



MARCELO MANCINI

**BEYOND PREDICTION: GAINING PEDOLOGIC KNOWLEDGE
FROM SENSORS AND MACHINE-LEARNING MODELS**

LAVRAS – MG

2026

MARCELO MANCINI

**BEYOND PREDICTION: GAINING PEDOLOGIC KNOWLEDGE FROM SENSORS
AND MACHINE-LEARNING MODELS**

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós- Graduação em Ciência do Solo, área de concentração em Recursos Ambientais e Uso da Terra, para a obtenção do título de Doutor.

Prof. Dr. Nilton Curi

Orientador

Prof. Dr. Sérgio Henrique Godinho Silva

Co-orientador

LAVRAS – MG

2026

**Ficha Catalográfica elaborada pelo Sistema de Geração
de Ficha Catalográfica da Biblioteca Universitária da UFLA, com
dados informados pelo(a) próprio(a) autor(a).**

Mancini, Marcelo.

Beyond prediction : Gaining pedologic knowledge from sensors and
machine-learning models / Marcelo Mancini. - 2026.
170 p. : il.

Orientador: Nilton Curi

Coorientador: Sérgio Henrique Godinho Silva

Tese (Doutorado) - Universidade Federal de Lavras, 2026.
Bibliografia.

1. Pedometrics. 2. Soil mapping. 3. Soil sensing. 4. Pedology. 5. Explainable AI. I. Curi, Nilton. II. Silva, Sérgio Henrique Godinho. III. Universidade Federal de Lavras. IV. Título.

MARCELO MANCINI

**PARA ALÉM DA PREDIÇÃO: OBTENDO CONHECIMENTO PEDOLÓGICO À
PARTIR DE SENSORES E MODELOS DE APRENDIZADO DE MÁQUINA**

**BEYOND PREDICTION: GAINING PEDOLOGIC KNOWLEDGE FROM SENSORS
AND MACHINE-LEARNING MODELS**

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós- Graduação em Ciência do Solo, área de concentração em Recursos Ambientais e Uso da Terra, para a obtenção do título de Doutor.

APROVADA em 29 de maio de 2026.

Dr. Nilton Curi UFLA-DCS

Dr. Renata Andrade Reis UFLA-DCS

Dr. Carlos Rogério de Mello UFLA-DRH

Dr. Marcelo Ribeiro Viola UFLA-DRH

Dr. Zamir Libohova USDA-ARS

Prof. Dr. Nilton Curi

Orientador

Prof. Dr. Sérgio Henrique Godinho Silva

Co-orientador

LAVRAS – MG

2026

AGRADECIMENTOS

Aos meus pais, pelo incansável apoio ao longo de toda a vida, sem o qual jamais teria chegado até aqui.

Aos meus orientadores Sérgio e Nilton por todo companheirismo ao longo deste caminho e pelos ricos ensinamentos que sempre compartilharam comigo.

Aos integrantes do Departamento de Ciência do Solo por todo acolhimento sempre. São as fontes de todo material-base para trabalhos como este, desde a coleta de amostras e análises laboratoriais, até o compartilhamento do conhecimento necessário para interpretação de seus resultados.

A todos os professores e colegas do Programa de Pós-Graduação em Ciência do Solo que nunca faltaram em fornecer apoio e amizade ao longo do laborioso período de pós-graduação.

Aos colegas da USDA que muito bem me acolheram em minha estadia nos Estados Unidos e que, mesmo após minha volta ao Brasil, continuaram a cultivar nossa amizade e parceria profissional.

À Universidade Federal de Lavras por prover o ambiente fraterno de aprendizado no qual todas as minhas conquistas foram possíveis.

Ao Conselho Nacional de Desenvolvimento Científico (CNPq), à Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro.

A Deus por invariavelmente iluminar o meu caminho e o de todos à minha volta.

ABSTRACT

Soils are heterogeneous and vary greatly through space and time. Recently, soil sensing has become popular, facilitating the assessment of soils and their properties, especially when combined with machine-learning models. This thesis explores the application of proximal and remote sensors in the characterization of soil profiles, optimization of soil sampling, and in the estimation of soil properties. In addition, the reliability of black-box models is assessed using machine-learning interpretation tools. Portable X-ray fluorescence (pXRF), visible and near-infrared diffuse reflectance spectroscopy (Vis-NIR), mid-infrared spectroscopy (MIR), the Nix colorimeter, and Sentinel-2 imagery were used in several studies aiming to characterize soils and estimate their properties. Partial differential plots, individual conditional expectation, permutation feature importance, and Shapley values were applied to investigate the reliability of soil texture and organic carbon estimation models. When applied to the characterization of soil profiles, Vis-NIR and MIR were able to identify variations in soil organic matter and mineralogy, whereas pXRF unveiled lithologic discontinuities in soil profiles. Temporal trends extracted from Sentinel-2 imagery acted as proxies of soil moisture and vegetation variations that allowed for the optimization of soil sampling. Sampling grids guided by remotely sensed temporal trends coupled with conditioned Latin hypercube sampling represented soil variability with fewer samples compared to conventional random and grid sampling strategies. When using 1749 soil samples from five countries, pXRF was capable of estimating sand, silt, and clay contents accurately ($R^2 > 0.84$), as well as soil organic carbon ($R^2 = 0.74$), regardless of the very different climate, parent materials, and weathering intensities. Machine-learning model interpretation techniques indicated that models learned physically sound relationships between covariates and soil texture. Si was mostly related to quartz contents in the sand fraction of soils, while Fe and Al acted as fingerprints of clay content. Silt was particularly difficult to model using pXRF data, likely because it lacks characteristic minerals present only in the silt fraction. Si, Fe, and Al were the most relevant chemical elements for estimating soil texture. The use of machine-learning interpretation methods is advised to diagnose potential inconsistencies in models and to learn about the relationships between covariates and soil properties. Soil sensing has come a long way and has proven reliable. Sensors and machine-learning combined can reliably be applied to soil characterization and mapping if coupled with sound modeling techniques.

KEYWORDS: pedometrics; soil mapping; soil sensing; proximal sensing; remote sensing; machine-learning; explainable AI; pXRF; Vis-NIR; MIR; soil organic carbon; soil texture.

RESUMO

Solos são heterogêneos e variam muito no espaço e no tempo. Recentemente, sensoriamento de solos tornou-se uma técnica popular, facilitando a análise de solos e suas propriedades, especialmente quando combinada com modelos de aprendizado de máquina. Essa tese explora a aplicação de sensores próximos e remotos na caracterização de perfis de solo, na otimização da amostragem de solos, e na estimação de atributos do solo. Adicionalmente, a confiabilidade de modelos caixa-preta é avaliada usando métodos de interpretação de modelos de aprendizado de máquina. O espectrômetro portátil de fluorescência de raios-X (pXRF), a espectroscopia nas faixas do visível e infravermelho próximo (Vis-NIR), espectroscopia na faixa do infravermelho médio (MIR), o colorímetro Nix, e imagens do satélite Sentinel-2 foram usados em estudos com o objetivo de caracterizar solos e estimar seus atributos. Gráficos de dependência parcial, esperança condicional individual, importância de covariáveis por permutação, e valores Shapley foram utilizados para investigar a confiabilidade de modelos preditivos de textura do solo e carbono orgânico. Quando aplicados à caracterização de perfis de solos, Vis-NIR e MIR foram capazes de identificar variações no teor de matéria orgânica e mineralogia, enquanto o pXRF desvendou descontinuidades litológicas ao longo dos perfis. Tendências temporais extraídas de imagens de Sentinel-2 serviram como proxies de umidade do solo e vegetação que permitiram a otimização da amostragem do solo. Malhas de amostragem definidas com base em tendências temporais extraídas de sensoriamento remoto combinadas com hipercubo latino condicionado representaram a variabilidade do solo com menos amostras quando comparadas à estratégias de amostragem convencionais, como as malhas aleatória e regular. Utilizando 1749 amostras de solo de cinco países, o pXRF foi capaz de estimar os teores de areia, silte, e argila com exatidão ($R^2 > 0.84$), assim como carbono orgânico ($R^2 = 0.74$), mesmo com as grandes diferenças no clima, material de origem, e intensidades de intemperismo. Técnicas de interpretação de modelos de aprendizado de máquina indicaram que os modelos aprenderam correlações físicas adequadas entre covariáveis e a textura do solo. Si foi associado com o teor de quartzo na fração areais dos solos, enquanto o Fe e o Al agiram como identificadores da fração argila. A fração silte foi particularmente difícil de modelar utilizando dados de pXRF, provavelmente devido à ausência de minerais característicos presentes apenas na fração silte e não em outras frações. Si, Fe, e Al foram os elementos químicos mais relevantes para a estimação da textura do solo. O uso de métodos de interpretação de aprendizado de máquina é recomendado para diagnosticar potenciais inconsistências em modelos e para aprender sobre as relações entre covariáveis ambientais e atributos do solo. Sensoriamento de solos percorreu um longo caminho nas últimas décadas e se provou um método confiável. A combinação de

sensores e aprendizado de máquina pode ser seguramente aplicada à caracterização e ao mapeamento de solos se utilizados em conjunto com técnicas adequadas de modelagem.

PALAVRAS-CHAVE: pedometria; mapeamento de solos; sensoriamento de solos; sensoriamento próximo; sensoriamento remoto; aprendizado de máquina; AI explicável; pXRF; Vis-NIR; MIR; carbono orgânico; textura do solo.

IMPACT INDICATORS

Sensors can facilitate the analysis of soils samples, increasing data availability and hence improving the accuracy of maps and the completeness of soil profile characterizations. Soil sensing also contributes to lower the cost of soil analyses, making them more accessible for small farmer, incentivizing sustainable, adequate soil management. Complete soil profile characterizations and accurate soil maps can better inform crop management, environmental monitoring, and urban planning. The adoption of proximal and remote sensors reduces the necessity of laboratory analysis, decreasing the generation of chemical waste. The accurate modeling of soil organic carbon can promote a more reliable quantification of carbon stocks, land use impacts, and soil health. Texture-related properties such as structure and mineralogy affect the capacity of soils to retain water and nutrients; hence, accurate sensor-driven estimations of soil texture can guide adequate crop management. Proper assessment of machine-learning models avoids misinterpretations and can diagnose biases often neglected by modelers when using black-box models. The use of interpretation methods such as partial dependence plots and Shapley values makes models more transparent, increasing the trust of end users and decreasing the risk of misuse. More transparent machine-learning models can also lead to new scientific discoveries by unveiling the effects of environmental covariates on key soil properties.

INDICADORES DE IMPACTO

Sensores podem facilitar a análise de amostras de solo, aumentando a disponibilidade de dados e portanto melhorando a acurácia de mapas e a completude da caracterização de perfis de solo. Sensoriamento de solos também contribui para a redução de custos da análise de solos, tornando-as mais acessíveis à pequenos produtores, incentivando um manejo de solo mais sustentável e adequado. A caracterização mais completa de perfis de solos e mapas de solos mais acurados podem guiar um melhor manejo de culturas, monitoramento ambiental, e planejamento urbano. A adoção de sensores próximos e remotos reduzem a necessidade de análises de laboratório, diminuindo a geração de rejeitos químicos. A modelagem acurada de carbono orgânico do solo pode promover a quantificação mais confiável de estoques de carbono, de impacto do uso do solo, e da saúde do solo. A textura e a mineralogia do solo afetam a capacidade dos solos de reter e disponibilizar água e nutrientes; portanto, estimativas acuradas da textura do solo via sensores podem guiar o manejo adequado do solo. A avaliação correta de modelos de aprendizado de máquina evita interpretações equivocadas e pode diagnosticar a presença de vies comumente negligenciada por cientistas quando utilizam modelos caixa-preta. A utilização de métodos de interpretação como gráficos de dependência parcial e valores Shapley tornam os modelos mais transparentes, aumentando a confiança dos usuários e diminuindo o risco de uso inadequado. Modelos mais transparentes de aprendizado de máquina podem também levar à novas descobertas científicas, desvendando os efeitos de covariáveis ambientais em importantes atributos do solo.

