diffuse reflectance spectroscopy: Promise versus reality. *European Journal of Soil Science*, *73*(1), e13192. https://doi.org/10.1111/ejss.13192

McDowell, M. L., Bruland, G. L., Deenik, J. L., Grunwald, S., & Knox, N. M. (2012). Soil total carbon analysis in Hawaiian soils with visible, near-infrared and mid-infrared diffuse reflectance spectroscopy. *Geoderma*, *189–190*, 312–320. https://doi.org/10.1016/j.geoderma.2012.06.009

Meyer, H., Kühnlein, M., Appelhans, T., & Nauss, T. (2016). Comparison of four machine learning algorithms for their applicability in satellite-based optical

rainfall retrievals. *Atmospheric Research*, *169*, 424–433. https://doi.org/10.1016/j.atmosres.2015.09.021

Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., & Nauss, T. (2018). Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. *Environmental Modelling and Software*, *101*, 1–9. https://doi.org/10.1016/j.envsoft.2017.12.001

Misiuk, B., & Brown, C. J. (2023). Improved environmental mapping and validation using bagging models with spatially clustered data. *Ecological Informatics*, *77*, 102181. https://doi.org/10.1016/j.ecoinf.2023.102181

Morellos, A., Pantazi, X.-E., Moshou, D., Alexandridis, T., Whetton, R., Tziotzios, G., Wiebensohn, J., Bill, R., & Mouazen, A. M. (2016). Machine learning based prediction of soil total nitrogen, organic carbon and moisture content by using VIS-NIR spectroscopy. *Biosystems Engineering*, *152*, 104–116. https://doi.org/10.1016/j.biosystemseng.2016.04.018

Moura-Bueno, J. M., Dalmolin, R. S. D., ten Caten, A., Dotto, A. C., & Demattê, J. A. M. (2019). Stratification of a local VIS-NIR-SWIR spectral library by homogeneity criteria yields more accurate soil organic carbon predictions. *Geoderma*, *337*, 565–581. https://doi.org/10.1016/j.geoderma.2018.10.015

Nguyen, T., Janik, L., & Raupach, M. (1991). Diffuse reflectance infrared Fourier transform (DRIFT) spectroscopy in soil studies. *Soil Research*, *29*(1), 49. https://doi.org/10.1071/SR9910049

Nocita, M., Stevens, A., van Wesemael, B., Aitkenhead, M., Bachmann, M., Barthès, B., Ben Dor, E., Brown, D. J., Clairotte, M., Csorba, A., Dardenne, P., Demattê, J. A. M., Genot, V., Guerrero, C., Knadel, M., Montanarella, L., Noon, C., Ramirez-Lopez, L., Robertson, J., … Wetterlind, J. (2015). Soil spectroscopy: An alternative to wet chemistry for soil monitoring. *Advances in Agronomy*, *132*, 139–159. https://doi.org/10.1016/bs.agron.2015.02.002

Onyutha, C. (2020). From R-squared to coefficient of model accuracy for assessing "goodness-offit". *Geoscientific Model Development Discussions*, 1–25. https://doi.org/10.5194/gmd-2020-51

Paiva, A. F. da S., Poppiel, R. R., Rosin, N. A., Greschuk, L. T., Rosas, J. T. F., & Demattê, J. A. M. (2022). The Brazilian Program of soil analysis via spectroscopy (ProBASE): Combining spectroscopy and wet laboratories to understand new technologies. *Geoderma*, *421*, 115905. https://doi.org/10.1016/j.geoderma.2022.115905

Paustian, K., Collier, S., Baldock, J., Burgess, R., Creque, J., DeLonge, M., Dungait, J., Ellert, B., Frank, S., Goddard, T., Govaerts, B., Grundy, M., Henning, M., Izaurralde, R. C., Madaras, M., McConkey, B., Porzig, E., Rice, C., Searle, R., … Jahn, M. (2019). Quantifying carbon for agricultural soil management: From the current status toward a global soil information system. *Carbon Management*, *10*(6), 567–587. https://doi.org/10.1080/17583004.2019.1633231

Pechanec, V., Purkyt, J., Benc, A., Nwaogu, C., Štěrbová, L., & Cudlín, P. (2018). Modelling of the carbon sequestration and its prediction under climate change. *Ecological Informatics*, *47*, 50–54. https://doi.org/10.1016/j.ecoinf.2017.08.006

Piikki, K., Wetterlind, J., Söderström, M., & Stenberg, B. (2021). Perspectives on validation in digital soil mapping of continuous attributes—A review. *Soil Use and Management*, *37*(1), 7–21. https://doi.org/10.1111/sum.12694

Poggio, M., Brown, D. J., & Bricklemyer, R. S. (2017). Comparison of Vis–NIR on in situ, intact core and dried, sieved soil to estimate clay content at field to regional scales. *European Journal of Soil Science*, *68*(4), 434–448. https://doi.org/10.1111/ejss.12434

R Core Team, R. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org/

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*(8), 913–929. https://doi.org/10.1111/ECOG.02881

Roudier, P., Hedley, C. B., Lobsey, C. R., Viscarra Rossel, R. A., & Leroux, C. (2017). Evaluation of two methods to eliminate the effect of water from soil vis–NIR spectra for predictions of organic carbon. *Geoderma*, *296*, 98–107. https://doi.org/10.1016/j.geoderma.2017.02.014

Santos, U. J. dos, Demattê, J. A. de M., Menezes, R. S. C., Dotto, A. C., Guimarães, C. C. B., Alves, B. J. R., Primo, D. C., & Sampaio, E. V. de S. B. (2020). Predicting carbon and nitrogen by visible near-infrared (Vis-NIR) and mid-infrared (MIR) spectroscopy in soils of Northeast Brazil. *Geoderma Regional*, *23*, e00333. https://doi.org/10.1016/j.geodrs.2020.e00333

Savage, N. (2022). Breaking into the black box of artificial intelligence. *Nature*. https://doi.org/10.1038/d41586-022-00858-1

Schlesinger, W. H. (1977). Carbon balance in terrestrial detritus. *Annual Review of Ecology and Systematics*, *8*(1), 51–81. https://doi.org/10.1146/annurev.es.08.110177.000411

Seidel, M., Hutengs, C., Ludwig, B., Thiele-Bruhn, S., & Vohland, M. (2019). Strategies for the efficient estimation of soil organic carbon at the field scale with vis-NIR spectroscopy: Spectral libraries and spiking vs. local calibrations. *Geoderma*, *354*, 113856. https://doi.org/10.1016/j.geoderma.2019.07.014

Silvero, N. E. Q., Di Raimo, L. A. D. L., Pereira, G. S., Magalhães, L. P. de, Terra, F. da S., Dassan, M. A. A., Salazar, D. F. U., & Demattê, J. A. M. (2020). Effects of water, organic matter, and iron forms in mid-IR spectra of soils: Assessments from laboratory to satellite-simulated data. *Geoderma*, *375*, 114480. https://doi.org/10.1016/j.geoderma.2020.114480

Siqueira, R. G., Veloso, G. V., Fernandes-Filho, E. I., Francelino, M. R., Schaefer, C. E. G. R., & Corrêa, G. R. (2022). Evaluation of machine learning algorithms to

classify and map landforms in Antarctica. *Earth Surface Processes and Landforms*, *47*(2), 367–382. https://doi.org/10.1002/esp.5253

Smith, P., Soussana, J., Angers, D., Schipper, L., Chenu, C., Rasse, D. P., Batjes, N. H., Egmond, F., McNeill, S., Kuhnert, M., Arias-Navarro, C., Olesen, J. E., Chirinda, N., Fornara, D., Wollenberg, E., Álvaro-Fuentes, J., Sanz-Cobena, A., & Klumpp, K. (2020). How to measure, report and verify soil carbon change to realize the potential of soil carbon sequestration for atmospheric greenhouse gas removal. *Global Change Biology*, *26*(1), 219–241. https://doi.org/10.1111/gcb.14815

Soriano-Disla, J. M., Janik, L. J., Viscarra Rossel, R. A., Macdonald, L. M., & McLaughlin, M. J. (2014). The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties. *Applied Spectroscopy Reviews*, *49*(2), 139–186. https://doi.org/10.1080/05704928.2013.811081

Stevens, A., Nocita, M., Tóth, G., Montanarella, L., & van Wesemael, B. (2013). Prediction of soil organic carbon at the European scale by visible and near infrared reflectance spectroscopy. *PLoS One*, *8*(6), e66409. https://doi.org/10.1371/journal.pone.0066409

Stuart, B. H. (2004). *Infrared spectroscopy: Fundamentals and applications*. John Wiley & Sons, Ltd. https://doi.org/10.1002/0470011149

Tay, J. K., Narasimhan, B., & Hastie, T. (2023). Elastic net regularization paths for all generalized linear models. *Journal of Statistical Software*, *106*, 1–31. https://doi.org/10.18637/jss.v106.i01

Terra, F. S., Viscarra Rossel, R. A., & Demattê, J. A. M. (2019). Spectral fusion by Outer Product Analysis (OPA) to improve predictions of soil organic C. *Geoderma*, *335*, 35–46. https://doi.org/10.1016/j.geoderma.2018.08.005

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tibshirani, R. (1997). The LASSO method for variable selection in the Cox model. *Statistics in Medicine*, *16*(4), 385–395. https://doi.org/10.1002/(SICI)1097-0258(19970228)16:4<385::AID-SIM380>3.0.CO;2-3

Tsai, F., & Philpot, W. (1998). Derivative analysis of hyperspectral data. *Remote Sensing of Environment*, *66*(1), 41–51. https://doi.org/10.1016/S0034-4257(98)00032-7

Vasques, G. M., Grunwald, S., & Harris, W. G. (2010). Spectroscopic models of soil organic carbon in Florida, USA. *Journal of Environmental Quality*, *39*(3), 923–934. https://doi.org/10.2134/jeq2009.0314

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., … Ji, W. (2016). A global spectral library to characterize the world's soil. *Earth-Science Reviews*, *155*, 198–230. https://doi.org/10.1016/j.earscirev.2016.01.012

Viscarra Rossel, R. A., Behrens, T., Ben-Dor, E., Chabrillat, S., Demattê, J. A. M., Ge, Y., Gomez, C., Guerrero, C., Peng, Y., Ramirez-Lopez, L., Shi, Z., Stenberg, B., Webster, R., Winowiecki, L., & Shen, Z. (2022). Diffuse reflectance spectroscopy for estimating soil properties: A technology for the 21st century. *European Journal of Soil Science*, *73*(4), e13271. https://doi.org/10.1111/ejss.13271

Wadoux, A. M. J.-C. (2023). Interpretable spectroscopic modelling of soil with machine learning. *European Journal of Soil Science*, *74*(3), e13370. https://doi.org/10.1111/ejss.13370

Wiesmeier, M., Urbanski, L., Hobley, E., Lang, B., von Lützow, M., Marin-Spiotta, E., van Wesemael, B., Rabot, E., Ließ, M., Garcia-Franco, N., Wollschläger, U., Vogel, H.-J., & Kögel-Knabner, I. (2019). Soil organic carbon storage as a key function of soils—A review of drivers and indicators at various scales. *Geoderma*, *333*, 149–162. https://doi.org/10.1016/j.geoderma.2018.07.026

Willmott, C. J. (1981). On the validation of models. *Physical Geography*, *2*(2), 184–194. https://doi.org/10.1080/02723646.1981.10642213

Willmott, C. J. (1982). Some comments on the evaluation of model performance. *Bulletin of the American Meteorological Society*, *63*(11), 1309–1313. https://doi.org/10.1175/1520-0477(1982)063<1309:SCOTEO>2.0.CO;2

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, *30*(1), 79–82. https://doi.org/10.3354/cr030079

Willmott, C. J., Robeson, S. M., & Matsuura, K. (2012). A refined index of model performance. *International Journal of Climatology*, *32*(13), 2088–2094. https://doi.org/10.1002/joc.2419

Yost, J. L., & Hartemink, A. E. (2019). Soil organic carbon in sandy soils: A review. Em *Advances in Agronomy*, *158*, 217–310. https://doi.org/10.1016/bs.agron.2019.07.004

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *68*(1), 49–67. https://doi.org/10.1111/j.1467-9868.2005.00532.x

Yun, Y.-H., Li, H.-D., Deng, B.-C., & Cao, D.-S. (2019). An overview of variable selection methods in multivariate analysis of near-infrared spectra. *TrAC Trends in Analytical Chemistry*, *113*, 102–115. https://doi.org/10.1016/j.trac.2019.01.018

Zambrano-Bigiarini, M. (2017). *HydroGOF: Goodness-of-fit functions for comparison of simulated and observed hydrological time series*. https://cran.r-project.org/package=hydroGOF

Zhao, P., & Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, *7*, 2541–2563.

2.2. Article 2 – Sentinel-1 imagery used for estimation of soil organic carbon by dual-polarization SAR vegetation indices [2]

Abstract: Despite optical remote sensing (and the spectral vegetation indices) contributions to digital soil-mapping studies of soil organic carbon (SOC), few studies have used active radar remote sensing mission data like that from synthetic aperture radar (SAR) sensors to predict SOC. Bearing in mind the importance of SOC mapping for agricultural, ecological, and climate interests and also the recently developed methods for vegetation monitoring using Sentinel-1 SAR data, in this work, we aimed to take advantage of the high operationality of Sentinel-1 imaging to test the accuracy of SOC prediction at different soil depths using machine learning systems. Using linear, nonlinear, and tree regression-based methods, it was possible to predict the SOC content of soils from western Bahia, Brazil, a region with predominantly sandy soils, using as explanatory variables the SAR vegetation indices. The models fed with SAR sensor polarizations and vegetation indices produced more accurate results for the topsoil layers (0 – 5 cm and 5 – 10 cm in depth). In these superficial layers, the models achieved an RMSE in the order of 5.0 g kg^{-1} and an R^2 ranging from 0.16 to 0.24, therefore explaining about 20% of SOC variability using only Sentinel-1 predictors.

Keywords: synthetic aperture radar; machine learning; Brazilian Cerrado; digital soil mapping.

---

[2] Article published in **Remote Sensing**, 2023, volume 15(23), p. 5464,
DOI: https://doi.org/10.3390/rs15235464

2.2.1. Introduction

Soil organic carbon (SOC) plays an important role in several ecological and agricultural processes related to physical, chemical, biological, and soil fertility properties and the biogeochemical carbon cycle in the Earth system [1]. Unfrozen soils are the third carbon pool in the Earth system [2] in the mineralized form ($CaCO_3$) and mainly in the organic form (SOC). Given the importance of SOC, efforts have been made to improve the mapping of the spatial variability of SOC [3–6], which aims to better understand its distribution in order to manage it as a finite natural resource.

In digital soil-mapping studies, the use of spectral vegetation indices is not new [4,7–9]. The SCORPAN factors themselves, which describe the variation of soil in the landscape as a function of the soil itself, climate, organisms, relief, parent material (lithology), soil age, and its spatial position [10], employ the optical vegetation indices as indicators of organisms since the main organisms that contribute to soil variation and formation in the landscape are vegetation and humans, although other organisms have driving effects on soil at the local scale [10,11].

Several approaches have used machine learning systems fed with images captured by sensors onboard aircraft or satellites to represent organisms (or the soil itself) and predict SOC contents and/or stocks. This is the case of studies such as those of Keskin et al. [8] and Odebiri et al. [9], which used images from Landsat 8, Landsat 7 ETM+, and MODIS satellite missions, and Guo et al. [7], who used an aircraft-embedded hyperspectral sensor. In these cases, the images captured one of two conditions: (1) the bare soil condition, e.g., Guo et al. [7], who predicted SOC in the topsoil layer using hyperspectral imagery, and (2) the vegetated land cover condition, e.g., in Odebiri et al. [9], who used vegetation indices and the multispectral bands to predict SOC. However, few studies have applied measurements from microwave sensors such as synthetic aperture radar (SAR) to predict SOC.

Few studies have examined the relationship between SAR sensor measurements of the Earth's surface with SOC. This may be due to several factors, including the complexity of SAR image processing and the difficulty in explaining the interactions of microwave electromagnetic radiation with the Earth's surface [12,13]. However, it may also be attributed to the limited availability of SAR data given that the number of operational SAR imaging missions has increased in the last decade [14].

Bartsch et al. [15] proposed the use of C-band measurements (with a wavelength of approximately 5.4 GHz) to quantify SOC stocks in circumpolar soils in the tundra biome. The authors exploited measurements from the Advanced SAR (ASAR) sensor onboard the ENVISAT orbital mission (mission finished in 2013), with the goal of improving SOC estimates and mapping. Bartsch et al. [15] concluded that near-surface SOC can be quantified with C-band SAR data for arctic and subarctic environments for non-peatland areas.

Following the work of Bartsch et al. [15], other works have also used SAR image polarizations to predict and/or map SOC in tropical and subtropical regions [16–20]. In

the cited studies, the authors mainly used images from the Sentinel-1 SAR orbital constellation (which has C-band SAR sensors), apart from the work by Ceddia et al. [16], which used images from the L-band ALOS PALSAR sensor (with wave frequency of ~ 1.2 GHz). The authors used the images separately and in conjunction with other environmental covariates but only reported using the dual polarization that the Sentinel-1 sensor provides (VH and VV polarizations).

SAR sensors are active and coherent sensors, meaning that they emit pulses of radiation that travel to the Earth's surface and are backscattered by targets, and sensors detect the backscatter at only one wavelength [12,13]. Since SAR sensors are monochromatic, interactions with different targets take place through the different polarizations the sensor is capable of measuring. The polarizations commonly adopted are HH (this emits and detects horizontally polarized waves), HV, VV, and VV. Due to the different backscattering mechanisms that may occur in a single scene in different polarizations, some authors have proposed vegetation indices for SAR images.

The first SAR vegetation index developed was the RVI (radar vegetation index) [21], which combines the fully polarimetric data (HH, HV, VV, and VH) of L-band images. The RVI was modified by Chang et al. [22], who proposed the polarized RVI (PRVI), which includes the metric degree of polarization (DoP). As for C-band sensors, mainly Sentinel-1, the first index proposed was the dual-polarization SAR vegetation index (DPSVI), which uses empirical relationships between Sentinel-1's VV and VH polarizations to quantify the biomass of crops and other vegetation landforms [23]. Considering C-band signal saturation in forest areas, dos Santos et al. [24] made modifications to DPSVI to improve the index sensitivity for areas of dense biomass, proposing the DPSVIm. In addition, Mandal et al. [25] proposed the dual-polarization RVI (DpRVI) to monitor annual crops' phenology using the DoP concept for C-band. Despite their characteristics, all these indices have in common the principle of measuring changes in the polarization of the radiation signal when interacting with volumetric and complex objects such as vegetation.

None of the cited SAR vegetation indices have been tested or used for SOC prediction, and the digital soil-mapping studies that used images from SAR sensors used only the original polarization. Considering that for C-band sensors, the radar signal interacts little with the soil under vegetation [12], this work aims to test the accuracy of SOC prediction for different land-use and land-cover classes and at different soil depths using different regression methods fed with the Sentinel-1 SAR dual-polarization images and their vegetation indices.

2.2.2. Materials and Methods

2.2.2.1. Study Area and Field Data Collection

The field soil data survey was conducted in the hydrographic basins of the rivers Grande, Corrente, and Carinhanha, in the western region of the state of Bahia, Brazil

(Figure 2.2.1). The study area, located in the Cerrado biome, has a tropical climate with dry winter (climate Aw, in Köppen climate typology, Alvares et al. [26]) (Figure S1 of the Supplementary Material). The mean rainfall precipitation in the region (1980 to 2015) is 1060 mm $year^{-1}$, with great seasonality: There is low precipitation in the driest months (10 mm $month^{-1}$ in June, July, and August) and high precipitation in December and January (150 to 200 mm $month^{-1}$) [27]. The study area is located in a transition zone between the Cerrado biome (with annual precipitation greater than 1200 mm $year^{-1}$) and the Caatinga biome (with precipitation < 800 mm $year^{-1}$) [27].

The relief is predominantly flat (Figure S2). On the western edge of the basins, elevation reaches more than 1000 m above sea level, while at the mouth of the basin (easternmost region), the elevation is around 380 m. The landscape in the region, especially on the western border, is composed of formations known as Chapadões [28]. These relief characteristics favor agricultural mechanization and the use of chemical inputs in agriculture, so western Bahia is an important producing region of temporary crops (corn, soybeans, cotton, etc.) and constitutes the most recent Brazilian agricultural frontier, the MATOPIBA (acronym formed from the abbreviations of the states of Maranhão, Tocantins, Piauí, and Bahia) [27].

The soils of the region are formed on the geological formation of the Urucuia Group (dated to the Upper Cretaceous), whose composition includes quartz arenites, sandstones, and argillites [29]. The soils under study are predominantly sandy (with clay contents lower than 40% and total sand contents higher than 70%; see Figure 2.2.2b). According to the soil mapping of Brazil (Figure S3), following the criteria of the Brazilian Soil Classification System [30] and its correspondence with the World Reference Base for Soil Resources, there is a predominance in the region of Ferralsols, Arenosols, and Leptosols.

Soil-profile samples were collected from six different land-use and land-cover classes (LULC, observed on site), these being Cerrado (CDO: 23 sampled profiles), forest formation (FOR, 19 profiles), rainfed agriculture (RAG, 20 profiles), irrigated agriculture (IRR, 20 profiles), pastureland (PAST, 21 profiles), and area of recently suppressed vegetation for agricultural purposes (ASV, 20 profiles) [28], totaling 123 soil profiles. RAG and IRR areas are used to grow soybean (*Glycine max*), corn (*Zea mays*), cotton (*Gossypium* spp.), and beans (*Phaseolus vulgaris*). The CDO class is represented by Cerrado sensu stricto (savanna) and grasslands, while the FOR class areas are represented by forest and Cerradão (dense savanna forest) strata. The PAST class includes well-managed cultivated pastures and degraded pastures. Finally, the ASV class represents areas with recently suppressed vegetation. Images of the LULC classes of the sample points can be seen in [31].

In the IRR class, the samples were collected in the middle of the growing season in July 2017, while the RAG class samples were collected at the beginning of the growing season in November and December 2017. For the CDO, FOR, PAST, and ASV classes, the samples were also collected in November and December 2017.

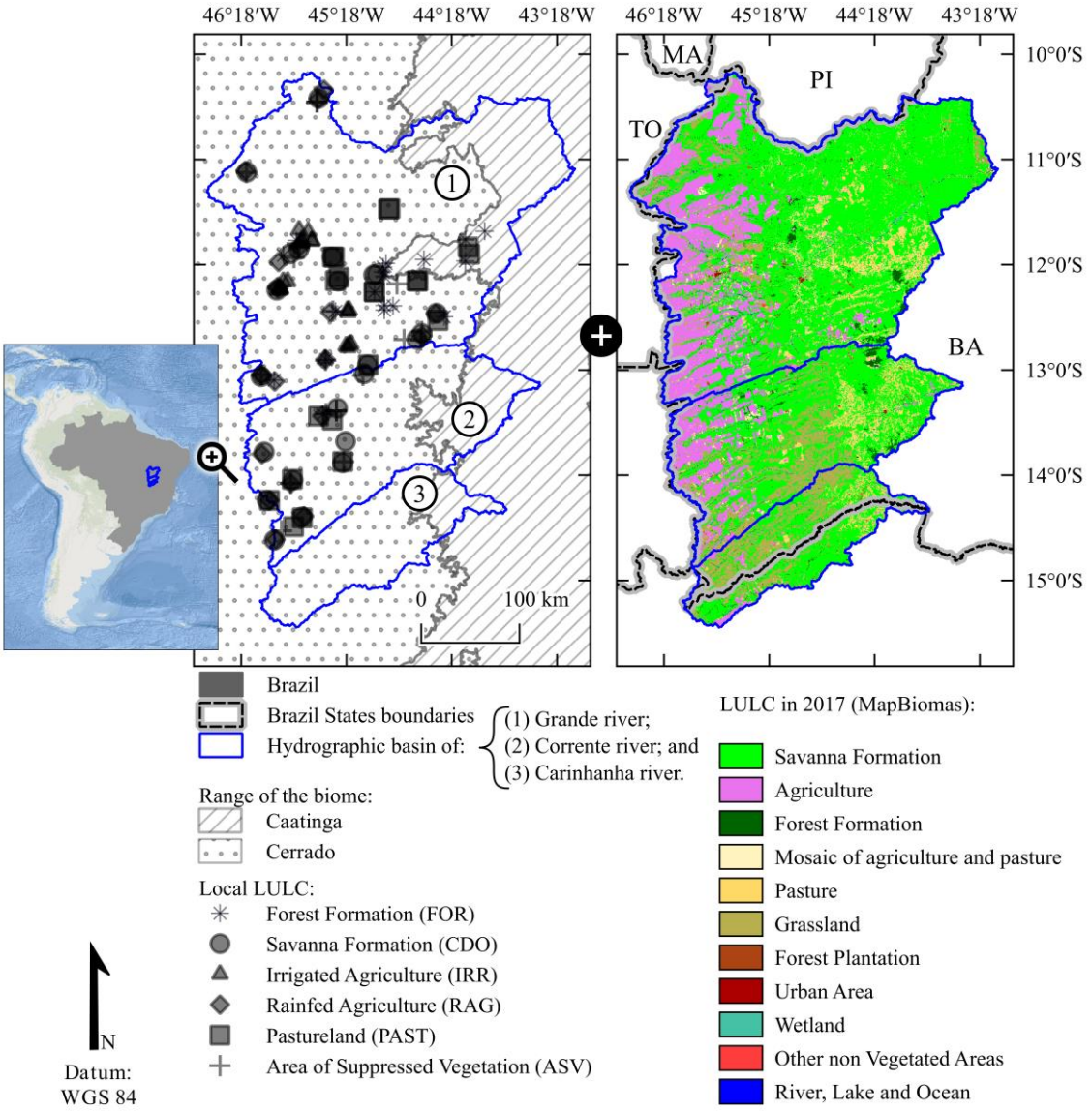


Figure 2.2.1. Location of the sampled soil profiles in the study area and their respective LULC classes observed on site, accompanied by the LULC map of the basins (MapBiomas Collection 7.0 [32]). Hydrographic basins boundaries: Brazilian National Agency for Water and Basic Sanitation (ANA, https://dadosabertos.ana.gov.br/ (accessed on 24 April 2023)); geopolitical divisions: Brazilian Institute for Geography and Statistics (IBGE, https://portaldemapas.ibge.gov.br/ (accessed on 24 April 2023)).

Deformed soil samples were collected at seven depths in the 123 soil profiles: 0–5, 5–10, 10–15, 15–20, 20–40, 40–60, and 60–100 cm, totaling 861 samples. Each sample was composed of three subsamples, which were homogenized in the field. These samples were analyzed to quantify particle size fractions (fine sand, coarse sand, clay, and silt contents (g kg^{-1}) of the superficial layer) and soil organic carbon (SOC (g kg^{-1}) of all 861 samples), whose analytical determination was carried out by the Walkley–Black colorimetric method [33]. The histogram of the SOC distribution and the texture

distribution of the studied soils can be seen in Figure 2.2.2a and Figure 2.2.2b, respectively.

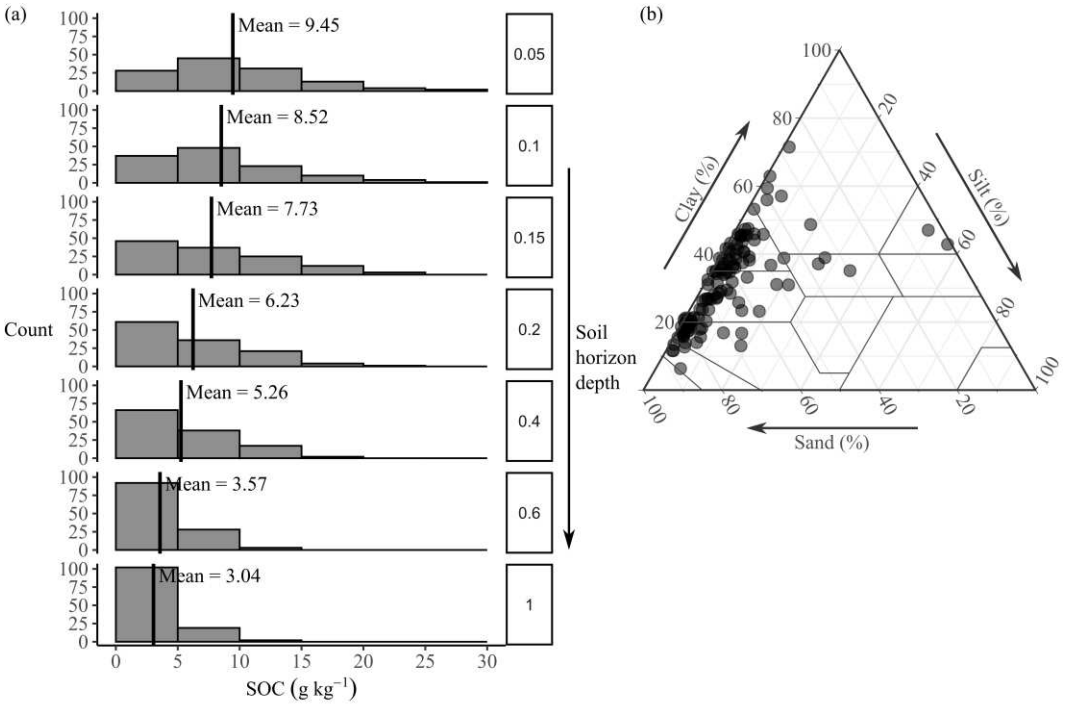


Figure 2.2.2. Histograms of soil organic carbon (SOC) contents in each soil layer (**a**) and soil texture diagram displaying the texture distribution of the surface layer (0–5 cm) of the sampled points (**b**).

2.2.2.2. Remote Sensing Data Acquisition and Processing

The remote sensing data used in this study were derived from images from the SAR sensor onboard the orbital platform of the Sentinel-1A mission, the first satellite of the Sentinel-1 constellation of the European Space Agency (ESA). The Sentinel-1 mission satellites operate with SAR-type imaging radar sensors, a category of active sensors, and in the case of the Sentinel-1 satellites, the SAR sensors carry a C-band radar ($\lambda \cong 5.4$ GHz) [34].

Dual-polarization images from the interferometric wide swath (IW) beam mode, which are images preprocessed with only the observed wave amplitude information, called GRD (ground range detected) products, were used. The Sentinel-1 bands with the greatest global coverage are the dual-polarization VH (where the sensor emits a pulse of radiation in vertical polarization and measures the detected reflectivity in horizontal polarization) and VV (emission and detection in vertical polarization). Sentinel-1 IW GRD images are formed after the sensor scans the Earth's surface in three sub-swaths, i.e., IW1, IW2, and IW3, with ellipsoid incidence angles of 32.9, 28.3, and 43.1 and azimuthal resolutions (spatial resolution concerning the direction of flight of the satellite) of 22.4, 22.5, and 22.6 m, respectively [35].

The Sentinel-1 IW GRD images used in this work are detailed in Table 2.2.1 and were obtained from the Alaska Satellite Facility (ASF) portal (https://asf.alaska.edu/, [36]). Although the products available on the ASF are not analysis-ready data like the Sentinel-1 imagery available via Google Earth Engine, the time series of radar imagery available on the ASF are readily available for download, which is no longer the case for the official ESA portal (the Copernicus Open-Access Hub) since the implementation of the long-term archive policy.

Table 2.2.1. Inventory of Sentinel-1 IW GRD images used in the study, generated by the SAR sensor of the Sentinel-1A satellite.

Acquisition Date	Product Unique Identifier	Relative Orbit Number
6 July 2017	B8D5	126
6 July 2017	0966	126
1 November 2017	82D6	126
1 November 2017	D608	126
1 November 2017	1ADD	126
1 November 2017	3E97	126
8 November 2017	3366	24

The processing steps (of the Table 2.2.1 products) and the algorithms used were as follows: (1) apply orbit file to obtain accurate satellite orbit and velocity vectors and generate accurate georeferencing of the images; (2) thermal noise removal to remove thermal antenna noise affecting the images; (3) border noise removal to remove noise on the edges of the images; (4) radiometric calibration to normalize the amplitude observed in each band for a radar cross-section and obtain the backscattering coefficient (reflectivity per unit area) in β^0 (radar reflectance coefficient necessary to perform radiometric terrain corrections); (5) application of Speckle noise filter, in which the Lee filter was applied with a 5×5 pixels window; (6) the application of the radiometric terrain flattening algorithm attenuate geometrical distortions in the backscattering that are likely to occur due to the presence of relief artifacts (slopes, hills, etc.) and the operating geometry of SAR sensors (side-looking type) [37], and in this step, the backscattering coefficient is transformed from β^0 to γ^0 (radar reflectance after terrain flattening); and (7) orthorectification of the images using the range-Doppler terrain correction algorithm. Details of the cited processing steps can be obtained in the texts by Filipponi [38] and dos Santos et al. [24].

Having the images with the VH and VV polarizations calibrated for the backscattering coefficient in γ^0, the next step was to compute the SAR vegetation indices. The SAR indices applied were as follows: RVIm (a proxy of the RVI for dual-polarized data) [39]; the normalized polarization index (Pol) [40], which calculates the normalized difference between VH and VV; the cross-ratio (CR), which is the ratio of VV by VH [41]; the dual-polarization SAR vegetation index (DPSVI) [23]; and the modified DPSVI (DPSVIm) index [24]. In addition to the aforementioned indices, the dual-polarization

RVI for GRD products (DpRVIc) were also calculated [42]. The mathematical notation of each index used is presented in Table 2.2.2.

Table 2.2.2. Description of synthetic aperture radar vegetation indices calculated with Sentinel-1 IW GRD images.

Vegetation Index		Equation	Theoretical Bounds	Source
DPSVI	Dual-polarization SAR Vegetation Index	$DPSVI_{(i,j)} = \frac{VH_{(i,j)}\left[\begin{matrix}(VV_{max}\cdot VH_{(i,j)} - VV_{(i,j)}\cdot VH_{(i,j)} + VH_{(i,j)}{}^2) + \\ (VV_{max}\cdot VV_{(i,j)} - VV_{(i,j)}{}^2 + VH_{(i,j)}\cdot VV_{(i,j)})\end{matrix}\right]}{\sqrt{2}\cdot VV_{(i,j)}}$	$DPSVI \geq 0$	[23]
DPSVIm	modified DPSVI	$DPSVIm_{(i,j)} = \frac{VV_{(i,j)}^2 + VV_{(i,j)}\cdot VH_{(i,j)}}{\sqrt{2}}$	$DPSVIm \geq 0$	[24]
CR	Cross-ratio	$CR_{(i,j)} = \frac{VV_{(i,j)}}{VH_{(i,j)}}$	$CR \geq 1$	[41]
Pol	Normalized polarization	$Pol_{(i,j)} = \frac{VH_{(i,j)} - VV_{(i,j)}}{VH_{(i,j)} + VV_{(i,j)}}$	$-1 \leq Pol \leq 1$	[40]
RVIm	Modified Radar Vegetation Index	$RVIm_{(i,j)} = \frac{4\,VH_{(i,j)}}{VH_{(i,j)} + VV_{(i,j)}}$	Unreported	[39]
DpRVIc	Dual-polarization Radar Vegetation Index for Sentinel-1 GRD products	$DpRVIc_{(i,j)} = \frac{q_{(i,j)}\cdot(q_{(i,j)}+3)}{(1+q_{(i,j)})^2}$; where $q_{(i,j)} = \frac{VH_{(i,j)}}{VV_{(i,j)}}$	$0 \leq DpRVIc \leq 1$	[42]

Note: VV(i, j) and VH(i, j) correspond to the backscattering coefficient of the VV and VH polarizations at pixel (i, j).

The presented indices can be divided into cross-pol ratios (CR, RVIm, and Pol) and based on the degree of depolarization of the radar signal (DPSVI, DPSVIm, and DpRVIc). The Pol index is a normalized difference between the vertical dual polarization that was already employed to monitor crops in the study region by Filgueiras et al. [43], while the RVI is an adaptation of the original RVI index to be used with Sentinel-1 dual-polarization images. CR, on the other hand, is a ratio between the co-polarized image (VV) and the cross-polarized image (VH) and is commonly employed as an indicator of the presence of vegetation [13,41].

The DPSVI and DPSVIm indices use the concept of signal depolarization and quantify this depolarization using contrasts between the backscattering of the VH and VV polarizations and even separate bare soil and water surface (which always appear with values closer to zero) from vegetation pixels [23,24]. The DpRVIc index incorporates polarimetric descriptors and the degree of polarization (known as DoP; degree of polarization) measure from the original RVI to detect vegetation structure (branches, leaves, etc.) and has been tested to differentiate phenological stages of annual crops [42].

The DPSVI, DPSVIm, and DpRVIc indices were calculated with the backscattering coefficient of the VV and VH polarizations in linear power units (dimensionless), while the CR, Pol, and RVIm indices were calculated with the backscattering coefficient transformed into the physical unit (decibel; dB), using Equation (1).

$$\gamma^0(in\ dB) = 10 \cdot \log_{10} \gamma^0 \quad (1)$$

where γ^0 represents the backscattering coefficient of the VV and VH polarizations in unit linear power units, and $\gamma^0(in\ dB)$ is the same coefficient transformed to dB.