Summary

PART I	11
Introduction	12
PART II.....	16
Chapter I: Digital morphometrics and genesis of soils with buried horizons and lithological discontinuities in Southeastern Brazil	17
Chapter II: Pixel-based spatiotemporal statistics from remotely sensed imagery improves spatial soil predictions and sampling strategies	49
Chapter III: Multinational prediction of soil organic carbon and texture via proximal sensors.....	81
Chapter IV: Interpreting machine-learning models trained to estimate soil texture and organic carbon in soils of Southeast Brazil	112
Chapter V: Machine-learning indicates which chemical elements are most influential in explaining soil texture in Brazilian biomes.	143
Part III	169
Closing Remarks.....	170

PART I

Introduction

In the 21st century, pedology has experienced an exponential increase in scientific research about the use of sensors and machine-learning (Padarian et al., 2020). As always happens with the advent of new technology, much was learned during the trial-and-error phases that ensued. Inconsistencies in maps, emphasis on accuracy instead of interpretation, abstraction of field knowledge in favor of numbers and models (Arrouyas et al., 2020; Rossiter et al., 2022; Wadoux et al., 2020a), and many other challenges. But the lessons learned have shown that there is space for gaining scientific knowledge from sensors and machine-learning. This thesis is a series of studies showing that the combination of sensors and machine-learning models is not only accurate, but also consistent with our accumulated knowledge about soils and can be used to advance our scientific understanding about them. Chapter I describes how sensors can be used in soil characterization via soil sensing; Chapter II investigates the optimization of soil sampling using remote sensing; Chapter III highlights that soil property models trained soil sensing data can be used in very distinct soils across of the world, but will not always be transferable between regions; and Chapter IV and V explore methods to interpret and learn from black-box soil models.

One of the biggest challenges in soil studies is sampling (Biswas and Zhang, 2018). Soil variability needs to be represented across landscapes and its depth, requiring scientists to dig soil pits and trenches, to sample and analyze many depths per point, and across many points throughout space. Sensors are used to mitigate this labor. Sensors facilitate soil analyses by reducing the cost and time needed to analyze samples, making the analysis of more samples viable, helping in obtaining better representations of the spatial variability of soils (Silva et al., 2025). Among the many sensors that have been used in soil studies, this work focused on studying three proximal sensors: the portable X-ray fluorescence (pXRF), the visible and near infrared diffuse reflectance spectroscope (Vis-NIR), and the Nix colorimeter. PXRF detects the fluorescence emitted by atoms and uses this information to estimate the chemical composition of samples (Silva et al., 2021). Vis-NIR can detect the electromagnetic spectrum reflected by soil, indicating the presence of minerals and other chemical compounds contained in samples (Teixeira et al., 2022). Lastly, Nix is a lightweight, inexpensive colorimeter capable of objectively measuring the color of soils within seconds (Faria et al., 2022). These three sensors require less than one minute to provide data and do not destroy the samples or contaminate the environment with the use of chemicals in laboratorial analyzes. The obtained chemical, spectral, and color data tend to correlate with key physical and chemical soil properties and can be used to enhance soil characterization.

Digital morphometrics, discussed in Chapter I, is a good example of how these sensors can be useful in soil characterization (Hartemink and Minasny, 2014). It consists in the morphological characterization and quantification of soil properties using digital tools such as sensors to complement the traditional methods. Soil characterization was commonly performed visually and with the aid of mechanical tools such as knives, pedologic and geologic hammers, magnets, and by feeling the soil with the pedologists' bare hands. Sensors, however, can provide insights invisible to the human eye. Spectral data can indicate the presence of minerals or organic compounds, pXRF can detect trace elements across soil profiles, and Nix can provide an objective and easy way of quantifying soil color. Digital morphometrics combines traditional methods and digital resources to offer a more complete and reliable characterization of soil profiles.

Sensors can also facilitate studies about the spatial variability of soils. As is discussed in Chapter II, remote sensing can help optimize soil sampling, allowing for the representation of soils' spatial variability using fewer samples, reducing cost and time of sampling campaigns. Publicly available satellite imagery provided by Sentinel-2 can be used to characterize spatiotemporal trends on Earth's surface. The reflectance data provided by Sentinel-2 vary with soil moisture and vegetation changes across time and space, unveiling trends that depend on soil properties, such as the retention of moisture throughout the year and the release and retention or leaching of plant nutrients. These spatiotemporal trends can be used to stratify soils and better inform sampling, optimizing fieldwork while not compromising the representativeness of samples.

In addition to supporting profile characterization and sampling optimization, sensor data can also be used to directly estimate soil properties. Statistical and machine-learning models use sensor outputs as input for estimating soil properties across space and time. Chapter III discusses machine-learning modeling techniques that use easy-to-obtain pXRF data to predict soil organic matter (SOM) and texture in soil samples from five different countries, attesting the reliability of pXRF data as a proxy for soil properties. However, although machine-learning models excel at finding complex relationships between covariates and the modeled variables, the complexity hinders the interpretation of these models by humans. Such black-box models may be accurate but do not necessarily represent the real physical and chemical relationships required to fully explain the phenomena studied. Many studies illustrate the pitfalls of blind trust in black-box models, proving that even pseudo-variables can be used to create models that appear accurate (Fourcade et al., 2018; Wadoux et al., 2020b). Chapters IV and V dive deep

into exploring methods of interpreting machine-learning models to validate them and allow scientists to extract pedologic knowledge from patterns learned by models.

The goal of this thesis is to provide sufficient evidence that proximal and remote sensing can be reliably used in soil assessment, and that the correct use of machine-learning can enrich our understanding about the physical and chemical processes that take place in soils, especially those related to texture, mineralogy, and organic carbon. The leading hypothesis is that the combination of sensors and machine-learning offers a powerful and reliable solution to increase the availability of soil data at reduced cost, making the onerous tasks of sampling and analyzing soils easier, enhancing our capacity of studying soils across space and time.

References

- Arrouays, D., McBratney, A., Bouma, J., Libohova, Z., Richer-de-Forges, A.C., Morgan, C.L.S., Roudier, P., Poggio, L., Mulder, V.L., 2020. Impressions of digital soil maps: The good, the not so good, and making them ever better. *Geoderma Regional* 20, e00255. <https://doi.org/10.1016/j.geodrs.2020.e00255>
- Biswas, A., Zhang, Y., 2018. Sampling Designs for Validating Digital Soil Maps: A Review. *Pedosphere* 28, 1–15. [https://doi.org/10.1016/S1002-0160\(18\)60001-3](https://doi.org/10.1016/S1002-0160(18)60001-3)
- Faria, A.J.G., Silva, S.H.G., Andrade, R., Mancini, M., Melo, L.C.A., Weindorf, D.C., Guilherme, L.R.G., Curi, N., 2022. Prediction of soil organic matter content by combining data from Nix ProTM color sensor and portable X-ray fluorescence spectrometry in tropical soils. *Geoderma Regional* 28, e00461. <https://doi.org/10.1016/j.geodrs.2021.e00461>
- Fourcade, Y., Besnard, A.G., Secondi, J., 2018. Paintings predict the distribution of species, or the challenge of selecting environmental predictors and evaluation statistics. *Global Ecology and Biogeography* 27, 245–256. <https://doi.org/10.1111/geb.12684>
- Hartemink, A.E., Minasny, B., 2014. Towards digital soil morphometrics. *Geoderma* 230–231, 305–317. <https://doi.org/10.1016/j.geoderma.2014.03.008>
- Padarian, J., Minasny, B., McBratney, A.B., 2020. Machine learning and soil sciences: a review aided by machine learning tools. *SOIL* 6, 35–52. <https://doi.org/10.5194/soil-6-35-2020>
- Rossiter, D.G., Poggio, L., Beaudette, D., Libohova, Z., 2022. How well does digital soil mapping represent soil geography? An investigation from the USA. *SOIL* 8, 559–586. <https://doi.org/10.5194/soil-8-559-2022>
- Silva, S.H.G., Mancini, M., Teixeira, A.F.S., Andrade, R., Curi, N., 2025. Chapter 5 - Applications of machine learning and artificial intelligence in soil science, in: Weindorf, D.C., Chakraborty, S., Li, B. (Eds.), *Unlocking the Secrets of Soil*. Elsevier, pp. 155–179. <https://doi.org/10.1016/B978-0-443-29879-0.00004-6>
- Silva, S.H.G., Ribeiro, B.T., Guerra, M.B.B., de Carvalho, H.W.P., Lopes, G., Carvalho, G.S., Guilherme, L.R.G., Resende, M., Mancini, M., Curi, N., Rafael, R.B.A., Cardelli, V., Cocco, S., Corti, G., Chakraborty, S., Li, B., Weindorf, D.C., 2021. Chapter One - pXRF in tropical soils: Methodology, applications, achievements and challenges, in: Sparks, D.L. (Ed.), *Advances in Agronomy*. Academic Press, pp. 1–62. <https://doi.org/10.1016/bs.agron.2020.12.001>
- Teixeira, A.F. dos S., Andrade, R., Mancini, M., Silva, S.H.G., Weindorf, D.C., Chakraborty, S., Guilherme, L.R.G., Curi, N., 2022. Proximal sensor data fusion for tropical soil property prediction: Soil fertility properties. *Journal of South American Earth Sciences* 116, 103873. <https://doi.org/10.1016/j.jsames.2022.103873>
- Wadoux, A.M.J.-C., Minasny, B., McBratney, A.B., 2020a. Machine learning for digital soil mapping: Applications, challenges and suggested solutions. *Earth-Science Reviews* 210, 103359. <https://doi.org/10.1016/j.earscirev.2020.103359>
- Wadoux, A.M.J.-C., Samuel-Rosa, A., Poggio, L., Mulder, V.L., 2020b. A note on knowledge discovery and machine learning in digital soil mapping. *European Journal of Soil Science* 71, 133–136. <https://doi.org/10.1111/ejss.12909>

PART II

Chapter I: Digital morphometrics and genesis of soils with buried horizons and lithological discontinuities in Southeastern Brazil

Accepted in Geoderma Regional on 23 January 2023.

<https://doi.org/10.1016/j.geodrs.2023.e00612>

Marcelo Mancini^a, Sérgio Henrique Godinho Silva^a, Junior Cesar Avanzi^a, Alfred E. Hartemink^b, Alberto Vasconcellos Inda^c, José A.M. Demattê^d, Nilton Curi^{a,*}

^aDept. of Soil Science, Federal University of Lavras, P.O. Box 3037, 37200-900, Lavras, State of Minas Gerais, Brazil

^bDept. of Soil Science, University of Wisconsin-Madison, FD Hole Soils Lab, 1525 Observatory Drive, Madison 53706, USA

^cDept. of Soils, Federal University of Rio Grande Do Sul, Porto Alegre, State of Rio Grande do Sul, Brazil

^dDept. of Soil Science, Luiz de Queiroz College of Agriculture, University of São Paulo, Av. Pádua Dias 11, 13418-260, Piracicaba, State of São Paulo, Brazil

*Corresponding author: niltcuri@ufla.br

Abbreviations: TU: Typic Udifolists; TD: Typic Dystrustepts; TE: Typic Endoaquents; PXRF: portable X-ray fluorescence spectrometry; Vis-NIR: visible and near-infrared spectroscopy; MIR: mid-infrared spectroscopy; XRD: X-ray diffraction; SOM: soil organic matter content; HDI: horizon difference index.