The download of the images, the processing, and the calculation of the vegetation indices was performed with Python programming language resources, using the SNAP (Sentinel Application Platform, version 9.0.6) software algorithms, and the raster sampling using the geographic coordinates of the soil profiles was performed with R programming language [44]. The codes built to process the Sentinel-1 IW GRD images from Table 2.2.1 can be found in the repository: < https://github.com/eupassarinho/sentinel-1-SAR-vegetation-indices.git >.

2.2.2.3. Modeling Soil Organic Carbon by Machine Learning Methods

In the processing of modeling SOC (g kg^{-1}), the vegetation indices of Table 2.2.2 plus the VV and VH polarizations (both in linear power unit as well as in dB) were used as predictors of SOC. To predict the SOC contents, three regression methods were employed: the least absolute shrinkage and selection operator (LASSO) [45], the support vector machine (SVM) [46] for regression (SVR), and the random forest (RF) [47]. These were chosen to encompass the modeling process methods based on linear (LASSO), nonlinear (SVR), and tree (RF) regression.

RF and SVM methods are already widely employed in soil attribute prediction and in digital soil mapping [17,48,49] and can be applied to both regression and

classification problems. RF for regression works by building a collection of M regression trees that are random to each other. Then, the average prediction of all trees is taken to predict a value [47,50]. On the other hand, the support vector machine admits predictions with a tolerable error, controlled by the modeling support vectors and defined by the hyperparameter C (cost) [51]. SVR becomes a nonlinear regression method when a nonlinear kernel function is used to transform the covariates as a preprocessing step for the prediction [51]. In this study, the kernel function used was the radial basis function (RBF).

Unlike RF and SVM, LASSO is a regression-only method, which selects predictors by penalty and also deals with the collinearity among them. LASSO fits linear regressions between the dependent variable and the predictors via ordinary least squares while adjusting the penalty parameter (λ) [52]. The hyperparameter λ is used by the algorithm to force the parameter β_1 of each predictor to tend towards zero. If the β_1 of a predictor equals zero, then that predictor is not used in the prediction.

Regression methods are sensitive to the correlation that exists between the predictors [53,54]. This is the case for linear regression, SVM, and RF [17,48,49,55]. Therefore, before splitting the data for training the models, a correlation filter was applied with a threshold equal to 0.80. That is, of the covariables (Gamma0_VV, Gamma0_VH, Gamma0_VV_dB, Gamma0_VH_dB, Pol, RVIm, CR, DPSVI, DPSVIm, and DpRVIc), any with a correlation $|r| \geq 0.80$ [1,20] were eliminated from the modeling. This step, however, was employed only for modeling with the SVR-RBF and RF methods since LASSO deals with multicollinearity.

The original dataset (n = 861), containing SOC samples and the ten radar predictors, was randomly divided into the training and test (holdout) subsets, containing 70% and 30%, respectively, of the original data set [56]. However, since there is only one value of each covariate per soil profile, the models were trained for each soil layer [18,57], which means that each model was trained on 86 samples and tested on another 37 samples.

The models were then trained using k-fold cross-validation as the validation method for fitting the hyperparameters of the regression algorithms and 10 folds were defined [56]. For RF, the adjusted hyperparameter was the mtry (the number of predictor covariables to be used in each regression tree), which was set as 1/3 of the predictors (after filtering for correlation). For SVR-RBF, the C (cost) and sigma hyperparameters were tuned using a search grid ($C = \{0.1, 0.2, 0.4, 0.6, 0.8, 1.0, and\ 10.0\}$, $sigma = \{0.0001, 0.001, 0.01, 1/5\ (or\ 1/n\ of\ predictors), \text{and } 1.0\}$). For the LASSO models, the hyperparameter λ was tuned via search grid. The search grid for λ was defined as an arithmetic progression of decimal numbers increasing from 0.0 to 2.0 every 0.2. For SVR-RBF and LASSO, preprocessing steps of the predictors were applied: centering (using the mean) and scaling (by the standard deviation).

For each regression method and soil layer, 100 models (or repetitions) were generated; for example, in the 0–5 cm layer, 100 models were generated with the SVR-RBF method and so on. This was carried out by setting, in each iteration of the loop, a new randomization seed coming from a truly random number. This methodology allows

that in the sub-sampling process of the cross-validation, at each repetition, different data are in the k-folds and are used to adjust the hyperparameters of the model. Thus, it is possible to assess the degree of uncertainty of the models to the input data [48,54,57]. Since there are seven soil layers and three regression methods, this means that 2100 SOC prediction models were fitted. The modeling methodology is displayed in Figure 2.2.3.

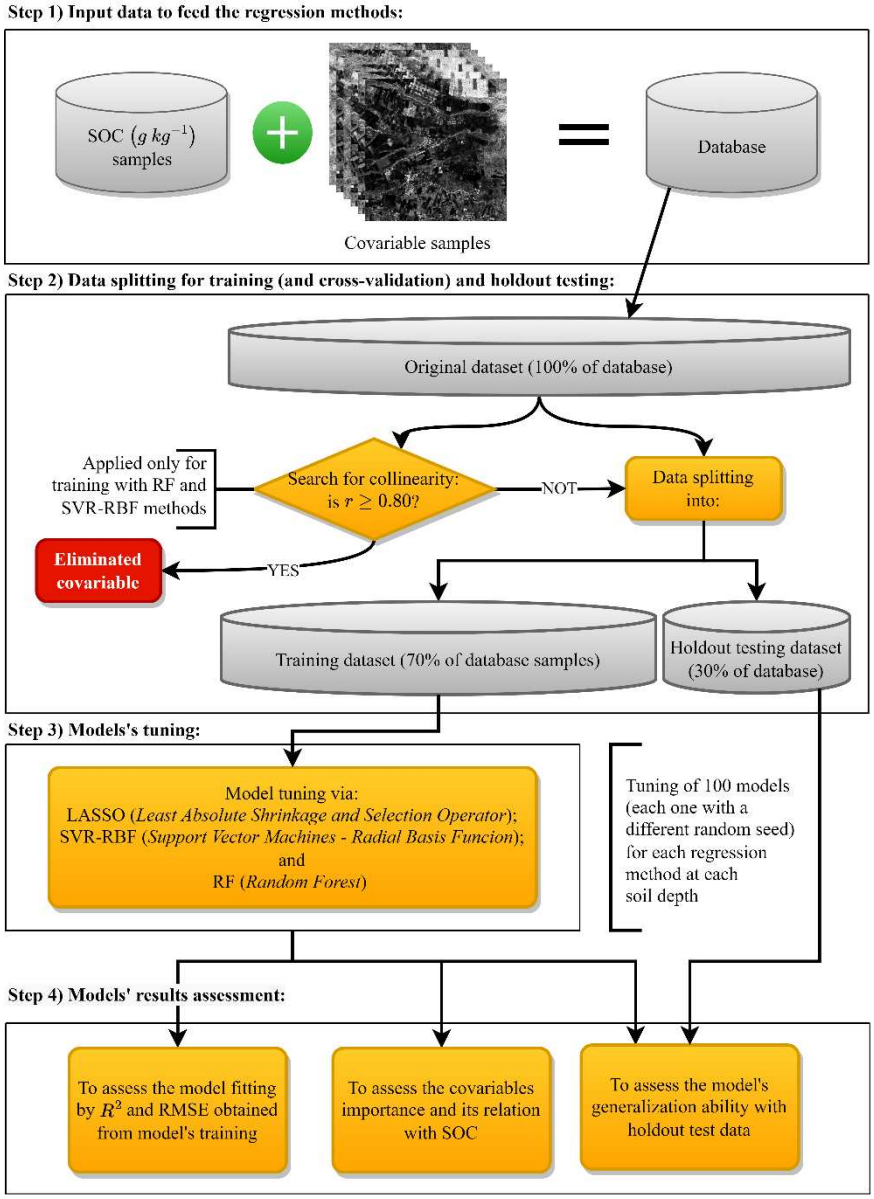


Figure 2.2.3. Schematic of the soil organic carbon (SOC) modeling steps using covariables derived from Sentinel-1 radar imagery and the machine learning regression methods.

All modeling steps were performed with R programming language resources, using functions from the tidymodels and caret (classification and regression training) packages, whose kernels for the regression methods were the glmnet (for LASSO), kernlab (for SVR-RBF), and randomForest (for RF) packages.

2.2.2.4. Results Assessment

The 2100 models were evaluated in both their training and prediction of samples from the test data set, and model performance analysis in both stages was performed to evaluate the generalization ability of the models. The evaluation of model performance in the training stage was carried out using the RMSE (root mean squared error) and R^2 (determination coefficient) metrics, which are standard **caret** accuracy and correlation metrics, respectively. To evaluate the predictions in the holdout test, other metrics were included: Willmott's concordance index (d), MBE (mean bias error), MAE (mean absolute error), and CCC (Lin's concordance correlation coefficient).

The 21 model architectures (e.g., RF for soil layer 0–5 cm, etc.) were compared to each other using non-parametric statistical tests. Since each model architecture generated 100 values for each statistical metric mentioned (MBE, RMSE, MAE, R^2, CCC, and d), the goal of this step was to test for statistically significant differences in the performance of models trained with different architectures (e.g., RMSE of RF versus RMSE of SVR-RBF for the 0–5 cm soil layer, etc.). For this, the Kruskal–Wallis's test (non-parametric test for three or more groups of continuous variables) and Dunn's test (post hoc pairwise test of the Kruskal–Wallis's test) [58] were used, adopting a 95% confidence interval ($P = 0.05$).

To assess the importance of the predictors in SOC modeling, the variable importance plots (VIP) method was used. The importance of the covariables used in the LASSO models consists of the normalization (from 0 to 100) of the slope coefficient (β_1) adjusted for the covariables used by the method. For the SVM and RF methods, the importance is obtained by permuting a covariable in the model and evaluating the loss of model accuracy.

2.2.3. Results

2.2.3.1. Accuracy of Soil Organic Carbon Prediction

For predicting soil organic carbon (SOC) contents, not all covariables obtained from the processing of Sentinel-1 images were used. The linear correlation diagram between the covariables was obtained (Figure 2.2.4), where we noticed that although each covariable comes from a different equation, some are highly correlated. This is the case for the covariables CR, Pol, and RVIm, which are ratios between the radar polarizations, in addition to the VV and VH polarizations (both in linear power units and in dB). Therefore, after filtering covariables for multicollinearity ($r \geq 0.80$) to feed the SVR-RBF and RF regression methods, the predictors DPSVIm, DPSVI, Gamma0_VH, DpRVIc, and CR were selected. For the LASSO method, all covariables in Figure 2.2.4 were employed.

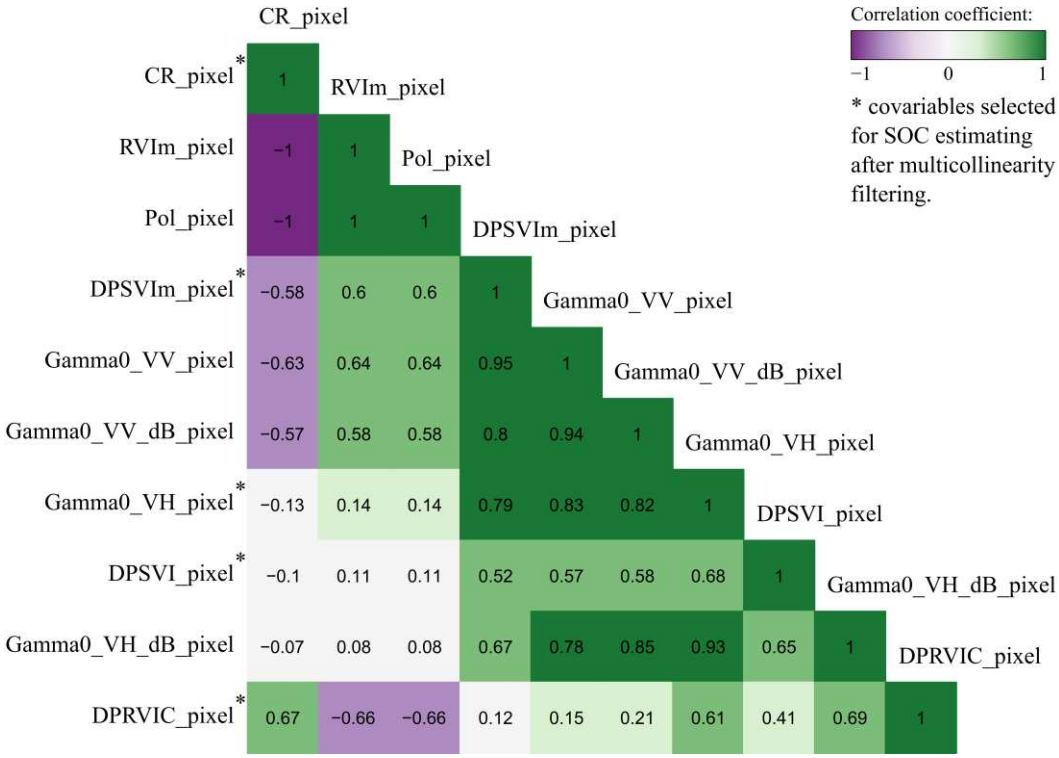


Figure 2.2.4. Linear correlation diagram between the covariables obtained from Sentinel-1A images: highlighted with an asterisk (*) are those covariables selected after filtering by correlation to feed the SVR-RBF and RF methods.

Having filtered out the covariables to be used in modeling, the prediction of SOC by the different regression methods was evaluated in two steps: the performance of the models in the cross-validation step and in the holdout test (with samples not used to train the models).

Figure 2.2.5 exhibits the results of the goodness-of-fit metrics (RMSE and R^2) obtained in the cross-validation of each model in each soil layer; in other words, the optimization results of the hyperparameters of each model are reported. In general, for both RMSE and R^2, there are significant differences between the fitting of each soil layer and each regression method, according to the non-parametric Kruskal–Wallis test. In addition, Dunn's pairwise test indicates that the models of the deeper layers (>20 to 100 cm) and the topsoil layers (0 to 20 cm depth) are more similar to each other. Detailed results of the statistical tests can be found in the Supplementary Material.

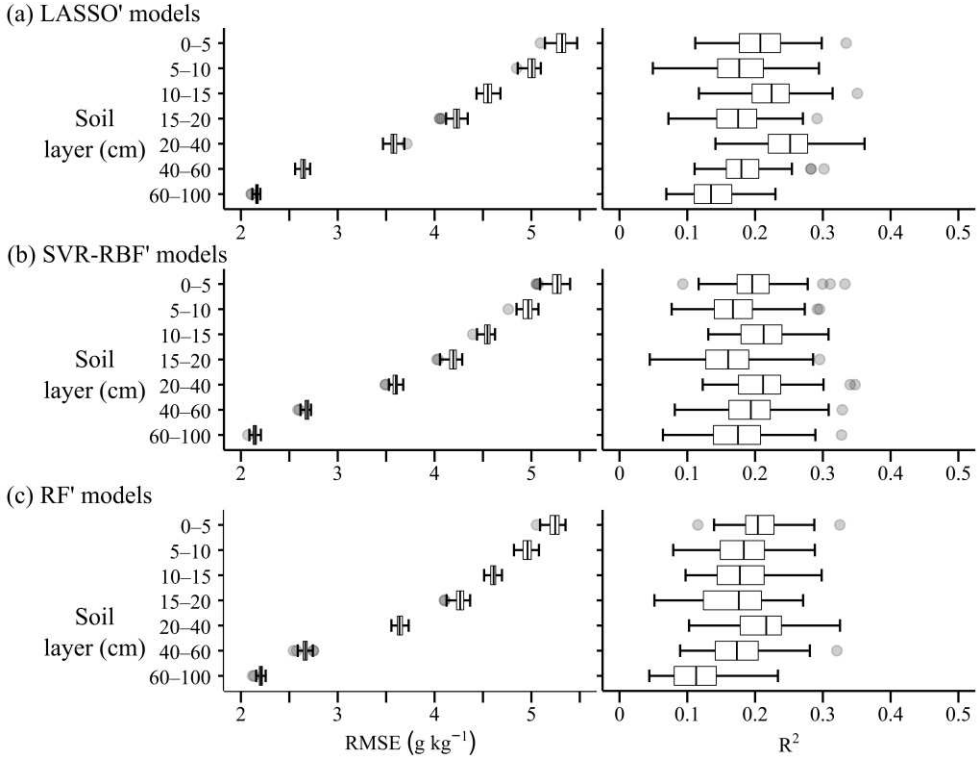


Figure 2.2.5. Cross-validation results on the fitting of the LASSO (subgraphs (**a**)), SVR-RBF (in (**b**)), and RF (in (**c**)) models: accuracy and correlation of the estimates with the observed SOC values are denoted by RMSE and R^2, respectively.

We observe that the RMSE obtained in the SOC modeling is of the order of 5.5 g kg^{-1} in the topsoil layers and decreases as the soil depth increases. With R^2, median values of around 0.2 were obtained in the topsoil layers, tending to decrease in deeper layers.

The RMSE of the deeper soil layers, of the order of 2.0 to 3.5 g kg^{-1}, suggests that as the depth increases, the models become more accurate. However, this is mainly due to the smaller amplitude of the data in these layers, as we observe the distribution of SOC (g kg^{-1}) for the layers 20–40, 40–60, and 60–100 cm in Figure 2.2.2. In addition, the median R^2 of the deeper layers, close to 0.1 for the LASSO and RF methods, indicates that the set of predictors explains less SOC variation than in the topsoil layers (as can be consulted in Figure B5 and Figure B6).

The loss of correlation between SOC estimates and observations as soil depth increases could also be evidenced in the test results obtained from new soil samples. Table 2.2.3 is a summary of the models' performance in the holdout test, and it shows the median value (Md) of the statistical metrics: MBE, RMSE, MAE, R^2, CCC, and d. From Table 2.2.3, we notice that the best correlation (R^2) and agreement (CCC and d) values were obtained with the models for the 0–5 cm and 5–10 cm (topsoil) layers.

Among the metrics R^2, CCC, and d, the d index shows the highest values of agreement between SOC estimates and observations (Table 2.2.3). However, the results obtained with R^2 and CCC metrics indicate that around 20% of the SOC variability was explained with the radar covariables, mainly in the topsoil layers: 0–5 cm and 5–10 cm. For these layers, whose results are highlighted in Table 2.2.3, the best results were

obtained with the SVR-RBF and LASSO regression methods, which showed better generalization ability than the models trained with RF.

Table 2.2.3. Model performance results in the holdout test: median (Md) of MBE (mean bias error), RMSE (root mean squared error), MAE (mean absolute error), R^2 (coefficient of determination), CCC (Lin's concordance correlation coefficient), and d (Willmott's concordance index).

Regression Method	Soil Layer	Md(MBE)	Md(RMSE)	Md(MAE)	Md(R^2)	Md(CCC)	Md(d)
LASSO	0–5 cm	0.869	4.864	3.914	0.243	0.231	0.442
	5–10 cm	−1.248	5.135	3.869	0.167	0.127	0.345
	10–15 cm	−0.824	5.557	4.532	0.027	0.079	0.297
	15–20 cm	−0.727	4.099	3.489	0.059	0.083	0.291
	20–40 cm	0.670	3.912	3.278	0.007	0.050	0.314
	40–60 cm	−0.005	2.414	1.956	0.001	0.017	0.318
	60–100 cm	0.093	2.322	2.015	0.000	−0.004	0.167
SVR-RBF	0–5 cm	0.899	4.959	3.942	0.238	0.196	0.400
	5–10 cm	−0.989	4.953	3.686	0.172	0.212	0.452
	10–15 cm	−1.123	5.556	4.582	0.039	0.086	0.325
	15–20 cm	−0.812	4.213	3.513	0.015	0.053	0.362
	20–40 cm	0.824	3.876	3.307	0.002	0.021	0.265
	40–60 cm	−0.035	2.306	1.928	0.040	0.042	0.193
	60–100 cm	0.285	2.436	2.151	0.011	−0.043	0.135
RF	0–5 cm	0.777	4.955	3.898	0.208	0.184	0.360
	5–10 cm	−1.196	5.082	3.829	0.180	0.151	0.372
	10–15 cm	−1.084	5.670	4.724	0.003	0.016	0.236
	15–20 cm	−0.666	4.187	3.567	0.015	0.046	0.294
	20–40 cm	0.740	3.811	3.219	0.010	0.037	0.272
	40–60 cm	−0.064	2.304	1.899	0.034	0.067	0.287
	60–100 cm	0.110	2.297	2.014	0.003	0.007	0.129

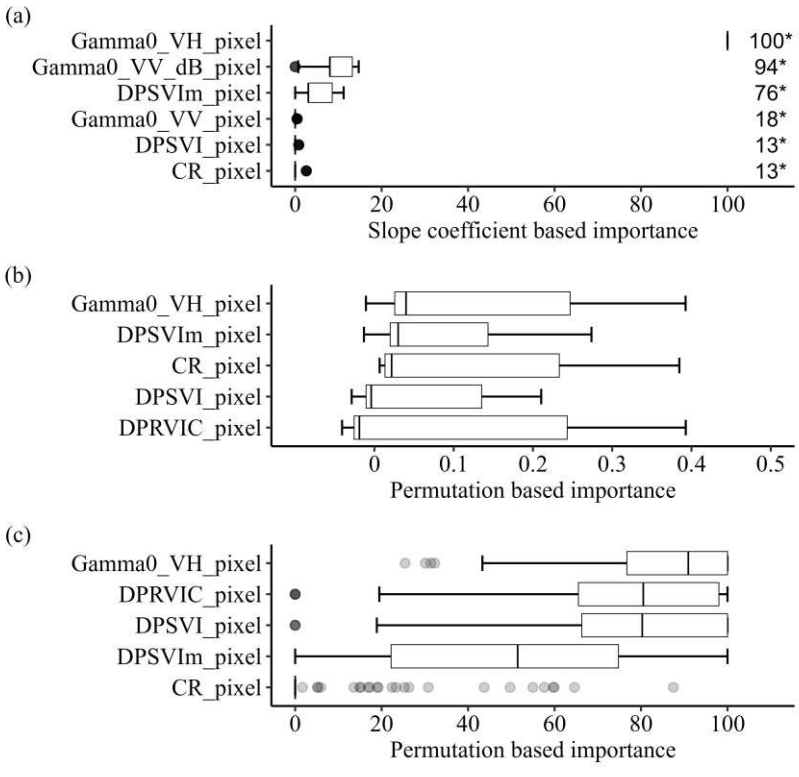
2.2.3.2. Covariables' Importance and Their Relationship to Soil Organic Carbon

As demonstrated by the analysis of Figure 2.2.5 and Table 2.2.3, the predictive ability of the radar covariates was better for the topsoil layers (0–5 cm and 5–10 cm). Considering the similarity of accuracy and correlation for both layers, the analysis of the importance of the covariates is presented only for the 0–5 cm layer and for each model architecture (Figure 2.2.6).

Figure 2.2.6 displays boxplots with the importance of each covariable used in the SOC modeling. In the graph in Figure 2.2.6a, each covariable has as many importance values as the number of times it was used by the LASSO method. On the other hand, in Figure 2.2.6b and Figure 2.2.6c, each covariable has 100 importance values from the 100 SVR-RBF and RF models, respectively.

For the linear method, LASSO, the most important covariable, the VH polarization of the Sentinel-1 sensor was present in almost all models. In turn, the VV polarization (in dB) was also selected for most models but with less importance and more similar to the DPSVIm index. Other covariables were used in some models: the VV polarization in linear power units, the DPSVI index, and the CR ratio.

The VH polarization was also the most important covariable for the SVR-RBF and RF models. In the case of these models, the method of estimating importance is based on the loss of prediction accuracy when a particular covariable is removed from the model. This means that although the VH band is the covariable that most contributes to the prediction accuracy, the vegetation indices also contributed to the SOC estimates.



* the number of times the covariable was used by the LASSO regression method

Figure 2.2.6. Importance of covariables used to predict SOC at the soil layer of 0–5 cm for the models: (**a**) LASSO, (**b**) SVR-RBF, and (**c**) RF.

For the 0–5 cm layer, it is possible to differentiate the distribution of SOC content between each land-use and land-cover class (LULC) where the samples were collected. In Figure 2.2.7, we observe that the SOC values tend to increase from the pasture class (PAST) towards the forest formations (FLO). The lowest SOC contents are observed in the PAST and rainfed agriculture (RAG) classes, intermediate SOC values are seen in the Cerrado savanna formation (CDO) and recently suppressed area (ASV) classes, while the highest SOC contents are in the irrigated agriculture (IRR) and forestry (FLO) classes.

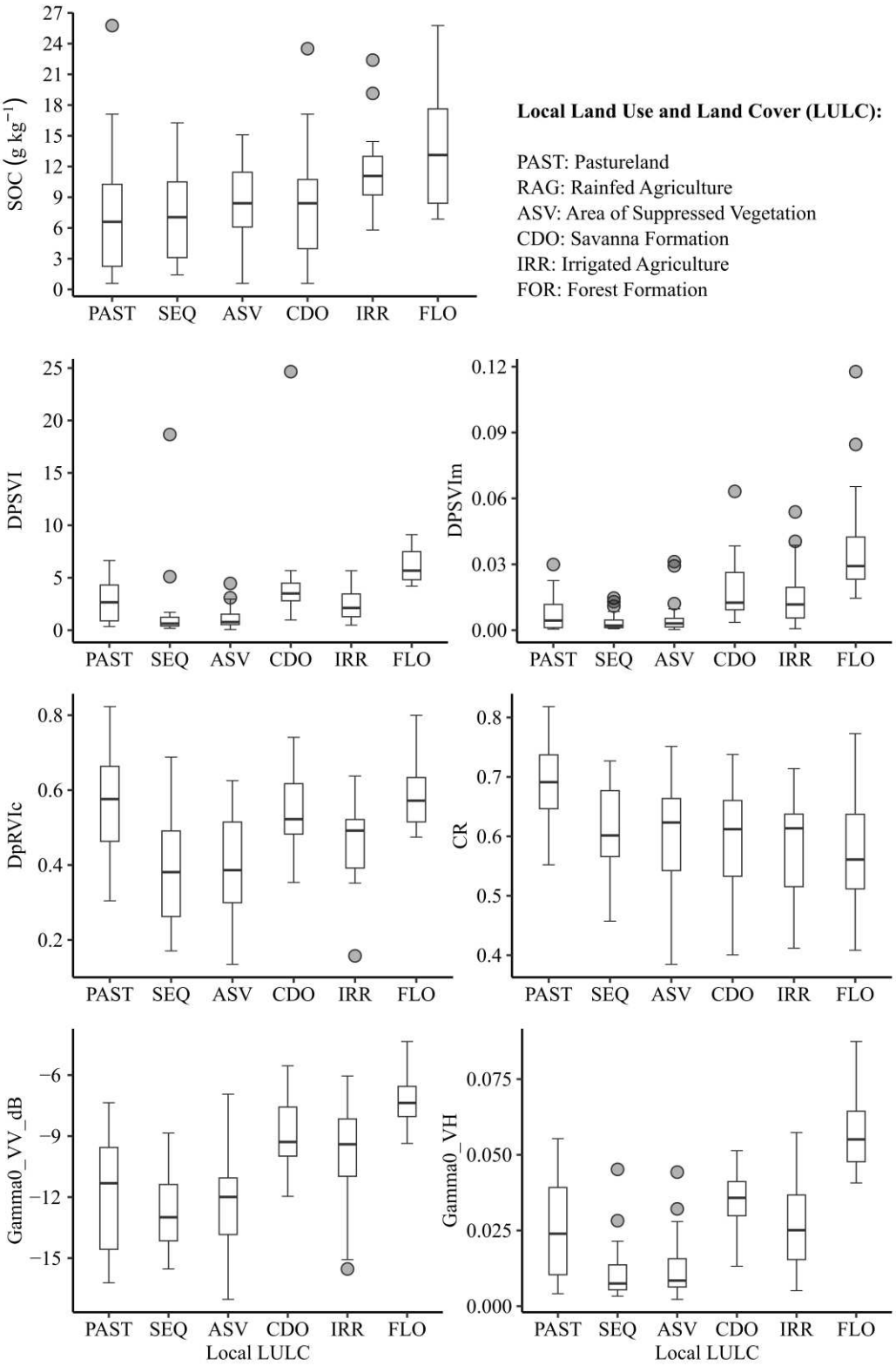


Figure 2.2.7. Distribution of the values of each covariable selected by the regression methods, including soil organic carbon (SOC) itself of the 0–5 cm soil layer in the different land-use and land-cover (LULC) classes.

The same behavior was observed in SOC contents among LULC classes, with contents increasing from the PAST class to FLO class, was observed in other covariables. The

covariables DPSVI, DPSVIm, Gamma0_VH, and Gamma0_VV_dB also showed similar behavior to the SOC contents, and in the case of the DPSVIm index, we observed that it tends to separate the FLO (with the highest SOC contents) class from the others. The CR index has an inversely proportional behavior: The values tend to decrease from the PAST to the FLO classes, although the only class that differed from the others was the PAST class. Also, in the case of DpRVIc, there was no behavior similar to SOC among the LULC classes.

2.2.4. Discussion

Zhou, Geng, Chen, Pan, et al. [20] performed SOC prediction (with contents ranging from 4.70 to 439.10 g kg^{-1}) in a central European region (in the countries of Slovenia, Austria, and Italy), in which one of the experiments consisted of predicting SOC contents using as predictors only the VV and VH polarizations of Sentinel-1 IW GRD images. The modeling developed by the authors used soil samples from the 0 to 20 cm layer. Although the dimensional accuracy metrics (RMSE and MAE) employed by the authors cannot be compared since they applied transformations to the modeled SOC samples, for both regression methods (SVM and RF), the authors obtained an R^2 of 0.16.

In similar modeling work, Zhou, Geng, Chen, Liu, et al. [19] obtained a goodness-of-fit $R^2 = 0.19$ in SOC predicting. In this work, the authors also used Sentinel-1 IW GRD imagery (also only the backscattering coefficients) and the RF regression method to predict SOC (with contents ranging from 1.75 to 139.83 g kg^{-1}) in a hydrographic basin in China in high-altitude and low-temperature terrain. In the present work, we achieved R^2 values between predictions and observations of SOC contents (with contents ranging from 0.59 to 26.91 g kg^{-1}) in testing the models of the order of $R^2 \sim 0.24$ using the LASSO and SVR-RBF algorithms in models of the topsoil layers (see Figure 2.2.5 5 and Table 2.2.3).

In addition, the models achieved RMSE values between 4 and 6 g kg^{-1} (Table 2.2.3). These accuracy values found are comparable to regression models that use the spectral signature of soil samples to predict SOC. In these types of studies, authors found RMSE values ranging from 1.0 to 7.0 g kg^{-1}, as can be seen in the work of Soriano-Disla et al. [59] and Santos et al. [60], even though the values of R^2 in studies at the spectroscopic scale are of the order of two to three times higher than the values of R^2 found in our study with SAR remote sensing. Similarly, Moura-Bueno et al. [61] found RMSE values between 4.0 and 12 g kg^{-1} in the prediction of SOC in southern Brazil by the utilizing spectral signatures (in the visible and near-infrared spectral regions) of soil samples. According to the authors, these variations in accuracy were related to pedological and environmental characteristics, including soil texture and type of land use and land cover.

The best-performing regression methods were the linear method, LASSO, and nonlinear SVR-RBF, to the detriment of RF (see Table 2.2.3), a behavior also observed by Shafizadeh-Moghadam et al. [17]. This suggests that there is a linear (or quasi-linear) relationship between the covariables and SOC contents. In this sense, it is important to

highlight how the radar covariables and SOC contents behave in the different LULC classes in which the soil samples were collected, as the influence that the type of LULC has on SOC variations, especially in the topsoil layers, is well known [62].

At landscape scales (and larger scales), vegetation type and LULC class are factors that control SOC (along with other factors), and this is due to the control that organisms exert over the rates of organic matter input and decomposition [62,63]. In the topsoil layer, we noticed that SOC contents vary according to the type of vegetation (Figure 2.2.7). In Figure 2.2.6, we note the relative importance that the variables DPSVI, DPSVIm, and the VH polarization have for predicting SOC contents, and in Figure 2.2.7, we note that these same radar covariates showed similar behavior to the SOC contents in each LULC class. The ability of the Sentinel-1 covariables to stratify the different vegetation types indicates why the topsoil layer models were able to explain around 20% of the variation in SOC contents in the studied soils in contrast to the models from deeper layers (Table 2.2.3).

The capability of Sentinel-1 covariates to account for approximately 20% of the variation in soil organic carbon (SOC) contents across landscapes may be perceived as limited. However, it is important to consider that, for the purposes of digital soil mapping, additional covariates representing soil-forming factors are required to achieve more accurate predictions [10]. Take, for example, the SCORPAN framework, in which various elements such as the soil properties itself, climate, living organisms, relief descriptors, parent material, age, and geographical location are integrated to model complex phenomena contributing to spatial variations in soil properties [10].

Nonetheless, studies that focus on isolated groups of predictors (e.g., those representing only organisms in a specific location, as in the present study) tend to yield low R^2 values [4,17,19,20]. It is important to note, however, that a low R^2 value does not necessarily indicate a lack of predictive ability. The concept of "low" accuracy in digital soil mapping is relative given that the pedosphere is an intricate Earth system with landscape variations that are challenging to capture, even when considering all the covariates within the SCORPAN framework.

The covariables obtained from the SAR images contributed differently to the SOC prediction modeling (Figure 2.2.6). The VH polarization was the most-used covariable for the LASSO method and the one with the greatest relative importance for the SVR-RBF and RF methods. Although the intensity of the backscattered radar signal for SAR sensors is not a direct measure of aboveground plant biomass [64], the backscatter observed in HV or VH cross-polarization bands is directly associated with aboveground biomass [12,65–67]. This is because for cross-polarization bands, only surface elements that change the polarization of the electromagnetic wave reflected to the sensor are detected with higher brightness, and this is the case for vegetation [68]. Vegetation is a type of target in which microwaves, with wavelengths ranging from ~2 cm to 1 m, interact and undergo a change in their polarization, a mechanism known as volumetric backscattering [13].

The contribution of the SAR vegetation indices to the SOC prediction models is due to the purpose of each of the indices when they were proposed. DpRVI is an

adaptation for GRD products of the DpRVI index of Mandal et al. [25], which in turn is an RVI-based index whose formulation is conceptually based on the degree of polarization of microwaves as they interact with vegetation [25,42]. The degree of polarization measures how much of the total energy backscattered by the targets has had its polarization changed. Both DpRVI and DpRVIc have been successful in discretizing phenological stages in annual crops such as canola, wheat, corn, etc.

DPSVI is also based on the depolarization of the microwave signal, but its structure also seeks to distinguish areas of water bodies and bare-ground surfaces. To this end, the DPSVI takes in its formulation the Euclidean distance relationships between the VV and VH backscattering to distinguish these conditions, also incorporating the VH polarization to graduate different levels of biomass in vegetation [23], and was originally tested on crops such as cassava and corn. Taking advantage of the concept of signal degree of depolarization, dos Santos et al. [24] proposed modifications in the DPSVI model to make the index more sensitive to different levels of biomass in forest-like areas. Among the modifications made, dos Santos et al. [24] incorporated the CR index, which facilitates the separation of different biomass levels in forest areas.

Since the soil samples used were collected in different LULC situations, from pastures to forested areas in the Brazilian Cerrado, and considering the applicability of the different SAR indices, we can affirm that there is a contribution of the different indices for SOC prediction in all regression methods in the topsoil layers. In addition to this, the prediction of SOC using Sentinel-1 IW GRD imagery products is feasible due to the ability to monitor the surface plant biomass and not the soil itself because, as shown by El Hajj et al. [69] and Saatchi [12], there is low C-band microwave penetration in vegetated areas.

SOC modeling in the landscape has historically had the contribution of optical vegetation indices to directly represent land cover and indirectly represent the condition of use and cover by the stratum of that vegetation in addition to other remote sensing products (such as net primary productivity) that denote the input of organic carbon to the soil from plant organs [4,18,57]. Optical vegetation indices represent the biophysical, biochemical, and physiological properties of the mapped vegetation, as shown in the systematic review by de Zeng et al. [70]. On the other hand, SAR vegetation indices tend to represent the vegetation structure, i.e., its geometry according to the vegetation type. Therefore, they can contribute to digital soil-mapping studies—even taking advantage of the continuity of operational space missions such as Sentinel-1—of missions that have recently become operational, such as the SAOCOM (Satélite Argentino de Observación Con Microondas) satellite constellation, which has L-band SAR sensors in the twin satellites (SAOCOM 1A and 1B), as well as planned missions such as NISAR (NASA-ISRO SAR), with L- and S-band sensors.

Furthermore, more accurate predictions of SOC content can be obtained by combining SAR-derived variables with covariables related to other SCORPAN components. In this regard, other components of SCORPAN, more precisely the relief, have already been described in studies of SOC modeling using SAR sensor images [71].

The use of SAR remote sensing data has already been proven to be an important alternative to optical remote sensing and has one major advantage over optical remote sensing. Imaging with SAR sensors is much less influenced by atmospheric conditions such as the presence of clouds [43,72] since microwaves (with $\lambda > 2$ cm) barely interact with atmospheric particles [13,73]. This advantage ensures operationality when employing SAR imagery in digital soil-mapping works in cloudy or rainy environments.

2.2.5. Conclusions

It was possible to predict the soil organic carbon (SOC) content of a region with different land-use and land-cover classes and predominantly sandy soils, using as explanatory variables the SAR (synthetic aperture radar) vegetation indices for Sentinel-1 satellite dual-polarization images.

The models fed with SAR sensor polarizations and their vegetation indices produced more accurate results in the topsoil layers (0–5 cm and 5–10 cm). In these superficial layers, the LASSO (least absolute shrinkage and selection operator), SVM (support vector machine), and RF (random forest) methods generated models with statistical metrics: RMSE of the order of 5.0 g kg^{-1}; MAE of the order of 3.9 g kg^{-1}; R^2 ranging from 0.16 to 0.24; CCC ranging from 0.12 to 0.23; and the d index ranging from 0.34 to 0.45. In deeper soil layers, although more accurate (looking at RMSE and MAE), the covariables lost their ability to explain SOC variability.

We conclude that SAR covariables alone are insufficient for predicting and mapping SOC contents, especially in deeper soil layers. It is essential to include additional covariables that represent other soil-forming factors beyond vegetation.

Supplementary Materials (APPENDIX B: supplementary material for the article 2): The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/rs15235464/s1, Figure S1 (Figure B1 of APPENDIX B): Study area location: highlights for the location of soil profiles over the Köppen's climatic typology for the region. Climate data source: Alvares et al. [26]; Figure S2 (Figure B2): Digital elevation and slope models of the study area. Elevation data source (NASADEM): NASA JPL [74]; Figure S3 (Figure B3): Dominant soil classes, at the third categorical level of the Brazilian Soil Classification System (SiBCS) [30], in the study area. Data source: Map of Brazil' Soils at the compatible scale of 1:5,000,000 [75]; Table S1 (Table B1): Kruskal-Wallis hypothesis test results for the model groups at the training step. Note that: X^2 is the chi-square statistic of the test; GL: degrees of freedom; and (*) indicates significant difference at P = 0.05; Table S2 (Table B2): Pairwise Dunn test results for the model groups at the training step: (*) indicates significant difference at P = 0.05; Figure S4 (Figure B4): Holdout testing results on the fitted LASSO (subgraphs a), SVR-RBF (in b), and RF (in c) models: accuracy and correlation of the estimates with the observed SOC (soil organic carbon) values are denoted by: MBE (Mean Bias Error), RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), R^2 (determination coefficient) and CCC (Lin's concordance and correlation coefficient), respectively; Table

S3 (Table B3): Pairwise Dunn test results for the model groups at the testing step: (*) indicates significant difference at P = 0.05; Figure B5: Scatterplots between observed and predicted soil organic carbon (SOC) contents in each topsoil layers for the holdout testing samples. In the first column, the predictions were made by LASSO (least absolute shrinkage and selection operator), in the second column by SVR-RBF (support vector regression with a radial basis function), and in the third column by the RF (random forest) regression method; Figure B6: Scatterplots between observed and predicted soil organic carbon (SOC) contents in each deeper soil layers (soil depth from 20 up to 100 cm) for the holdout testing samples. In the first column, the predictions were made by LASSO (least absolute shrinkage and selection operator), in the second column by SVR-RBF (support vector regression with a radial basis function), and in the third column by the RF (random forest) regression method.

2.2.6. References

1. Lombardo, L.; Saia, S.; Schillaci, C.; Mai, P.M.; Huser, R. Modeling Soil Organic Carbon with Quantile Regression: Dissecting Predictors' Effects on Carbon Stocks. Geoderma 2018, 318, 148–159. https://doi.org/10.1016/j.geoderma.2017.12.011.
2. Yost, J.L.; Hartemink, A.E. Soil Organic Carbon in Sandy Soils: A Review. In Advances in Agronomy; Academic Press Inc.: London, UK, 2019; Volume 158, pp. 217–310. ISBN 978-0-12-817412-8.
3. FAO; ITPS. Global Soil Organic Carbon Map (GSOCmap) Version 1.5; FAO: Rome, Italy, 2020; ISBN 978-92-5-132144-7.
4. Kunkel, V.R.; Wells, T.; Hancock, G.R. Modelling Soil Organic Carbon Using Vegetation Indices across Large Catchments in Eastern Australia. Sci. Total Environ. 2022, 817, 152690. https://doi.org/10.1016/J.SCITOTENV.2021.152690.
5. Padarian, J.; Minasny, B.; McBratney, A.; Smith, P. Soil Carbon Sequestration Potential in Global Croplands. PeerJ 2022, 10, e13740. https://doi.org/10.7717/PEERJ.13740.
6. Poggio, L.; de Sousa, L.M.; Batjes, N.H.; Heuvelink, G.B.M.; Kempen, B.; Ribeiro, E.; Rossiter, D. SoilGrids 2.0: Producing Soil Information for the Globe with Quantified Spatial Uncertainty. Soil 2021, 7, 217–240. https://doi.org/10.5194/soil-7-217-2021.
7. Guo, L.; Zhang, H.; Shi, T.; Chen, Y.; Jiang, Q.; Linderman, M. Prediction of Soil Organic Carbon Stock by Laboratory Spectral Data and Airborne Hyperspectral Images. *Geoderma* **2019**, *337*, 32–41. https://doi.org/10.1016/j.geoderma.2018.09.003.
8. Keskin, H.; Grunwald, S.; Harris, W.G. Digital Mapping of Soil Carbon Fractions with Machine Learning. *Geoderma* **2019**, *339*, 40–58. https://doi.org/10.1016/j.geoderma.2018.12.037.

9. Odebiri, O.; Mutanga, O.; Odindi, J.; Peerbhay, K.; Dovey, S. Predicting Soil Organic Carbon Stocks under Commercial Forest Plantations in KwaZulu-Natal Province, South Africa Using Remotely Sensed Data. *GIScience Remote Sens.* **2020**, *57*, 450–463. https://doi.org/10.1080/15481603.2020.1731108.
10. McBratney, A.B.; Mendonça Santos, M.L.; Minasny, B. On Digital Soil Mapping. *Geoderma* **2003**, *117*, 3–52. https://doi.org/10.1016/S0016-7061(03)00223-4.
11. Hole, F.D. Effects of Animals on Soil. *Geoderma* **1981**, *25*, 75–112. https://doi.org/10.1016/0016-7061(81)90008-2.
12. Saatchi, S. SAR Methods for Mapping and Monitoring Forest Biomass. In *The Synthetic Aperture Radar (SAR) Handbook: Comprehensive Methodologies for Forest Monitoring and Biomass Estimation*; Flores-Anderson, A.I., Herndon, K.E., Thapa, R.B., Cherrington, E., Eds.; NASA: Huntsville, AL, USA, 2019.
13. Woodhouse, I.H. *Introduction to Microwave Remote Sensing*; CRC Press: Boca Raton, FL, USA, 2006; ISBN 0-415-27123-1.
14. Paradella, W.R.; Mura, J.C.; Gama, F.F. *Monitoramento DInSAR Para Mineração e Geotecnia*; Oficina de Textos: São Paulo, Brazil, 2021; ISBN 978-65-86235-19-7.
15. Bartsch, A.; Widhalm, B.; Kuhry, P.; Hugelius, G.; Palmtag, J.; Siewert, M.B. Can C-Band Synthetic Aperture Radar Be Used to Estimate Soil Organic Carbon Storage in Tundra? *Biogeosciences* **2016**, *13*, 5453–5470. https://doi.org/10.5194/bg-13-5453-2016.
16. Ceddia, M.B.; Gomes, A.S.; Vasques, G.M.; Pinheiro, É.F.M. Soil Carbon Stock and Particle Size Fractions in the Central Amazon Predicted from Remotely Sensed Relief, Multispectral and Radar Data. *Remote Sens.* **2017**, *9*, 124. https://doi.org/10.3390/RS9020124.
17. Shafizadeh-Moghadam, H.; Minaei, F.; Talebi-khiyavi, H.; Xu, T.; Homaee, M. Synergetic Use of Multi-Temporal Sentinel-1, Sentinel-2, NDVI, and Topographic Factors for Estimating Soil Organic Carbon. *Catena* **2022**, *212*, 106077. https://doi.org/10.1016/J.CATENA.2022.106077.
18. Sothe, C.; Gonsamo, A.; Arabian, J.; Snider, J. Large Scale Mapping of Soil Organic Carbon Concentration with 3D Machine Learning and Satellite Observations. *Geoderma* **2022**, *405*, 115402. https://doi.org/10.1016/j.geoderma.2021.115402.
19. Zhou, T.; Geng, Y.; Chen, J.; Liu, M.; Haase, D.; Lausch, A. Mapping Soil Organic Carbon Content Using Multi-Source Remote Sensing Variables in the Heihe River Basin in China. *Ecol. Indic.* **2020**, *114*, 106288. https://doi.org/10.1016/J.ECOLIND.2020.106288.
20. Zhou, T.; Geng, Y.; Chen, J.; Pan, J.; Haase, D.; Lausch, A. High-Resolution Digital Mapping of Soil Organic Carbon and Soil Total Nitrogen Using DEM Derivatives, Sentinel-1 and Sentinel-2 Data Based on Machine Learning Algorithms. *Sci. Total Environ.* **2020**, *729*, 138244. https://doi.org/10.1016/J.SCITOTENV.2020.138244.

21. Kim, Y.; van Zyl, J. On the Relationship between Polarimetric Parameters. In Proceedings of the IGARSS 2000. IEEE 2000 International Geoscience and Remote Sensing Symposium. Taking the Pulse of the Planet: The Role of Remote Sensing in Managing the Environment. Proceedings (Cat. No.00CH37120), Honolulu, HI, USA, 24–28 July 2000; IEEE: New York, NY, USA, 2000; Volume 3, pp. 1298–1300.
22. Chang, J.G.; Shoshany, M.; Oh, Y. Polarimetric Radar Vegetation Index for Biomass Estimation in Desert Fringe Ecosystems. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 7102–7108. https://doi.org/10.1109/TGRS.2018.2848285.
23. Periasamy, S. Significance of Dual Polarimetric Synthetic Aperture Radar in Biomass Retrieval: An Attempt on Sentinel-1. *Remote Sens. Environ.* **2018**, *217*, 537–549. https://doi.org/10.1016/j.rse.2018.09.003.
24. dos Santos, E.P.; da Silva, D.D.; do Amaral, C.H. Vegetation Cover Monitoring in Tropical Regions Using SAR-C Dual-Polarization Index: Seasonal and Spatial Influences. *Int. J. Remote Sens.* **2021**, *42*, 7581–7609. https://doi.org/10.1080/01431161.2021.1959955.
25. Mandal, D.; Kumar, V.; Ratha, D.; Dey, S.; Bhattacharya, A.; Lopez-Sanchez, J.M.; McNairn, H.; Rao, Y.S. Dual Polarimetric Radar Vegetation Index for Crop Growth Monitoring Using Sentinel-1 SAR Data. *Remote Sens. Environ.* **2020**, *247*, 111954. https://doi.org/10.1016/j.rse.2020.111954.
26. Alvares, C.A.; Stape, J.L.; Sentelhas, P.C.; de Moraes Gonçalves, J.L.; Sparovek, G. Köppen's Climate Classification Map for Brazil. *Meteorol. Z.* **2013**, *22*, 711–728. https://doi.org/10.1127/0941-2948/2013/0507.
27. Pousa, R.; Costa, M.H.; Pimenta, F.M.; Fontes, V.C.; Castro, M. Climate Change and Intense Irrigation Growth in Western Bahia, Brazil: The Urgent Need for Hydroclimatic Monitoring. *Water* **2019**, *11*, 933. https://doi.org/10.3390/W11050933.
28. Dionizio, E.A.; Pimenta, F.M.; Lima, L.B.; Costa, M.H. Carbon Stocks and Dynamics of Different Land Uses on the Cerrado Agricultural Frontier. *PLoS ONE* **2020**, *15*, e0241637. https://doi.org/10.1371/journal.pone.0241637.
29. SGB. *GeoSGB*; Serviço Geológico do Brasil: Brasília, Brazil, 2022.
30. dos Santos, H.G.; Jacomine, P.K.T.; dos Anjos, L.H.C.; de Oliveira, V.Á.; Lumbreras, J.F.; Coelho, M.R.; de Almeida, J.A.; de Araújo-Filho, J.C.; Cunha, T.J.F. *Brazilian Soil Classification System*, 5th ed.; Embrapa: Brasília, Brazil, 2018; ISBN 978-85-7035-800-4.
31. Dionizio, E.A.; Costa, M.H. Influence of Land Use and Land Cover on Hydraulic and Physical Soil Properties at the Cerrado Agricultural Frontier. *Agriculture* **2019**, *9*, 24. https://doi.org/10.3390/AGRICULTURE9010024.
32. Souza, C.M.; Z. Shimbo, J.; Rosa, M.R.; Parente, L.L.; A. Alencar, A.; Rudorff, B.F.T.; Hasenack, H.; Matsumoto, M.; G. Ferreira, L.; Souza-Filho, P.W.M.; et al. Reconstructing Three Decades of Land Use and Land Cover Changes in Brazilian Biomes with Landsat Archive and Earth Engine. *Remote Sens.* **2020**, *12*, 2735. https://doi.org/10.3390/rs12172735.

33. Walkley, A.; Black, I.A. An Examination of the Degtjareff Method for Determining Soil Organic Matter, and a Proposed Modification of the Chromic Acid Titration Method. *Soil Sci.* **1934**, *37*, 29–38. https://doi.org/10.1097/00010694-193401000-00003.
34. ESA. *Sentinel-1: ESA's Radar Observatory Mission for GMES Operational Services*; Fletcher, K., Ed.; European Space Agency: Paris, France, 2012.
35. ESA. Sentinel-1 SAR Technical Guide Available online: https://sentinel.esa.int/web/sentinel/technical-guides/sentinel-1-sar (accessed on 18 November 2022).
36. ASF Copernicus Sentinel Data 2017, 2018, and 2019. Retrieved from ASF DAAC, Processed by ESA. Available online: https://asf.alaska.edu/ (accessed on 17 November 2022).
37. Small, D. Flattening Gamma: Radiometric Terrain Correction for SAR Imagery. *IEEE Trans. Geosci. Remote Sens.* **2011**, *49*, 3081–3093. https://doi.org/10.1109/TGRS.2011.2120616.
38. Filipponi, F. Supplementary Materials: Sentinel-1 GRD Preprocessing Workflow. *Proceedings* **2019**, *18*, 11. https://doi.org/10.3390/ecrs-3-06201.
39. Nasirzadehdizaji, R.; Balik Sanli, F.; Abdikan, S.; Cakir, Z.; Sekertekin, A.; Ustuner, M. Sensitivity Analysis of Multi-Temporal Sentinel-1 SAR Parameters to Crop Height and Canopy Coverage. *Appl. Sci.* **2019**, *9*, 655. https://doi.org/10.3390/app9040655.
40. Hird, J.; DeLancey, E.; McDermid, G.; Kariyeva, J. Google Earth Engine, Open-Access Satellite Data, and Machine Learning in Support of Large-Area Probabilistic Wetland Mapping. *Remote Sens.* **2017**, *9*, 1315. https://doi.org/10.3390/rs9121315.
41. Frison, P.-L.; Fruneau, B.; Kmiha, S.; Soudani, K.; Dufrêne, E.; Toan, T.L.; Koleck, T.; Villard, L.; Mougin, E.; Rudant, J.-P. Potential of Sentinel-1 Data for Monitoring Temperate Mixed Forest Phenology. *Remote Sens.* **2018**, *10*, 2049. https://doi.org/10.3390/rs10122049.
42. Bhogapurapu, N.; Dey, S.; Mandal, D.; Bhattacharya, A.; Karthikeyan, L.; McNairn, H.; Rao, Y.S. Soil Moisture Retrieval over Croplands Using Dual-Pol L-Band GRD SAR Data. *Remote Sens. Environ.* **2022**, *271*, 112900. https://doi.org/10.1016/J.RSE.2022.112900.
43. Filgueiras, R.; Mantovani, E.C.; Althoff, D.; Fernandes Filho, E.I.; Cunha, F.F. da Crop NDVI Monitoring Based on Sentinel 1. *Remote Sens.* **2019**, *11*, 1441. https://doi.org/10.3390/rs11121441.
44. R Core Team, R. R: A Language and Environment for Statistical Computing. 2023.
45. Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *J. R. Stat. Soc. Ser. B Methodol.* **1996**, *58*, 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.
46. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297. https://doi.org/10.1023/A:1022627411411.

47. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. https://doi.org/10.1023/A:1010933404324.
48. Mishra, U.; Yeo, K.; Adhikari, K.; Riley, W.J.; Hoffman, F.M.; Hudson, C.; Gautam, S. Empirical Relationships between Environmental Factors and Soil Organic Carbon Produce Comparable Prediction Accuracy to Machine Learning. *Soil Sci. Soc. Am. J.* **2022**, *86*, 1611–1624. https://doi.org/10.1002/saj2.20453.
49. Xiao, Y.; Xue, J.; Zhang, X.; Wang, N.; Hong, Y.; Jiang, Y.; Zhou, Y.; Teng, H.; Hu, B.; Lugato, E.; et al. Improving Pedotransfer Functions for Predicting Soil Mineral Associated Organic Carbon by Ensemble Machine Learning. *Geoderma* **2022**, *428*, 116208. https://doi.org/10.1016/J.GEODERMA.2022.116208.
50. Biau, G.; Scornet, E. A Random Forest Guided Tour. *Test* **2016**, *25*, 197–227. https://doi.org/10.1007/s11749-016-0481-7.
51. Smola, A.J.; Schölkopf, B. A Tutorial on Support Vector Regression. *Stat. Comput.* **2004**, *14*, 199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88.
52. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning*; Springer Texts in Statistics; Springer: New York, NY, USA, 2013; Volume 103. ISBN 978-1-4614-7137-0.
53. Boehmke, B.; Greenwell, B. *Hands-On Machine Learning with R*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2019; ISBN 978-0-367-81637-7.
54. Kuhn, M.; Johnson, K. *Applied Predictive Modeling*; Springer: New York, NY, USA, 2013; ISBN 978-1-4614-6848-6.
55. Moura-Bueno, J.M.; Dalmolin, R.S.D.; ten Caten, A.; Dotto, A.C.; Demattê, J.A.M. Stratification of a Local VIS-NIR-SWIR Spectral Library by Homogeneity Criteria Yields More Accurate Soil Organic Carbon Predictions. *Geoderma* **2019**, *337*, 565–581. https://doi.org/10.1016/j.geoderma.2018.10.015.
56. Brus, D.J.; Kempen, B.; Heuvelink, G.B.M. Sampling for Validation of Digital Soil Maps. *Eur. J. Soil Sci.* **2011**, *62*, 394–407. https://doi.org/10.1111/j.1365-2389.2011.01364.x.
57. Gomes, L.C.; Faria, R.M.; de Souza, E.; Veloso, G.V.; Schaefer, C.E.G.R.; Filho, E.I.F. Modelling and Mapping Soil Organic Carbon Stocks in Brazil. *Geoderma* **2019**, *340*, 337–350. https://doi.org/10.1016/j.geoderma.2019.01.007.
58. McKight, P.E.; Najab, J. Kruskal-Wallis Test. In *The Corsini Encyclopedia of Psychology*; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 2010; p. 1. ISBN 978-0-470-47921-6.
59. Soriano-Disla, J.M.; Janik, L.J.; Viscarra Rossel, R.A.; Macdonald, L.M.; McLaughlin, M.J. The Performance of Visible, Near-, and Mid-Infrared Reflectance Spectroscopy for Prediction of Soil Physical, Chemical, and Biological Properties. *Appl. Spectrosc. Rev.* **2014**, *49*, 139–186. https://doi.org/10.1080/05704928.2013.811081.
60. dos Santos, U.J.; de Demattê, J.A.M.; Menezes, R.S.C.; Dotto, A.C.; Guimarães, C.C.B.; Alves, B.J.R.; Primo, D.C.; Sampaio, E.V.d.S.B. Predicting Carbon and Nitrogen by Visible Near-Infrared (Vis-NIR) and Mid-Infrared (MIR)

Spectroscopy in Soils of Northeast Brazil. *Geoderma Reg.* **2020**, *23*, e00333. https://doi.org/10.1016/j.geodrs.2020.e00333.

61. Moura-Bueno, J.M.; Dalmolin, R.S.D.; Horst-Heinen, T.Z.; ten Caten, A.; Vasques, G.M.; Dotto, A.C.; Grunwald, S. When Does Stratification of a Subtropical Soil Spectral Library Improve Predictions of Soil Organic Carbon Content? *Sci. Total Environ.* **2020**, *737*, 139895. https://doi.org/10.1016/j.scitotenv.2020.139895.
62. Wiesmeier, M.; Urbanski, L.; Hobley, E.; Lang, B.; von Lützow, M.; Marin-Spiotta, E.; van Wesemael, B.; Rabot, E.; Ließ, M.; Garcia-Franco, N.; et al. Soil Organic Carbon Storage as a Key Function of Soils—A Review of Drivers and Indicators at Various Scales. *Geoderma* **2019**, *333*, 149–162. https://doi.org/10.1016/j.geoderma.2018.07.026.
63. Guo, L.B.; Gifford, R.M. Soil Carbon Stocks and Land Use Change: A Meta Analysis. *Glob. Change Biol.* **2002**, *8*, 345–360. https://doi.org/10.1046/J.1354-1013.2002.00486.X.
64. Woodhouse, I.H.; Mitchard, E.T.A.; Brolly, M.; Maniatis, D.; Ryan, C.M. Radar Backscatter Is Not a "direct Measure" of Forest Biomass. *Nat. Clim. Change* **2012**, *2*, 556–557. https://doi.org/10.1038/nclimate1601.
65. Bispo, P.d.C.; Rodríguez-Veiga, P.; Zimbres, B.; do Couto de Miranda, S.; Henrique Giusti Cezare, C.; Fleming, S.; Baldacchino, F.; Louis, V.; Rains, D.; Garcia, M.; et al. Woody Aboveground Biomass Mapping of the Brazilian Savanna with a Multi-Sensor and Machine Learning Approach. *Remote Sens.* **2020**, *12*, 2685. https://doi.org/10.3390/rs12172685.
66. Joshi, N.; Mitchard, E.T.A.; Brolly, M.; Schumacher, J.; Fernández-Landa, A.; Johannsen, V.K.; Marchamalo, M.; Fensholt, R. Understanding "saturation" of Radar Signals over Forests. *Sci. Rep.* **2017**, *7*, 3505. https://doi.org/10.1038/s41598-017-03469-3.
67. Santoro, M.; Cartus, O.; Carvalhais, N.; Rozendaal, D.M.A.; Avitabile, V.; Araza, A.; de Bruin, S.; Herold, M.; Quegan, S.; Rodríguez-Veiga, P.; et al. The Global Forest Above-Ground Biomass Pool for 2010 Estimated from High-Resolution Satellite Observations. Earth Syst. Sci. Data 2021, 13, 3927–3950. https://doi.org/10.5194/essd-13-3927-2021.
68. Mitchard, E.T.A.; Saatchi, S.S.; Lewis, S.L.; Feldpausch, T.R.; Woodhouse, I.H.; Sonké, B.; Rowland, C.; Meir, P. Measuring Biomass Changes Due to Woody Encroachment and Deforestation/Degradation in a Forest–Savanna Boundary Region of Central Africa Using Multi-Temporal L-Band Radar Backscatter. Remote Sens. Environ. 2011, 115, 2861–2873. https://doi.org/10.1016/j.rse.2010.02.022.
69. El Hajj, M.; Baghdadi, N.; Bazzi, H.; Zribi, M. Penetration Analysis of SAR Signals in the C and L Bands for Wheat, Maize, and Grasslands. Remote Sens. 2018, 11, 31. https://doi.org/10.3390/rs11010031.
70. Zeng, Y.; Hao, D.; Huete, A.; Dechant, B.; Berry, J.; Chen, J.M.; Joiner, J.; Frankenberg, C.; Bond-Lamberty, B.; Ryu, Y.; et al. Optical Vegetation Indices

for Monitoring Terrestrial Ecosystems Globally. Nat. Rev. Earth Environ. 2022, 3, 477–493. https://doi.org/10.1038/s43017-022-00298-5.

71. Ferreira, A.C.S.; Pinheiro, É.F.M.; Costa, E.M.; Ceddia, M.B. Predicting Soil Carbon Stock in Remote Areas of the Central Amazon Region Using Machine Learning Techniques. Geoderma Reg. 2023, 32, e00614. https://doi.org/10.1016/j.geodrs.2023.e00614.
72. dos Santos, E.P.; da Silva, D.D.; do Amaral, C.H.; Fernandes-Filho, E.I.; Dias, R.L.S. A Machine Learning Approach to Reconstruct Cloudy Affected Vegetation Indices Imagery via Data Fusion from Sentinel-1 and Landsat 8. Comput. Electron. Agric. 2022, 194, 106753. https://doi.org/10.1016/j.compag.2022.106753.
73. Flores-Anderson, A.I.; Herndon, K.E.; Thapa, R.B.; Cherrington, E. (Eds.) The Synthetic Aperture Radar (SAR) Handbook: Comprehensive Methodologies for Forest Monitoring and Biomass Estimation; NASA: Huntsville, AL, USA, 2019.
74. NASA JPL NASADEM Merged DEM Global 1 Arc Second V001 [Data Set]. NASA EOSDIS Land Processes DAAC Available online: https://doi.org/10.5067/MEaSUREs/NASADEM/NASADEM_HGT.001 (accessed on 10 September 2022).
75. Embrapa Mapa de Solos Do Brasil; Empresa Brasileira de Pesquisa Agropecuária, 2011.

2.3. Article 3 – Integrating satellite radar vegetation indices and environmental descriptors with visible-infrared soil spectroscopy improved organic carbon prediction in soils of semi-arid Brazil [3]

Abstract: Soil Organic Carbon (SOC) is a paramount soil attribute for climate regulation, soil fertility, and agricultural productivity. The global demand for SOC testing came in response to expanding soil management practices aimed at ensuring soil health. This study explores enhanced accuracy in predicting SOC using soil spectroscopy (proximal sensing). A Soil Spectral Library (SSL), made from 127 soil profiles in Northeast Brazil, mainly by using soils from a semi-arid region, was used. Four modeling scenarios were employed, incorporating distinct covariable sets: 1) diffuse reflectance from laboratory spectroscopy (SSL); 2) diffuse reflectance and radar vegetation indices from all-weather and globally available Sentinel-1 satellite data; 3) diffuse reflectance and environmental factors; 4) all covariables. Integration of radar vegetation indices and environmental factors significantly improved SOC estimates by soil spectroscopy. Predicting SOC solely from SSL reflectance data yielded an average RMSE of 4.52 g kg^{-1} and R^2 of 0.62. However, by using all covariables significantly reduced RMSE by approximately 13% (to 3.93 g kg^{-1}) and increased R^2 by 15% (to 0.72). This comprehensive approach, combining SSL, satellite radar vegetation indices, and environmental variables, substantially advances SOC spectroscopic prediction accuracy, offering valuable insights for applications in agriculture and environmental monitoring. These findings contribute to the reliability of proximal and remote sensing methodologies in soil testing.

Keywords: soil health; soil spectral library; agriculture; forest; carbon science; carbon credits.

---

[3] Article to be submitted as manuscript to the journal **CATENA** (https://www.sciencedirect.com/journal/catena)

2.3.1. Introduction

The soil has ecosystem functions important for the entire biosphere. Among the main ecological functions of soils (besides technical and cultural functions), we have: biomass production, water filtration and storage, nutrient storage and recycling, habitat for biological activity, and carbon storage (Wiesmeier et al., 2019). Soil organic carbon (SOC), which is the main fraction of soil organic matter, is a key soil attribute due to soil carbon storage being important not only for climate regulation but also for affecting all of the soil functions mentioned above (Wiesmeier et al., 2019) including controlling soil fertility and agricultural production (Jobbágy and Jackson, 2000).

The knowledge of the importance of SOC for soil conservation and maintenance of its functions for the quality of ecosystems and the entire biosphere has led to the development and adoption of sustainable practices as a significant strategy for mitigating greenhouse gases (FAO, 2020; Lal et al., 2018; Paustian et al., 2019; Smith et al., 2020). One conservation practice is no-tillage. The development and high adoption of sustainable land use practices increase the demand for soil monitoring, including the SOC attribute (FAO, 2020). Monitoring, reporting, and verifying SOC is also necessary for implementing site-specific soil management practices (Angelopoulou et al., 2020).

However, the high global demand for soil analysis consumes a lot of chemical reagents used in the analytical determination of SOC, such as dichromate, ferrous ammonium sulfate, and sulfuric acid (Demattê et al., 2019a). Also, traditional chemical soil analysis methods are time-consuming and expensive. As a result, soil spectroscopy in the visible and near-infrared regions of the spectra is an alternative and complementary method to traditional chemical determinations of SOC. Spectroscopy is a relatively fast, low-cost, non-destructive, proximal soil sensing method (Bellon-Maurel and McBratney 2011; Demattê et al. 2019b; FAO 2020; Viscarra Rossel et al. 2016).

Soil spectroscopy consists of measuring the diffuse reflectance of soil samples, commonly manipulated in the laboratory, and building soil spectral libraries (SSL) (Demattê et al. 2019b; Viscarra Rossel, McBratney, and Minasny 2010). The method consists of quantifying the reflectance of samples at wavelengths of visible (Vis: 350 to 700 nm), near-infrared (NIR: 700 to 1100 nm), and short-wave infrared (SWIR: 1100 to 2500 nm). Although mid-infrared (MIR: 4000 to 600 cm^{-1}) can also be measured, it is less common than other spectral regions (Mendes et al., 2022). SSL contains the reflectance of the samples in the visible and infrared spectral bands, and the chemical and/or physical attributes of the soil samples. With an SSL, the attributes of interest can be modeled with the spectral readings through chemometrics models (Pudełko and Chodak, 2020; Soriano-Disla et al., 2014). Typically Machine Learning (ML) methodologies are used to build these models, by using regression methods such as Partial Least Squares, Support Vector Machine, Random Forest, and others (Ben-Dor et al., 2009; Mendes et al., 2022; Moura-Bueno et al., 2019; Santos et al., 2020; Soriano-Disla et al., 2014).

Spectroscopy is said to be a promising methodology to increase efficiency and help monitor soil attributes (Nocita et al., 2015) and thus is in constant evolution to improve the accuracy and reliability of estimates. Among the advances in soil

spectroscopy, we highlight the use of modeling strategies with independent samples from the same SSL to improve the generalization capacity of the models (Brown et al., 2005; dos Santos et al., 2023; McBride, 2022; Viscarra Rossel et al., 2022). Ways to improve the understanding of soil attribute prediction models with SSL are also necessary and have been studied (McBride, 2022; Viscarra Rossel et al., 2022; Wadoux, 2023). Furthermore, recently, the use of other environmental variables (added to SSL) stood out as a way of increasing the accuracy and precision of models, especially for SOC prediction models (Adi et al., 2019; Moura-Bueno et al., 2021; Sabetizade et al., 2021; Wang et al., 2022).

In this context, Moura-Bueno et al. (2020, 2019) observed that stratifying an SSL into subsets based on homogeneity criteria defined by other environmental variables produced more accurate estimates of SOC. However, the downside of stratifying was the reduced number of samples to calibrate ML models (Moura-Bueno et al., 2020). Then the Cubist regression method, which had already been used in modeling with global SSL (Viscarra Rossel et al., 2016), started to be used (Adi et al., 2019; Moura-Bueno et al., 2021; Sabetizade et al., 2021; Wang et al., 2022).

The advantage of Cubist is the ability to automatically build subsets of samples. Continuous and/or categorical variables are used to create sample division rules through the creation of decision trees. Furthermore, continuous variables are used in internal linear regression models to make predictions (Kuhn and Johnson, 2013; Kuhn and Quinlan, 2023; Quinlan, 1992). Notably, the environmental variables added to SSL must be related to soil formation factors and the variability of SOC content in the landscape.

Although there are still no criteria for choosing environmental variables specifically for SSL, to achieve SOC values, such variables must indicate soil formation factors (Moura-Bueno et al., 2021). The possible and available environmental variables must be SCORPAN factors, which describe soil variations in the landscape such as the soil itself, climate, organisms, relief, parental material, age, and spatial position (McBratney et al., 2003; Minasny and McBratney, 2016). Therefore, variables such as soil texture, climate classification, vegetation indices by orbital remote sensing (ORS), classifications of land use and cover, mineralogy, elevation, and others have been used (Adi et al., 2019; Meng et al., 2022; Moura-Bueno et al., 2021, 2020, 2019; Sabetizade et al., 2021; Wang et al., 2022).

Specifically, the vegetation indices by ORS indicated by Sabetizade et al. (2021) for modeling SOC with a national SSL have the advantage of representing both the presence and amount of vegetation (Zeng et al., 2022), which is the main input of organic matter into the soil after deposition and decomposition (Wiesmeier et al., 2019). These variables are continuous and can be used both to stratify an SSL and to predict SOC levels. Unfortunately, the problem with optical satellite products is the uncertainty regarding the availability of scenes due to cloud cover, mainly in tropical regions (Asner, 2001; dos Santos et al., 2022).

Hence, radar ORS is important because it is much less influenced by cloud cover than optical sensors (Flores-Anderson et al., 2019; Woodhouse, 2006). In this context, the vegetation indices for synthetic aperture radar (SAR) sensors emerged to put the high operability of radars to good use as an alternative to spectral vegetation indices through

optical sensing. SAR vegetation indices aim to represent the amount of aboveground plant biomass or phenology (Bhogapurapu et al., 2022; dos Santos et al., 2021; Frison et al., 2018; Mandal et al., 2020b, 2020a; Periasamy, 2018). Therefore, they can be alternatives to optical ORS products in modeling SOC through soil spectroscopy, requiring only tests to prove their feasibility.

Therefore, in this work, we seek to contribute to spectroscopy as an alternative method for quantifying SOC. The originality of the study lies in proposing SAR vegetation indices (obtained by ORS) and other environmental variables to obtain more accurate and precise SOC estimates. The proposal to use radar ORS consists of trying to enhance the operability of vegetation indices to be used in SSL. At the same time, we seek to adopt assumptions to better understand the models created and the independence of the soil samples used to model SOC levels to improve the reliability of model estimates.

The work hypothesizes that the accuracy and precision of SOC prediction models using diffuse reflectance spectroscopy in the laboratory in the Vis-NIR-SWIR spectrum tends to improve when environmental variables are added to the modeling in Cubist models. To test this, we used a regional SSL surveyed in soils from Northeastern Brazil and evaluated the prediction of SOC using the spectral signature of soil samples in the Vis-NIR-SWIR spectrum by adding covariates derived from radar remote sensors (SAR vegetation indices) and other environmental variables to the models.

2.3.2. Methodology

2.3.2.1. Study area description

The soil data was surveyed in the State of Pernambuco, Northeast Brazil. The studied soil profiles are between meridians 34° 48’ and 38° 48’ W and between latitudes 7° 18’ and 9° 18’ S (Figure 2.3.1). The surveys were carried out in campaigns from 2011 to 2013.

The studied soil profiles are in three different climatic regions, according to the Köppen climate typology for Brazil (Alvares et al., 2013). From east to west, the first region is the coastal zone, which has an Am climate (tropical zone with monsoons), whose climate is hot and humid, with average annual temperatures varying between 24 and 26 ºC and total annual rainfall between 1600 and 2500 mm $year^{-1}$; the second zone is subhumid and classified as As climate (tropical zone with dry summer), it is a hot and dry region (26 – 28 ºC and 600 – 1000 mm $year^{-1}$); further west is the semi-arid region, which has an Bsh climate (dry zone, semiarid with low latitude and altitude), with a hot climate (27 – 29 ºC) with low (400 – 800 mm $year^{-1}$) and irregular rainfall (Santos et al., 2020). Only one soil profile falls under the climate class Csa (humid subtropical zone with hot and dry summer). This class is only found in this region due to its elevation close to 1,000m.

127 soil profiles were surveyed in different classes of land use and land cover. The profiles were collected in Acrisols (35 profiles), Luvisols (12), Leptsols (23), Arenosols (11), Regosols (12), Planosols (23), and Ferralsols (11).

Regarding the different classes of land use and land cover (LULC) in the Caatinga and Atlantic Forest biomes: Agriculture (31 profiles), Pasture (27), Other agricultural uses (5), Open Caatinga (31), Dense Caatinga (27) and Forest (forest formation, 6 profiles). The Agriculture class corresponds to rainfed agricultural cultivation, mainly subsistence agriculture with the cultivation (mostly) of corn (*Zea mays* L.), beans (*Phaseolus vulgaris* L.), cowpea beans (*Vigna unquiculata* L. Walp.), and cassava (*Manihot esculenta* Crantz.). The Pasture class represents areas grazed and covered by native or introduced African grasses and other herbaceous plants. The dense Caatinga class represents areas of native Caatinga forest that had substrate and tree shrub covering between 60 and 80% of the soil surface. The open Caatinga class has less soil covered by trees and shrubs, between 40 and 60%, due to natural causes or human interference.