Abstract

Soil morphometrics aims to measure attributes observed in soil profiles. The use of digital technologies in the morphological study of soil profiles has yet to be further explored by the scientific community. Particularly, soils with buried horizons, which are common in lowlands and floodplains throughout the world, have never been examined using digital morphometrics. This study aimed to examine the morphology and genesis of three soil profiles with buried horizons in Brazil using digital soil morphometrics combined with wet laboratory and mineralogical analyses (including X-ray diffraction – XRD). Digital analyzes used three proximal sensors – portable X-ray fluorescence spectrometry (pXRF), visible near-infrared (Vis-NIR) and mid infrared (MIR) diffuse reflectance spectrometry. The soils had buried horizons and were classified as Typic Udifolists (TU), Typic Dystrustepts (TD) and Typic Endoaquents (TE). Each soil profile was sampled in a 15 x 15 cm grid, with 40 samples per profile. All sensors identified the buried soil horizons. Vis-NIR and MIR identified the presence of several key minerals and variations in soil organic matter (SOM) that were not observed in the soil pits. PXRF captured discontinuities in the parent material, not only between deposited and buried horizons, but also within horizons. The deposited material above TU and TE were derived from neighboring Fe-rich Oxisols and Ultisols, as shown by high Fe content and absorption features of Fe oxides. Traditional X-ray techniques are difficult and costly. Sensor analyses were quick and environmentally friendly. Proximal sensor data enhanced morphological assessment of soil profiles, and provided insights otherwise impossible with only traditional and visual methods.

Keywords: Histosol; Inceptisol; Entisol; Cambisol; Gleysol; pedogenesis; parent material discontinuity; morphometry; digital soil mapping; weathering; spectroscopy; PXRF; Vis-NIR; MIR

1. Introduction

The universal presence of soils in human life is unquestionable. Their matrix is the basis of food production, foundation of civil constructions, source of materials for human activities and responsible for a myriad of ecosystem services (Young et al., 2001). The complexity of soil formation is as large as the importance of soils for humankind. Soils are the product of an incessant and intricate interaction of climate, time, parent material, relief and biota (Jenny, 1941). The temporal and spatial heterogeneity of soil is a recurring challenge in pedology.

Ever since soils were recognized as distinct natural bodies, scientists seek to identify and quantify their variations in order to group and classify their features (Hartemink et al., 2020). The investigation of vertical cross-sections –soil profiles– in order to quantify and map their attributes with the aid of new technologies is referred to as digital soil morphometrics (Hartemink and Minasny, 2014). Although technological advancements have been incorporated in the last couple of years, only a handful of works have applied these techniques to study the variation of soil properties within and in-between horizons (Mancini et al., 2021; Silva et al., 2018; Zhang and Hartemink, 2019; Sun et al., 2020).

Among the digital equipment that can be used in pedological studies beyond the visible spectrum, some proximal sensors have received special attention, including the portable X-ray fluorescence spectrometer (PXRF), used to assess the elemental contents of samples (Weindorf et al., 2014; Silva et al., 2021), and those that use diffuse reflectance spectroscopy (DRS) to infer interactions between soil samples and radiation in the visible (350–780 nm, Vis), near-infrared (780–2500 nm, NIR) and mid-infrared ranges (2500–25000 nm or 4,000–600 waves per centimeter, cm^{-1} , MIR). Such sensors have been extensively applied to soil-related studies (O'Rourke et al., 2016; Weindorf et al., 2016; Gozukara et al., 2021; Chen et al., 2021; Gozukara et al., 2022)

DRS has been used for the prediction of soil chemical and physical attributes (Andrade et al., 2022; Faria et al., 2020), the investigation of pedogenetic processes (Stockmann et al., 2016), the study of parent material distribution (Mancini et al., 2019a, 2019b) and the assessment of contaminants in soils (Horta et al., 2015; Rouillon and Taylor, 2016). Parallely, DRS methods have been also applied to the prediction of physical and chemical properties (Benedet et al., 2020; Kandpal et al., 2022; Zhang and Hartemink, 2020), horizon distinction and detection of minerals (Coblinski et al., 2021), biological properties (Zhang et al., 2022) and soil pollutants (Corradini et al., 2019; Li et al., 2021).

Soils with buried horizons or lithologic discontinuities have not been widely studied using digital soil morphometrics. These soils are common in lowlands and floodplains in many

regions of the world, including Brazil. Soils with buried horizons have deposition of altered material, and if the superficial horizons and the underlying material have different ages and properties, the soils are classified as buried soils (USDA, 2010), presenting horizons that were formed prior to the depositional event. Understanding how to apply sensor technologies to buried horizons could greatly improve soil characterization methods.

Here, the goal was to use digital soil morphometrics to characterize soils with buried horizons; i.e., soils with lithologic discontinuities due to recent deposition of material. Digital soil morphometrics was done using PXRF, Vis-NIR and MIR, combined with laboratory analyses and mineralogical identification via X-ray diffractometry (DRX). Specifically, the objectives were to: i) investigate the morphology and genesis of three soil profiles with buried horizons using proximal sensors; ii) compare data provided by sensors with those of traditional laboratory and mineralogical assessment methods; iii) identify the contributions of each sensor in characterizing key soil attributes. We hypothesized that the proximal sensors would provide key insights about the variation of soils attributes and help to enhance the characterization of the deposited and buried horizons, complementing and improving soil morphological assessment methods.

2. Material and Methods

2.1. Study area

The study was conducted in the municipality of Lavras, in the State of Minas Gerais, Brazil (Figure 1). The municipality is situated between the latitudes 21.45 and 21.10°S, and longitudes 45.29 and 44.79°W (WGS84). The mean annual temperature in the region is 20.4 °C and the mean annual precipitation is 1460 mm (Alvares et al., 2013). The altitude of the municipality varies between 760 and 1260 m asl. The dominant vegetation in the region is the semi-perennial forest (Curi et al., 2017).

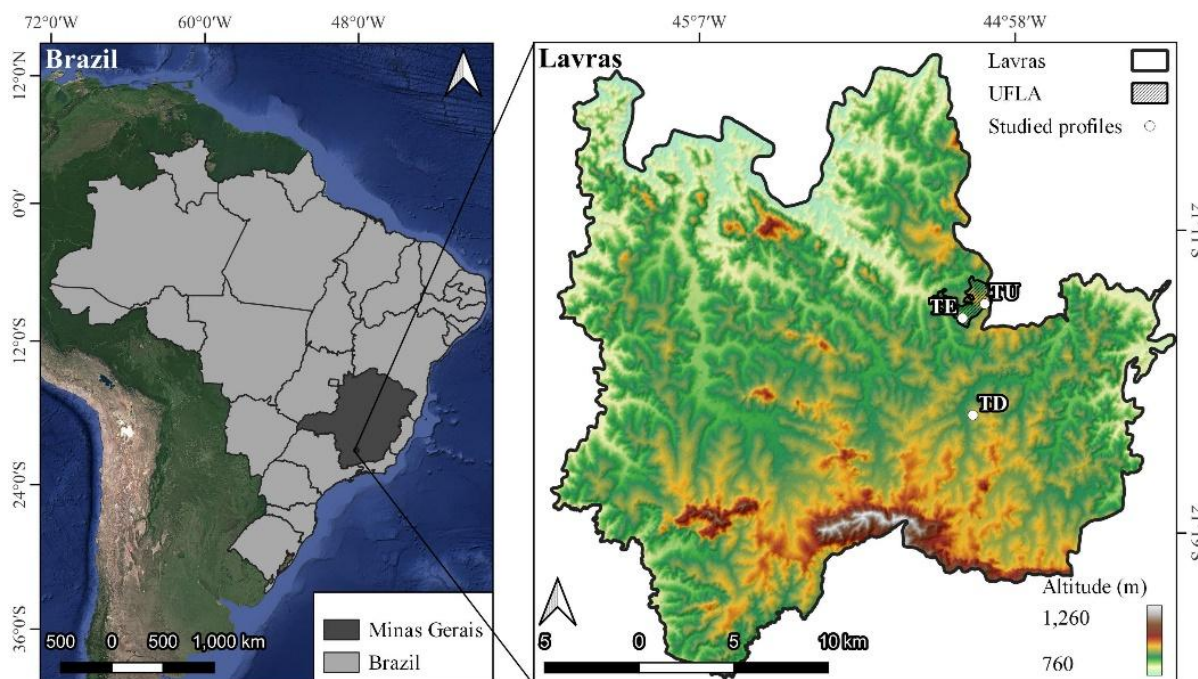


Figure. 1. Study area, localization of the studied soil profiles and elevation map of the municipality of Lavras, situated in the State of Minas Gerais, Brazil. Digital elevation model from Alos Palsar, with spatial resolution of 12.5 m (ASF DAAC, 2015). TU – Typic Udifolists, TD – Typic Dystrustepts, TE – Typic Endoaquents, UFLA – Federal University of Lavras.

2.2. Soil profile classification

Three soil profiles were studied and classified according to the Soil Taxonomy as: Typic Udifolists (TU), Typic Dystrustepts (TD) and Typic Endoaquents (TE) (Table 1; Figure 2) (Soil Survey Staff, 2014). Classifications according to the World Reference Base for Soil Resources (WRB) (FAO, 2014) and the Brazilian Soil Classification System (SiBCS) (Santos et al., 2018) are also presented in Table 1. Their position in the landscape and neighboring upslope soils are shown in Figure 2. Correlations with other classification systems, parent material and topographic information can be found in Table 1.

The three soils have buried horizons, i.e., horizons of different age deposited above a previously developed soil. Horizons below the deposited material are indicated with the letter “b” (Figure 2). To be classified as a buried soil, the deposited mantle must be of unaltered material (C or Bw horizon) and must be thicker than 50 cm (USDA, 2010). If classified as a buried soil, the classification is based in the material deposited above; otherwise, only horizons below the deposited material are considered in the classification. Accordingly, the only soil considered buried is TD, hence classified based on the horizons above Ab (Figure 2). The other two soils have mantles thinner than 50 cm and thus were classified according to the horizons below the deposited mantle.

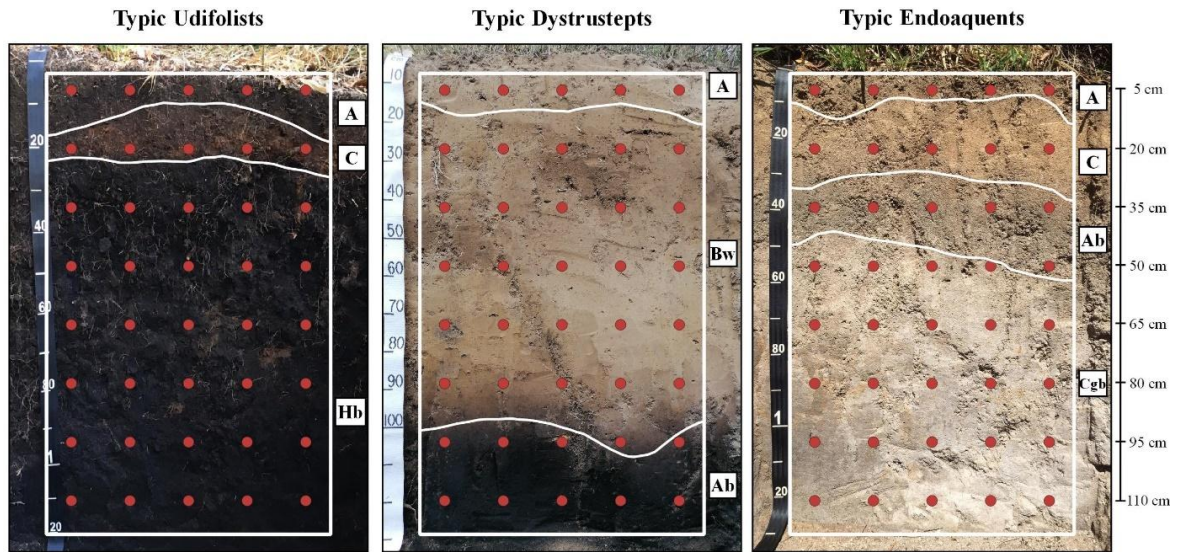


Figure 2. Studied soil profiles, horizon delineation and sampling grid design (15 x 15 cm), in Lavras, Minas Gerais, Brazil. Letters indicate horizon designations and white lines indicate horizon boundaries.