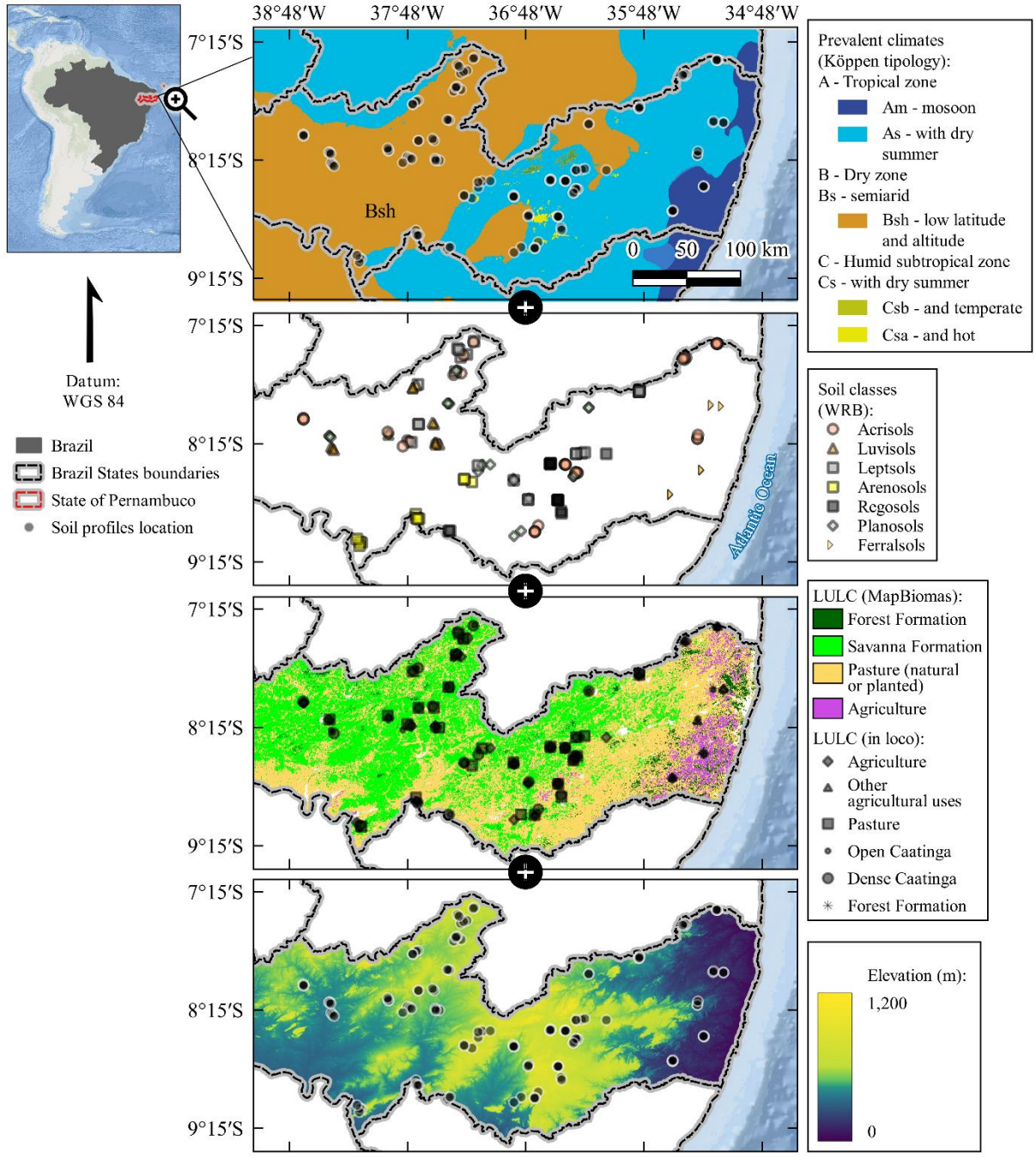


Figure 2.3.1. Location of soil profiles sampled in the study area and their distribution according to the Köppen climate classification (Alvares et al., 2013), their respective soil classes, and land use and land cover classification (LULC) surveyed by MapBiomas (Coleção 6.0 do MapBiomas (Souza et al., 2020)), as well as the terrain elevation (NASA JPL, 2020).

2.3.2.2. Collection and analysis of soil samples: analytical and spectroscopic determination

Soil samples were collected in trenches of 0.7 x 0.7 m in area in the soil profiles mentioned previously. After removing the superficial litter, the trenches were dug. The samples were collected in seven layers at standard depths (cm): 0 to 10, 10 – 20, 20 – 30, 30 – 40, 40 – 60, 60 – 80 and 80 – 100. Since not all profiles were deeper than or equal to 1 m, a total of 701 samples were collected (n = 701).

The samples underwent an analytical determination of their chemical and physical properties in the laboratory. The samples were dried and sieved through a 2 mm mesh. Soil organic carbon contents (SOC, g kg^{-1}) were determined with subsamples of approximately 10 mg via dry determination, using the TruSpec CHN-analyzer (LECO® 2006, St. Joseph, EUA).

To create the regional spectral library (R-SSL), soil samples were prepared in the laboratory for spectral readings in visible, near-infrared, and short-wave wavelengths (Vis-NIR-SWIR). Hence, the FieldSpec 3 spectroradiometer (Analytical Spectral Devices Inc., Boulder, CO, USA) was used. The Fieldspec 3 measures radiance from 350 to 2500 nm and has spectral resolutions of 1 nm (from 350 to 700 nm), 3 nm (700 to 1400 nm), and 10 nm (1400 to 2500 nm). The sampling intervals of the output data are 1 nm with 2151 spectral bands. 20 g of each soil sample was placed in Petri dishes and distributed homogeneously on a flat surface. The source of electromagnetic radiation was two halogen lamps (50 W) (both non-collimated and with a zenith angle of 30º) positioned 35 cm from the sample with an angle of 90º between them. A fiber-optic cable, located 8 cm from the center of the sample surface, captured the reflected energy from an area of approximately 2 cm^2. For each sample, the average reflectance was calculated from three repetitions of readings in different positions, decreasing the shading effect. Each repetition consisted of 100 sensor readings, to maximize the signal-to-noise ratio. The instrument was calibrated before sample readings and every 20 minutes thereafter, using a white *Spectralon* plate.

2.3.2.3. Spectral library pre-processing

The original reflectance, without any transformation, of all R-SSL samples was subjected to two pre-processing steps. The first step consisted of smoothing the spectral curves using a moving average filter, for this purpose a convolution function with a band of 4 nm was used. The second step consisted of calculating the normalized reflectance from the smoothed curves by using the continuum removal algorithm of Clark and Roush (1984).

The normalized reflectance spectral curves in Vis-NIR-SWIR were used to model and predict SOC contents. This was defined because the continuum removal technique highlights light absorption features by organic compounds (Viscarra Rossel et al., 2016) and performed best in modeling SOC contents with the same R-SSL (dos Santos et al., 2023; Santos et al., 2020).

2.3.2.4. Model's environmental variables

In addition to the normalized reflectance in Vis-NIR-SWIR, other environmental variables were used to predict SOC levels. The selected environmental variables are related to the formation and variation factors of SOC in the landscape, as recommended by (Moura-Bueno et al., 2021).

The set of environmental covariates can be divided between categorical and continuous. The categorical covariates used were the LULC classes and soil types (obtained from the field survey, Figure 2.3.1), and the climate class of each soil profile. The climate classification was obtained from the Köppen classification mapping for Brazil (Alvares et al., 2013). The continuous covariates are elevation, which was obtained from the NASADEM digital elevation model (NASA JPL, 2020), and SAR vegetation indices.

2.3.2.5. Obtaining SAR vegetation indices from the Sentinel-1 mission

In this study, orbital remote sensing products were also used to predict SOC contents. The satellite data used in this study came from images of the SAR (Synthetic Aperture Radar) sensor on board the orbital platforms of the European Space Agency (ESA) Sentinel-1 mission. The Sentinel-1 images were used to obtain vegetation indices proposed in the literature for this sensor.

The Sentinel-1 mission satellites operate with SAR-type imaging radar sensors. These are active sensors, which in the case of Sentinel-1 satellites, operate in the C band (with a wave frequency of $\cong 5.4\ GHz$) (ESA, 2012). Dual polarization images from the *Interferometric Wide Swath* (IW) imaging mode were used, which are pre-processed images with only the observed wave amplitude information, called GRD (Ground Range Detected) products.

Since the Sentinel-1 orbital mission started in 2014 (ESA, 2022), the Sentinel-1 IW GRD images obtained for the study were multi-temporal to the soil surveys. Although the survey was carried out from 2011 to 2013, the images that covered the entire study area (first selection criteria and scenes) were from 2017. Therefore, 28 products were used, from the relative orbits of numbers 9 and 82. The images were selected from dates (month) compatible with the dates of the survey (second criterion). This choice was based on studies that used products from remote sensing platforms to model SOC (mainly Sentinel satellites) and used observations from different dates than the soil surveys for SOC analysis (Kunkel et al., 2022; Shafizadeh-Moghadam et al., 2022; Sothe et al., 2022; Zhou et al., 2020a, 2020b). This methodology is common and valid when there are no significant changes in the factors that control SOC levels in the soil, such as changes in land use and cover (Wiesmeier et al., 2019), which was not observed for the study area.

Details of the Sentinel-1 IW GRD images used in the study can be found in the Supplementary Material (Table C1).

Sentinel-1 IW GRD images have two wave polarizations with greater global coverage. These polarizations are the VH, whose sensor emits a pulse of radiation in vertical polarization and measures the reflectivity detected in horizontal polarization, and VV, whose emission and detection are in vertical polarization. Dual polarization Sentinel-1 IW GRD images are required to determine Sentinel-1 SAR vegetation indices.

Sentinel-1 IW GRD images are formed after the sensor scans the Earth's surface over a wide swath of 250 km. This broad band is made up of three subbands (IW1, IW2, and IW3) obtained by the TOPSAR method – Terrain Observation with Progressive Scans SAR (De Zan and Guarnieri, 2006). Subbands IW1, IW2, and IW3 are scanned with incidence angles of 32.9º, 28.3º, and 43.1°, respectively. Furthermore, the spatial resolution of the images (in the range x azimuth directions) of the IW1, IW2, and IW3 subranges are 20.4 m x 22.5 m, 20.3 x 22.6, 20.5 x 22.6, respectively (ESA, 2022). The scenes used are from the Alaska Satellite Facility (ASF, 2022) portal due to their accessibility.

After the acquisition, the Sentinel-1 IW GRD products were processed using the following algorithms:

1) *Apply Orbit File*: obtains accurate satellite orbit and velocity vectors and generates accurate georeferenced images;
2) *Thermal Noise Removal*: removes antenna thermal noise that affects images;
3) *Border Noise Removal*: removes noise at the images' edges;
4) *Radiometric Calibration*: normalizes the amplitude of each polarization for a Radar Cross Section (RCS) and obtains the backscatter coefficient (reflectivity per unit area) in β^0 (RCS required to perform terrain corrections);
5) *Despeckling*: applies the Speckle noise filter using the Lee Sigma filter with a window of 11 x 11 pixels ($sigma = 0.9$);
6) *Radiometric Terrain Flattening (RTF)*: mitigates distortions in backscatter that are likely to occur due to the relief (slopes, hills, etc.) and the operating geometry of SAR sensors (of the side-looking type) (Small, 2011). At this stage, the digital elevation model used to represent the relief was the Copernicus 30m Global. After the RTF algorithm, the symbology of the backscattering coefficient (reflectance for radar) is transformed from β^0 to γ^0 (more details can be found in the methodology of Small (2011)).
7) *Range-Doppler Terrain Correction*: orthorectification of the images from the Copernicus 30m Global.

The described steps of the digital processing of Sentinel-1 images are necessary to transform the wave amplitude detected into a backscatter coefficient. Although the IW images have different spatial resolutions in terms of range and azimuth, after orthorectification the output image has a pixel size of 10 x 10 m. More details of the Sentinel-1 IW GRD image processing steps can be obtained in the texts Filipponi (2019)

and dos Santos et al. (2021). Once we had the images with VH and VV polarizations calibrated to the backscatter coefficient at γ^0, the next step was to calculate the SAR vegetation indices.

Following the methodology of Santos et al. (2023), five products from Sentinel-1 images were used to predict SOC contents, such as the VH polarization, which is itself a polarimetric indicator of the presence of vegetation and the amount of aboveground biomass (Bispo et al., 2020; Joshi et al., 2017; Saatchi, 2019; Santoro et al., 2021; Woodhouse, 2006), and four other vegetation indices. The vegetation indices calculated were: the Cross-ratio (CR), the Dual-polarization SAR Vegetation Index (DPSVI), the modified DPSVI (DPSVIm), and the Dual-polarization Radar Vegetation Index for GRD products (DpRVIc). Table 2.3.1 provides details about each of the indexes mentioned.

Table 2.3.1. Description of vegetation indices and polarimetric descriptors calculated from Sentinel-1 IW GRD images.

Vegetation index	Formula	Theoretical bounds	Bibliographic reference
DPSVI	$DPSVI_{(i,j)} = \frac{VH_{(i,j)}\left[\begin{matrix}(VV_{max}\cdot VH_{(i,j)} - VV_{(i,j)}\cdot VH_{(i,j)} + VH_{(i,j)}^2)+ \\ (VV_{max}\cdot VV_{(i,j)} - VV_{(i,j)}^2 + VH_{(i,j)}\cdot VV_{(i,j)})\end{matrix}\right]}{\sqrt{2}\cdot VV_{(i,j)}}$	$0 \leq DPSVI$	(Periasamy, 2018)
DPSVIm	$DPSVIm_{(i,j)} = \frac{VV_{(i,j)}^2 + VV_{(i,j)}\cdot VH_{(i,j)}}{\sqrt{2}}$	$0 \leq DPSVIm$	(dos Santos et al., 2021)
CR	$CR_{(i,j)} = \frac{VV_{(i,j)}}{VH_{(i,j)}}$	$1 \leq CR$	(Frison et al., 2018)
DpRVI*c*	$DpRVIc_{(i,j)} = \frac{q_{(i,j)}\cdot(q_{(i,j)}+3)}{(1+q_{(i,j)})^2}$; in which $q_{(i,j)} = \frac{VH_{(i,j)}}{VV_{(i,j)}}$	$0 \leq DpRVIc \leq 1$	(Bhogapurapu et al., 2022)

Note: $VV_{(i,j)}$ and $VH_{(i,j)}$ correspond to the backscatter coefficient of polarizations VV e VH in pixel (i, j).

Python programming language resources, using SNAP (**Sentinel Application Platform**) software algorithms were used to download and process images and calculate vegetation indices. The algorithms used to process Sentinel-1 images were from the **SNAP Sentinel-1 toolbox** module, made available by ESA. **R** programming language resources (R Core Team, 2023) were used to sample the images using the geographic coordinates of the soil profiles. The codes used to process and sample the Sentinel-1 IW GRD images can be checked in the repository: <https://github.com/eupassarinho/sentinel-1-SAR-vegetation-indices.git> (Santos et al., 2023).

2.3.2.6. Soil organic carbon content modeling approaches

Four scenarios were defined for modeling: *model set 1, model set 2, model set 3,* and *model set 4* (Figure 2.3.2). In *model set 1,* only the Vis-NIR-SWIR normalized reflectance spectral bands (obtained by proximal sensing) were used to predict SOC contents. In *model set 2*, in addition to the Vis-NIR-SWIR spectral bands, the SAR satellite vegetation indices (obtained by remote sensing) were used (polarization VH, CR, DPSVI, DPSVIm, and DpRVIc). In *model set 3*, the Vis-NIR-SWIR spectral bands and the covariates LULC, soil type, climate, and elevation were used. Finally, *model set 4* used all covariates: Vis-NIR-SWIR, vegetation indices SAR, LULC, soil type, climate, and elevation. The four scenarios followed the modeling method described in the following topics:

2.3.2.6.1. Regression methods and covariate selection

Cubist was the regression method used in all four scenarios to predict SOC levels. Cubist is a regression method based on regression trees. It divides the training data into homogeneous partitions for the covariates used. A series of rules using "if-then conditions" define the partitions. When a partition is created, at the end of the trees (final leaves) a linear regression model (ordinary least squares) is used to predict the soil attribute. Continuous or categorical variables can be used to define conditions, but only numerical variables are used in the regression equations. Details about how Cubist works can be found in Quinlan (1992) and Kuhn and Johnson (2013). Cubist was chosen for its ability to handle spectral data and heterogeneous datasets such as R-SSL (Demattê et al. 2019b; Moura-Bueno et al. 2021; Viscarra Rossel et al. 2016).

The **R** language implementation of **Cubist** has two hyperparameters that can be used to optimize Cubist models: the number of *committees* and *neighbors*. When setting up a combination of committees, Cubist adopts a boosting-like scheme that creates iterative model trees in sequence, which means that the argument *committees* controls the number of model trees. The argument *neighbors* controls the number of similar samples (from the training data with defined rules) that are used to adjust to predict a new sample (Kuhn and Johnson, 2013; Kuhn and Quinlan, 2023). In all *model sets*, the committees and neighbors arguments were tuned via the *Search grid* declared in the *train* function of the **caret** library (Classification and Regression Training for **R**, (Kuhn 2020)). The number of declared committees was 50 and 100, and the number of neighbors was 5 and 9.

To optimize model training by Cubist, a Vis-NIR-SWIR covariate selection step was added to the modeling. The Vis-NIR-SWIR spectral covariates to predict SOC levels were selected using the LASSO (Least Absolute Shrinkage and Selection Operator) regression method, following the methodology of dos Santos et al. (2023). LASSO is a regression and selection method based on the principle of parsimony, in which unimportant and/or highly correlated covariates are eliminated. LASSO fits a multiple linear regression model (using the ordinary least squares method) that has a covariate penalty parameter: when the slope coefficient of a covariate is equal to zero, then the covariate is eliminated from the model (James et al., 2013; Tibshirani, 1996). Covariate

selection with LASSO was applied to Vis-NIR-SWIR covariates in all modeling scenarios.

The LASSO implementation provided by the **glmnet** package in the R language was used (Friedman et al., 2010). LASSO's penalizing hyperparameter is λ, which was also tuned via the *Search grid* using a vector with numbers ranging from 0 to 2 in intervals of 0.05. In addition to the hyperparameter λ, the hyperparameter *alpha* was set up, which is the *elastic net mixing parameter*. To fit LASSO models, *alpha* was kept constant at one (1) (Friedman et al., 2010; Tay et al., 2023). Furthermore, as a requirement of the LASSO method, to correctly penalize the covariates, they must be on the same scale. Therefore, as pre-processing steps for training data in the **caret** *train* function, the following was set up: centering and scaling by the mean and standard deviation of the covariates, respectively.

2.3.2.6.2. Dividing data to train and test (holdout) the models and cross-validation

The database was divided into a training set (for training Cubist models) and a holdout test using the proportion of 70% and 30% of the data, respectively (Brus et al., 2011). Although the division was done randomly, it was ensured that all samples from the same soil profile were in just one data set. This was done to ensure that the samples used to calibrate were independent of the ones used to test the models, following recommendations from (Brown et al., 2005; dos Santos et al., 2023; Malmir et al., 2019; Poggio et al., 2017). Therefore, from the 127 soil profiles studied, 88 samples (n = 487, 70% of the total) were used to train the models. The remaining 39 soil profiles (n = 214, 30% of the total) were used to test the models.

The models were trained using 88 soil profiles (and their 487 samples). The cross-validation method used to optimize the models' hyperparameters was Leave-Soil-Profile-Out (LSPO CV). LSPO CV is an object-oriented k-Fold cross-validation that subsamples entire soil profiles for each of the calibration and validation partitions (dos Santos et al., 2023). LSPO CV also aims to use independent soil samples to calibrate and validate the model (Brown et al., 2005), especially when the dependent variable is SOC, which may have a spatial dependence structure between nearby soil layers (dos Santos et al., 2023). In the LSPO cross-validation, 10-Folds were configured. The *CreateSpacetimeFolds* method, from the **CAST** library (Meyer, 2021; Meyer et al., 2018), was used to randomly select which soil profiles were allocated to each of the *Folds*.

The division of the data into training and testing was done only once, but the division of the training data into different subsets of the LSPO cross-validation was done 100 times. Therefore, soil profiles were randomly drawn 100 times to select different samples to calibrate and validate the models. Hence, for each modeling scenario, 100 models were adjusted. The objective of repeating the modeling process is to evaluate the variability of model uncertainty when different data are used to train them (Gomes et al., 2019; Kuhn and Johnson, 2013; Mishra et al., 2022).

2.3.2.6.3. Model assessment

The models were adjusted to minimize the RMSE (Root Mean Squared Error) in training. However, the accuracy and correlation of the models' predictions on the test data were measured by the statistical metrics: RMSE, MAE (Mean Absolute Error), MSE (Mean Squared Error), coefficient of determination (R^2), Lin's concordance correlation coefficient (CCC) and Nash-Sutcliffe model efficiency (NSE).

The different *model sets* (e.g. *model set 1* versus *model set 3*) were compared using non-parametric statistical tests. The results of the correlation and error statistics, between the predicted and observed SOC values (in the test data), from each *model set* were statistically compared to verify whether the statistical values differed significantly, that is, whether they came from different distributions. For this, the Kruskal-Wallis test was applied, a non-parametric test for three or more groups of continuous variables. If a statistical difference was verified in this, the Dunn test was applied (test for paired groups, after the Kruskal-Wallis test) (McKight and Najab, 2010). In both cases, a 95% confidence interval (P=0.05) was adopted.

The relative importance of the covariates for model prediction was obtained by Variable Importance Plots – VIP (Greenwell et al., 2020; Greenwell and Boehmke, 2020; Kuhn and Johnson, 2013). This was done to identify and evaluate the relative importance of different sources of covariates (spectral covariates, SAR vegetation indices, and other environmental covariates) in the Cubist model. Regarding Cubist, the relative importance of covariates consists of the average between the percentage of times that a given covariate was used: to create a rule (in decision trees) and in an internal linear regression. Figure 2.3.2 displays a general scheme of the modeling process, mainly the modeling scenarios, for SOC contents.

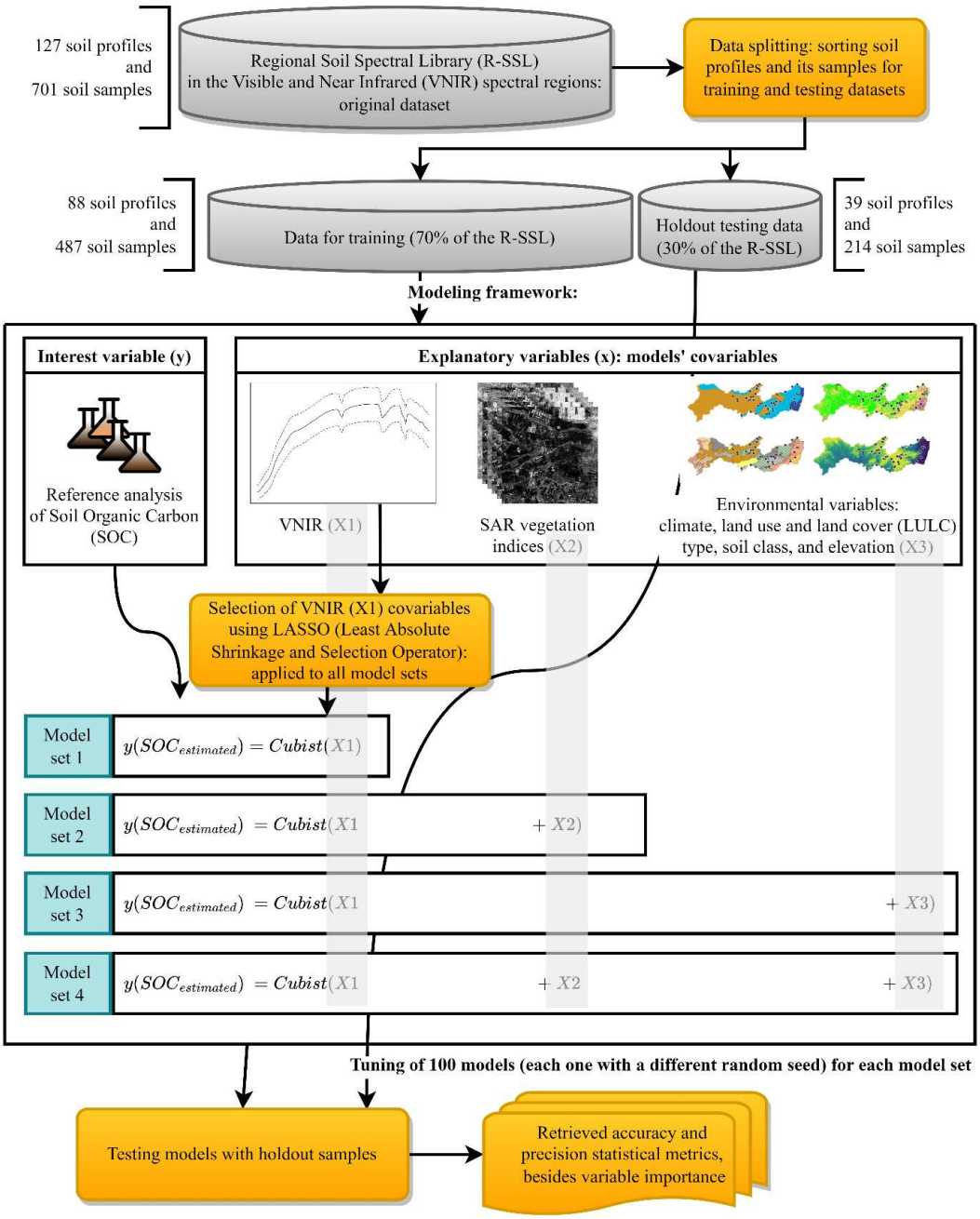


Figure 2.3.2. Soil organic carbon (SOC) content modeling scheme using different predictor variables. SAR, synthetic aperture radar.

2.3.3. Results

2.3.3.1. Environmental variables and SOC levels

The highest levels of SOC were found in forest and other agricultural uses, followed by dense Caatinga whereas the lowest levels were in pasture and agriculture areas (Figure 2.3.3a). Regarding the SOC levels as a function of climate, in Figure 2.3.3b, the lowest SOC levels were in the semi-arid climate (BSh). The highest SOC levels were

found in the Am and Csa climate types, in a soil profile in the Dense Caatinga area (with an elevation of around 970 m).

According to the local elevation (Figure 2.3.3c), the highest SOC levels were observed at altitudes < 200 m, a region where the Am and As climates predominate. In the other elevation ranges, the SOC levels are similar. Finally, regarding soil classes (Figure 2.3.3d), the highest SOC levels were observed in Leptosols, Ferralsols, and Acrisols, and the lowest ones in Regosols.

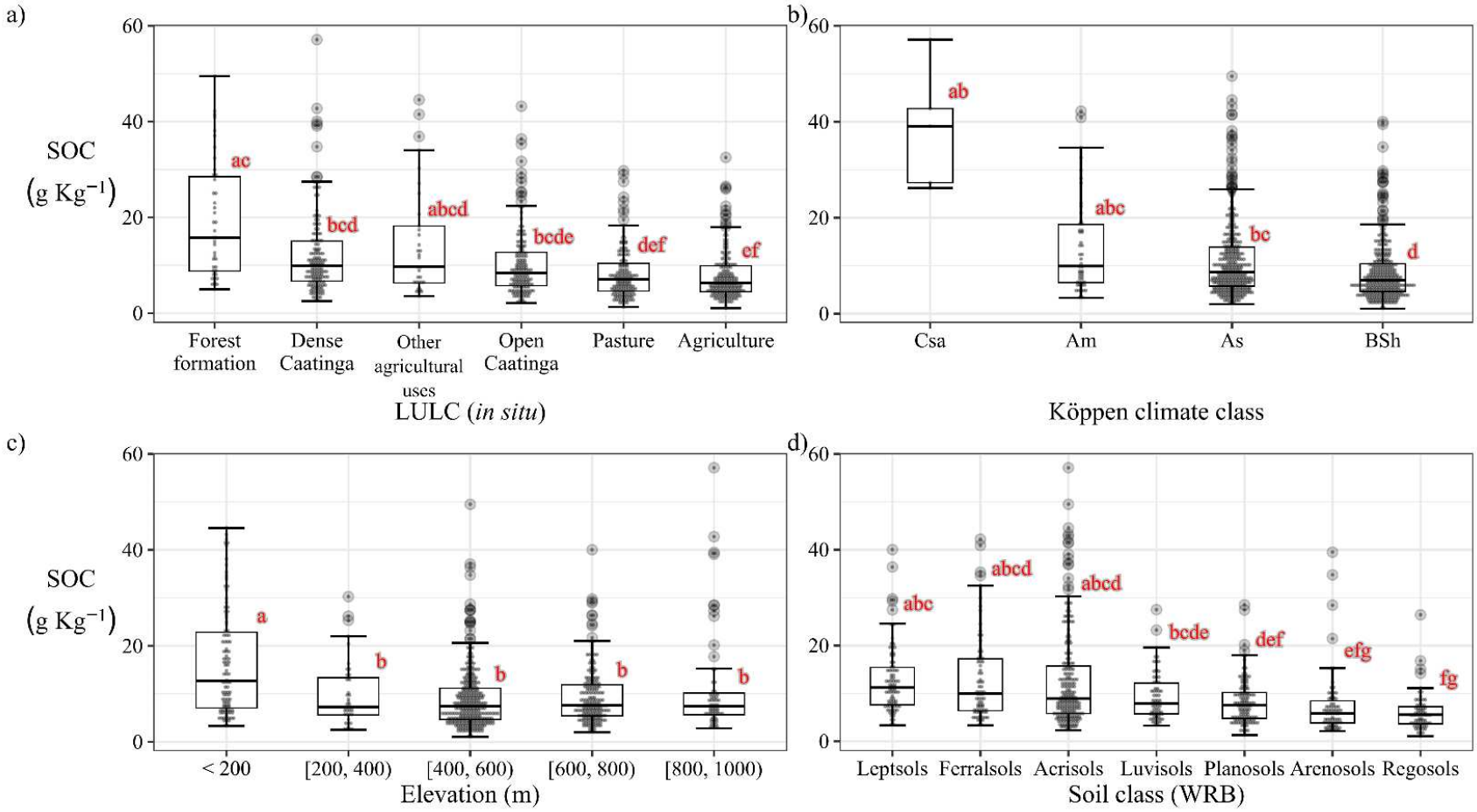


Figure 2.3.3. Distribution of SOC (Soil Organic Carbon) levels in classes of landscape attributes. The following are displayed: a) SOC levels depending on land use and land cover (LULC); b) SOC levels in different climates; c) SOC levels at elevation ranges; d) the SOC levels in each soil class. Groups of distributions accompanied by different letters differ statistically from each other according to the non-parametric Kruskal-Wallis and Dunn tests (P = 0.05).

SAR vegetation indices indicate the amount of vegetation, which can be correlated with the amount of aboveground biomass, the leaf area index, or another biological parameter. Therefore, Figure 2.3.4 shows the distributions of SOC levels and SAR vegetation index values in each LULC class.

The graphs from Figure 2.3.4 show that the DPSVI and DPSVIm indices follow the same pattern of SOC: the highest levels are concentrated in forest areas, dense Caatinga, and other agricultural uses, decreasing values up to the pasture and agriculture classes.

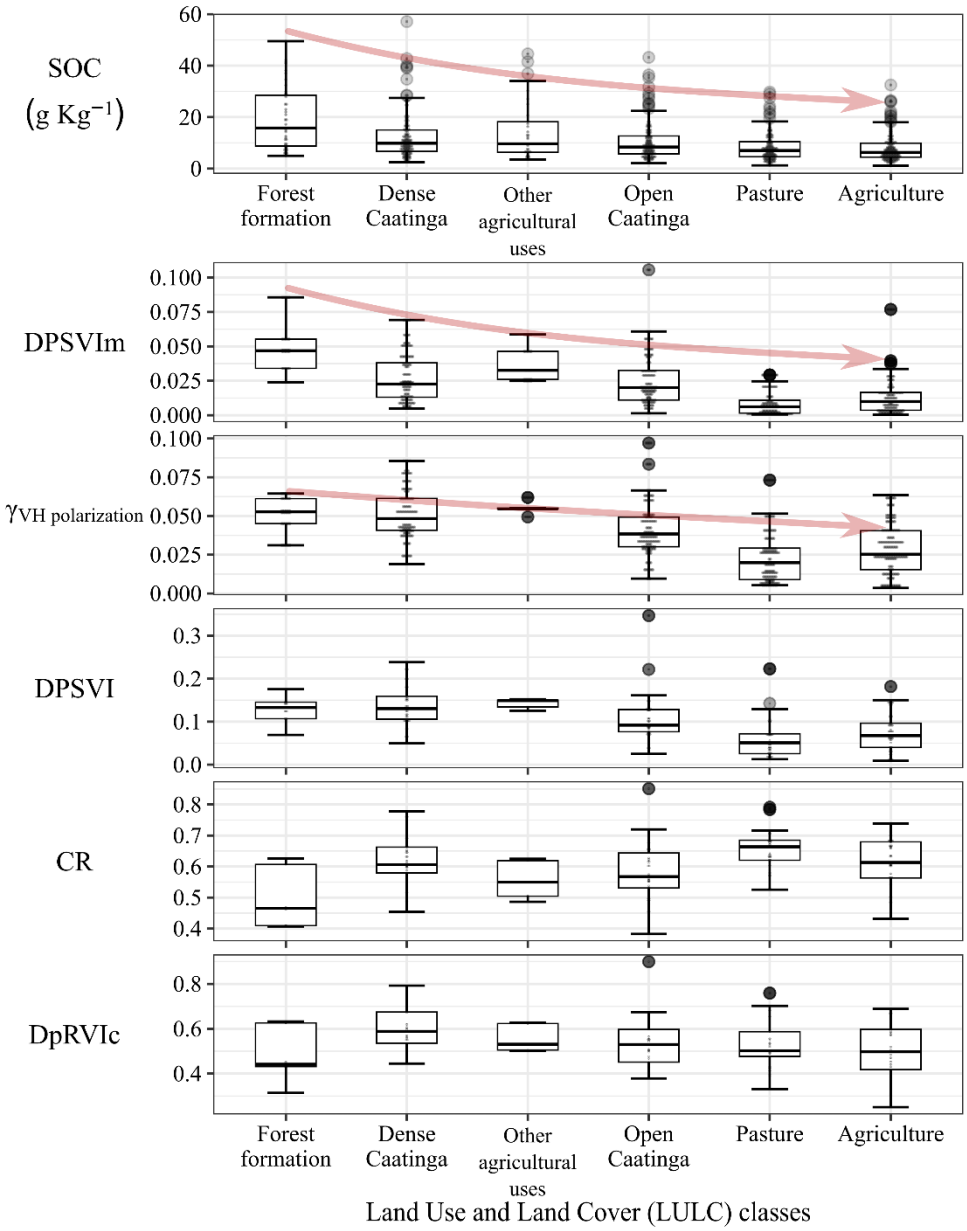


Figure 2.3.4. Distribution of SOC (Soil Organic Carbon) levels and SAR vegetation indices as a function of land use and land cover (LULC). DPSVI, Dual-polarization SAR vegetation index; DPSVIm, modified DPSVI; CR, cross-ratio; DpRVIc, Dual-polarization Radar Vegetation index for Sentinel-1 GRD products.

2.3.3.2. Performance of the models in predicting SOC levels with different covariates

Figure 2.3.5 displays the results of the accuracy and correlation statistics of the predictions of the four *model sets*. Model predictions were obtained using the test dataset, with samples independent from the model training. Furthermore, the results are expressed in boxplots since for each of the four *model sets*, 100 values of each metric were obtained, totaling 400 trained models.

In the RMSE, MAE, and MSE error metrics of each *model set*, it is observed that the more explanatory variables are added to the modeling, the more accurate predictions become (Figure 2.3.5a). The average RMSE in *model set 1* was 4.52 g kg^{-1}. Adding the SAR vegetation indices, the average RMSE was 6.6% lower, equal to 4.23 g kg^{-1}. Using spectral covariates plus environmental covariates, the average RMSE was 4.16 g kg^{-1}, 8% lower than in *model set 1*. Finally, in *model set 4*, which used all available covariates to model SOC levels, the average RMSE was 3.93 g kg^{-1}, 13% lower than *model set 1*. The same behavior was noted for the MAE and MSE metrics.

Figure 2.3.5b) displays the results of the R^2, CCC, and NSE metrics for each *model set*. The result obtained with this set of correlation metrics between predictions and observations corroborates the pattern observed in Figure 2.3.5a): adding environmental covariates improved the model performance, making them more efficient and accurate. The average R^2 obtained in *model set 1* was 0.62, but when radar and environmental covariates were added (*model set 4*) this value increased by 15% to 0.72. The same pattern of improving predictions could be noticed in the other metrics of agreement (CCC) and model efficiency (NSE).