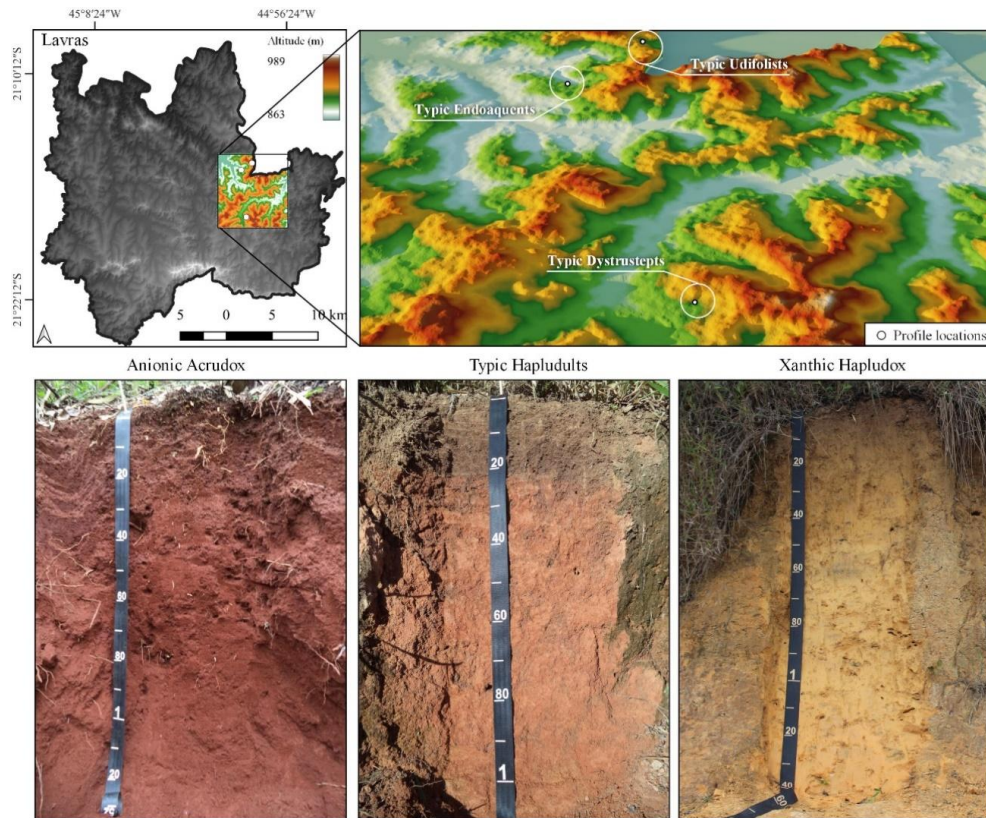


Figure 3. Three-dimensional diagram showing the position of studied soils (top) and examples of neighboring upslope soils derived from gabbro, gneiss and quartzite (bottom, from left to right), in Lavras, Minas Gerais, Brazil. Digital elevation model from ALOS PALSAR, with pixel resolution of 12.5 m (ASF DAAC, 2015).

2.3. Sampling and laboratory analyses

A sampling grid of 8 rows and 5 columns (15 x 15 cm) was utilized to collect the samples (Figure 2). A total of 120 samples were collected, 40 per profile. The first row of samples was collected at a depth of 5 cm.

Table 1. Topographic data, soil classification and parent material of three soil profiles with buried horizons in Lavras, Minas Gerais, Brazil. Soil classifications provided according to the Soil Taxonomy (Soil Survey Staff, 2014), the World Reference Base for Soil Resources (WRB) (FAO, 2014) and the Brazilian Soil Classification System (SiBCS) (Santos et al., 2018).

Soil classification			Altitude (m)	Slope (%)	Parent material
Soil Taxonomy	WRB	SiBCS			
Typic Udifolists (TU)	Folic Histosol	Organossolo Háplico	931	5.1	Sediments (surrounding upslope soils derived from gabbro and gneiss)
Typic Dystrustepts (TD)	Dystric Cambiso 1	Cambissolo Háplico	932	11.3	Quartzite + sediments
Typic Endoaquents (TE)	Dystric Gleysol	Gleissolo Háplico	876	17.6	Gneiss + sediments

Samples were air-dried and sieved at 2 mm (air-dried fine earth – ADFE). The following chemical analyses were performed: exchangeable Ca^{2+} , Mg^{2+} and Al^{3+} extracted with KCl (1 mol L^{-1}) (McLean et al., 1958); available K and P extracted with Mehlich-1 (Mehlich, 1984); potential acidity ($\text{H}^+ + \text{Al}^{3+}$) extracted via calcium acetate and titration with NaOH (Vettori, 1969); SOM by oxidation using potassium dichromate in sulfuric solution (Walkley and Black, 1934). Texture was obtained by the pipette method (Gee and Bauder, 1986). Texture was not analyzed in Histosol's horizons below 20 cm due to the high organic matter content, as the strong aggregation of soil particles by organic compounds could lead to inaccurate results. Using the sum of bases (SB) ($\text{K}^+ + \text{Ca}^{2+} + \text{Mg}^{2+}$), Al saturation (m) and base saturation (V) were calculated as follows: $m = \text{Al}^{3+} / \text{SB} + \text{Al}^{3+}$; $V = 100 \times \text{SB} / (\text{SB} + \text{H}^+ + \text{Al}^{3+})$.

2.4. Sensor analyses

2.4.1. Portable X-ray fluorescence spectrometer

Samples were analyzed by a Delta Professional Olympus (MA, US). The equipment has a silver anode and an X-ray tube of 4 W. The samples were analyzed in triplicate using the “soil” mode; 60 s each sample. The mean of the triplicates was used.

The equipment was calibrated using four reference samples certified by the National Institute of Standards and Technology (NIST): SRM2710a, containing high content of trace-elements; SRM711a, containing moderate content of trace elements; a metallic alloy 316; and a sample of pure SiO_2 .

The detection capability of PXRF ranges from Mg to U in the Periodic Table. From those, the elements Al, Ca, Fe, Mg, Si, Ti and Zr were selected for this study, as they are commonly used to compare soil pedogenic processes and weathering (Stockmann et al., 2016).

Al, Fe and Si were chosen for their ubiquitous presence in soil mineralogy; Ca and Mg for their relevance in fertility attributes; and Ti and Zr for their role as parent material fingerprints due to their high stability in soils (Schaetzl and Anderson, 2005).

2.4.2. Visible and near-infrared diffuse reflectance spectroscopy

Samples were homogeneously spread in petri dishes. A Fieldspec 3 (Analytical Spectral Devices, ASD, Boulder, CO) spectroradiometer was utilized. The equipment detects the electromagnetic spectrum ranges between 350 nm and 2500 nm, with resolution of 1 nm between 350 and 700 nm; 3 nm between 700 and 1400 nm; and 10 nm between 1400 and 2500 nm. All ranges were interpolated to 1 nm resolution, resulting in 2,151 data points.

The sensor was positioned at 8 cm of distance from the samples. Samples were illuminated by two halogen light sources positioned at a distance of 35 cm (non-collimated beams; zenithal angle of 30°), forming an angle of 90° between each other. Every 20 min a white Spectralon plate was utilized to calibrate the equipment. Samples were analyzed in duplicates; the sensor was rotated by 180° each repetition. The mean of the duplicates was used. Each individual sample analysis consisted of 100 scans during a period of 10 s.

The resulting spectra were pre-processed using the continuum removal technique in order to visualize the data (Curcio et al., 2013). The method consists in creating a hull surrounding each spectrum that serves as a reference and receives value 1. The variations in the spectrum are then analyzed in relation to the new reference.

2.4.3. Mid-infrared spectroscopy

The methodology followed Mendes et al. (2022). Sample preparation was done by milling ADFE samples and sieving them to 0.149 mm. Next, samples were analyzed using a Fourier Transform Infrared (FT-IR) Alpha spectroradiometer (Bruker Optics Corporation, Billerica, MA 01821, USA), equipped with a DRIFT accessory. The sensor's light source is a HeNe laser. A KBr beam splitter permits the capture of large radiation amplitudes. The scanning process comprised 32 readings per second per spectrum, with a spectral resolution of 1 cm⁻¹. Calibration was performed using a gold plate.

2.5. Horizon difference index

To evaluate the capability of sensor data to identify horizon discontinuities, principal component analysis (PCA) was used. PCA is a linear transformation that reduces the redundancy of features. The transformed sets of scores (principal components, PC) represent

data variability in fewer values, avoiding correlated features from being repeatedly considered. Following Weindorf et al. (2012; 2015), an index was calculated using PC scores to express differences between horizons (Eq. 1).

$$HDI_n = \sqrt{\sum_{i=1}^F (L_{i(n-1)} - L_{in})^2} \quad (1)$$

Where HDIn stands for the horizon difference index calculated in horizon n; F is the number of PCs used; Li values are the ith PC scores of the horizon above (n-1) and the current horizon (n).

In all datasets, the accumulated variance explained of the first four PCs was close to 100%. For this reason, four PCs were used for all datasets (F in Eq. 1). Data was standardized prior to PCA using the z-score method (difference from the mean divided by the standard deviation).

In the PCAs, every detected PXRF element with non-zero variance was used. Utilized laboratory included K^+ , Ca^{2+} , Al^{3+} , Mg^{2+} , Al^{3+} , base and aluminum saturation, and soil organic matter content. Vis-NIR and MIR PCAs included all available spectra.

2.6. X-ray diffractometry

One sample from the deposited mantle and one from a buried horizon was submitted to X-ray diffractometry (XRD) analysis for each studied profile. A and Hb horizons were sampled for the Histosol; Bw and Ab for the Inceptisol; and A and Cgb for Aquent (Figure 2).

A Bruker D2 PHASER diffractometer was used, containing a LYNXEYETM fast linear module. The operation was done by the software DIFFRAC SUITETM. Cuk α radiation ($\lambda = 1.541838$) and a Ni filter were used. Operations were made under 30 kVA with a current of 10 mA. Sample holders with 0.6 mm bearing non-oriented powder were prepared by grinding each sample with an agate mortar. Next, samples were irradiated over a 7-50°2 θ range at steps of 0.02 θ per second. Sample preparation was done by front-load (top-load). Patterns were interpreted following Brindley and Brown (1980).

2.7. Statistical and spatial analyses

The manipulation and visualization of data were performed using the R environment and the software QGIS and SAGA GIS (Conrad et al., 2015; QGIS, 2022; R Core Team, 2022). The utilized R packages were: ‘tidyverse’ (Wickham et al., 2019), ‘ggpubr’ (Kassambara, 2020), ‘prospectr’ (Stevens and Ramirez-Lopez, 2022), ‘scales’ (Wickham and Seidel, 2020) e ‘RColorBrewer’ (Neuwirth, 2014). 3D modelling was done via the Blender software (BDT, 2022).

The PXRF analyses results were spatialized across the studied profiles to obtain the two-dimensional distribution of each element. The Multilevel B-Spline algorithm was utilized for that end (Conrad et al., 2015).

The visualization of depth functions of the elemental contents obtained via PXRF, Vis-NIR and laboratory data was done by calculating the mean of each row of sampling points. Hence, the results in each depth (5, 20, 35, ..., 110 cm) consist in the mean of 5 points (Figure 2).

3. Results

3.1. Physical and chemical characterization

Results of chemical analyses for TU showed a clear change between 5-35 cm and the rest of the profile (Figure 4). In the deposited material (A and C horizons), contents of exchangeable/available K^+ , Ca^{2+} and Mg^{2+} were high. The same could be observed for V (Figure 4). SOM, exchangeable Al^{3+} and m showed the opposite trend: their contents were lower in A and C horizons, but overall higher in Hb. In TU, texture analysis presented high content of clay in the C horizon (Figure 5), indicating that the source of the material lies in clay-rich soils, which are common in the neighboring upslope areas (Figure 3) (Mancini et al., 2019b; Curi et al., 1990).