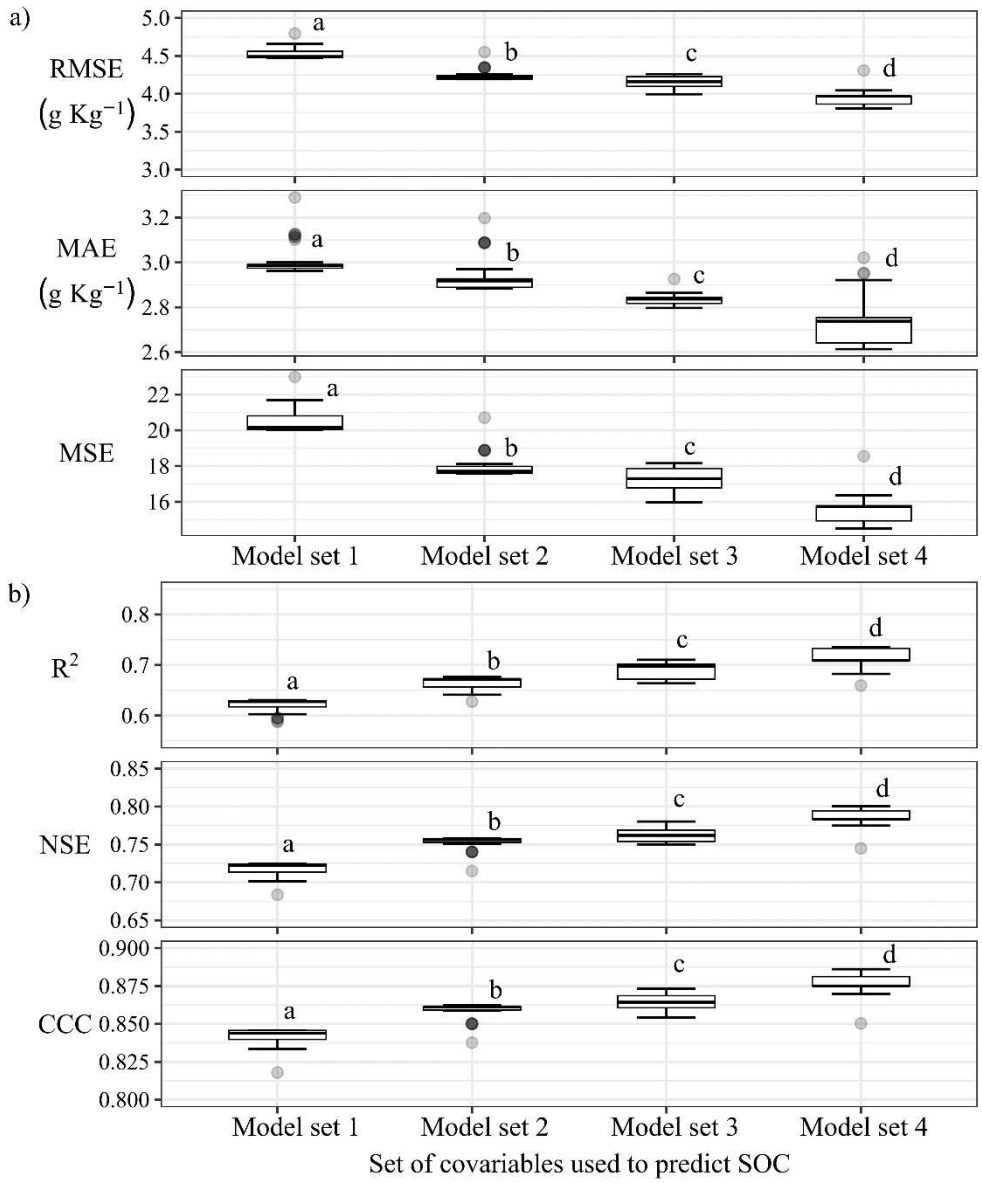


Figure 2.3.5. Distribution of the 100 error and correlation values for each metric (Root Mean Squared Error [RMSE], Mean Absolute Error [MAE], Mean Squared Error [MSE], coefficient of determination [R^2], Lin's concordance correlation coefficient [CCC], and Nash- Sutcliffe model efficiency [NSE]) from the SOC (Soil Organic Carbon) predictions of the four *model sets*. Different lowercase letters indicate significant statistical differences between groups of the same statistical metric (P = 0.05).

It is important to highlight that there was a significant statistical difference between the RMSE, MAE, MSE, R^2, CCC, and NSE values obtained in the four *model sets*. The results of the Kruskal-Wallis non-parametric test can be seen in Table 2.3.2. Furthermore, the results of the paired test (for the metric groups of each *model set*) and Dunn's non-parametric test (after the Kruskal-Wallis test) can be checked on the Supplementary Material (Table C2).

Table 2.3.2. Result of the Kruskal-Wallis test for each statistical metric (Root Mean Squared Error [RMSE], Mean Absolute Error [MAE], Mean Squared Error [MSE],

coefficient of determination [R2], Lin's concordance correlation coefficient [CCC], and Nash- Sutcliffe model efficiency [NSE]): χ^2 is the chi-square statistic of the test, df is the degrees of freedom and (*) indicates a significant statistical difference (P = 0.05) between the model sets.

Tested statistical metric (groups of *model sets*):	χ^2	**df**	**Adjusted *p*-value**
RMSE	344.81	3	0 *
MAE	336.01	3	0 *
MSE	344.76	3	0 *
R^2	351.84	3	0 *
CCC	344.34	3	0 *
NSE	344.41	3	0 *

Note: the asterisk (*) next to the adjusted p-value indicates that the null hypothesis (which establishes that the compared groups are statistically equal) was rejected and the adjusted p-value is statistically significant at P = 0.05.

Numerically, we identified that by adding predictor covariates to the Cubist models the models' error dropped. Although SOC predictions using only Vis-NIR-SWIR bands in Cubist models are already very good, by comparing the predicted and observed SOC values, as shown in Figure 2.3.6, the predictions became even better and more accurate.

Figure 2.3.6 shows the scatterplots between observed and predicted SOC values. The plotted points represent the average value of the predictions of the 100 models, while the vertical bars represent the range of the 100 predicted values for each sample, from the 100 training sessions of each *model set*. In Figure 2.3.6a), which predicted SOC samples using only the Vis-NIR-SWIR bands, the samples are more dispersed (further away from the 1:1 ratio) than in Figure 2.3.6d). In Figure 2.3.6d) SOC samples were predicted using the Vis-NIR-SWIR bands plus all other predictor variables.

Furthermore, comparing Figure 2.3.6a) with Figure 2.3.6d), we noticed that the amplitude of the 100 predictions for each SOC sample decreased. In other words, predictions became more accurate when more variables were used.

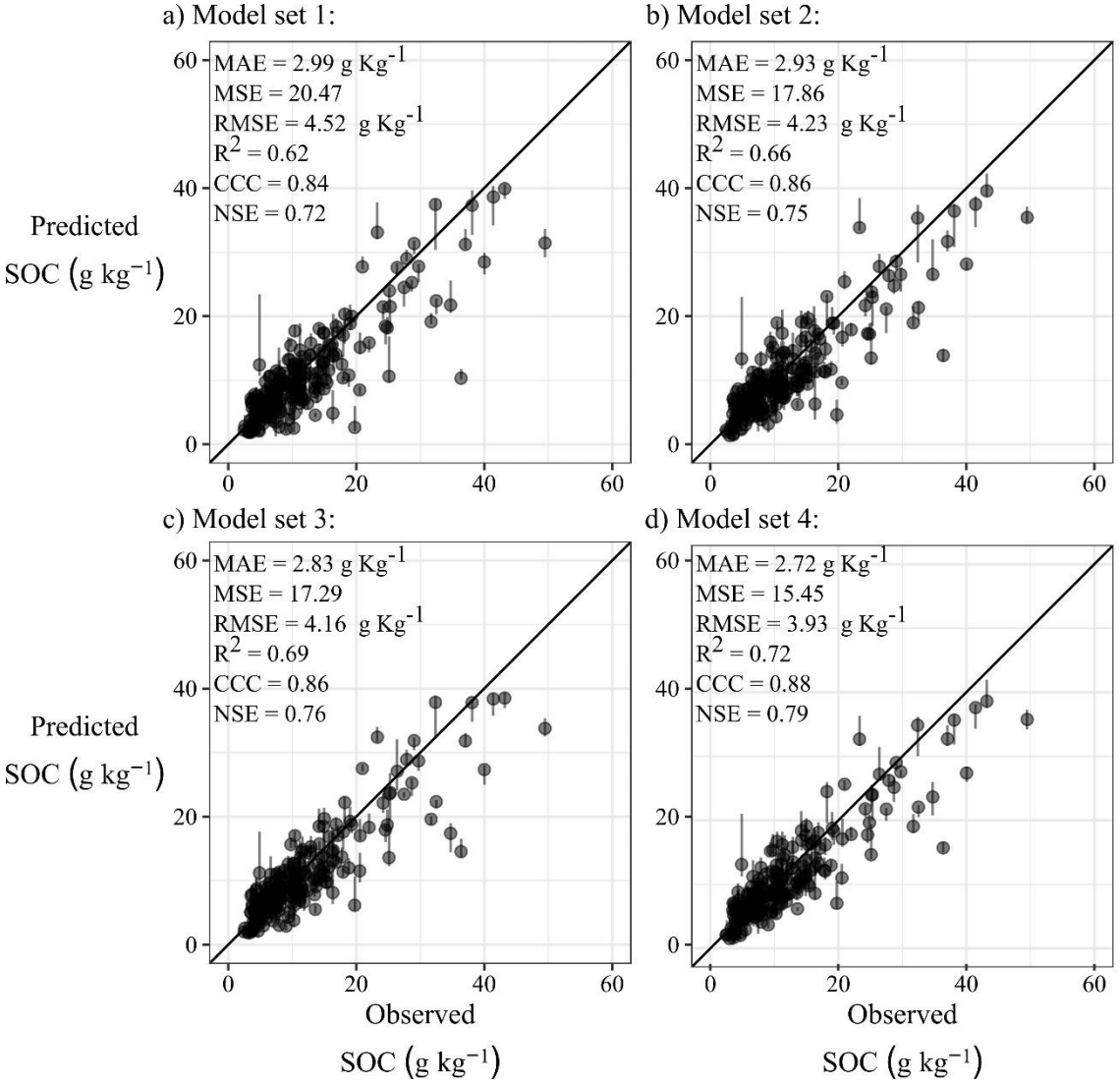


Figure 2.3.6. Scatterplots between observed and predicted soil organic carbon (SOC) values in the four model sets: a) using only spectral variables (Vis-NIR-SWIR); b) Vis-NIR-SWIR bands plus SAR vegetation indices; c) Vis-NIR-SWIR plus environmental covariates; d) all predictors. The plotted points represent the average of the predictions of the 100 models, while the vertical bars represent the range of the 100 predicted values for a given sample.

2.3.3.3. Contribution of different covariates in estimating SOC levels

We observed, in the previous results, improvements in the accuracy and precision of SOC content estimates when covariates were added to the spectroscopy models. However, it is necessary to better understand the contribution of these environmental variables (vegetation types and indices, climate classes, soil in addition to elevation) used for the estimates. Thus, Figure 2.3.7 displays the relative importance of these covariates.

Figure 2.3.7a) shows the relative importance of individual spectral bands selected by the LASSO method for Cubist regression. In this graph, the plotted points represent the highest importance measured, while the vertical bars represent the amplitude of the measured importance for each band. In the graphs in Figure 2.3.7b), the relative

importance of the covariates was grouped into boxplots to aid the comparison between the variable groups. The sub-graphs in Figure 2.3.7b) were divided according to the model's settings.

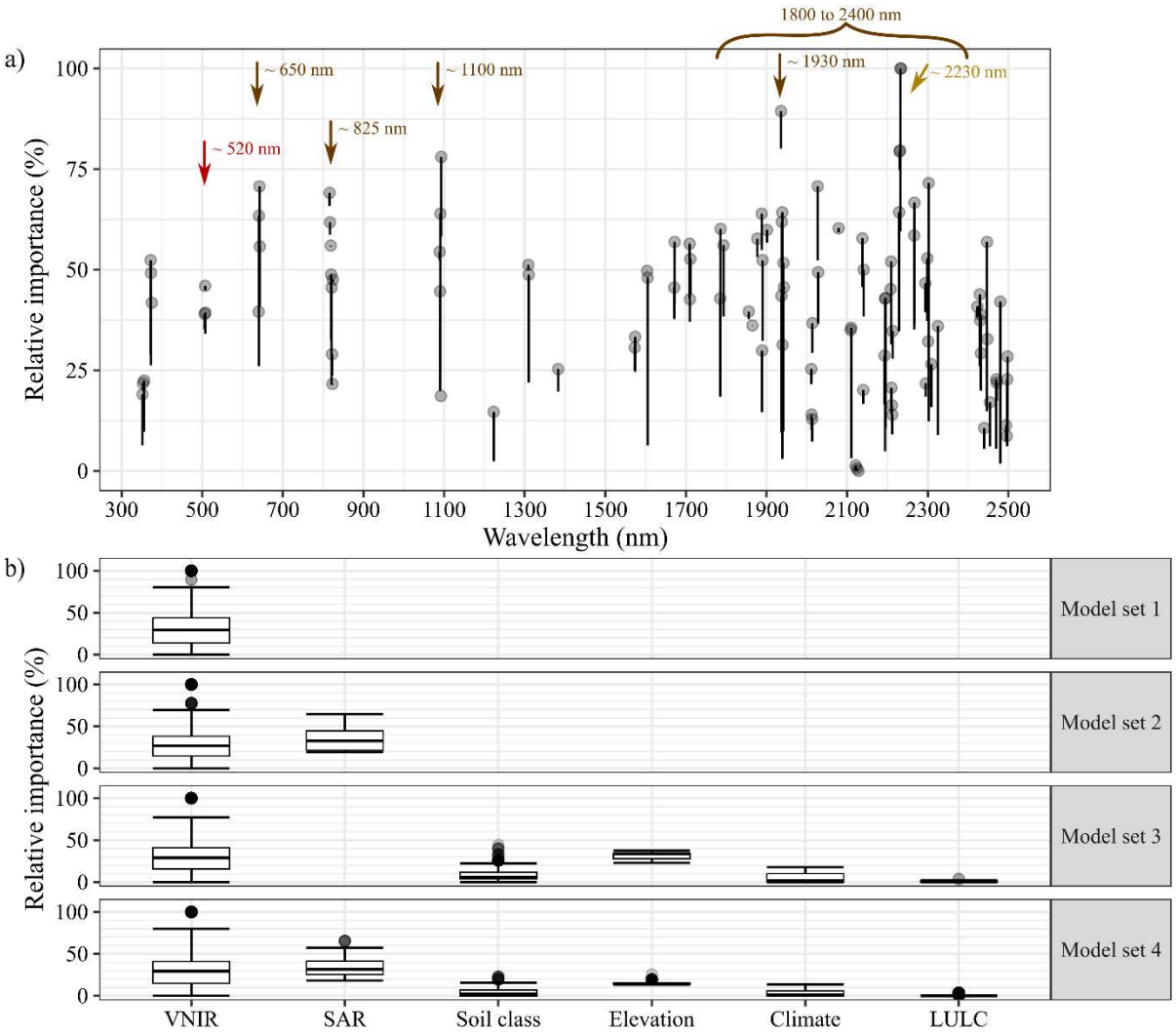


Figure 2.3.7. Relative importance graphs of different predictor variables for modeling soil organic carbon (SOC) content using the Cubist regression method. The importance of the Vis-NIR-SWIR spectral bands for *model set 1* is shown in a), while graph b) displays the relative importance of the other covariates for all *model sets*.

Different spectral bands were selected by LASSO for SOC modeling via Cubist. Figure 2.3.7a) shows the importance and indicates specific wavelengths at which the measured reflectance is intrinsically related to soil organic particles. The spectral bands had the highest importance for SOC prediction, being used by Cubist in its internal decision trees, as well as in internal linear models.

After the spectral bands, the highest importance values were given to the covariates derived from the Sentinel-1 SAR: the importance of these variables ranged from 19 to 65%. It is important to highlight that the importance of covariates for Cubist

reflects the average number of times a variable was used for data division and prediction (in decision trees and linear models, respectively).

2.3.4. Discussion

2.3.4.1. Relation of environmental variables with SOC levels

There are variations in the SOC levels of the R-SSL samples that can be associated with different environmental (LULC, climate, and elevation) and pedological (soil classes) covariates. Regarding the LULC's vegetation classes, SOC levels were higher in places where the soil receives a greater contribution of plant residues and is not managed (such as the disturbance of surface layers). The input of organic plant residues is the main source of soil organic matter, and the vegetation stratum is an indicator of SOC content in the soil (Jobbágy and Jackson, 2000; Wiesmeier et al., 2019). Although it is known that in areas occupied by pastures, SOC levels may exceed those from original forest conditions (Guo and Gifford, 2002), including in the Caatinga biome (Ferreira et al., 2016), from where the majority of R-SSL samples in this study come from, this also depends on the management of the pastures (Medeiros et al., 2020), including the degree of degradation, if any. Hence, the differences in SOC levels between the classes with the highest amount of above-ground biomass (Forest Formation and Dense Caatinga) and those with the lowest level of plant biomass (Pasture and Agriculture) (Figure 2.3.3a).

Climate is also a source of variation in SOC levels in R-SSL samples. Temperature and rainfall, both meteorological variables used in existing climate typologies, are the main factors in soil SOC storage at regional scales. Temperature and rainfall control the primary vegetation production rates (input of organic matter into the soil). Furthermore, temperature influences soil microbial activity, microorganisms responsible for soil respiration, and decomposition of organic matter (Guo and Gifford, 2002; Wiesmeier et al., 2019).

The lowest SOC levels can be found in Figure 2.3.3b) in areas with a semi-arid climate, with an average SOC of 8.63 g kg^{-1}. In these areas, temperatures are higher, and annual rainfall is lower, which increases the rate of decomposition of organic matter and reduces the input of biomass, compared to areas under other climate typologies. The highest levels (average SOC of 38.48 g kg^{-1}) were observed in a profile under a humid subtropical climate with a dry and hot summer (Csa). These climate types occur in the semi-arid region of Brazil mainly due to the relief in plateau regions. In this case, it is an area with an elevation between 800 and 1,000 m in the Borborema Plateau, in the State of Pernambuco (Alvares et al., 2013; de Souza et al., 2022).

There was no significant difference between the SOC levels found in elevations higher than 200 m above sea level (Figure 2.3.3c). There were, however, higher SOC levels in elevations lower than 200 m, as well as outliers in elevations greater than 800 m. We can say that in both cases, the SOC levels can be related to the climate. In the first case, in elevations lower than 200 m, there is a predominance of tropical climates with monsoons and dry summers, where precipitation is also higher (1,000 to 2,500 mm $year^{-1}$). The average SOC level with elevation < 200 m was 16.45 g kg^{-1}. In the second case,

in elevations greater than 800 m, although the average SOC level was 12.25 g kg^{-1}, very high SOC values (> 30 g kg^{-1}) were found under Csa climate.

A possible explanation for this SOC behavior regarding elevation is that elevation affects temperature, and consequently, the climate. In other words, on a regional scale, the effect of elevation on SOC variations is related to the influence of elevation on environmental parameters that affect other soil formation factors, notably average annual temperature and rainfall (Hobley et al., 2015, 2016; Wiesmeier et al., 2019). Areas with a humid subtropical climate in the Brazilian semi-arid region, for example, have elevation as the main factor for their occurrence, as the aforementioned Borborema Plateau (Alvares et al., 2013; de Souza et al., 2022).

Soil type is also strongly associated with organic carbon storage. In many cases and classification systems, SOC content is part of the classification criteria (Mayes et al., 2014). However, the soil type only reflects soil properties that influence the supply and storage of organic carbon.

During pedogenesis, weathering reactions lead to changes in soil mineralogy, which strongly influence the surface area of mineral reactivity and carbon storage (Wiesmeier et al., 2019). Therefore, SOC levels may be higher in soils with a greater propensity for physical protection of particles, whether in soil aggregates or by clay and silt particles (Six et al., 2002; Stewart et al., 2008; Yu et al., 2019). Furthermore, metal sesquioxides (such as Fe and Al oxides) also contribute to the stabilization of SOC through the affinity between Fe and Al oxides and organic compounds (Baldock and Skjemstad, 2000; Wiesmeier et al., 2019; Yu et al., 2019). This is an indicator of what happens with the Ferralsols and Acrisols classes, which appear in Figure 2.3.3d) with the highest SOC levels. On the other hand, the Arenosols (sandy soils) and Regosols appear among the classes with the lowest SOC levels.

2.3.4.2. Accuracy and precision improvement in modeling strategies

Specifically dealing with *model set 1*, whose SOC predictors were only the spectral bands in Vis-NIR-SWIR, the accuracy and precision values obtained are compatible with those found in the literature. The average RMSE was about 4.5 g kg^{-1} while the accuracies reported in the literature range from 0.3 to 25 g kg^{-1} (Demattê et al. 2019b; Soriano-Disla et al. 2014). This error amplitude found in the literature, however, depends on the size, location, representation scale of the spectral library used, and amplitude of the SOC values. Furthermore, a similar RMSE value was found by dos Santos et al. (2023), around 4.2 g kg^{-1}, applying the same validation/test methodology on the same R-SSL.

Noteworthy, the spectral bands selected by LASSO for input into the Cubist models are intrinsically related to either the SOC levels or the presence of clay minerals and iron oxides. In the Vis-NIR spectra (400 to 1,000 nm) the spectral bands can be related to iron oxides, especially in the peaks around 520 and 650 nm (Viscarra Rossel and Behrens, 2010), as well as to the presence of organic compounds (650 and 825 nm peaks) (Ben-Dor, 1997; Nocita et al., 2015; Stevens et al., 2013; Viscarra Rossel and Behrens, 2010), as highlighted in Figure 2.3.7a. The peaks highlighted in the NIR region

(around 1100 nm) are associated with water and organic compounds (Viscarra Rossel and Behrens, 2010). In the infrared shortwave, specifically from 1800 to 2400 nm, the selected spectral bands can be associated with different organic compounds, such as amines, alkyls, carboxylic acids (especially at 1930 nm), amides, aliphatic compounds, methyls, phenolic compounds, polysaccharides and carbohydrates (Coates, 2006; Knadel et al., 2015; Vasques et al., 2010; Viscarra Rossel and Behrens, 2010).

Among the selected spectral bands, there are also clay mineral signatures. This association can be made with the relative importance of the bands around 2230 nm (also highlighted in Figure 2.3.7a), which indicates the presence of aluminum hydroxides, Al–OH (Viscarra Rossel and Behrens, 2010; Zheng et al., 2016).

The same selected spectral bands (Figure 2.3.7a) were used in all model configurations, from *model sets 1* to *4*. Therefore, the results of the statistical performance metrics of the *model sets* (Figure 2.3.5 and Figure 2.3.6) show that both accuracy and precision improved in the Vis-NIR-SWIR spectroscopy predictions as environmental variables were added to the model. This corroborates the results found by Moura-Bueno et al. (2021), in subtropical soils from southern Brazil, and Wang et al. (2022), in soils in northern China.

When adding SAR vegetation indices to the Vis-NIR-SWIR spectral bands, the average RMSE dropped about 6.6 %, from 4.52 to 4.23 g kg^{-1} (see Figure 2.3.6a and Figure 2.3.6b). Also, the average R^2 increased 6.3%, from 0.62 to 0.66. Furthermore, we observed in Figure 2.3.7b that the SAR vegetation indices have relative importance ranging from 20 to 65%, both in *model set 2* and in *model set 4*.

SAR vegetation indices are good indicators of the amount of aboveground biomass (AGB). Although the backscatter from the vegetated surface for SAR sensors is not a direct measure of AGB (Woodhouse et al., 2012), the backscatter observed over these areas in cross-polarizations (VH or HV) is directly associated with the AGB content of that area (Bispo et al., 2020; dos Santos et al., 2021; Joshi et al., 2017; Saatchi, 2019; Santoro et al., 2021). The plant organs that most interact with microwave radiation usually depend on the wavelength used. While radiation in the C band (with a wavelength of approximately 6 cm) tends to interact more with leaves and branches, in areas with denser vegetation, radiation in the L band ($\lambda \cong 23$ cm) can interact with tree trunks (Flores-Anderson et al., 2019). In any case, in crossed polarizations, the surface elements that change the polarization state of the electromagnetic wave that reflects to the sensor appear brighter (Mitchard et al., 2011) in a wave reflection mechanism known as volumetric backscattering (Woodhouse, 2006).

The vegetation indices employed use the VV and VH polarizations of the Sentinel-1 images, to measure and represent the biophysical parameters of the detected vegetation (Santos et al., 2023). DpRVIc is an index based on the polarization degree that vegetation causes in microwave radiation (Bhogapurapu et al., 2022; Mandal et al., 2020a). The polarization degree measures how much of the total energy backscattered by the targets has had its polarization modified.

The DPSVI and DPSVIm indices are also based on the depolarization degree of the signal, however, they seek to distinguish surfaces of water bodies and bare soil from

vegetated areas. The values of the indices are close to zero for water bodies and bare soil and increase with the amount of AGB (dos Santos et al., 2021; Periasamy, 2018). The difference between DPSVI and DPSVIm is that the latter is more sensitive to different levels of biomass in forest areas (dos Santos et al., 2021). For this reason, DPSVIm has incorporated the CR index (Frison et al., 2018) to ease the separation of different AGB levels in these areas.

The SAR indices with the greatest relative importance in SOC modeling were the VH polarization and the DPSVIm index, whose relative importance ranged from 45 to 65%. Considering the characteristics of the SAR vegetation indices, we can conclude that they contribute to the prediction of SOC levels because they are capable of representing the plant residue intake in the soil, which is more correlated with the SOC contents found in the surface layers of the soil (Guo and Gifford, 2002; Hobley et al., 2015; Wiesmeier et al., 2019).

By adding soil formation factors (elevation, soil type, vegetation type, and climate) to R-SSL to predict SOC levels the model's performance also improved. So, from *model set 1* to *model set 3*, the average RMSE fell about 8%, from 4.52 to 4.16 g kg^{-1}. The average R^2 obtained in *model set* 3 was 0.69, 11% higher than in *model set 1*. This can be seen in Figure 2.3.6a and Figure 2.3.6c.

The relative importance of the covariates elevation, soil type, vegetation type, and climate for modeling with R-SSL ranged from 0 to 40% (Figure 2.3.7b). Except for elevation, a numerical variable that was used for both stages: decision and regression; all other covariates are categorical, and were used only in the decision trees of the Cubist model to separate the samples based on similar SOC levels.

Using this set of categorical variables, Cubist builds spectral sub-libraries with samples grouped by similarity to reduce the SOC variance as a function of the classes of environmental variables. Moura-Bueno et al. (2020, 2019) studied the effect of SSL stratification on SOC prediction performance. The authors tested the hypothesis that reducing SOC variance, after stratifying SSL based on environmental, pedological, and spectral class criteria/variables, could improve the accuracy of SOC estimates. Moura-Bueno et al. (2020) concluded in their study with SSL of subtropical soils in southern Brazil that stratification may increase accuracy as long as there is a significant difference in SOC between classes of environmental variables. Hence, the stratification with the categorical LULC and Physiographic Region produced better predictions of SOC. However, the disadvantage found in manual stratification was the reduction in the number of samples available for model calibration (Moura-Bueno et al., 2020).

For *model set 3*, after the Vis-NIR-SWIR normalized reflectance bands, the relative importance given to elevation ranged from 22 to 37%. Although there is no statistically significant difference between most elevation classes (Figure 2.3.3c), in the Northeast region of Brazil, elevation is closely related to the occurrence of different climates (Alvares et al., 2013). Where the elevation was less than 200 m, there were wetter and rainier areas and higher SOC levels (see Figure 2.3.3c).

In turn, the climate, which is defined by the normal pattern of rainfall and temperature, will influence both the rate of primary production of vegetation (organic

carbon input), such as decomposition and emission through microbial respiration (carbon output). So much so that in *model set 3* the BSh climate class, with higher temperatures and lower rainfall, had the greatest importance (relative importance of around 17%) compared to the other classes in the stratification of the SOC samples for Cubist.

Hobley et al. (2015, 2016) studied SOC variations in Western Australian soils as a function of different indicators and drivers (SOC input/output modulators). The authors identified that climate is an important driver of SOC, with higher rainfall associated with higher proportions of organic carbon in humus (Hobley et al., 2016). For this reason, when predicting SOC, numerical variables, mainly average annual rainfall, had greater importance for the results. Corroborating the relationship that rainfall has on primary production (Michaletz et al., 2014) and SOC storage mainly in surface layers (Jobbágy and Jackson, 2000; Wiesmeier et al., 2019).

In *model set 3*, the pedological variables, essentially the dummy variables for the Arenosols and Regosols classes, had greater relative importance (about 44%), after the Vis-NIR-SWIR covariates. The stratification of samples of these two soil types of soil, which are mainly high sandy soils, among the others, was important to improve the modeling accuracy.

This result corroborates the conclusions of the study by Jaconi et al. (2017). These authors studied SSL stratification strategies in the NIR spectrum, at a country scale (in Germany), created from soil samples under agricultural and pasture use. The best strategy for the accuracy of SOC estimates was to separate samples of sandy soils from samples of other textural classes, calibrating two models separately (Jaconi et al., 2017). The basis for this is the already discussed relationship of greater SOC storage capacity in soils with higher clay and silt contents, due to organic matter protection mechanisms.

Although the relationship between the LULC classes and SOC variations is known, as can be seen in Figure 2.3.3b) discussed in the previous topic, the lowest relative importance in *model set 3* was given to these variables by Cubist. This does not necessarily mean that these variables are not important for predicting SOC. Variations in uses and coverage are considered important indicators and drivers of SOC at different scales, mainly in the surface layers of the soil (Guo and Gifford, 2002; Hobley et al., 2015, 2016; Moura-Bueno et al., 2021; Wiesmeier et al., 2019). However, for the Cubist models of *model set 3* with R-SSL, stratification of samples using climate, elevation, and pedology were more important.

The combination of all predictors, *model set 4*: Vis-NIR-SWIR spectral bands, SAR vegetation indices, elevation, soil type, climate, and vegetation cover; produced the best SOC estimates. From *model set 1* to *model set 4*, the average RMSE dropped about 13%, from 4.52 to 3.93 g kg^{-1}. Following, the average R^2 increased by 15%, going from 0.62 (*model set 1*) to 0.72 (in *model set 4*). The other statistical metrics followed the same trends (see Figure 2.3.6).

Moura-Bueno et al. (2021) obtained, by comparing a model with only Vis-NIR-SWIR spectral signatures to another with spectral bands plus a set of auxiliary covariates, a reduction in RMSE of about 22%, and an R^2 increasing from 0.76 to 0.85. In a similar approach, but to predict soil organic matter content, Wang et al. (2022) obtained an RMSE

of 3.85 g kg^{-1} and R^2 of 0.85 with the addition of auxiliary environmental covariates and spectral classification to the Vis-NIR-SWIR spectral signatures. When using only spectral signatures the RMSE and R^2 were 4.30 g kg^{-1} and 0.76, respectively.

For more precise applications, in which it is necessary to make as few mistakes as possible, more reliable estimates are required. This is the case with precision agriculture applications, whose goal sometimes is to map small variations in soil properties on a more detailed scale (Camargo et al., 2022). Other applications of SOC estimates by spectroscopy include fertilization recommendations (Barra et al., 2021; Rosin et al., 2020), and soil monitoring to implement good soil management practices (Angelopoulou et al., 2020). In these situations, reducing the error when using auxiliary variables is justified.

It is necessary to take into account that there is a cost with the adoption of auxiliary remote sensing products for SSLs, which is the need for more data processing. In this case, the processing of satellite images to generate the vegetation indices commonly used in digital soil mapping. However, the advantage is that high-resolution products are needed for estimates at smaller scales, from watershed (Kunkel et al., 2022) to field scale for prediction in agricultural areas (Nguyen et al., 2022).

An important issue related to SAR vegetation indices is that indices derived from optical orbital sensors are already widely used for digital mapping of soil organic carbon/organic matter. However, the use of radar remote sensing has three major advantages, inherent to microwave systems and the Sentinel-1 mission, which can provide operability for monitoring soil, agriculture, forestry, and environmental activities with SSLs. The first advantage is the low (or almost zero) cloud interference in imaging, which is a problem in optical remote sensing, mainly for tropical regions (Asner, 2001; Carrasco et al., 2019; dos Santos et al., 2022; Flores-Anderson et al., 2019; Roy and Yan, 2020). The second advantage is the high temporal and spatial resolution since the high operability of the Sentinel-1 mission with the constellation of two satellites (Sentinel-1A, -1B [inoperative since 2022] and -1C [planned for launch]) allows an almost weekly revisit rate (for some regions) (Mandal et al., 2020a; Periasamy, 2018). Furthermore, the GRD products, for which the vegetation indices used were designed, are distributed in global coverage and at high spatial resolution (ESA, 2012, 2022). Therefore, satellite products can be applied at different soil monitoring scales, and when obtained in time series, they can represent vegetation dynamics and/or land cover management (Kunkel et al., 2022).

In addition to high-resolution remote sensing, future work at local scales may benefit from other relief parameters that aid SSLs. This is the case with the topographic wetness index (TWI). The TWI has a good capacity to help explain SOC variations in the landscape on small scales, as it is an indicator of water movement and availability in the soil (Hobley et al., 2015; Sørensen et al., 2006; Wiesmeier et al., 2019).

2.3.5. Conclusions

Adding SAR vegetation indices (obtained by orbital remote sensing) to the R-SSL in Vis-NIR-SWIR (obtained by laboratory proximal sensing) improved SOC spectral estimates. The combination of categorical and continuous variables describing the environment and soil formation factors (climate, soil type, land use and land cover class, and elevation) to the R-SSL in Vis-NIR-SWIR resulted in a significant improvement in the estimates of SOC.

2.3.6. References

Adi, S.H., Grunwald, S., Tafakresnanto, C., 2019. Fusing environmental variables into soil spectroscopy modeling using a novel two-step regression method. IOP Conf. Ser. Earth Environ. Sci. 393, 012100. https://doi.org/10.1088/1755-1315/393/1/012100

Alvares, C.A., Stape, J.L., Sentelhas, P.C., de Moraes Gonçalves, J.L., Sparovek, G., 2013. Köppen's climate classification map for Brazil. Meteorol. Z. 22, 711–728. https://doi.org/10.1127/0941-2948/2013/0507

Angelopoulou, T., Balafoutis, A., Zalidis, G., Bochtis, D., 2020. From Laboratory to Proximal Sensing Spectroscopy for Soil Organic Carbon Estimation—A Review. Sustainability 12, 443. https://doi.org/10.3390/su12020443

ASF, 2022. Copernicus Sentinel data 2017, 2018, and 2019. Retrieved from ASF DAAC, processed by ESA. [WWW Document]. URL https://asf.alaska.edu/ (accessed 11.17.22).