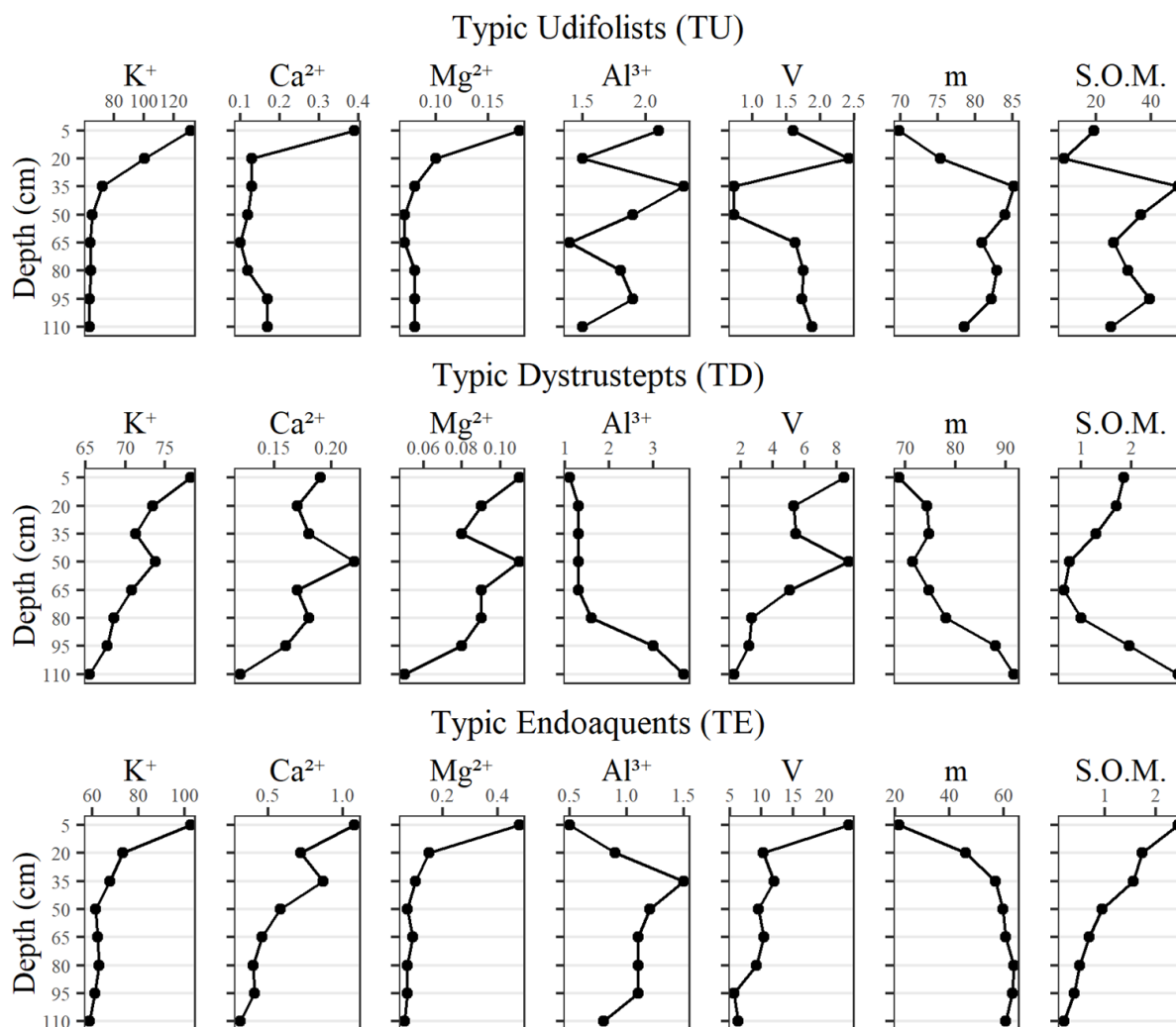


Figure 4. Depth functions of chemical attributes of three soil profiles with buried horizons situated in Lavras, Minas Gerais, Brazil. K^+ in mg dm⁻³; Ca^{2+} , Mg^{2+} and Al^{3+} in cmolc dm⁻³. V – base saturation (%); m – aluminum saturation (%); SOM – soil organic matter (%).

The TD profile also had higher base content (K^+ , Ca^{2+} and Mg^{2+}) closer to the surface (Figure 4), with content peaks at 50 cm. Below 65 cm, there was a sharp decrease of Al^{3+} content, coinciding with the Ab horizon. SOM increased with depth and was responsible for the darker color of the Ab horizon (Figure 2). Silt and clay contents increased below 65 cm (Figure 5), a sign of parent material discontinuity.

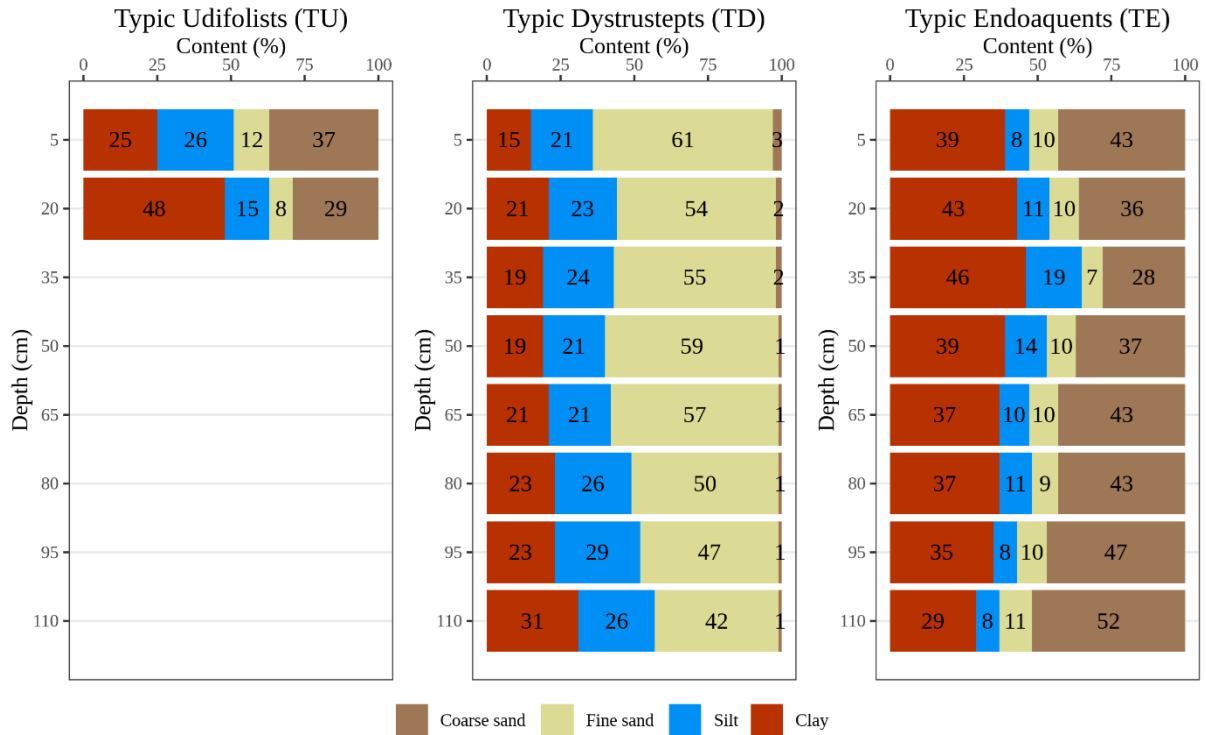


Figure 5. Depth functions of texture analysis of three soil profiles with buried horizons situated in Lavras, Minas Gerais, Brazil.

High base content was observed in the surface of TE as well (Figure 4). SOM was higher in the surface and, after a slight decrease, it remained constant in C and Ab horizons (20-35 cm), but sharply decreased in Cgb. The content of exchangeable Al^{3+} abruptly increased in the Ab horizon (35 cm), coinciding with the boundary between the deposited material (A and C horizons) and the buried horizons (Ab and Cgb horizons). Clay content increased between 5 and 35 cm and gradually decreased below (Figure 5), expressing the differences between the deposited and buried horizons.

3.2. Visible and near-infrared spectra

Table 2 summarizes the correspondence between absorption bands and soil constituents in the Vis-NIR spectra. The TU profile presented very distinct absorption features between 400 and 700 nm (Figure 6). The wide concavity in this range below 35 cm that continues until ~1400 nm is related to organic compounds (Thomas et al., 2022). Absorption valleys close to 480 nm in 5 and 20 cm are related to goethite (Fang et al., 2018). The shoulder in 650 nm, observed only at 20 cm, is associated with hematite (Rossel and Behrens, 2010), corroborating the red color of the C horizon (Figure 2). Peaks related to water close to 1900 nm were more accentuated in 35 cm. Features related to kaolinite and gibbsite (~2200 and 2265 nm,

respectively) (Benedet et al., 2022; Terra et al., 2018) were more accentuated in 20 cm of depth (Figure 6).

Table 2. Relationships between Vis-NIR spectral region and soil constituents found in the literature.

Wavelength (nm)	Soil constituent	References
400-700	Organic matter	Thomas et al. (2010)
480	Goethite	Fang et al. (2018); Vohland et al. (2014)
650	Hematite	Fang et al. (2018); Viscarra Rossel and Behrens (2010)
~ 1400	O-H (1:1 or 2:1 minerals)	Coblinski et al. (2022); Fang et al. (2018)
~1900	H ₂ O, 2:1 minerals	Coblinski et al. (2022); Fang et al. (2018)
~2200	Al-OH (Kaolinite)	Benedet et al. (2022); Fang et al. (2018)
~2265	Gibbsite	Madeira Netto (1996); Terra et al. (2018)
>2300	Fe-OH, Mg-OH (Illite and Chlorite)	Coblinski et al. (2022); Fang et al. (2018)

In TD, common Fe oxide features could be clearly distinguished until 65 cm (Figure 6). Below 65 cm, goethite and hematite absorption valleys (~480 and 650 nm, respectively) were no longer distinguishable. As depth increased, SOM attenuated the features of other minerals and increased absorption in the visible region, resulting in wide concavities (~400-700 nm) similar to those observed in Hb of the TU profile (Figure 6). Absorption features related to water increased with depth, reaching their lowest reflectance values in the Ab horizon.