Asner, G.P., 2001. Cloud cover in Landsat observations of the Brazilian Amazon. Int. J. Remote Sens. 22, 3855–3862. https://doi.org/10.1080/01431160010006926

Baldock, J.A., Skjemstad, J.O., 2000. Role of the soil matrix and minerals in protecting natural organic materials against biological attack. Org. Geochem. 31, 697–710. https://doi.org/10.1016/S0146-6380(00)00049-8

Barra, I., Haefele, S.M., Sakrabani, R., Kebede, F., 2021. Soil spectroscopy with the use of chemometrics, machine learning and pre-processing techniques in soil diagnosis: Recent advances–A review. TrAC Trends Anal. Chem. 135, 116166. https://doi.org/10.1016/j.trac.2020.116166

Bellon-Maurel, V., McBratney, A., 2011. Near-infrared (NIR) and mid-infrared (MIR) spectroscopic techniques for assessing the amount of carbon stock in soils – Critical review and research perspectives. Soil Biol. Biochem. 43, 1398–1410. https://doi.org/10.1016/j.soilbio.2011.02.019

Ben-Dor, E., 1997. The reflectance spectra of organic matter in the visible near-infrared and short wave infrared region (400–2500 nm) during a controlled decomposition process. Remote Sens. Environ. 61, 1–15. https://doi.org/10.1016/S0034-4257(96)00120-4

Ben-Dor, E., Chabrillat, S., Demattê, J.A.M., Taylor, G.R., Hill, J., Whiting, M.L., Sommer, S., 2009. Using Imaging Spectroscopy to study soil properties. Remote Sens. Environ. 113, S38–S55. https://doi.org/10.1016/j.rse.2008.09.019

Bhogapurapu, N., Dey, S., Mandal, D., Bhattacharya, A., Karthikeyan, L., McNairn, H., Rao, Y.S., 2022. Soil moisture retrieval over croplands using dual-pol L-band GRD SAR data. Remote Sens. Environ. 271. https://doi.org/10.1016/J.RSE.2022.112900

Bispo, P. da C., Rodríguez-Veiga, P., Zimbres, B., do Couto de Miranda, S., Henrique Giusti Cezare, C., Fleming, S., Baldacchino, F., Louis, V., Rains, D., Garcia, M., Del Bon Espírito-Santo, F., Roitman, I., Pacheco-Pascagaza, A.M., Gou, Y., Roberts, J., Barrett, K., Ferreira, L.G., Shimbo, J.Z., Alencar, A., Bustamante, M., Woodhouse, I.H., Eyji Sano, E., Ometto, J.P., Tansey, K., Balzter, H., 2020. Woody Aboveground Biomass Mapping of the Brazilian Savanna with a Multi-Sensor and Machine Learning Approach. Remote Sens. 12, 2685. https://doi.org/10.3390/rs12172685

Brown, D.J., Bricklemyer, R.S., Miller, P.R., 2005. Validation requirements for diffuse reflectance soil characterization models with a case study of VNIR soil C prediction in Montana. Geoderma 129, 251–267. https://doi.org/10.1016/j.geoderma.2005.01.001

Brus, D. j., Kempen, B., Heuvelink, G. b. m., 2011. Sampling for validation of digital soil maps. Eur. J. Soil Sci. 62, 394–407. https://doi.org/10.1111/j.1365-2389.2011.01364.x

Camargo, L.A., do Amaral, L.R., dos Reis, A.A., Brasco, T.L., Magalhães, P.S.G., 2022. Improving soil organic carbon mapping with a field-specific calibration approach through diffuse reflectance spectroscopy and machine learning algorithms. Soil Use Manag. 38, 292–303. https://doi.org/10.1111/sum.12775

Carrasco, L., O'Neil, A., Morton, R., Rowland, C., 2019. Evaluating Combinations of Temporally Aggregated Sentinel-1, Sentinel-2 and Landsat 8 for Land Cover Mapping with Google Earth Engine. Remote Sens. 11, 288. https://doi.org/10.3390/rs11030288

Clark, R.N., Roush, T.L., 1984. Reflectance spectroscopy: Quantitative analysis techniques for remote sensing applications. J. Geophys. Res. Solid Earth 89, 6329–6340. https://doi.org/10.1029/JB089iB07p06329

Coates, J., 2006. Interpretation of Infrared Spectra, A Practical Approach, in: Encyclopedia of Analytical Chemistry. John Wiley & Sons, Ltd, Chichester, UK. https://doi.org/10.1002/9780470027318.a5606

de Souza, J.J.L.L., Souza, B.I., Xavier, R.A., Cardoso, E.C.M., de Medeiros, J.R., da Fonseca, C.F., Schaefer, C.E.G.R., 2022. Organic carbon rich-soils in the brazilian semiarid region and paleoenvironmental implications. CATENA 212, 106101. https://doi.org/10.1016/j.catena.2022.106101

De Zan, F., Guarnieri, A.M., 2006. TOPSAR: Terrain observation by Progressive Scans. IEEE Trans. Geosci. Remote Sens. https://doi.org/10.1109/TGRS.2006.873853

Demattê, José Alexandre M., Dotto, A.C., Bedin, L.G., Sayão, V.M., Souza, A.B. e, 2019a. Soil analytical quality control by traditional and spectroscopy techniques: Constructing the future of a hybrid laboratory for low environmental impact. Geoderma 337, 111–121. https://doi.org/10.1016/j.geoderma.2018.09.010

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

dos Santos, E.P., da Silva, D.D., do Amaral, C.H., 2021. Vegetation cover monitoring in tropical regions using SAR-C dual-polarization index: seasonal and spatial influences. Int. J. Remote Sens. 42, 7581–7609. https://doi.org/10.1080/01431161.2021.1959955

dos Santos, E.P., da Silva, D.D., do Amaral, C.H., Fernandes-Filho, E.I., Dias, R.L.S., 2022. A Machine Learning approach to reconstruct cloudy affected vegetation indices imagery via data fusion from Sentinel-1 and Landsat 8. Comput. Electron. Agric. 194, 106753. https://doi.org/10.1016/j.compag.2022.106753

dos Santos, E.P., Moreira, M.C., Fernandes-Filho, E.I., Demattê, J.A.M., Santos, U.J. dos, da Silva, D.D., Cruz, R.R.P., Moura-Bueno, J.M., Santos, I.C., Sampaio, E.V. de S.B., 2023. Improving the generalization error and transparency of regression models to estimate soil organic carbon using soil reflectance data. Ecol. Inform. 77, 102240. https://doi.org/10.1016/j.ecoinf.2023.102240

ESA, 2012. Sentinel-1: ESA's Radar Observatory Mission for GMES Operational Services. European Space Agency.

ESA, E.S.A., 2022. Sentinel-1 SAR Technical Guide [WWW Document]. URL https://sentinel.esa.int/web/sentinel/technical-guides/sentinel-1-sar (accessed 11.18.22).

FAO, 2020. A protocol for measurement, monitoring, reporting and verification of soil organic carbon in agricultural landscapes: GSOC-MRV Protocol. FAO, Rome, Italy. https://doi.org/10.4060/cb0509en

Ferreira, A.C.C., Leite, L.F.C., de Araújo, A.S.F., Eisenhauer, N., 2016. Land-Use Type Effects on Soil Organic Carbon and Microbial Properties in a Semi-arid Region of Northeast Brazil. Land Degrad. Dev. 27, 171–178. https://doi.org/10.1002/ldr.2282

Filipponi, F., 2019. Sentinel-1 GRD Preprocessing Workflow. Proceedings 18, 11. https://doi.org/10.3390/ECRS-3-06201

Flores-Anderson, A.I., Herndon, K.E., Thapa, R.B., Cherrington, E. (Eds.), 2019. The Synthetic Aperture Radar (SAR) Handbook: Comprehensive Methodologies for Forest Monitoring and Biomass Estimation. NASA, Huntsville. https://doi.org/10.25966/nr2c-s697

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. J. Stat. Softw. 33, 1–22. https://doi.org/10.18637/jss.v033.i01

Frison, P.-L., Fruneau, B., Kmiha, S., Soudani, K., Dufrêne, E., Toan, T.L., Koleck, T., Villard, L., Mougin, E., Rudant, J.-P., 2018. Potential of Sentinel-1 Data for Monitoring Temperate Mixed Forest Phenology. Remote Sens. 10, 2049. https://doi.org/10.3390/rs10122049

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

Greenwell, B., Boehmke, B., Gray, B., 2020. vip: Variable Importance Plots.

Greenwell, B.M., Boehmke, B.C., 2020. Variable Importance Plots—An Introduction to the vip Package. R J. 12, 343. https://doi.org/10.32614/RJ-2020-013

Guo, L.B., Gifford, R.M., 2002. Soil carbon stocks and land use change: a meta analysis. Glob. Change Biol. 8, 345–360. https://doi.org/10.1046/J.1354-1013.2002.00486.X

Hobley, E., Wilson, B., Wilkie, A., Gray, J., Koen, T., 2015. Drivers of soil organic carbon storage and vertical distribution in Eastern Australia. Plant Soil 390, 111–127. https://doi.org/10.1007/s11104-015-2380-1

Hobley, E.U., Baldock, J., Wilson, B., 2016. Environmental and human influences on organic carbon fractions down the soil profile. Agric. Ecosyst. Environ. 223, 152–166. https://doi.org/10.1016/j.agee.2016.03.004

Jaconi, A., Don, A., Freibauer, A., 2017. Prediction of soil organic carbon at the country scale: stratification strategies for near-infrared data. Eur. J. Soil Sci. 68, 919–929. https://doi.org/10.1111/ejss.12485

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning, Performance Evaluation, Springer Texts in Statistics. Springer New York, New York, NY. https://doi.org/10.1007/978-1-4614-7138-7

Jobbágy, E.G., Jackson, R.B., 2000. The vertical distribution of soil organic carbon and its relation to climate and vegetation. Ecol. Appl. 10, 423–436. https://doi.org/10.1890/1051-0761(2000)010[0423:TVDOSO]2.0.CO;2

Joshi, N., Mitchard, E.T.A., Brolly, M., Schumacher, J., Fernández-Landa, A., Johannsen, V.K., Marchamalo, M., Fensholt, R., 2017. Understanding “saturation” of radar signals over forests. Sci. Rep. 7, 1–11. https://doi.org/10.1038/s41598-017-03469-3

Knadel, M., Thomsen, A., Schelde, K., Greve, M.H., 2015. Soil organic carbon and particle sizes mapping using vis–NIR, EC and temperature mobile sensor platform. Comput. Electron. Agric. 114, 134–144. https://doi.org/10.1016/j.compag.2015.03.013

Kuhn, M., Johnson, K., 2013. Applied Predictive Modeling. Springer New York, New York, NY. https://doi.org/10.1007/978-1-4614-6849-3

Kuhn, M., Quinlan, R., 2023. Cubist: Rule- and Instance-Based Regression Modeling.

Kunkel, V.R., Wells, T., Hancock, G.R., 2022. Modelling soil organic carbon using vegetation indices across large catchments in eastern Australia. Sci. Total Environ. 817, 152690. https://doi.org/10.1016/J.SCITOTENV.2021.152690

Lal, R., Smith, P., Jungkunst, H.F., Mitsch, W.J., Lehmann, J., Nair, P.K.R., McBratney, A.B., Sá, J.C. de M., Schneider, J., Zinn, Y.L., Skorupa, A.L.A., Zhang, H.-L., Minasny, B., Srinivasrao, C., Ravindranath, N.H., 2018. The carbon sequestration potential of terrestrial ecosystems. J. Soil Water Conserv. 73, 145A-152A. https://doi.org/10.2489/jswc.73.6.145A

Li, T., Xia, A., McLaren, T.I., Pandey, R., Xu, Z., Liu, H., Manning, S., Madgett, O., Duncan, S., Rasmussen, P., Ruhnke, F., Yüzügüllü, O., Fajraoui, N., Beniwal, D., Chapman, S., Tsiminis, G., Smith, C., Dalal, R.C., Dang, Y.P., 2023. Preliminary Results in Innovative Solutions for Soil Carbon Estimation: Integrating Remote Sensing, Machine Learning, and Proximal Sensing Spectroscopy. Remote Sens. 15, 5571. https://doi.org/10.3390/rs15235571

Malmir, M., Tahmasbian, I., Xu, Z., Farrar, M.B., Bai, S.H., 2019. Prediction of soil macro- and micro-elements in sieved and ground air-dried soils using laboratory-based hyperspectral imaging technique. Geoderma 340, 70–80. https://doi.org/10.1016/j.geoderma.2018.12.049

Mandal, D., Kumar, V., Ratha, D., Dey, S., Bhattacharya, A., Lopez-Sanchez, J.M., McNairn, H., Rao, Y.S., 2020a. Dual polarimetric radar vegetation index for crop growth monitoring using sentinel-1 SAR data. Remote Sens. Environ. 247, 111954. https://doi.org/10.1016/j.rse.2020.111954

Mandal, D., Ratha, D., Bhattacharya, A., Kumar, V., McNairn, H., Rao, Y.S., Frery, A.C., 2020b. A Radar Vegetation Index for Crop Monitoring Using Compact Polarimetric SAR Data. IEEE Trans. Geosci. Remote Sens. 58, 6321–6335. https://doi.org/10.1109/TGRS.2020.2976661

Mayes, M., Marin-Spiotta, E., Szymanski, L., Akif Erdoğan, M., Ozdoğan, M., Clayton, M., 2014. Soil type mediates effects of land use on soil carbon and nitrogen in the Konya Basin, Turkey. Geoderma 232–234, 517–527. https://doi.org/10.1016/j.geoderma.2014.06.002

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

McBride, M.B., 2022. Estimating soil chemical properties by diffuse reflectance spectroscopy: Promise versus reality. Eur. J. Soil Sci. 73, e13192. https://doi.org/10.1111/ejss.13192

McKight, P.E., Najab, J., 2010. Kruskal-Wallis Test, in: The Corsini Encyclopedia of Psychology. John Wiley & Sons, Ltd, pp. 1–1. https://doi.org/10.1002/9780470479216.corpsy0491

Medeiros, A. de S., Maia, S.M.F., Santos, T.C. dos, Gomes, T.C. de A., 2020. Losses and gains of soil organic carbon in grasslands in the Brazilian semi-arid region. Sci. Agric. 78, e20190076. https://doi.org/10.1590/1678-992X-2019-0076

Mendes, W. de S., Demattê, J.A.M., Rosin, N.A., Terra, F. da S., Poppiel, R.R., Urbina-Salazar, D.F., Boechat, C.L., Silva, E.B., Curi, N., Silva, S.H.G., José dos Santos, U., Souza Valladares, G., 2022. The Brazilian soil Mid-infrared Spectral Library: The Power of the Fundamental Range. Geoderma 415, 115776. https://doi.org/10.1016/j.geoderma.2022.115776

Meng, X., Bao, Y., Zhang, X., Wang, X., Liu, H., 2022. Prediction of soil organic matter using different soil classification hierarchical level stratification strategies and spectral characteristic parameters. Geoderma 411, 115696. https://doi.org/10.1016/j.geoderma.2022.115696

Meyer, H., 2021. CAST: "caret" Applications for Spatial-Temporal Models.

Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., Nauss, T., 2018. Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. Environ. Model. Softw. 101, 1–9. https://doi.org/10.1016/j.envsoft.2017.12.001

Michaletz, S.T., Cheng, D., Kerkhoff, A.J., Enquist, B.J., 2014. Convergence of terrestrial plant production across global climate gradients. Nature 512, 39–43. https://doi.org/10.1038/nature13470

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

Mishra, U., Yeo, K., Adhikari, K., Riley, W.J., Hoffman, F.M., Hudson, C., Gautam, S., 2022. Empirical relationships between environmental factors and soil organic carbon produce comparable prediction accuracy to machine learning. Soil Sci. Soc. Am. J. https://doi.org/10.1002/saj2.20453

Mitchard, E.T.A., Saatchi, S.S., Lewis, S.L., Feldpausch, T.R., Woodhouse, I.H., Sonké, B., Rowland, C., Meir, P., 2011. Measuring biomass changes due to woody encroachment and deforestation/degradation in a forest–savanna boundary region of central Africa using multi-temporal L-band radar backscatter. Remote Sens. Environ., DESDynI VEG-3D Special Issue 115, 2861–2873. https://doi.org/10.1016/j.rse.2010.02.022

Moura-Bueno, J.M., Dalmolin, R.S.D., Horst-Heinen, T.Z., Grunwald, S., ten Caten, A., 2021. Environmental covariates improve the spectral predictions of organic carbon in subtropical soils in southern Brazil. Geoderma 393, 114981. https://doi.org/10.1016/j.geoderma.2021.114981

Moura-Bueno, J.M., Dalmolin, R.S.D., Horst-Heinen, T.Z., ten Caten, A., Vasques, G.M., Dotto, A.C., Grunwald, S., 2020. When does stratification of a subtropical soil spectral library improve predictions of soil organic carbon content? Sci. Total Environ. 737, 139895. https://doi.org/10.1016/j.scitotenv.2020.139895

Moura-Bueno, J.M., Dalmolin, R.S.D., ten Caten, A., Dotto, A.C., Demattê, J.A.M., 2019. Stratification of a local VIS-NIR-SWIR spectral library by homogeneity

criteria yields more accurate soil organic carbon predictions. Geoderma 337, 565–581. https://doi.org/10.1016/j.geoderma.2018.10.015

NASA JPL, 2020. NASADEM Merged DEM Global 1 arc second V001 [Data set]. NASA EOSDIS Land Processes DAAC [WWW Document]. https://doi.org/10.5067/MEaSUREs/NASADEM/NASADEM_HGT.001

Nguyen, T.T., Pham, T.D., Nguyen, C.T., Delfos, J., Archibald, R., Dang, K.B., Hoang, N.B., Guo, W., Ngo, H.H., 2022. A novel intelligence approach based active and ensemble learning for agricultural soil organic carbon prediction using multispectral and SAR data fusion. Sci. Total Environ. 804, 150187. https://doi.org/10.1016/j.scitotenv.2021.150187

Nocita, M., Stevens, A., van Wesemael, B., Aitkenhead, M., Bachmann, M., Barthès, B., Ben Dor, E., Brown, D.J., Clairotte, M., Csorba, A., Dardenne, P., Demattê, J.A.M., Genot, V., Guerrero, C., Knadel, M., Montanarella, L., Noon, C., Ramirez-Lopez, L., Robertson, J., Sakai, H., Soriano-Disla, J.M., Shepherd, K.D., Stenberg, B., Towett, E.K., Vargas, R., Wetterlind, J., 2015. Soil Spectroscopy: An Alternative to Wet Chemistry for Soil Monitoring, in: Advances in Agronomy. pp. 139–159. https://doi.org/10.1016/bs.agron.2015.02.002

Paustian, K., Collier, S., Baldock, J., Burgess, R., Creque, J., DeLonge, M., Dungait, J., Ellert, B., Frank, S., Goddard, T., Govaerts, B., Grundy, M., Henning, M., Izaurralde, R.C., Madaras, M., McConkey, B., Porzig, E., Rice, C., Searle, R., Seavy, N., Skalsky, R., Mulhern, W., Jahn, M., 2019. Quantifying carbon for agricultural soil management: from the current status toward a global soil information system. Carbon Manag. 10, 567–587. https://doi.org/10.1080/17583004.2019.1633231

Periasamy, S., 2018. Significance of dual polarimetric synthetic aperture radar in biomass retrieval: An attempt on Sentinel-1. Remote Sens. Environ. 217, 537–549. https://doi.org/10.1016/j.rse.2018.09.003

Poggio, M., Brown, D.J., Bricklemyer, R.S., 2017. Comparison of Vis–NIR on in situ, intact core and dried, sieved soil to estimate clay content at field to regional scales. Eur. J. Soil Sci. 68, 434–448. https://doi.org/10.1111/ejss.12434

Pudełko, A., Chodak, M., 2020. Estimation of total nitrogen and organic carbon contents in mine soils with NIR reflectance spectroscopy and various chemometric methods. Geoderma 368, 114306. https://doi.org/10.1016/j.geoderma.2020.114306

Quinlan, J., 1992. Learning with continuous classes, in: Proceedings AI'92. Presented at the 5th Australian Conference on Artificial Intelligence, World Scientific, Singapore.

R Core Team, R., 2023. R: A Language and Environment for Statistical Computing.

Rosin, N.A., Dalmolin, R.S.D., Horst-Heinen, T.Z., Moura-Bueno, J.M., Silva-Sangoi, D.V. da, Silva, L.S. da, 2020. Diffuse reflectance spectroscopy for estimating soil organic carbon and make nitrogen recommendations. Sci. Agric. 78, e20190246. https://doi.org/10.1590/1678-992X-2019-0246

Roy, D.P., Yan, L., 2020. Robust Landsat-based crop time series modelling. Remote Sens. Environ. 238, 110810. https://doi.org/10.1016/j.rse.2018.06.038

Saatchi, S., 2019. SAR Methods for Mapping and Monitoring Forest Biomass, in: Flores-Anderson, A.I., Herndon, K.E., Thapa, R.B., Cherrington, E. (Eds.), The Synthetic Aperture Radar (SAR) Handbook: Comprehensive Methodologies for Forest Monitoring and Biomass Estimation. NASA, Huntsville. https://doi.org/10.25966/nr2c-s697

Sabetizade, M., Gorji, M., Roudier, P., Zolfaghari, A.A., Keshavarzi, A., 2021. Combination of MIR spectroscopy and environmental covariates to predict soil organic carbon in a semi-arid region. CATENA 196, 104844. https://doi.org/10.1016/j.catena.2020.104844

Santoro, M., Cartus, O., Carvalhais, N., Rozendaal, D.M.A., Avitabile, V., Araza, A., de Bruin, S., Herold, M., Quegan, S., Rodríguez-Veiga, P., Balzter, H., Carreiras, J., Schepaschenko, D., Korets, M., Shimada, M., Itoh, T., Moreno Martínez, Á., Cavlovic, J., Cazzolla Gatti, R., da Conceição Bispo, P., Dewnath, N., Labrière, N., Liang, J., Lindsell, J., Mitchard, E.T.A., Morel, A., Pacheco Pascagaza, A.M., Ryan, C.M., Slik, F., Vaglio Laurin, G., Verbeeck, H., Wijaya, A., Willcock, S., 2021. The global forest above-ground biomass pool for 2010 estimated from high-resolution satellite observations. Earth Syst. Sci. Data 13, 3927–3950. https://doi.org/10.5194/essd-13-3927-2021

Santos, E.P. dos, Moreira, M.C., Fernandes-Filho, E.I., Demattê, J.A.M., Dionizio, E.A., Silva, D.D. da, Cruz, R.R.P., Moura-Bueno, J.M., Santos, U.J. dos, Costa, M.H., 2023. Sentinel-1 Imagery Used for Estimation of Soil Organic Carbon by Dual-Polarization SAR Vegetation Indices. Remote Sens. 15, 5464. https://doi.org/10.3390/rs15235464

Santos, U.J. dos, Demattê, J.A. de M., Menezes, R.S.C., Dotto, A.C., Guimarães, C.C.B., Alves, B.J.R., Primo, D.C., Sampaio, E.V. de S.B., 2020. Predicting carbon and nitrogen by visible near-infrared (Vis-NIR) and mid-infrared (MIR) spectroscopy in soils of Northeast Brazil. Geoderma Reg. 23, e00333. https://doi.org/10.1016/j.geodrs.2020.e00333

Shafizadeh-Moghadam, H., Minaei, F., Talebi-khiyavi, H., Xu, T., Homaee, M., 2022. Synergetic use of multi-temporal Sentinel-1, Sentinel-2, NDVI, and topographic factors for estimating soil organic carbon. CATENA 212, 106077. https://doi.org/10.1016/J.CATENA.2022.106077

Six, J., Conant, R.T., Paul, E.A., Paustian, K., 2002. Stabilization mechanisms of soil organic matter: Implications for C-saturation of soils. Plant Soil 241, 155–176. https://doi.org/10.1023/A:1016125726789

Small, D., 2011. Flattening Gamma: Radiometric Terrain Correction for SAR Imagery. IEEE Trans. Geosci. Remote Sens. 49, 3081–3093. https://doi.org/10.1109/TGRS.2011.2120616

Smith, P., Soussana, J., Angers, D., Schipper, L., Chenu, C., Rasse, D.P., Batjes, N.H., Egmond, F., McNeill, S., Kuhnert, M., Arias-Navarro, C., Olesen, J.E., Chirinda, N., Fornara, D., Wollenberg, E., Álvaro-Fuentes, J., Sanz-Cobena, A., Klumpp, K., 2020. How to measure, report and verify soil carbon change to realize the

potential of soil carbon sequestration for atmospheric greenhouse gas removal. Glob. Change Biol. 26, 219–241. https://doi.org/10.1111/gcb.14815

Sørensen, R., Zinko, U., Seibert, J., 2006. On the calculation of the topographic wetness index: evaluation of different methods based on field observations. Hydrol. Earth Syst. Sci. 10, 101–112. https://doi.org/10.5194/hess-10-101-2006

Soriano-Disla, J.M., Janik, L.J., Viscarra Rossel, R.A., Macdonald, L.M., McLaughlin, M.J., 2014. The Performance of Visible, Near-, and Mid-Infrared Reflectance Spectroscopy for Prediction of Soil Physical, Chemical, and Biological Properties. Appl. Spectrosc. Rev. 49, 139–186. https://doi.org/10.1080/05704928.2013.811081

Sothe, C., Gonsamo, A., Arabian, J., Snider, J., 2022. Large scale mapping of soil organic carbon concentration with 3D machine learning and satellite observations. Geoderma 405, 115402. https://doi.org/10.1016/j.geoderma.2021.115402

Souza, C.M., Z. Shimbo, J., Rosa, M.R., Parente, L.L., A. Alencar, A., Rudorff, B.F.T., Hasenack, H., Matsumoto, M., G. Ferreira, L., Souza-Filho, P.W.M., de Oliveira, S.W., Rocha, W.F., Fonseca, A.V., Marques, C.B., Diniz, C.G., Costa, D., Monteiro, D., Rosa, E.R., Vélez-Martin, E., Weber, E.J., Lenti, F.E.B., Paternost, F.F., Pareyn, F.G.C., Siqueira, J.V., Viera, J.L., Neto, L.C.F., Saraiva, M.M., Sales, M.H., Salgado, M.P.G., Vasconcelos, R., Galano, S., Mesquita, V.V., Azevedo, T., 2020. Reconstructing Three Decades of Land Use and Land Cover Changes in Brazilian Biomes with Landsat Archive and Earth Engine. Remote Sens. 12, 2735. https://doi.org/10.3390/rs12172735

Stevens, A., Nocita, M., Tóth, G., Montanarella, L., van Wesemael, B., 2013. Prediction of Soil Organic Carbon at the European Scale by Visible and Near InfraRed Reflectance Spectroscopy. PLoS ONE 8, e66409. https://doi.org/10.1371/journal.pone.0066409

Stewart, C.E., Plante, A.F., Paustian, K., Conant, R.T., Six, J., 2008. Soil Carbon Saturation: Linking Concept and Measurable Carbon Pools. Soil Sci. Soc. Am. J. 72, 379–392. https://doi.org/10.2136/sssaj2007.0104

Tay, J.K., Narasimhan, B., Hastie, T., 2023. Elastic Net Regularization Paths for All Generalized Linear Models. J. Stat. Softw. 106, 1–31. https://doi.org/10.18637/jss.v106.i01

Tibshirani, R., 1996. Regression Shrinkage and Selection Via the Lasso. J. R. Stat. Soc. Ser. B Methodol. 58, 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Vasques, G.M., Grunwald, S., Harris, W.G., 2010. Spectroscopic Models of Soil Organic Carbon in Florida, USA. J. Environ. Qual. 39, 923–934. https://doi.org/10.2134/jeq2009.0314

Viscarra Rossel, R.A., Behrens, T., 2010. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma, Diffuse reflectance spectroscopy in soil science and land resource assessment 158, 46–54. https://doi.org/10.1016/j.geoderma.2009.12.025

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., 2016. A global spectral library to characterize the world's soil. Earth-Sci. Rev. 155, 198–230. https://doi.org/10.1016/j.earscirev.2016.01.012

Viscarra Rossel, R.A., Behrens, T., Ben-Dor, E., Chabrillat, S., Demattê, J.A.M., Ge, Y., Gomez, C., Guerrero, C., Peng, Y., Ramirez-Lopez, L., Shi, Z., Stenberg, B., Webster, R., Winowiecki, L., Shen, Z., 2022. Diffuse reflectance spectroscopy for estimating soil properties: A technology for the 21st century. Eur. J. Soil Sci. 73, e13271. https://doi.org/10.1111/ejss.13271

Viscarra Rossel, R.A., McBratney, A.B., Minasny, B. (Eds.), 2010. Proximal Soil Sensing. Springer Netherlands, Dordrecht. https://doi.org/10.1007/978-90-481-8859-8

Wadoux, A.M.J.-C., 2023. Interpretable spectroscopic modelling of soil with machine learning. Eur. J. Soil Sci. 74, e13370. https://doi.org/10.1111/ejss.13370

Wang, Z., Ding, J., Zhang, Z., 2022. Estimation of Soil Organic Matter in Arid Zones with Coupled Environmental Variables and Spectral Features. Sensors 22, 1194. https://doi.org/10.3390/s22031194

Wiesmeier, M., Urbanski, L., Hobley, E., Lang, B., von Lützow, M., Marin-Spiotta, E., van Wesemael, B., Rabot, E., Ließ, M., Garcia-Franco, N., Wollschläger, U., Vogel, H.-J., Kögel-Knabner, I., 2019. Soil organic carbon storage as a key function of soils - A review of drivers and indicators at various scales. Geoderma 333, 149–162. https://doi.org/10.1016/j.geoderma.2018.07.026

Woodhouse, I.H., 2006. Introduction to Microwave Remote Sensing. CRC Press, Boca Raton.

Woodhouse, I.H., Mitchard, E.T.A., Brolly, M., Maniatis, D., Ryan, C.M., 2012. Radar backscatter is not a "direct measure" of forest biomass. Nat. Clim. Change 2, 556–557. https://doi.org/10.1038/nclimate1601

Yu, M., Wang, Y., Jiang, J., Wang, C., Zhou, G., Yan, J., 2019. Soil Organic Carbon Stabilization in the Three Subtropical Forests: Importance of Clay and Metal Oxides. J. Geophys. Res. Biogeosciences 124, 2976–2990. https://doi.org/10.1029/2018JG004995

Zayani, H., Fouad, Y., Michot, D., Kassouk, Z., Baghdadi, N., Vaudour, E., Lili-Chabaane, Z., Walter, C., 2023. Using Machine-Learning Algorithms to Predict Soil Organic Carbon Content from Combined Remote Sensing Imagery and Laboratory Vis-NIR Spectral Datasets. Remote Sens. 15, 4264. https://doi.org/10.3390/rs15174264

Zeng, Y., Hao, D., Huete, A., Dechant, B., Berry, J., Chen, J.M., Joiner, J., Frankenberg, C., Bond-Lamberty, B., Ryu, Y., Xiao, J., Asrar, G.R., Chen, M., 2022. Optical

vegetation indices for monitoring terrestrial ecosystems globally. Nat. Rev. Earth Environ. https://doi.org/10.1038/s43017-022-00298-5

Zheng, G., Jiao, C., Zhou, S., Shang, G., 2016. Analysis of soil chronosequence studies using reflectance spectroscopy. Int. J. Remote Sens. 37, 1881–1901. https://doi.org/10.1080/01431161.2016.1163751

Zhou, T., Geng, Y., Chen, J., Liu, M., Haase, D., Lausch, A., 2020a. Mapping soil organic carbon content using multi-source remote sensing variables in the Heihe River Basin in China. Ecol. Indic. 114, 106288. https://doi.org/10.1016/J.ECOLIND.2020.106288

Zhou, T., Geng, Y., Chen, J., Pan, J., Haase, D., Lausch, A., 2020b. High-resolution digital mapping of soil organic carbon and soil total nitrogen using DEM derivatives, Sentinel-1 and Sentinel-2 data based on machine learning algorithms. Sci. Total Environ. 729, 138244. https://doi.org/10.1016/J.SCITOTENV.2020.138244

3. GENERAL CONCLUSIONS

The thesis made significant contributions to proximal and remote sensing applied to tropical soils. The primary contributions were demonstrating the efficiency of diffuse reflectance in predicting soil organic carbon (SOC) using transparent machine learning methods; developing methodologies for SOC prediction using all-weather and globally available radar satellite images; and establishing a methodology for combining remote and proximal sensing for SOC modeling. Throughout the development of the thesis, the following conclusions were drawn at each stage:

In Article 1 – Improving the generalization error and transparency of regression models to estimate soil organic carbon using soil reflectance data, the importance of the Partial Least Squares (PLS) regression method for proximal soil sensing and modeling SOC content was considered. However, PLS alone does not meet the current demands for transparency in machine learning models for estimating soil properties. On the other hand, the Least Absolute Shrinkage and Selection Operator (LASSO) method produced better accuracy than PLS in SOC prediction and proved to be a transparent method, facilitating the examination of the relationship between spectral bands – visible, near-infrared, short-wave, and mid-infrared – and SOC. Additionally, it was evident that ensuring the spatial independence of soil samples used in model calibration is crucial when the soil spectral library (SSL) consists of samples from soil profiles.

In Article 2 – Sentinel-1 imagery used for estimation of soil organic carbon by dual-polarization SAR vegetation indices, the accuracy of microwave vegetation indices in predicting SOC content at different soil depths was explored. The vegetation indices employed were derived from Sentinel-1 synthetic aperture radar (SAR) satellite data, obtained using dual-polarization sensor images. The study was conducted in a region with varying land use and cover classes and predominantly sandy soils. Models using vegetation indices produced more accurate results for topsoil layers (0–5 cm and 5–10 cm). In these layers, linear models (LASSO) performed as accurately as models trained with Support Vector Machine and Random Forest, suggesting a linear relationship between SAR vegetation indices and surface SOC content. However, in deeper soil layers (up to 1 m), models trained with SAR vegetation indices lost their ability to explain SOC variability.

Finally, in Article 3 – Integrating satellite radar vegetation indices and environmental descriptors with visible-infrared soil spectroscopy improved organic

carbon prediction in soils of semi-arid Brazil, combining SAR vegetation indices with the spectral library in the Vis-NIR-SWIR spectra improved SOC predictions. Additionally, incorporating categorical and continuous variables describing soil forming factors (land use and cover classes, soil types, climate, elevation) into the spectral library significantly enhanced SOC estimates. The most accurate and precise SOC predictions were obtained from models calibrated with the spectral library augmented with all auxiliary covariates.

4. APPENDIX A: supplementary material for the article 1

The Köppen climate classes for the study area is presented in Figure A1. In it the points of the sampled soil profiles are also displayed. The Kruskal-Wallis hypothesis test results for the model groups is presented in Table A1. Table A1 shows that all models, evaluated in each of the metrics (R^2, RMSE, and MAE), presented different performances (with significant statistical difference), which allowed us to compute Dunn's pairwise test (Kruskall-Wallis *post hoc* test). Dunn's test evaluated pairs of models in each of the variables (R^2, RMSE, and MAE). The performance of the models in the training can be consulted in Table A2 and in the holdout test in Table A3.

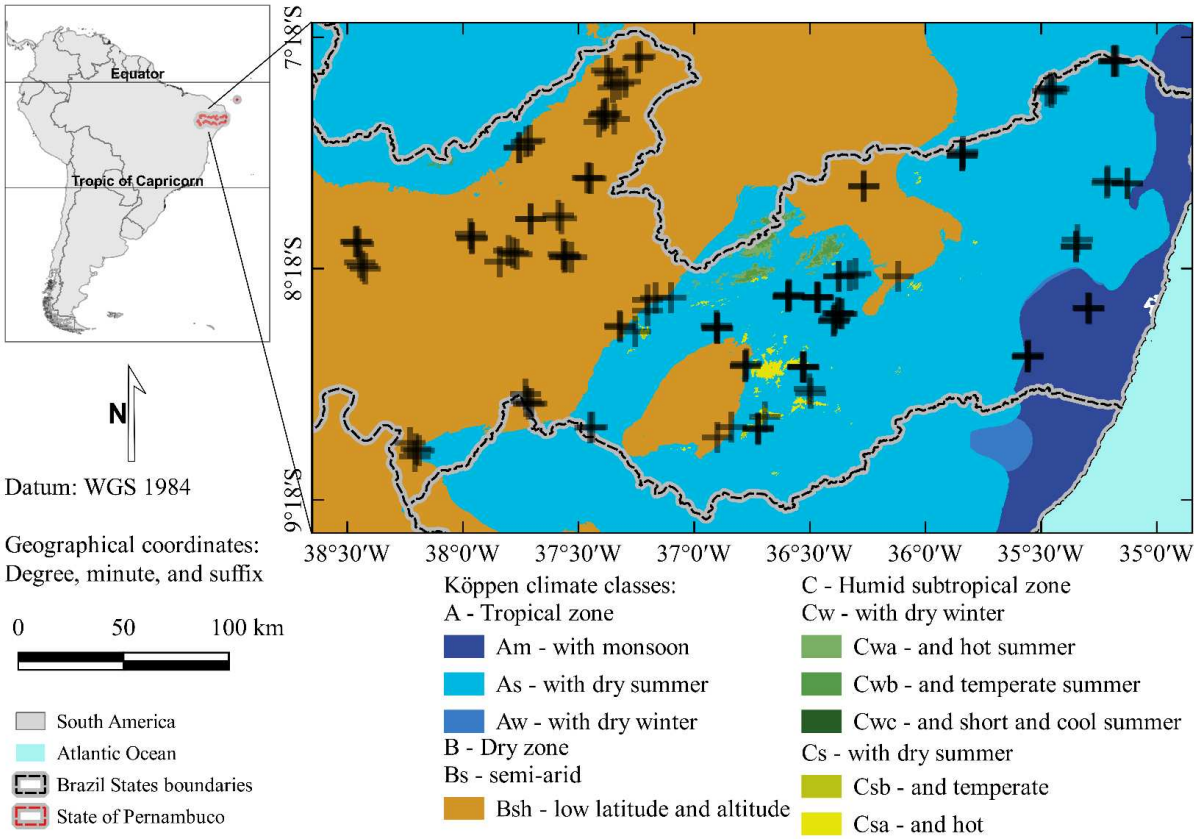


Figure A1. Location of the study area showing the location of soil profiles over the Köppen climate classes, spatialized for Brazil by Alvares et al. (2013).

Table A1. Kruskal-Wallis hypothesis test result for model groups.

Groups of models	**Tested variable**					
	R^2		RMSE		MAE	
	X^2	*p-value*	X^2	*p-value*	X^2	*p-value*
Training (cross-validation)	757.33	$< 2.2\ x\ 10^{-16}*$	758.73	$< 2.2\ x\ 10^{-16}*$	762.83	$< 2.2\ x\ 10^{-16}*$
Holdout testing	711.6	$< 2.2\ x\ 10^{-16}*$	718.83	$< 2.2\ x\ 10^{-16}*$	725.97	$< 2.2\ x\ 10^{-16}*$

Obs.: X^2 is the chi-square test statistic; and * indicates significant difference to P = 0.05.

Table A2. Dunn's pairwise test result for the model groups in the training.

Pairwised models		**Tested variable**					
		R^2		RMSE		MAE	
		Z	p	Z	p	Z	p
MIR LASSO k-Fold CV	MIR LASSO LSPO CV	4.803	0.000*	-2.651	0.112	-4.710	0.000*
MIR LASSO k-Fold CV	MIR PLS k-Fold CV	-3.850	0.002*	5.436	0.000*	3.502	0.007*
MIR LASSO LSPO CV	MIR PLS k-Folqd CV	-8.653	0.000*	8.087	0.000*	8.212	0.000*
MIR LASSO k-Fold CV	MIR PLS LSPO CV	1.805	0.995	1.957	0.705	-2.825	0.066
MIR LASSO LSPO CV	MIR PLS LSPO CV	-2.998	0.038*	4.608	0.000*	1.885	0.832
MIR PLS k-Fold CV	MIR PLS LSPO CV	5.655	0.000*	-3.480	0.007*	-6.327	0.000*
MIR LASSO k-Fold CV	VNIR LASSO k-Fold CV	11.257	0.000*	-9.490	0.000*	-9.089	0.000*
MIR LASSO LSPO CV	VNIR LASSO k-Fold CV	6.454	0.000*	-6.839	0.000*	-4.379	0.000*
MIR PLS k-Fold CV	VNIR LASSO k-Fold CV	15.107	0.000*	-14.926	0.000*	-12.591	0.000*
MIR PLS LSPO CV	VNIR LASSO k-Fold CV	9.452	0.000*	-11.446	0.000*	-6.264	0.000*

MIR LASSO k-Fold CV	VNIR LASSO LSPO CV	14.741	0.000*	-12.666	0.000*	-14.752	0.000*
MIR LASSO LSPO CV	VNIR LASSO LSPO CV	9.938	0.000*	-10.016	0.000*	-10.042	0.000*
MIR PLS k-Fold CV	VNIR LASSO LSPO CV	18.590	0.000*	-18.103	0.000*	-18.254	0.000*
MIR PLS LSPO CV	VNIR LASSO LSPO CV	12.936	0.000*	-14.623	0.000*	-11.927	0.000*
VNIR LASSO k-Fold CV	VNIR LASSO LSPO CV	3.484	0.007*	-3.177	0.021*	-5.663	0.000*
MIR LASSO k-Fold CV	VNIR PLS k-Fold CV	8.040	0.000*	-6.420	0.000*	-11.237	0.000*
MIR LASSO LSPO CV	VNIR PLS k-Fold CV	3.262	0.016*	-3.783	0.002*	-6.551	0.000*
MIR PLS k-Fold CV	VNIR PLS k-Fold CV	11.871	0.000*	-11.829	0.000*	-14.721	0.000*
MIR PLS LSPO CV	VNIR PLS k-Fold CV	6.244	0.000*	-8.367	0.000*	-8.426	0.000*
VNIR LASSO k-Fold CV	VNIR PLS k-Fold CV	-3.160	0.022*	3.022	0.035*	-2.193	0.396
VNIR LASSO LSPO CV	VNIR PLS k-Fold CV	-6.626	0.000*	6.182	0.000*	3.441	0.008*
MIR LASSO k-Fold CV	VNIR PLS LSPO CV	17.075	0.000*	-15.516	0.000*	-17.818	0.000*
MIR LASSO LSPO CV	VNIR PLS LSPO CV	12.272	0.000*	-12.865	0.000*	-13.108	0.000*
MIR PLS k-Fold CV	VNIR PLS LSPO CV	20.925	0.000*	-20.952	0.000*	-21.320	0.000*
MIR PLS LSPO CV	VNIR PLS LSPO CV	15.270	0.000*	-17.473	0.000*	-14.993	0.000*
VNIR LASSO k-Fold CV	VNIR PLS LSPO CV	5.818	0.000*	-6.026	0.000*	-8.729	0.000*
VNIR LASSO LSPO CV	VNIR PLS LSPO CV	2.334	0.274	-2.849	0.061	-3.066	0.030*
VNIR PLS k-Fold CV	VNIR PLS LSPO CV	8.948	0.000*	-9.017	0.000*	-6.491	0.000*

Obs.: Z is the Z statistic of the test, and p the p-value; * indicates significant difference to P = 0.05.

Table A3. Dunn's pairwise test result for the model groups in the holdout test.

Pairwised models		**Tested variable**					
		R^2		RMSE		MAE	
		Z	**p**	**Z**	**p**	**Z**	**p**
MIR LASSO k-Fold CV	MIR LASSO LSPO CV	-0.310	1.000	0.310	1.000	0.310	1.000
MIR LASSO k-Fold CV	MIR PLS k-Fold CV	6.368	0.000*	-6.508	0.000*	-5.661	0.000*
MIR LASSO LSPO CV	MIR PLS k-Fold CV	6.678	0.000*	-6.817	0.000*	-5.970	0.000*
MIR LASSO k-Fold CV	MIR PLS LSPO CV	5.210	0.000*	-5.070	0.000*	-6.413	0.000*
MIR LASSO LSPO CV	MIR PLS LSPO CV	5.520	0.000*	-5.380	0.000*	-6.722	0.000*
MIR PLS k-Fold CV	MIR PLS LSPO CV	-1.158	1.000	1.438	1.000	-0.752	1.000
MIR LASSO k-Fold CV	VNIR LASSO k-Fold CV	12.759	0.000*	-12.566	0.000*	-15.598	0.000*
MIR LASSO LSPO CV	VNIR LASSO k-Fold CV	13.068	0.000*	-12.876	0.000*	-15.908	0.000*
MIR PLS k-Fold CV	VNIR LASSO k-Fold CV	6.391	0.000*	-6.058	0.000*	-9.937	0.000*
MIR PLS LSPO CV	VNIR LASSO k-Fold CV	7.549	0.000*	-7.496	0.000*	-9.186	0.000*
MIR LASSO k-Fold CV	VNIR LASSO LSPO CV	15.846	0.000*	-15.296	0.000*	-12.443	0.000*
MIR LASSO LSPO CV	VNIR LASSO LSPO CV	16.156	0.000*	-15.605	0.000*	-12.752	0.000*
MIR PLS k-Fold CV	VNIR LASSO LSPO CV	9.478	0.000*	-8.788	0.000*	-6.782	0.000*
MIR PLS LSPO CV	VNIR LASSO LSPO CV	10.636	0.000*	-10.225	0.000*	-6.030	0.000*
VNIR LASSO k-Fold CV	VNIR LASSO LSPO CV	3.087	0.028*	-2.729	0.089	3.156	0.022*
MIR LASSO k-Fold CV	VNIR PLS k-Fold CV	13.996	0.000*	-14.226	0.000*	-14.419	0.000*
MIR LASSO LSPO CV	VNIR PLS k-Fold CV	14.304	0.000*	-14.534	0.000*	-14.727	0.000*
MIR PLS k-Fold CV	VNIR PLS k-Fold CV	7.660	0.000*	-7.751	0.000*	-8.786	0.000*
MIR PLS LSPO CV	VNIR PLS k-Fold CV	8.812	0.000*	-9.182	0.000*	-8.038	0.000*
VNIR LASSO k-Fold CV	VNIR PLS k-Fold CV	1.302	1.000	-1.724	1.000	1.101	1.000
VNIR LASSO LSPO CV	VNIR PLS k-Fold CV	-1.770	1.000	0.992	1.000	-2.039	0.581

MIR LASSO k-Fold CV	VNIR PLS LSPO CV	17.983	0.000*	-18.499	0.000*	-18.623	0.000*
MIR LASSO LSPO CV	VNIR PLS LSPO CV	18.292	0.000*	-18.809	0.000*	-18.933	0.000*
MIR PLS k-Fold CV	VNIR PLS LSPO CV	11.615	0.000*	-11.991	0.000*	-12.962	0.000*
MIR PLS LSPO CV	VNIR PLS LSPO CV	12.772	0.000*	-13.429	0.000*	-12.211	0.000*
VNIR LASSO k-Fold CV	VNIR PLS LSPO CV	5.224	0.000*	-5.933	0.000*	-3.025	0.035*
VNIR LASSO LSPO CV	VNIR PLS LSPO CV	2.137	0.457	-3.203	0.019*	-6.180	0.000*
VNIR PLS k-Fold CV	VNIR PLS LSPO CV	3.895	0.001*	-4.179	0.000*	-4.110	0.001*

Obs.: Z is the Z statistic of the test, and p the p-value; * indicates significant difference to P = 0.05.

4.1. References

Alvares, C. A., Stape, J. L., Sentelhas, P. C., de Moraes Gonçalves, J. L., & Sparovek, G. (2013). Köppen's climate classification map for Brazil. *Meteorologische Zeitschrift*, *22*(6), 711–728. https://doi.org/10.1127/0941-2948/2013/0507

5. APPENDIX B: supplementary material for the article 2

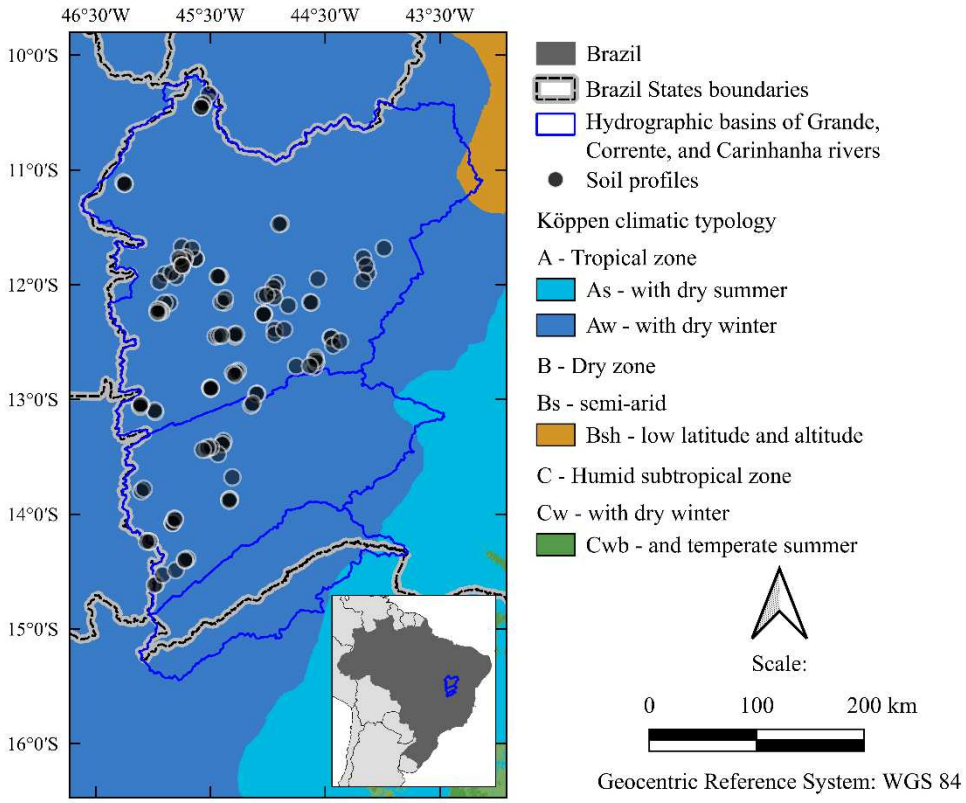


Figure B1. Study area location: highlights for the location of soil profiles over the Köppen's climatic typology for the region. Climate data source: Alvares et al. [1].

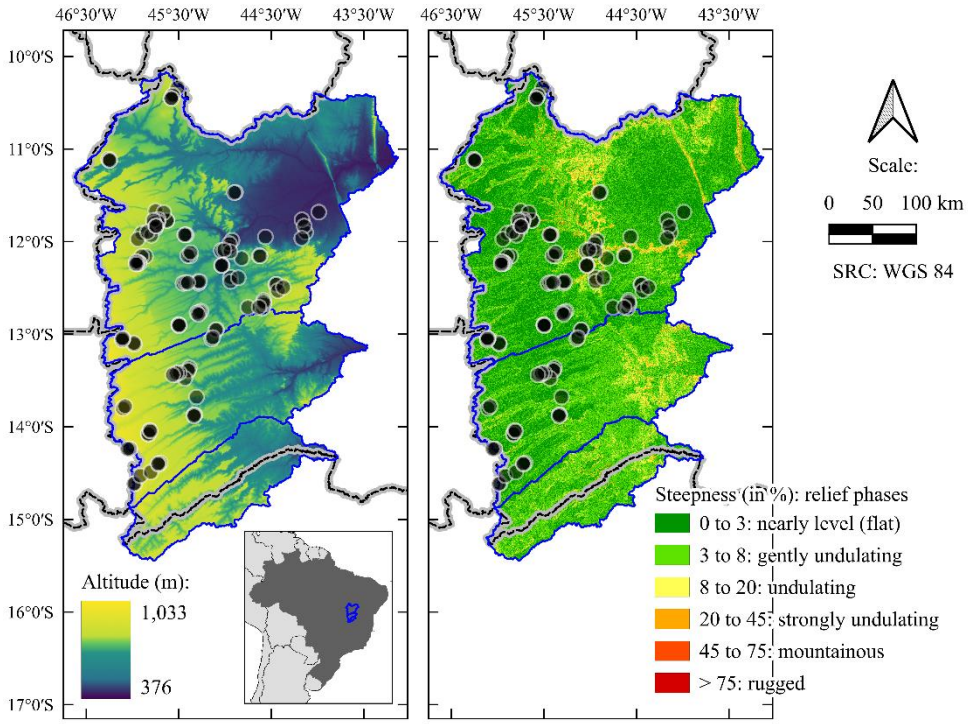


Figure B2. Digital elevation and slope models of the study area. Elevation data source (NASADEM): NASA JPL [2].

Figure B3 displays the soil class map of the study area according to the Brazilian Soil Classification System (SiBCS) [3]. The soil classes presented in Figure B3, at the SiBCS first second categorical levels, are: Cambissolos (corresponding to Cambisols in the World Reference Base for Soil Resources (WRB)), Plintossolos (WRB Plinthosols), Gleissolos (Gleysols), Latossolos (Ferralsols), Argissolos (Acrisols), Neossolos Litolicos (Leptosols), Neossolos Quartzarenicos (Arenosols), Neossolos Fluvicos (Fluvisols), Planossolos Nátricos (Solonetz), Planossolos Háplicos (Planosols), and Luvissolos (Luvisols).

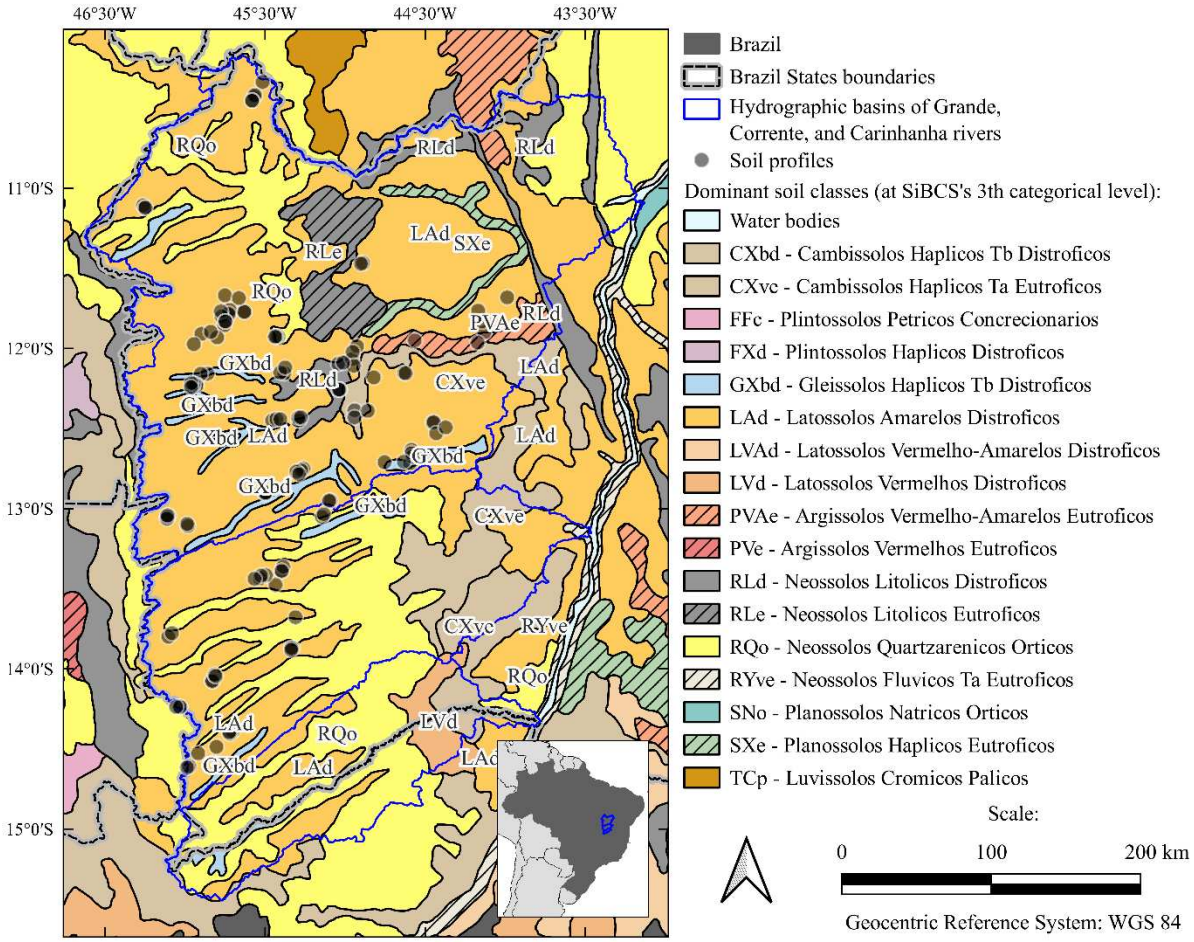


Figure B3. Dominant soil classes, at the third categorical level of the Brazilian Soil Classification System (SiBCS), in the study area. Data source: Map of Brazil' Soils at the compatible scale of 1:5,000,000 [4].

Table B1. Kruskal-Wallis hypothesis test results for the model groups at the training step. Note that: X² is the chi-square statistic of the test; GL: degrees of freedom; and (*) indicates significant difference at P = 0.05.

Tests for the variable: R^2				**Tests for the variable: RMSE**			
X^2	GL	*p-value*		X^2	GL	*p-value*	
539.08	20	$< 2.2\ x\ 10^{-16}$	*	2067.9	20	$< 2.2\ x\ 10^{-16}$	*

Table B2. Pairwise Dunn test results for the model groups at the training step: (*) (*) indicates significant difference at P = 0.05.

Compared model pairs		**Pairwise tests for the variable: R^2**			**Pairwise tests for the variable: RMSE**		
		Z statistic	**p-value**		**Z statistic**	**p-value**	
LASSO 00-05 cm	LASSO 05-10 cm	3.797239	0.0001	*	3.538786	0.0002	*
LASSO 00-05 cm	LASSO 10-15 cm	-2.047717	0.0203	*	7.841166	0	*
LASSO 05-10 cm	LASSO 10-15 cm	-5.606299	0	*	4.302379	0	*
LASSO 00-05 cm	LASSO 15-20 cm	4.538189	0	*	10.98346	0	*
LASSO 05-10 cm	LASSO 15-20 cm	0.29333	0.3846		7.444676	0	*
LASSO 10-15 cm	LASSO 15-20 cm	6.542793	0	*	3.142297	0.0008	*
LASSO 00-05 cm	LASSO 20-40 cm	-4.639443	0	*	15.11057	0	*
LASSO 05-10 cm	LASSO 20-40 cm	-7.895966	0	*	11.57178	0	*
LASSO 10-15 cm	LASSO 20-40 cm	-2.591726	0.0048	*	7.269404	0	*
LASSO 15-20 cm	LASSO 20-40 cm	-9.07995	0	*	4.127107	0	*
LASSO 00-05 cm	LASSO 40-60 cm	3.789052	0.0001	*	18.5458	0	*
LASSO 05-10 cm	LASSO 40-60 cm	-0.449792	0.3264		15.00701	0	*
LASSO 10-15 cm	LASSO 40-60 cm	5.83677	0	*	10.70463	0	*
LASSO 15-20 cm	LASSO 40-60 cm	-0.828915	0.2036		7.56234	0	*
LASSO 20-40 cm	LASSO 40-60 cm	8.428496	0	*	3.435233	0.0003	*
LASSO 00-05 cm	LASSO 60-100 cm	8.410486	0	*	21.71317	0	*

LASSO 05-10 cm	LASSO 60-100 cm	4.119156	0	*	18.17438	0	*
LASSO 10-15 cm	LASSO 60-100 cm	10.24476	0	*	13.872	0	*
LASSO 15-20 cm	LASSO 60-100 cm	4.185517	0	*	10.72971	0	*
LASSO 20-40 cm	LASSO 60-100 cm	12.56633	0	*	6.602602	0	*
LASSO 40-60 cm	LASSO 60-100 cm	5.016385	0	*	3.167369	0.0008	*
LASSO 00-05 cm	RF 00-05 cm	0.096115	0.4617		0.963003	0.1678	
LASSO 05-10 cm	RF 00-05 cm	-3.712325	0.0001	*	-2.575783	0.005	*
LASSO 10-15 cm	RF 00-05 cm	2.143833	0.016	*	-6.878162	0	*
LASSO 15-20 cm	RF 00-05 cm	-4.444098	0	*	-10.02045	0	*
LASSO 20-40 cm	RF 00-05 cm	4.735559	0	*	-14.14756	0	*
LASSO 40-60 cm	RF 00-05 cm	-3.692937	0.0001	*	-17.5828	0	*
LASSO 60-100 cm	RF 00-05 cm	-8.324389	0	*	-20.75017	0	*
LASSO 00-05 cm	RF 05-10 cm	3.733005	0.0001	*	4.27754	0	*
LASSO 05-10 cm	RF 05-10 cm	-0.499307	0.3088		0.738753	0.23	
LASSO 10-15 cm	RF 05-10 cm	5.780723	0	*	-3.563625	0.0002	*
LASSO 15-20 cm	RF 05-10 cm	-0.883782	0.1884		-6.705922	0	*
LASSO 20-40 cm	RF 05-10 cm	8.372449	0	*	-10.83303	0	*
LASSO 40-60 cm	RF 05-10 cm	-0.056047	0.4777		-14.26826	0	*
LASSO 60-100 cm	RF 05-10 cm	-5.06659	0	*	-17.43563	0	*
RF 00-05 cm	RF 05-10 cm	3.63689	0.0001	*	3.314536	0.0005	*
LASSO 00-05 cm	RF 10-15 cm	4.120977	0	*	6.71875	0	*
LASSO 05-10 cm	RF 10-15 cm	-0.156552	0.4378		3.179963	0.0007	*
LASSO 10-15 cm	RF 10-15 cm	6.168695	0	*	-1.122415	0.1308	
LASSO 15-20 cm	RF 10-15 cm	-0.503979	0.3071		-4.264712	0	*
LASSO 20-40 cm	RF 10-15 cm	8.760421	0	*	-8.39182	0	*
LASSO 40-60 cm	RF 10-15 cm	0.331924	0.37		-11.82705	0	*

LASSO 60-100 cm	RF 10-15 cm	-4.719058	0	*	-14.99442	0	*
RF 00-05 cm	RF 10-15 cm	4.024862	0	*	5.755747	0	*
RF 05-10 cm	RF 10-15 cm	0.387972	0.349		2.44121	0.0073	*
LASSO 00-05 cm	RF 15-20 cm	5.113968	0	*	10.39887	0	*
LASSO 05-10 cm	RF 15-20 cm	0.720707	0.2355		6.860087	0	*
LASSO 10-15 cm	RF 15-20 cm	7.161686	0	*	2.557708	0.0053	*
LASSO 15-20 cm	RF 15-20 cm	0.468103	0.3199		-0.584589	0.2794	
LASSO 20-40 cm	RF 15-20 cm	9.753412	0	*	-4.711696	0	*
LASSO 40-60 cm	RF 15-20 cm	1.324915	0.0926		-8.146929	0	*
LASSO 60-100 cm	RF 15-20 cm	-3.829571	0.0001	*	-11.31429	0	*
RF 00-05 cm	RF 15-20 cm	5.017852	0	*	9.43587	0	*
RF 05-10 cm	RF 15-20 cm	1.380962	0.0836		6.121333	0	*
RF 10-15 cm	RF 15-20 cm	0.99299	0.1604		3.680123	0.0001	*
LASSO 00-05 cm	RF 20-40 cm	-0.310135	0.3782		13.71539	0	*
LASSO 05-10 cm	RF 20-40 cm	-4.071229	0	*	10.1766	0	*
LASSO 10-15 cm	RF 20-40 cm	1.737582	0.0411		5.874227	0	*
LASSO 15-20 cm	RF 20-40 cm	-4.841795	0	*	2.73193	0.0031	*
LASSO 20-40 cm	RF 20-40 cm	4.329308	0	*	-1.395177	0.0815	
LASSO 40-60 cm	RF 20-40 cm	-4.099188	0	*	-4.83041	0	*
LASSO 60-100 cm	RF 20-40 cm	-8.688294	0	*	-7.997779	0	*
RF 00-05 cm	RF 20-40 cm	-0.40625	0.3423		12.75239	0	*
RF 05-10 cm	RF 20-40 cm	-4.043141	0	*	9.437853	0	*
RF 10-15 cm	RF 20-40 cm	-4.431113	0	*	6.996643	0	*
RF 15-20 cm	RF 20-40 cm	-5.424103	0	*	3.316519	0.0005	*
LASSO 00-05 cm	RF 40-60 cm	4.932753	0	*	17.98313	0	*
LASSO 05-10 cm	RF 40-60 cm	0.560612	0.2875		14.44435	0	*

LASSO 10-15 cm	RF 40-60 cm	6.980471	0	*	10.14197	0	*
LASSO 15-20 cm	RF 40-60 cm	0.290704	0.3856		6.999675	0	*
LASSO 20-40 cm	RF 40-60 cm	9.572197	0	*	2.872567	0.002	*
LASSO 40-60 cm	RF 40-60 cm	1.1437	0.1264		-0.562665	0.2868	
LASSO 60-100 cm	RF 40-60 cm	-3.991897	0	*	-3.730034	0.0001	*
RF 00-05 cm	RF 40-60 cm	4.836637	0	*	17.02013	0	*
RF 05-10 cm	RF 40-60 cm	1.199747	0.1151		13.70559	0	*
RF 10-15 cm	RF 40-60 cm	0.811775	0.2085		11.26438	0	*
RF 15-20 cm	RF 40-60 cm	-0.181215	0.4281		7.584264	0	*
RF 20-40 cm	RF 40-60 cm	5.242888	0	*	4.267744	0	*
LASSO 00-05 cm	RF 60-100 cm	11.49668	0	*	20.48673	0	*
LASSO 05-10 cm	RF 60-100 cm	6.359531	0	*	16.94795	0	*
LASSO 10-15 cm	RF 60-100 cm	13.5444	0	*	12.64557	0	*
LASSO 15-20 cm	RF 60-100 cm	6.716431	0	*	9.503274	0	*
LASSO 20-40 cm	RF 60-100 cm	16.13612	0	*	5.376166	0	*
LASSO 40-60 cm	RF 60-100 cm	7.707631	0	*	1.940933	0.0261	
LASSO 60-100 cm	RF 60-100 cm	1.887842	0.0295		-1.226435	0.11	
RF 00-05 cm	RF 60-100 cm	11.40056	0	*	19.52373	0	*
RF 05-10 cm	RF 60-100 cm	7.763678	0	*	16.20919	0	*
RF 10-15 cm	RF 60-100 cm	7.375706	0	*	13.76798	0	*
RF 15-20 cm	RF 60-100 cm	6.382715	0	*	10.08786	0	*
RF 20-40 cm	RF 60-100 cm	11.80681	0	*	6.771343	0	*
RF 40-60 cm	RF 60-100 cm	6.56393	0	*	2.503599	0.0061	*
LASSO 00-05 cm	SVR-RBF 00-05 cm	1.563267	0.059		0.621555	0.2671	
LASSO 05-10 cm	SVR-RBF 00-05 cm	-2.416167	0.0078	*	-2.917231	0.0018	*
LASSO 10-15 cm	SVR-RBF 00-05 cm	3.610985	0.0002	*	-7.21961	0	*

LASSO 15-20 cm	SVR-RBF 00-05 cm	-3.007837	0.0013	*	-10.3619	0	*
LASSO 20-40 cm	SVR-RBF 00-05 cm	6.202711	0	*	-14.48901	0	*
LASSO 40-60 cm	SVR-RBF 00-05 cm	-2.225785	0.013	*	-17.92424	0	*
LASSO 60-100 cm	SVR-RBF 00-05 cm	-7.010166	0	*	-21.09161	0	*
RF 00-05 cm	SVR-RBF 00-05 cm	1.467151	0.0712		-0.341447	0.3664	
RF 05-10 cm	SVR-RBF 00-05 cm	-2.169738	0.015	*	-3.655984	0.0001	*
RF 10-15 cm	SVR-RBF 00-05 cm	-2.55771	0.0053	*	-6.097194	0	*
RF 15-20 cm	SVR-RBF 00-05 cm	-3.550701	0.0002	*	-9.777318	0	*
RF 20-40 cm	SVR-RBF 00-05 cm	1.873402	0.0305		-13.09383	0	*
RF 40-60 cm	SVR-RBF 00-05 cm	-3.369486	0.0004	*	-17.36158	0	*
RF 60-100 cm	SVR-RBF 00-05 cm	-9.933416	0	*	-19.86518	0	*
LASSO 00-05 cm	SVR-RBF 05-10 cm	5.388998	0	*	4.238124	0	*
LASSO 05-10 cm	SVR-RBF 05-10 cm	0.963683	0.1676		0.699337	0.2422	
LASSO 10-15 cm	SVR-RBF 05-10 cm	7.436716	0	*	-3.603041	0.0002	*
LASSO 15-20 cm	SVR-RBF 05-10 cm	0.737343	0.2305		-6.745338	0	*
LASSO 20-40 cm	SVR-RBF 05-10 cm	10.02844	0	*	-10.87244	0	*
LASSO 40-60 cm	SVR-RBF 05-10 cm	1.599945	0.0548		-14.30767	0	*
LASSO 60-100 cm	SVR-RBF 05-10 cm	-3.583209	0.0002	*	-17.47504	0	*
RF 00-05 cm	SVR-RBF 05-10 cm	5.292883	0	*	3.275121	0.0005	*
RF 05-10 cm	SVR-RBF 05-10 cm	1.655993	0.0489		-0.039415	0.4843	
RF 10-15 cm	SVR-RBF 05-10 cm	1.268021	0.1024		-2.480625	0.0066	*
RF 15-20 cm	SVR-RBF 05-10 cm	0.27503	0.3916		-6.160749	0	*
RF 20-40 cm	SVR-RBF 05-10 cm	5.699134	0	*	-9.477269	0	*
RF 40-60 cm	SVR-RBF 05-10 cm	0.456245	0.3241		-13.74501	0	*
RF 60-100 cm	SVR-RBF 05-10 cm	-6.107685	0	*	-16.24861	0	*
SVR-RBF 00-05 cm	SVR-RBF 05-10 cm	3.825731	0.0001	*	3.616568	0.0001	*

LASSO 00-05 cm	SVR-RBF 10-15 cm	-0.656586	0.2557		8.00256	0	*
LASSO 05-10 cm	SVR-RBF 10-15 cm	-4.377302	0	*	4.463773	0	*
LASSO 10-15 cm	SVR-RBF 10-15 cm	1.391131	0.0821		0.161394	0.4359	
LASSO 15-20 cm	SVR-RBF 10-15 cm	-5.180952	0	*	-2.980902	0.0014	*
LASSO 20-40 cm	SVR-RBF 10-15 cm	3.982857	0	*	-7.10801	0	*
LASSO 40-60 cm	SVR-RBF 10-15 cm	-4.445639	0	*	-10.54324	0	*
LASSO 60-100 cm	SVR-RBF 10-15 cm	-8.998633	0	*	-13.71061	0	*
RF 00-05 cm	SVR-RBF 10-15 cm	-0.752702	0.2258		7.039557	0	*
RF 05-10 cm	SVR-RBF 10-15 cm	-4.389592	0	*	3.72502	0.0001	*
RF 10-15 cm	SVR-RBF 10-15 cm	-4.777564	0	*	1.28381	0.0996	
RF 15-20 cm	SVR-RBF 10-15 cm	-5.770555	0	*	-2.396313	0.0083	*
RF 20-40 cm	SVR-RBF 10-15 cm	-0.346451	0.3645		-5.712833	0	*
RF 40-60 cm	SVR-RBF 10-15 cm	-5.589339	0	*	-9.980577	0	*
RF 60-100 cm	SVR-RBF 10-15 cm	-12.15327	0	*	-12.48417	0	*
SVR-RBF 00-05 cm	SVR-RBF 10-15 cm	-2.219853	0.0132	*	7.381004	0	*
SVR-RBF 05-10 cm	SVR-RBF 10-15 cm	-6.045585	0	*	3.764436	0.0001	*
LASSO 00-05 cm	SVR-RBF 15-20 cm	6.355357	0	*	11.67545	0	*
LASSO 05-10 cm	SVR-RBF 15-20 cm	1.817415	0.0346		8.136667	0	*
LASSO 10-15 cm	SVR-RBF 15-20 cm	8.403075	0	*	3.834288	0.0001	*
LASSO 15-20 cm	SVR-RBF 15-20 cm	1.683355	0.0462		0.691991	0.2445	
LASSO 20-40 cm	SVR-RBF 15-20 cm	10.9948	0	*	-3.435116	0.0003	*
LASSO 40-60 cm	SVR-RBF 15-20 cm	2.566305	0.0051	*	-6.870349	0	*
LASSO 60-100 cm	SVR-RBF 15-20 cm	-2.717578	0.0033	*	-10.03771	0	*
RF 00-05 cm	SVR-RBF 15-20 cm	6.259242	0	*	10.71245	0	*
RF 05-10 cm	SVR-RBF 15-20 cm	2.622352	0.0044	*	7.397914	0	*
RF 10-15 cm	SVR-RBF 15-20 cm	2.23438	0.0127	*	4.956703	0	*

RF 15-20 cm	SVR-RBF 15-20 cm	1.241389	0.1072		1.27658	0.1009	
RF 20-40 cm	SVR-RBF 15-20 cm	6.665493	0	*	-2.039939	0.0207	*
RF 40-60 cm	SVR-RBF 15-20 cm	1.422604	0.0774		-6.307684	0	*
RF 60-100 cm	SVR-RBF 15-20 cm	-5.141326	0	*	-8.811283	0	*
SVR-RBF 00-05 cm	SVR-RBF 15-20 cm	4.79209	0	*	11.05389	0	*
SVR-RBF 05-10 cm	SVR-RBF 15-20 cm	0.966359	0.1669		7.437329	0	*
SVR-RBF 10-15 cm	SVR-RBF 15-20 cm	7.011944	0	*	3.672893	0.0001	*
LASSO 00-05 cm	SVR-RBF 20-40 cm	0.048541	0.4806		14.72714	0	*
LASSO 05-10 cm	SVR-RBF 20-40 cm	-3.754354	0.0001	*	11.18835	0	*
LASSO 10-15 cm	SVR-RBF 20-40 cm	2.096259	0.018	*	6.885975	0	*
LASSO 15-20 cm	SVR-RBF 20-40 cm	-4.49067	0	*	3.743678	0.0001	*
LASSO 20-40 cm	SVR-RBF 20-40 cm	4.687985	0	*	-0.383428	0.3507	
LASSO 40-60 cm	SVR-RBF 20-40 cm	-3.74051	0.0001	*	-3.818661	0.0001	*
LASSO 60-100 cm	SVR-RBF 20-40 cm	-8.367004	0	*	-6.986031	0	*
RF 00-05 cm	SVR-RBF 20-40 cm	-0.047573	0.481		13.76413	0	*
RF 05-10 cm	SVR-RBF 20-40 cm	-3.684463	0.0001	*	10.4496	0	*
RF 10-15 cm	SVR-RBF 20-40 cm	-4.072435	0	*	8.008391	0	*
RF 15-20 cm	SVR-RBF 20-40 cm	-5.065426	0	*	4.328267	0	*
RF 20-40 cm	SVR-RBF 20-40 cm	0.358677	0.3599		1.011748	0.1558	
RF 40-60 cm	SVR-RBF 20-40 cm	-4.884211	0	*	-3.255996	0.0006	*
RF 60-100 cm	SVR-RBF 20-40 cm	-11.44814	0	*	-5.759595	0	*
SVR-RBF 00-05 cm	SVR-RBF 20-40 cm	-1.514725	0.0649		14.10558	0	*
SVR-RBF 05-10 cm	SVR-RBF 20-40 cm	-5.340456	0	*	10.48901	0	*
SVR-RBF 10-15 cm	SVR-RBF 20-40 cm	0.705128	0.2404		6.724581	0	*
SVR-RBF 15-20 cm	SVR-RBF 20-40 cm	-6.306815	0	*	3.051687	0.0011	*
LASSO 00-05 cm	SVR-RBF 40-60 cm	2.381458	0.0086	*	17.51947	0	*