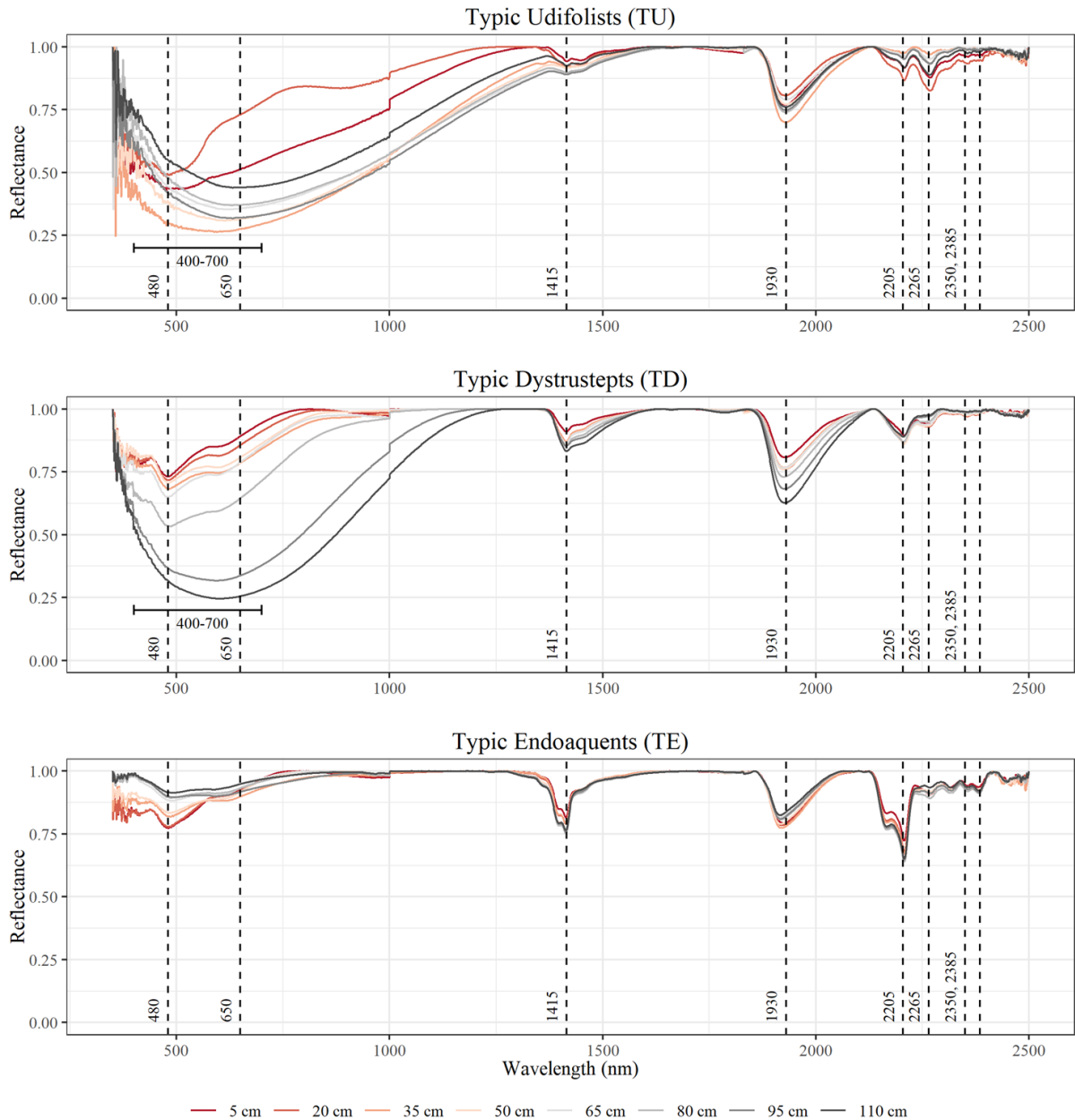


Figure 6. Continuum removal of the visible and near-infrared spectra for three soils with buried horizons in Lavras, Minas Gerais, Brazil.

Contrasting with TU and TD, TE spectra did not show distinct features of SOM and Fe oxides in the visible range (380-750 nm) (Figure 6). Comparatively, absorption valleys close to 1400 and above 2200 nm –associated with 2:1 clay minerals and water– were more accentuated, indicating different mineralogy. Vis-NIR spectra showed clear differences between deposited material (A and C horizons) and buried horizons (Ab and Cgb) (Figs. 2 and 5). Until 50 cm clear goethite features could be observed (~480 nm), but hematite absorption features are not evident (~650 nm), which was expected due to reduction conditions prevalent in this type of soil (Schaetzl and Anderson, 2005; Resende et al., 2014).

3.3. Mid-infrared spectra

Table 3 presents the relationships between MIR bands and soil constituents found in the literature. The MIR spectra of TU presented very high absorption across all bands (reflectance did not surpass 0.02) (Figure 7). Absorption features related to phyllosilicates (3694 and 3622 cm^{-1}) were more accentuated at 20 cm, followed by depths 5 and 110 cm (Figure 7). A peak related to Fe oxides close to 1050 cm^{-1} was more accentuated in 20 cm, followed by 5 cm (C and A horizons, respectively). The morphology of the spectra between 2900 and 2000 cm^{-1} is similar to that observed by Mendes et al. (2022) for soils rich in clay and SOM in Brazil. There were very few evidences of quartz regarding absorption features, peaks and intensities.

Table 3. Relationships between MIR bands and soil constituents found in the literature.

Wavenumber (cm^{-1})	Soil constituent	Feature type	References
800	Quartz	Peak	Mendes et al. (2022)
1050	Crystalline iron	Peak	Mendes et al. (2022)
1159	Polysaccharides, nucleic acids, proteins	Valley	Matamala et al. (2019); Vohland et al. (2014)
1220	Crystalline iron (clayey soils), quartz (sandy soils)	Peak	Mendes et al. (2022)
1527	Quartz	Valley	Mendes et al. (2022)
1788	Kaolinite	Valley	Matamala et al. (2019); Mendes et al. (2022)
1870	Kaolinite	Valley	Matamala et al. (2019); Mendes et al. (2022)
1993	Quartz	Valley	Matamala et al. (2019); Mendes et al. (2022)
2233	Quartz	Valley	Mendes et al. (2022)
2852, 2934	Aliphatic methyl and methylene	Valley	Matamala et al. (2019); Vohland et al. (2014)
3529	2:1 or 1:1 phyllosilicates, gibbsite	Valley	Mendes et al. (2022)
3622, 3694	(OH)-Al of 2:1 phyllosilicates, 1:1 phyllosilicates	Valley	Matamala et al. (2019); Mendes et al. (2022)

The TD profile showed clear features associated with quartz (Figure 7). The double peaks close to 800 cm^{-1} and the valleys at 1527, 1870, 1993 and 2233 cm^{-1} are examples of quartz-related features, which were more accentuated in TD compared to the other two profiles. The overall reflectance decreased with depth, following the increase of SOM (Figure 4), and reached the lowest values in the Ab horizon, similar to Vis-NIR spectra (Figs. 5 and 6).

The reflectance of TE's spectra tended to increase with depth in the first portion of the spectra (3800-2300 cm^{-1}) (Figure 7), following the decrease of SOM (Figure 4). Concomitantly, an increase of quartz was detected with depth in many absorption features and peaks after 2300 cm^{-1} . The absorption feature at 1870, associated with kaolinite, was more accentuated at 110

cm. Features between 3700 and 3500 cm^{-1} , related to 1:1 phyllosilicates (mostly kaolinite in Brazilian soils) and gibbsite were very accentuated for all depths. Absorption features related to SOM were more accentuated in horizons A, C and Ab (5-35 cm), especially in bands 2924 and 2852 cm^{-1} .

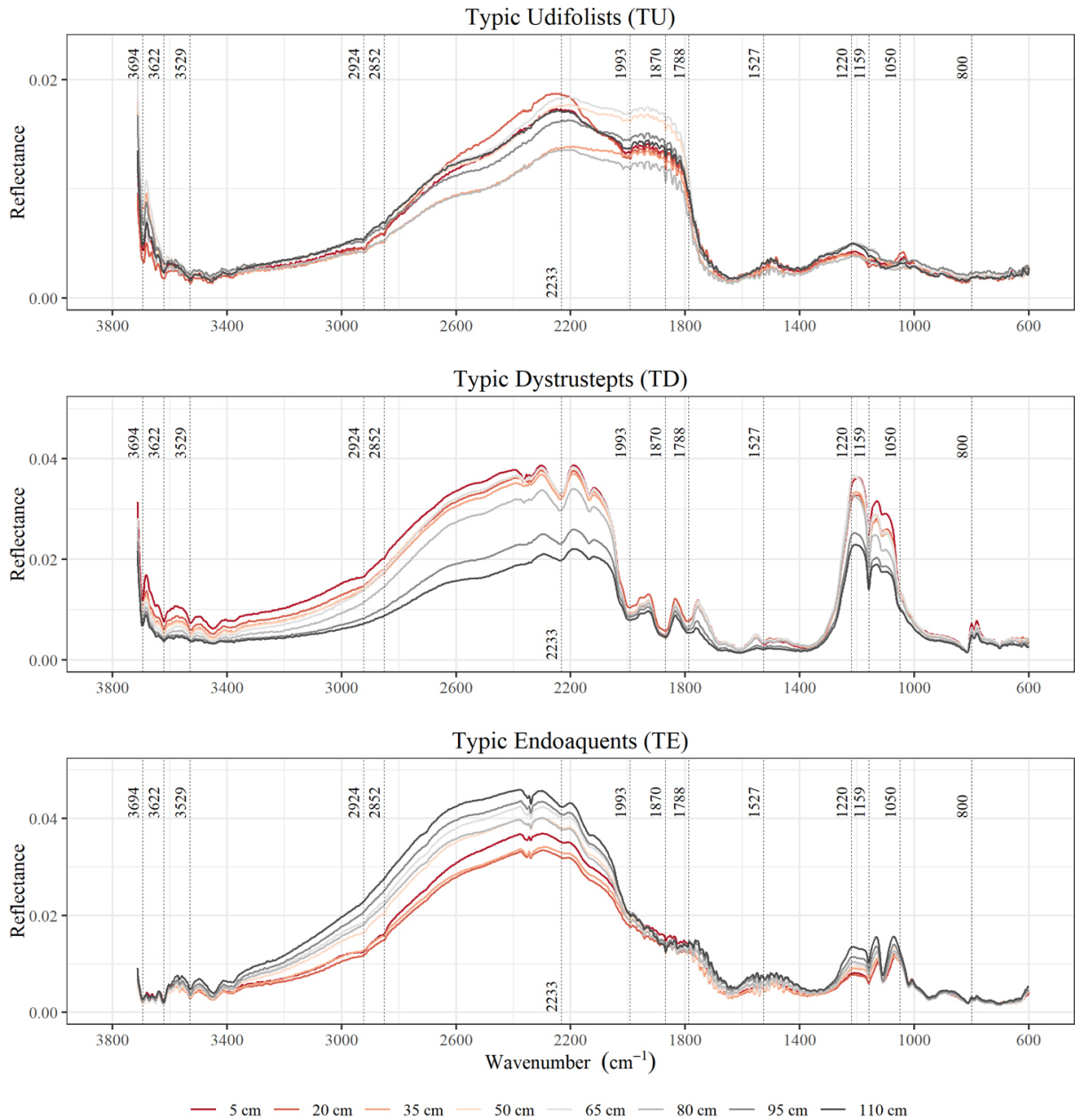


Figure 7. Mid-infrared spectra for three soils with buried horizons in Lavras, Minas Gerais, Brazil.

3.4. Variation of soil elemental contents

The elemental contents for TU were markedly distinct in the deposited material (horizons A and C) (Figs. 8 and 9). A and C horizons showed contrasting high contents of Fe, Ti and Zr, and low contents of Si. The Hb horizon presented high Si content and low Fe content.

Ca content was not detected in most of the profile. At 110 cm, high contents of Al, Ca, Ti and Zr were observed, suggesting parent material discontinuity (Figs. 7 and 8). The Ti/Zr ratio emphasized four discontinuous regions in TU: above 35 cm; at 35 cm; between 35-80 cm; and below 80 cm (Figure 9).

In TD, high contents of Ca, Fe and Si were observed in the Bw horizon (Figure 8). The Ab horizon showed contrasting high contents of Ti and Zr, and lower contents of Ca, Fe and Si compared to Bw (Figs. 8 and 9). A spot could be observed with high content of Mg and low contents of Ca, Fe, Si, Ti and Zr at 80 cm, indicating parent material discontinuity. This was evident in the Ti/Zr depth function, in the peak at 80 cm (Figure 9).

The TE profile had a very erratic distribution of elemental contents (Figure 8). Fe contents were distinctively higher in A, C and Ab horizons. Fe contents decreased exponentially below the Ab horizon (Figs. 8 and 9). The highs and lows of contents in Cgb highlighted the heterogeneity of the sedimentary material forming the gley horizon.