LASSO 05-10 cm	SVR-RBF 40-60 cm	-1.693334	0.0452		13.98069	0	*
LASSO 10-15 cm	SVR-RBF 40-60 cm	4.429176	0	*	9.678312	0	*
LASSO 15-20 cm	SVR-RBF 40-60 cm	-2.206872	0.0137	*	6.536015	0	*
LASSO 20-40 cm	SVR-RBF 40-60 cm	7.020902	0	*	2.408907	0.008	*
LASSO 40-60 cm	SVR-RBF 40-60 cm	-1.407594	0.0796		-1.026325	0.1524	
LASSO 60-100 cm	SVR-RBF 40-60 cm	-6.277258	0	*	-4.193694	0	*
RF 00-05 cm	SVR-RBF 40-60 cm	2.285343	0.0111	*	16.55647	0	*
RF 05-10 cm	SVR-RBF 40-60 cm	-1.351547	0.0883		13.24193	0	*
RF 10-15 cm	SVR-RBF 40-60 cm	-1.739519	0.041		10.80072	0	*
RF 15-20 cm	SVR-RBF 40-60 cm	-2.732509	0.0031	*	7.120604	0	*
RF 20-40 cm	SVR-RBF 40-60 cm	2.691594	0.0036	*	3.804085	0.0001	*
RF 40-60 cm	SVR-RBF 40-60 cm	-2.551294	0.0054	*	-0.463659	0.3214	
RF 60-100 cm	SVR-RBF 40-60 cm	-9.115225	0	*	-2.967258	0.0015	*
SVR-RBF 00-05 cm	SVR-RBF 40-60 cm	0.818191	0.2066		16.89792	0	*
SVR-RBF 05-10 cm	SVR-RBF 40-60 cm	-3.00754	0.0013	*	13.28135	0	*
SVR-RBF 10-15 cm	SVR-RBF 40-60 cm	3.038045	0.0012	*	9.516918	0	*
SVR-RBF 15-20 cm	SVR-RBF 40-60 cm	-3.973899	0	*	5.844024	0	*
SVR-RBF 20-40 cm	SVR-RBF 40-60 cm	2.332916	0.0098	*	2.792336	0.0026	*
LASSO 00-05 cm	SVR-RBF 60-100 cm	4.492244	0	*	22.34382	0	*
LASSO 05-10 cm	SVR-RBF 60-100 cm	0.171444	0.4319		18.80503	0	*
LASSO 10-15 cm	SVR-RBF 60-100 cm	6.539962	0	*	14.50265	0	*
LASSO 15-20 cm	SVR-RBF 60-100 cm	-0.140529	0.4441		11.36036	0	*
LASSO 20-40 cm	SVR-RBF 60-100 cm	9.131688	0	*	7.233254	0	*
LASSO 40-60 cm	SVR-RBF 60-100 cm	0.703191	0.241		3.798021	0.0001	*
LASSO 60-100 cm	SVR-RBF 60-100 cm	-4.38649	0	*	0.630651	0.2641	
RF 00-05 cm	SVR-RBF 60-100 cm	4.396129	0	*	21.38082	0	*

RF 05-10 cm	SVR-RBF 60-100 cm	0.759238	0.2239		18.06628	0	*
RF 10-15 cm	SVR-RBF 60-100 cm	0.371266	0.3552		15.62507	0	*
RF 15-20 cm	SVR-RBF 60-100 cm	-0.621723	0.2671		11.94495	0	*
RF 20-40 cm	SVR-RBF 60-100 cm	4.80238	0	*	8.628431	0	*
RF 40-60 cm	SVR-RBF 60-100 cm	-0.440508	0.3298		4.360686	0	*
RF 60-100 cm	SVR-RBF 60-100 cm	-7.004439	0	*	1.857087	0.0316	
SVR-RBF 00-05 cm	SVR-RBF 60-100 cm	2.928977	0.0017	*	21.72226	0	*
SVR-RBF 05-10 cm	SVR-RBF 60-100 cm	-0.896754	0.1849		18.1057	0	*
SVR-RBF 10-15 cm	SVR-RBF 60-100 cm	5.148831	0	*	14.34126	0	*
SVR-RBF 15-20 cm	SVR-RBF 60-100 cm	-1.863113	0.0312		10.66837	0	*
SVR-RBF 20-40 cm	SVR-RBF 60-100 cm	4.443702	0	*	7.616683	0	*
SVR-RBF 40-60 cm	SVR-RBF 60-100 cm	2.110785	0.0174	*	4.824346	0	*

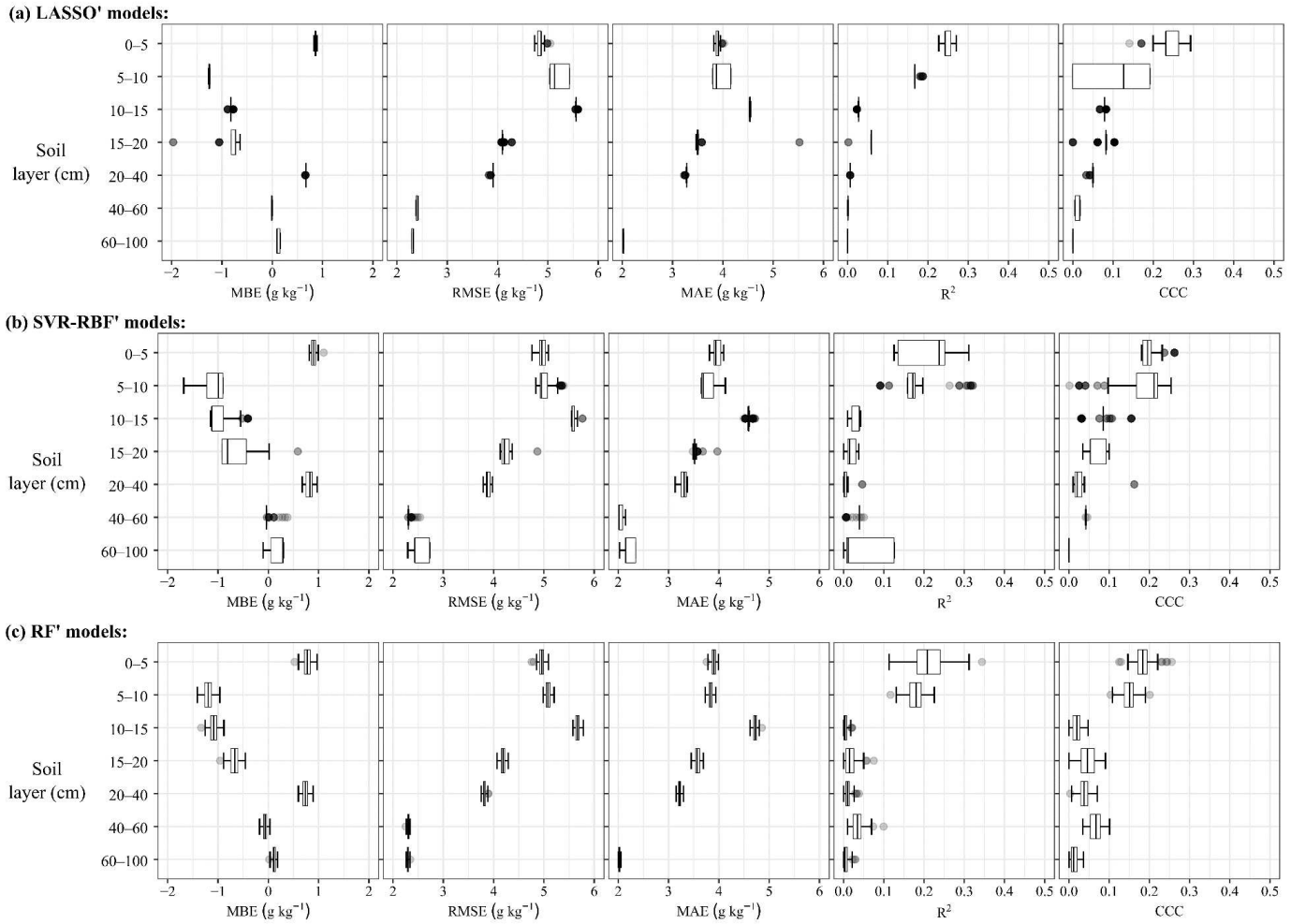


Figure B4. Holdout testing results on the fitted LASSO (subgraphs a), SVR-RBF (in b), and RF (in c) models: accuracy and correlation of the estimates with the observed SOC (soil organic carbon) values are denoted by: MBE (Mean Bias Error), RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), R^2 (determination coefficient) and CCC (Lin's concordance and correlation coefficient), respectively.

Table B3. Pairwise Dunn test results for the model groups at the testing step: (*) indicates significant difference at P = 0.05.

Compared model pairs		Pairwise tests for the variables:														
		RMSE			MAE			R^2			CCC			d		
		Z statistic	p-value		Z statistic	p-value		Z statistic	p-value		Z statistic	p-value		Z statistic	p-value	
LASSO 00-05 cm	LASSO 05-10 cm	-8.08595	0	*	0.143896	1		5.341844	0	*	11.62594	0	*	9.926394	0	*
LASSO 00-05 cm	LASSO 10-15 cm	-12.62347	0	*	-6.541872	0	*	15.01984	0	*	13.82911	0	*	14.06325	0	*
LASSO 05-10 cm	LASSO 10-15 cm	-4.537529	0.0002	*	-6.685768	0	*	7.732878	0	*	2.203171	0.9102		4.136864	0.0012	*
LASSO 00-05 cm	LASSO 15-20 cm	5.846679	0	*	10.92979	0	*	10.40899	0	*	13.23536	0	*	15.62672	0	*
LASSO 05-10 cm	LASSO 15-20 cm	13.93262	0	*	10.78589	0	*	3.851105	0.0039	*	1.609413	1		5.700333	0	*
LASSO 10-15 cm	LASSO 15-20 cm	18.47015	0	*	17.47166	0	*	-4.294608	0.0006	*	-0.593758	1		1.563468	1	
LASSO 00-05 cm	RF 00-05 cm	-2.97855	0.0956		0.125934	1		2.687161	0.2378		4.447778	0.0003	*	7.655894	0	*
LASSO 05-10 cm	RF 00-05 cm	5.107399	0	*	-0.017961	1		-3.002679	0.0883		-7.178168	0	*	-2.270499	0.7649	
LASSO 10-15 cm	RF 00-05 cm	9.644928	0	*	6.667807	0	*	-12.33268	0	*	-9.38134	0	*	-6.407364	0	*
LASSO 15-20 cm	RF 00-05 cm	-8.82523	0	*	-10.80385	0	*	-7.778412	0	*	-8.787582	0	*	-7.970833	0	*
LASSO 00-05 cm	RF 05-10 cm	-7.50567	0	*	2.729335	0.2094		5.106403	0	*	7.403302	0	*	6.212135	0	*
LASSO 05-10 cm	RF 05-10 cm	0.580279	1		2.585438	0.3209		-0.896738	1		-4.222644	0.0008	*	-3.714258	0.0067	*
LASSO 10-15 cm	RF 05-10 cm	5.117808	0	*	9.271207	0	*	-9.913443	0	*	-6.425816	0	*	-7.851123	0	*
LASSO 15-20 cm	RF 05-10 cm	-13.35235	0	*	-8.200455	0	*	-5.410107	0	*	-5.832057	0	*	-9.414592	0	*
RF 00-05 cm	RF 05-10 cm	-4.527119	0.0002	*	2.6034	0.3046		2.419242	0.5132		2.955524	0.103		-1.443759	1	
LASSO 00-05 cm	RF 10-15 cm	-15.33056	0	*	-10.28868	0	*	20.09544	0	*	20.04898	0	*	20.45174	0	*
LASSO 05-10 cm	RF 10-15 cm	-7.244616	0	*	-10.43258	0	*	12.15116	0	*	8.42304	0	*	10.52534	0	*
LASSO 10-15 cm	RF 10-15 cm	-2.707087	0.224		-3.746814	0.0059	*	5.075596	0	*	6.219868	0	*	6.388484	0	*
LASSO 15-20 cm	RF 10-15 cm	-21.17724	0	*	-21.21847	0	*	9.263337	0	*	6.813626	0	*	4.825015	0	*
RF 00-05 cm	RF 10-15 cm	-12.35201	0	*	-10.41462	0	*	17.40828	0	*	15.6012	0	*	12.79584	0	*
RF 05-10 cm	RF 10-15 cm	-7.824895	0	*	-13.01802	0	*	14.98903	0	*	12.64568	0	*	14.2396	0	*
LASSO 00-05 cm	RF 15-20 cm	4.418126	0.0003	*	8.524069	0	*	17.02616	0	*	17.24919	0	*	15.32464	0	*
LASSO 05-10 cm	RF 15-20 cm	12.50407	0	*	8.380172	0	*	9.479368	0	*	5.623252	0	*	5.398253	0	*
LASSO 10-15 cm	RF 15-20 cm	17.0416	0	*	15.06594	0	*	2.006314	1		3.42008	0.0207	*	1.261388	1	
LASSO 15-20 cm	RF 15-20 cm	-1.428553	1		-2.405721	0.5326		6.25868	0	*	4.013838	0.002	*	-0.30208	1	
RF 00-05 cm	RF 15-20 cm	7.396676	0	*	8.398134	0	*	14.339	0	*	12.80142	0	*	7.668753	0	*

RF 05-10 cm	RF 15-20 cm	11.92379	0	*	5.794734	0	*	11.91975	0	*	9.845896	0	*	9.112512	0	*
RF 10-15 cm	RF 15-20 cm	19.74869	0	*	18.81275	0	*	-3.069281	0.0708		-2.799787	0.1687		-5.127095	0	*
LASSO 00-05 cm	SVM-RBF 00-05 cm	-3.131632	0.0574		-1.53224	1		2.505397	0.4036		2.773457	0.183		2.708719	0.2229	
LASSO 05-10 cm	SVM-RBF 00-05 cm	4.954318	0	*	-1.676136	1		-3.160905	0.0519		-8.852489	0	*	-7.217674	0	*
LASSO 10-15 cm	SVM-RBF 00-05 cm	9.491847	0	*	5.009631	0	*	-12.51444	0	*	-11.05566	0	*	-11.35453	0	*
LASSO 15-20 cm	SVM-RBF 00-05 cm	-8.978311	0	*	-12.46203	0	*	-7.956349	0	*	-10.4619	0	*	-12.918	0	*
RF 00-05 cm	SVM-RBF 00-05 cm	-0.153081	1		-1.658175	1		-0.181764	1		-1.67432	1		-4.947174	0	*
RF 05-10 cm	SVM-RBF 00-05 cm	4.374038	0.0004	*	-4.261575	0.0007	*	-2.601006	0.3067		-4.629845	0.0001	*	-3.503415	0.0152	*
RF 10-15 cm	SVM-RBF 00-05 cm	12.19893	0	*	8.756446	0	*	-17.59004	0	*	-17.27552	0	*	-17.74302	0	*
RF 15-20 cm	SVM-RBF 00-05 cm	-7.549758	0	*	-10.0563	0	*	-14.52076	0	*	-14.47574	0	*	-12.61592	0	*
LASSO 00-05 cm	SVM-RBF 05-10 cm	-4.126863	0.0012	*	3.845194	0.004	*	4.842733	0	*	4.543302	0.0002	*	3.255933	0.0373	
LASSO 05-10 cm	SVM-RBF 05-10 cm	3.959086	0.0025	*	3.701298	0.0071	*	-1.126262	1		-7.082644	0	*	-6.67046	0	*
LASSO 10-15 cm	SVM-RBF 05-10 cm	8.496616	0	*	10.38706	0	*	-10.17711	0	*	-9.285816	0	*	-10.80732	0	*
LASSO 15-20 cm	SVM-RBF 05-10 cm	-9.973542	0	*	-7.084596	0	*	-5.668225	0	*	-8.692058	0	*	-12.37079	0	*
RF 00-05 cm	SVM-RBF 05-10 cm	-1.148312	1		3.719259	0.0066	*	2.155572	1		0.095523	1		-4.39996	0.0004	*
RF 05-10 cm	SVM-RBF 05-10 cm	3.378807	0.024	*	1.115859	1		-0.26367	1		-2.86	0.1398		-2.956201	0.1028	
RF 10-15 cm	SVM-RBF 05-10 cm	11.2037	0	*	14.13388	0	*	-15.2527	0	*	-15.50568	0	*	-17.1958	0	*
RF 15-20 cm	SVM-RBF 05-10 cm	-8.544989	0	*	-4.678874	0.0001	*	-12.18342	0	*	-12.70589	0	*	-12.06871	0	*
SVM-RBF 00-05 cm	SVM-RBF 05-10 cm	-0.995231	1		5.377434	0	*	2.337336	0.6409		1.769844	1		0.547214	1	
LASSO 00-05 cm	SVM-RBF 10-15 cm	-12.53652	0	*	-8.205558	0	*	13.88825	0	*	12.03886	0	*	11.48373	0	*
LASSO 05-10 cm	SVM-RBF 10-15 cm	-4.450579	0.0003	*	-8.349454	0	*	6.74783	0	*	0.412916	1		1.557345	1	
LASSO 10-15 cm	SVM-RBF 10-15 cm	0.08695	1		-1.663686	1		-1.131592	1		-1.790255	1		-2.579519	0.3265	
LASSO 15-20 cm	SVM-RBF 10-15 cm	-18.3832	0	*	-19.13534	0	*	3.18684	0.0475		-1.196497	1		-4.142988	0.0011	*
RF 00-05 cm	SVM-RBF 10-15 cm	-9.557978	0	*	-8.331493	0	*	11.20109	0	*	7.591084	0	*	3.827845	0.0043	*
RF 05-10 cm	SVM-RBF 10-15 cm	-5.030858	0	*	-10.93489	0	*	8.78185	0	*	4.63556	0.0001	*	5.271604	0	*
RF 10-15 cm	SVM-RBF 10-15 cm	2.794037	0.1718		2.083128	1		-6.207189	0	*	-8.010124	0	*	-8.968003	0	*
RF 15-20 cm	SVM-RBF 10-15 cm	-16.95465	0	*	-16.72962	0	*	-3.137907	0.0562		-5.210336	0	*	-3.840908	0.004	*
SVM-RBF 00-05 cm	SVM-RBF 10-15 cm	-9.404897	0	*	-6.673318	0	*	11.38285	0	*	9.265405	0	*	8.775019	0	*
SVM-RBF 05-10 cm	SVM-RBF 10-15 cm	-8.409665	0	*	-12.05075	0	*	9.04552	0	*	7.49556	0	*	8.227805	0	*
LASSO 00-05 cm	SVM-RBF 15-20 cm	3.575771	0.0115	*	9.609312	0	*	16.62863	0	*	14.5731	0	*	7.84449	0	*
LASSO 05-10 cm	SVM-RBF 15-20 cm	11.66172	0	*	9.465416	0	*	9.133325	0	*	2.947155	0.1058		-2.081903	1	

LASSO 10-15 cm	SVM-RBF 15-20 cm	16.19925	0	*	16.15118	0	*	1.608791	1		0.743983	1		-6.218768	0	*
LASSO 15-20 cm	SVM-RBF 15-20 cm	-2.270907	0.764		-1.320478	1		5.869526	0	*	1.337742	1		-7.782237	0	*
RF 00-05 cm	SVM-RBF 15-20 cm	6.554322	0	*	9.483377	0	*	13.94147	0	*	10.12532	0	*	0.188595	1	
RF 05-10 cm	SVM-RBF 15-20 cm	11.08144	0	*	6.879977	0	*	11.52223	0	*	7.169799	0	*	1.632355	1	
RF 10-15 cm	SVM-RBF 15-20 cm	18.90633	0	*	19.89799	0	*	-3.466804	0.0174	*	-5.475884	0	*	-12.60725	0	*
RF 15-20 cm	SVM-RBF 15-20 cm	-0.842354	1		1.085243	1		-0.397523	1		-2.676096	0.2458		-7.480157	0	*
SVM-RBF 00-05 cm	SVM-RBF 15-20 cm	6.707403	0	*	11.14155	0	*	14.12324	0	*	11.79964	0	*	5.13577	0	*
SVM-RBF 05-10 cm	SVM-RBF 15-20 cm	7.702635	0	*	5.764117	0	*	11.7859	0	*	10.0298	0	*	4.588556	0.0001	*
SVM-RBF 10-15 cm	SVM-RBF 15-20 cm	16.1123	0	*	17.81487	0	*	2.740384	0.2025		2.534239	0.3719		-3.639249	0.009	*

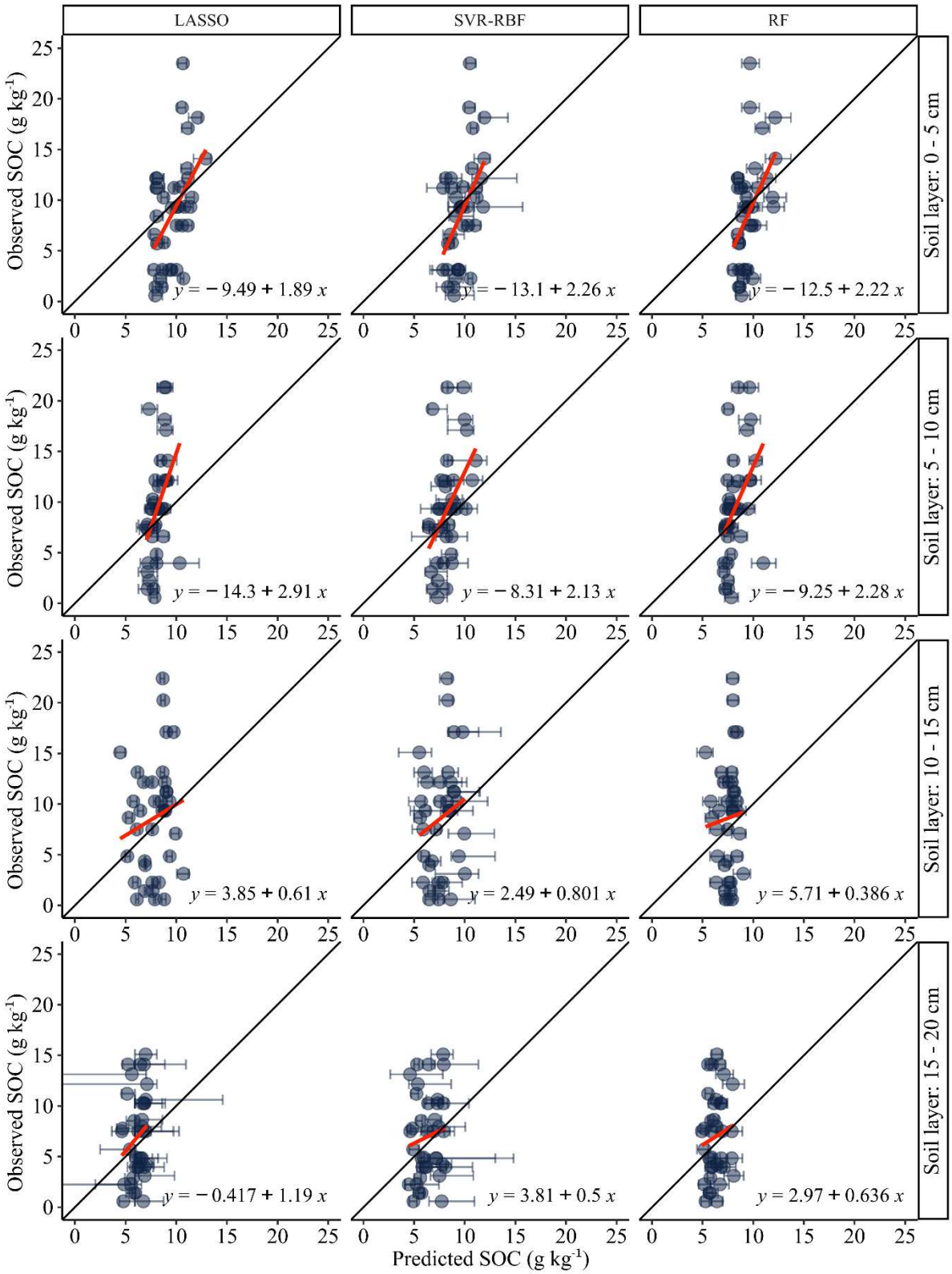


Figure B5. Scatterplots between observed and predicted soil organic carbon (SOC) contents in each topsoil layers for the holdout testing samples. In the first column, the predictions were made by LASSO (least absolute shrinkage and selection operator), in the second column by SVR-RBF (support vector regression with a radial basis function), and in the third column by the RF (random forest) regression method.

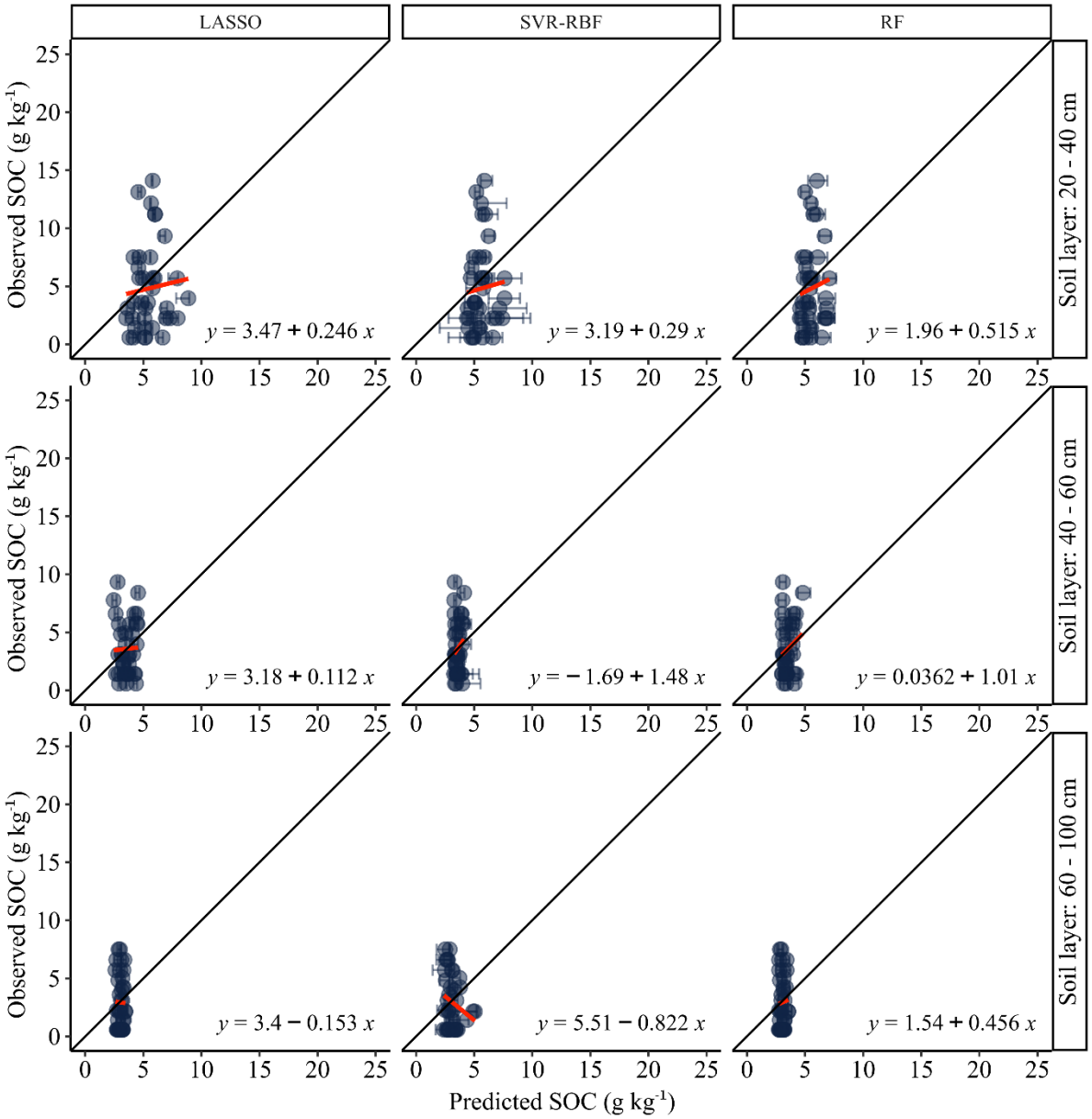


Figure B6. Scatterplots between observed and predicted soil organic carbon (SOC) contents in each deeper soil layers (soil depth from 20 up to 100 cm) for the holdout testing samples. In the first column, the predictions were made by LASSO (least absolute shrinkage and selection operator), in the second column by SVR-RBF (support vector regression with a radial basis function), and in the third column by the RF (random forest) regression method.

6. APPENDIX C: supplementary material for the article 3

The three Sankey diagrams presented in Figure C1 show the evolution of land use and land cover (LULC) at the locations of the soil profiles surveyed in the study. In order to construct those diagrams, data from the annual LULC classification carried out by the MapBiomas initiative (Souza et al., 2020), which generates collections of annual maps, was used. The 7.1 Collection data was used to construct the diagrams.

The LULC classes presented in Figure C1 are different from those established in the study, precisely because they come from different methodologies. Therefore, there is no exact correspondence between the classes observed in the field (LULC classes *in situ*: Agriculture, Other agricultural uses, Pasture, Open Caatinga, Dense Caatinga, and Forest Formation) and the classes of the MapBiomas methodology (Savanna Formation, Pasture, Mosaic of Uses, Other non-Vegetated Areas, Agriculture, Grassland, and Forest Formation). However, as MapBiomas generates annual collections, these were used precisely to facilitate analyzing the evolution of LULC classes over time.

Thus, in Figure C1 it is possible to notice little change in the LULC class between the profiles over the years. From the date of survey of the profiles (in the years 2011, 2012, and 2013), until the date of acquisition of the remote sensing measurements, most of the soil profiles remained under the same LULC class.

Even so, the few transitions that have occurred are susceptible to generalization errors in the models of the MapBiomas methodology. The Mosaic of Uses class is a good example because, in the MapBiomas methodology, this class was created to classify pixels whose algorithm could not distinguish between Agriculture and Pasture. There may be generalization errors in the model also at the time of sorting. Profiles under Pasture use (LULC *in situ*) were classified by MapBiomas as Other Non Vegetated Areas. These either remained under the Other Non Vegetated Areas class or evolved to Pasture (LULC MapBiomas).

Thus, there was no significant change in LULC class at the soil sampling sites.

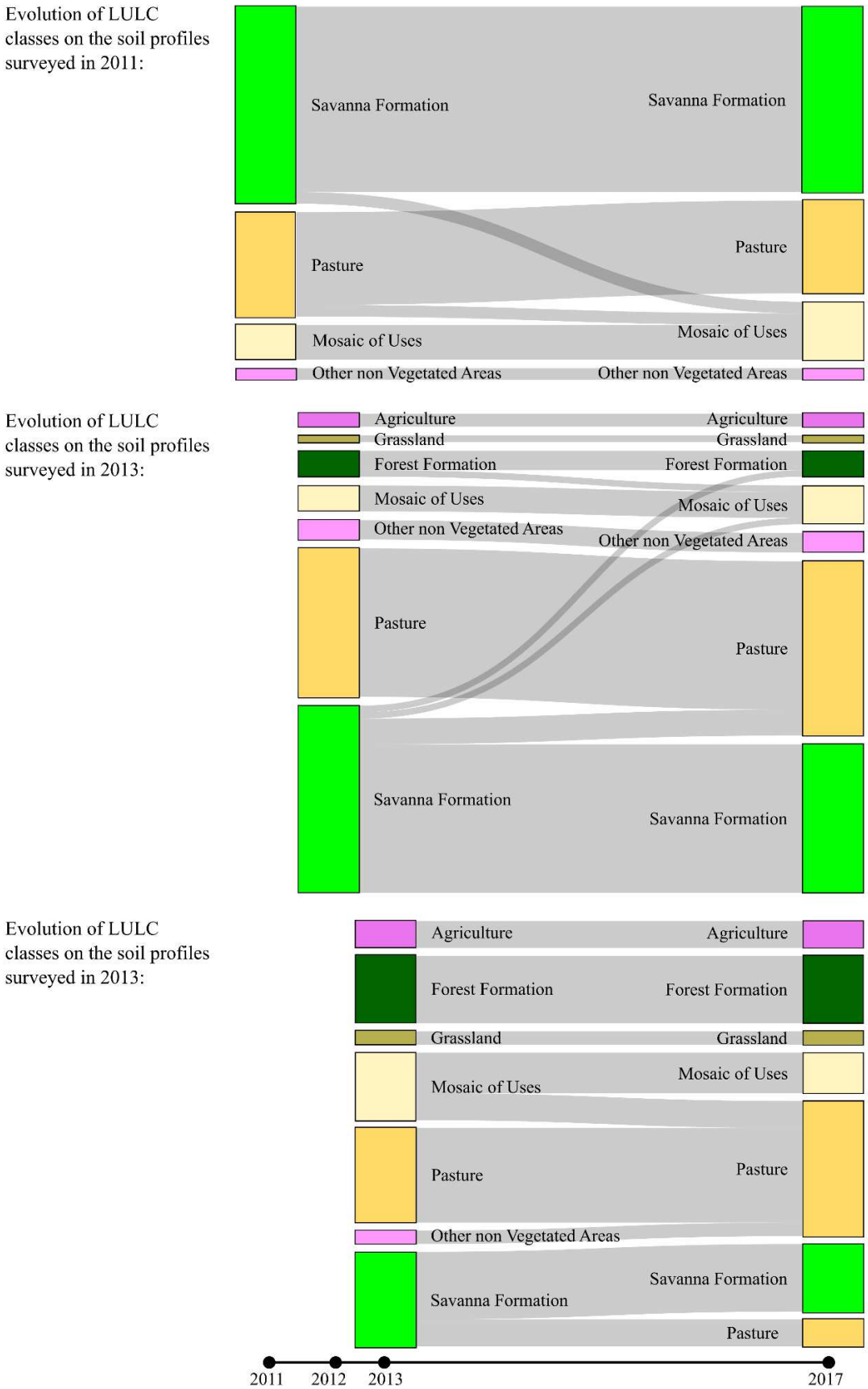


Figure C1. Evolution of Land Use and Land Cover (LULC) classes of soil profiles surveyed for the study: colors are for different LULC classes, and swath width represents amounts of soil profiles.

Table C1. Inventory of Sentinel-1 IW GRD images used in the study, generated by the SAR sensor of the Sentinel-1A satellite.

Acquisition date	Product unique identifier	Relative orbit number
23 January 2017	94C8	9
28 January 2017	5B3F	82
21 February 2017	7C09	82
28 February 2017	CB17	9
28 February 2017	E263	9
24 March 2017	BAE5	9
29 March 2017	C51B	82
29 March 2017	9E3F	82
22 April 2017	7C84	82
29 April 2017	256D	9
23 May 2017	40F0	9
23 May 2017	CC51	9
23 May 2017	69AD	9
28 June 2017	302D	9
28 June 2017	0899	9
22 July 2017	BDF5	9
22 July 2017	2ED7	9
27 July 2017	59F7	82
27 July 2017	ABDA	82
20 August 2017	D8E2	82
27 August 2017	8C1D	9
20 September 2017	5637	9
14 October 2017	6314	9
26 October 2017	33A2	9
26 October 2017	D3D9	9
19 November 2017	A01A	9
24 November 2017	FF23	82
24 November 2017	FD2D	82

Table C2. Pairwise Dunn test results for the model groups at the testing step: the asterisk aside adjusted p-value (*) indicates a significant difference at P = 0.05.

List of pairwise comparisons (groups)		Target variable	Z statistic	Adjusted p-value	
Model set 1	Model set 4	RMSE	18.155490	0.0000	*
Model set 1	Model set 3		11.155620	0.0000	*
Model set 4	Model set 3		-6.999871	0.0000	*
Model set 1	Model set 2		7.262932	0.0000	*
Model set 4	Model set 2		-10.892560	0.0000	*
Model set 3	Model set 2		-3.892692	0.0003	*
Model set 1	Model set 4	MAE	17.072920	0.0000	*
Model set 1	Model set 3		12.313000	0.0000	*
Model set 4	Model set 3		-4.759913	0.0000	*
Model set 1	Model set 2		5.671398	0.0000	*
Model set 4	Model set 2		-11.401520	0.0000	*
Model set 3	Model set 2		-6.641611	0.0000	*
Model set 1	Model set 4	MSE	18.154270	0.0000	*
Model set 1	Model set 3		11.154870	0.0000	*
Model set 4	Model set 3		-6.999402	0.0000	*
Model set 1	Model set 2		7.262445	0.0000	*
Model set 4	Model set 2		-10.891830	0.0000	*
Model set 3	Model set 2		-3.892431	0.0003	*
Model set 1	Model set 4	R^2	-18.020400	0.0000	*
Model set 1	Model set 3		-11.884080	0.0000	*
Model set 4	Model set 3		6.136325	0.0000	*
Model set 1	Model set 2		-6.693171	0.0000	*
Model set 4	Model set 2		11.327230	0.0000	*
Model set 3	Model set 2		5.190910	0.0000	*
Model set 1	Model set 4	CCC	-18.128350	0.0000	*
Model set 1	Model set 3		-11.157950	0.0000	*
Model set 4	Model set 3		6.970394	0.0000	*
Model set 1	Model set 2		-7.198039	0.0000	*
Model set 4	Model set 2		10.930310	0.0000	*
Model set 3	Model set 2		3.959916	0.0002	*
Model set 1	Model set 4	NSE	-18.162220	0.0000	*
Model set 1	Model set 3		-11.116910	0.0000	*
Model set 4	Model set 3		7.045305	0.0000	*
Model set 1	Model set 2		-7.308464	0.0000	*
Model set 4	Model set 2		10.853760	0.0000	*
Model set 3	Model set 2		3.808455	0.0004	*

6.1. References

Souza, C.M., Z. Shimbo, J., Rosa, M.R., Parente, L.L., A. Alencar, A., Rudorff, B.F.T., Hasenack, H., Matsumoto, M., G. Ferreira, L., Souza-Filho, P.W.M., de Oliveira, S.W., Rocha, W.F., Fonseca, A.V., Marques, C.B., Diniz, C.G., Costa, D., Monteiro, D., Rosa, E.R., Vélez-Martin, E., Weber, E.J., Lenti, F.E.B., Paternost, F.F., Pareyn, F.G.C., Siqueira, J.V., Viera, J.L., Neto, L.C.F., Saraiva, M.M., Sales, M.H., Salgado, M.P.G., Vasconcelos, R., Galano, S., Mesquita, V.V., Azevedo, T., 2020. Reconstructing Three Decades of Land Use and Land Cover Changes in Brazilian Biomes with Landsat Archive and Earth Engine. Remote Sens. 12, 2735. https://doi.org/10.3390/rs12172735