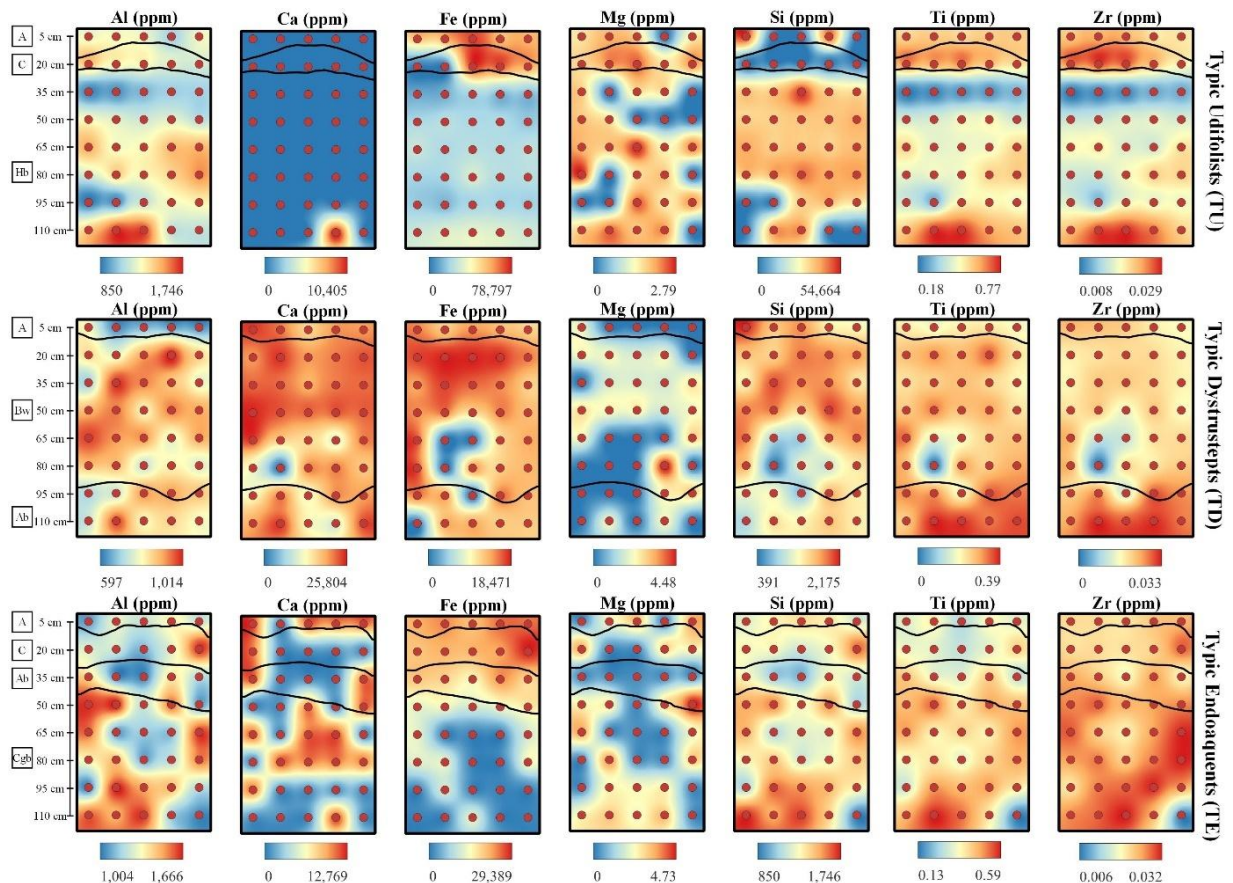


Figure 8. Spatialized total elemental contents provided by PXRF for three soils with buried horizons in Lavras, Minas Gerais, Brazil.

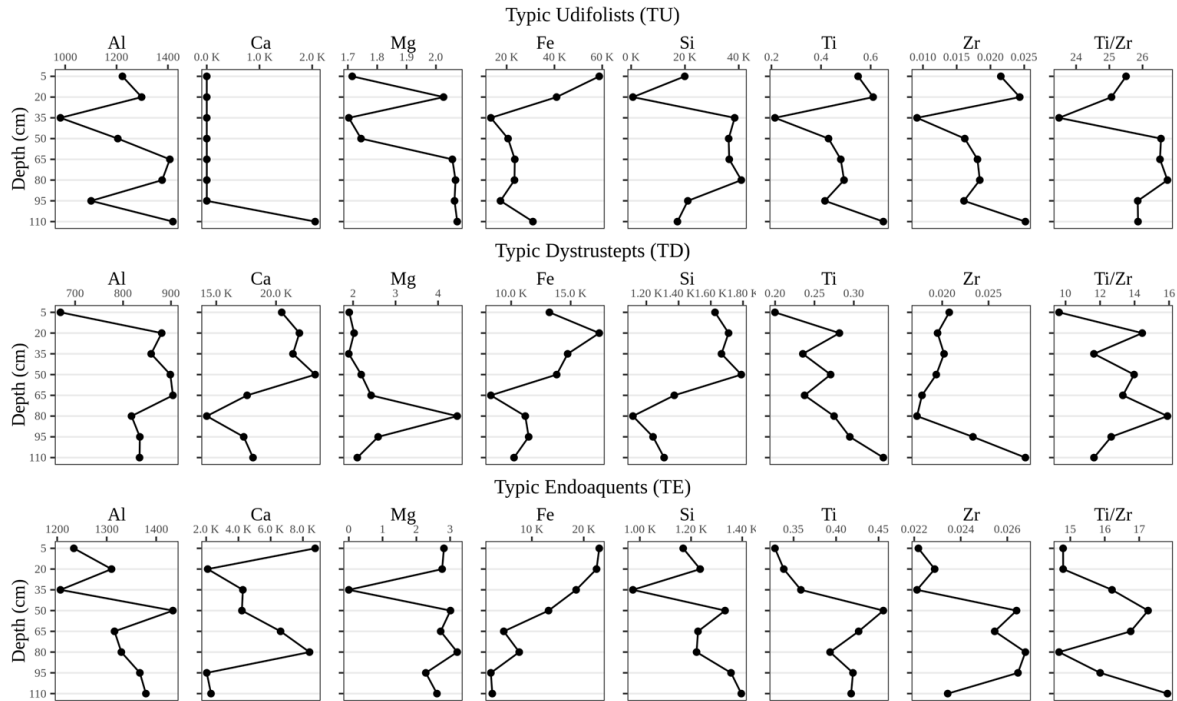


Figure 9. Depth functions of total elemental contents provided by PXRF for three soils with buried horizons in Lavras, Minas Gerais, Brazil. Values in ppm. K = 1,000.

3.5. Relative differences between horizons

Higher HDI values indicate changes in soil attributes compared to the overlying horizon. Laboratory data and sensors were capable of identifying the abrupt change between A and C horizons of TU (Figure 10). HDI peaks were observed at 20 and 35 cm (C and Hb horizons). Spectral data (Vis-NIR and MIR) identified a transition at 95 cm. A similar change was identified by PXRF, but at 110 cm. This possibly reflects differences in soil organic matter content and/or composition (Figure 4).

In TD, every dataset captured the transition between A-Bw and Bw-Ab (Figure 10). Laboratory and spectral data indicated another transition at 65 cm, not identified in the field. PXRF identified this transition at 80 cm. It might correspond to the spots with high Mg contents and low contents of Ca, Fe, Si, Ti and Zr at 80 cm observed in the elemental content spatialization (Figure 8).

Peak HDI values in TE were observed at 20 and 50 cm in all datasets, which corresponded to the transitions A-C and Ab-Cgb. High HDI values were also observed at 35 cm, indicating the transition between C and Ab horizons (Figs. 2, 8 and 10). Laboratory and PXRF data also highlighted an abrupt transition at 110 cm, possibly indicating a horizon not identified in the field.

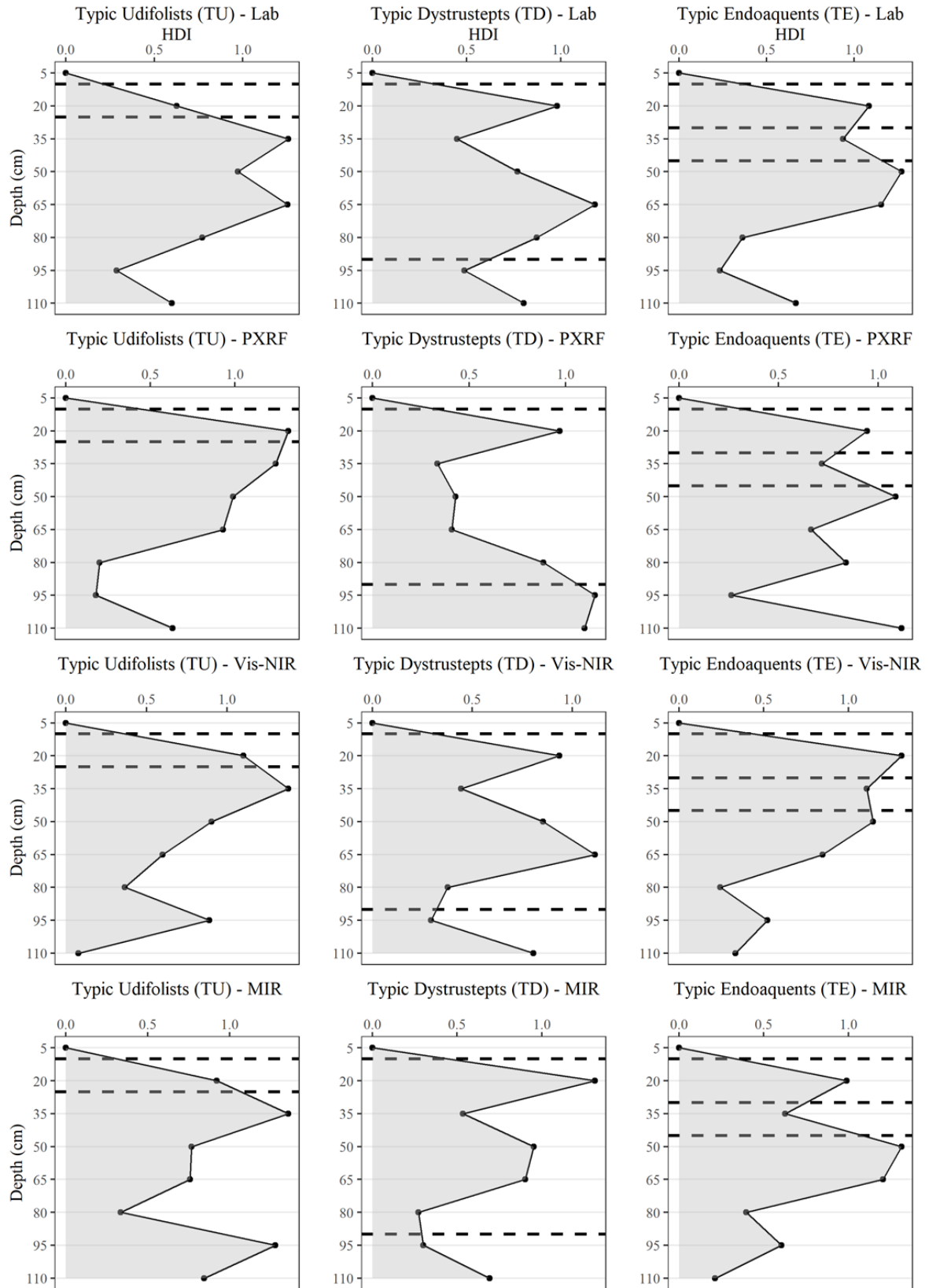


Figure 10. Depth functions of the horizon difference index (HDI) calculated using laboratory and proximal sensor data, according to Eq. 1. High HDI values mean more differences compared to the horizon above. Dashed lines indicate the approximate depth of horizons observed in the field (Figs. 2 and 8).

3.6. Soil mineralogy

The sand fraction of TU was mostly composed of quartz, the predominant mineral in this fraction in Brazilian soils (Figure 11) (Kämpf et al., 2012). Across all depths and fractions, the predominant minerals were gibbsite and kaolinite. Gibbsite, a mineral at the end of the Goldich weathering series, is very present in the clay fraction and indicates highly weathered material (Rezende et al., 2022). Gibbsite is commonly found in upslope soils surrounding TU (Anionic Acrudox in Figure 3) (Curi et al., 1990), indicating they could be the origin of the material deposited above TU.

Quartz was the predominant mineral in TD, especially in the sand and silt fractions (Figure 11). This is in accordance with its parent material, quartzite (Table 1), a rock mainly composed of quartz. The greater presence of 2:1 silicates evidenced the less weathered status of this profile. More accentuated peaks of mica, gibbsite and kaolinite in the silt fraction of the A horizon (Figure 11) showed a slight difference in the mineralogical composition of A and Ab horizons. Oxisols derived from quartzite surrounding TD (Xanthic Hapludox in Figure 3) have been reported as containing these minerals (Silva et al., 2019).

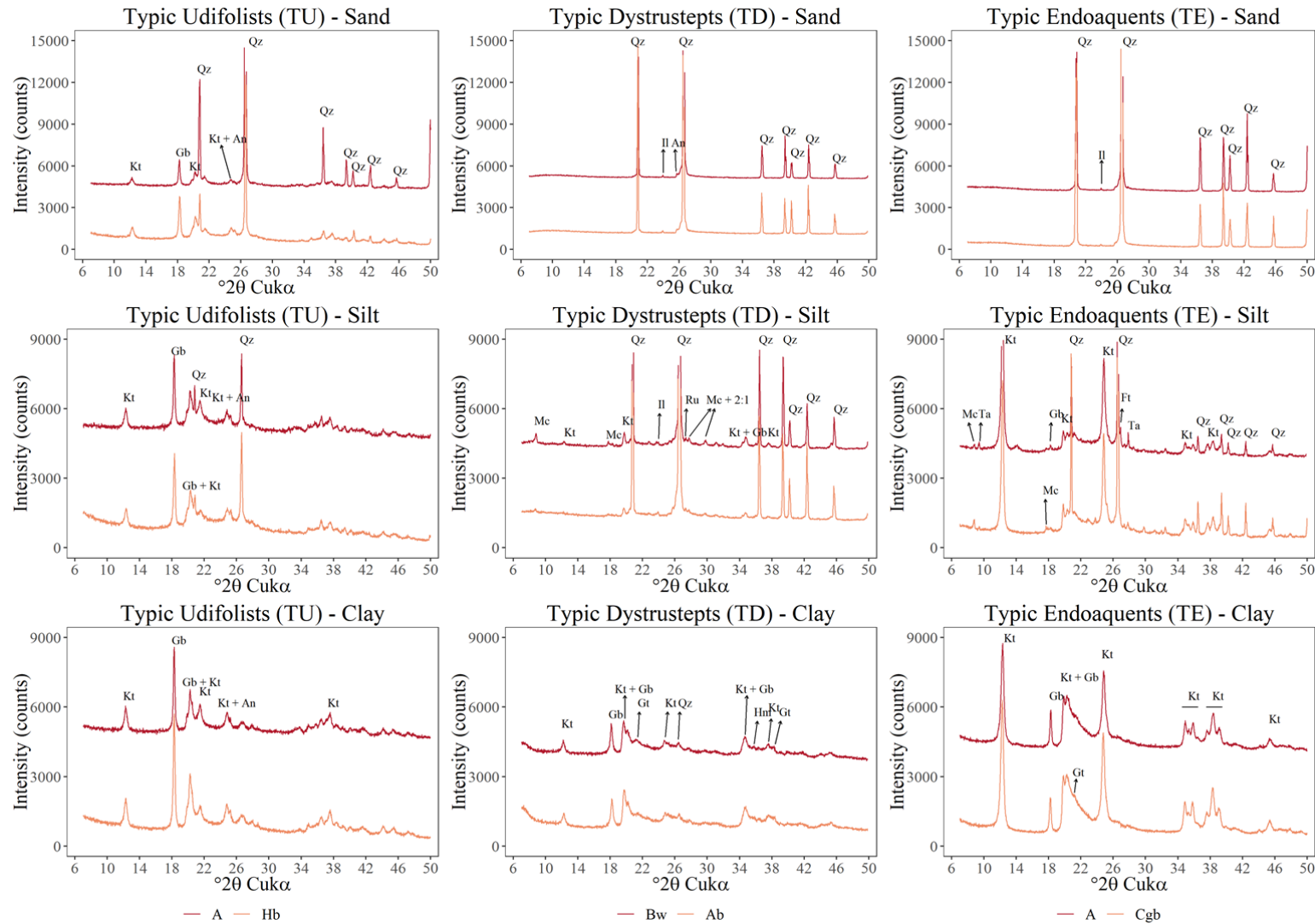


Figure 11. X-ray diffractograms for three soils with buried horizons in Lavras, Minas Gerais, Brazil. 2:1 – unidentified phyllosilicates; An – anatase; Ft – feldspar; Gb – gibbsite; Gt – goethite; Hm – hematite; Il – Ilmenite; Kt – kaolinite; Mc – mica; Qz – quartz; Ru – rutile; Ta – talc.

The TE profile presented high kaolinite content in the silt and clay fractions and high gibbsite content in the clay fraction (Figure 11). The silt fractions highlighted mineralogical differences between the A and Cgb horizons. In the Cgb horizon, higher peaks of mica were detected in the silt fraction. In the A horizon, higher peaks of talc were observed, a rare mineral in Brazilian soils due to its instability (Kämpf et al., 2012). Peaks of feldspar were observed in A horizon, but not in Cgb.

Overall, XRD data from the deposited material was not significantly different than that of the buried horizons. In each studied soil there were slight differences between buried and deposited horizons, but the key mineralogy was the same (Figure 11), as commonly found in most Brazilian soils.

4. Discussion

4.1. Typic Udifolists (TU)

Soils surrounding TU are mainly Oxisols and Ultisols (Figure 3). Mancini et al. (2019b) reported the presence of soils derived from gabbro and gneiss in the area surrounding the profile location. Accordingly, Vis-NIR and PXRF detected the Fe-rich nature of A and C horizons (Figures 6 and 8), confirming that the sediments composing TU came from oxide-rich soils in the area. A detailed analysis of Oxisols derived from gabbro and gneiss in the study area can be found in Benedet et al. (2022).

A discontinuity was observed in the Hb horizon at 110 cm, more visible in PXRF results (Figures 8 and 10), where contents of Al, Fe, Ti and Zr were higher; and contents of Si, lower. Similarly, Vis-NIR data captured this discontinuity: the absorption feature between 400 and 700 nm in the spectrum of 110 cm is less accentuated than others from the Hb horizon (Figure 6). This change was also shown by the HDI, but starting at 95 cm (Figure 10). In this range (400-700 nm), the morphology of the 110 cm spectrum was more similar to that of 5 and 20 cm spectra (A and C horizons) than to others of the Hb horizon (50-95 cm). This trend could be observed in MIR data in band 3694 cm^{-1} , related to silicate minerals (Table 3). Absorption in this band was higher for depths 5, 20 and 110 cm compared to other depths. Due to the immobility of Ti and Zr in soils (Zhang and Hartemink, 2019), the increase of elemental contents with depth is unlikely to be caused by translocation. Therefore, this is evidence that the sediments deposited therein in the past had a different chemical composition.

The Si contents detected in Hb were very high compared to other horizons and profiles (Figures 8 and 9). Usually, in highly weathered-leached soils, such as the neighboring Oxisols and Ultisols (Figure 3), most Si has been lost in pedogenesis (Rezende et al., 2022). High

contents of Si in Hb are an indication that past climatic conditions that preserved SOM in Hb (Ibrahim and Lal, 2014) have also hindered the leaching of Si and promoted its accumulation.

The uptake of Si by plant tissues can lead to its accumulation in the form of amorphous Si in environments of low weathering intensity (Alfredsson et al., 2015). Accordingly, high contents of Si in the A horizon are due to recent deposition of organic matter (Struyf et al., 2010) and biological recycling (uptake of Si by plants from underlying horizons) (Grau et al., 2017; Mancini et al., 2021).

Gibbsite and kaolinite contents were higher in Hb compared to A horizon (Figure 11). This is in accordance with the increase of Al content detected by PXRF (Figure 9), and agrees with the above-mentioned chemical differences perceived in the sediments that compose TU. Both of these minerals are at the end of the Goldich weathering series and demonstrate the advanced weathering status of the material that formed this sedimentary soil (Goldich, 1938; Rezende et al., 2022). This suggests that the high contents of Si in Hb are allochthonous, as Si is more prone to leaching than Al (gibbsite). It reinforces the hypothesis that Si was deposited, accumulated and preserved by past anoxic conditions.

4.2. Typic Dystrustepts (TD)

Because of its quartz-rich parent material (quartzite), TD showed high contents of quartz-rich sand (Figures 5 and 11). Soils with more sand (especially composed of quartz) are well-drained and have low cation exchange capacity (larger particles and higher macroporosity) (Bockheim et al., 2020). Hence, bases deposited in the A horizon, due to their mobile nature, were expected to be easily leached. Instead, their contents increased at around 50 cm (Figure 4). The increase of silt and clay contents below 65 cm could explain the peaks of base contents at 50 cm (Figures 3 and 4): as the soil becomes less sandy, the translocation of mobile elements becomes slower, leading to accumulation.

Following changes in texture towards the Ab horizon, elemental contents of Ti and Zr increased below 95 cm, signaling parent material changes. In the XRD diffractograms of the silt fraction, higher peaks of mica and kaolinite in the Bw horizon corroborated the discontinuities detected by PXRF. Total contents of Al, constituent of mica and kaolinite, decreased in the Ab horizon (Figures 8 and 9).

The reflectance values of quartz are high, as it is usually a light-colored mineral (Demattê et al., 2019). Organic matter has the opposite effect: it reduces reflectance, especially in the visible range (380-750 nm), due to the dark color of organic carbon (Benedet et al., 2022; Terra et al., 2018). Vis-NIR and MIR data clearly showed this trend, as overall reflectance

decreased with depth, towards the Ab horizon (Figures 5 and 6). Hutengs et al. (2019) reported that features in the 2000-1790 cm^{-1} range are related to the presence of quartz and correlate negatively with SOM. In fact, TD is the only profile that showed distinct absorption features in this range. The TU profile, contrarily, presented high reflectance in this region of the MIR spectrum.

Within the visible region, absorption features between 400-700 nm in the Vis-NIR spectra are related to chromophores (Yang et al., 2019). These compounds confer color to materials and indicate the presence of SOM (Thomas et al., 2022) and Fe oxyhydroxides. The later are mainly hematite and goethite, and are usually responsible for the red and yellow hues in soils, respectively (Mirabella and Carnicelli, 1992; Rezende et al., 2022).

In the Vis-NIR range, absorption features of hematite and goethite (480 and 650 nm, respectively), minerals detected in both Bw and Ab horizons by XRF analysis (Figure 11), were attenuated as SOM increased with depth. The lowering of albedo (overall reflectance values), which causes this effect, has been observed for a large dataset of soils in Brazil by Mendes et al. (2022) and Demattê et al. (2019).

PXRF data identified a local discontinuity at 80 cm in the parent material. This discontinuity was not visible in the field (Figure 2). It presented low contents of Ca, Fe, Si, Ti and Zr, and higher contents of Mg. This change was also detected by the HDI at depths 65-80 cm (Figure 10). This spot was not assessed by the XRD analysis, so its source is unknown. Still, it shows the potential of PXRF in detecting local variations otherwise imperceptible by usual mechanical and visual methods.

4.3. Typic Endoaquents (TE)

Endoaquents are hydromorphic wetland soils and hence receive a myriad of alluvial and colluvial sediments. The erratic distribution of total elemental contents (Figure 8) reflects the many kinds of materials carried by the river and the erosion of upper parts of the landscape (Stolt et al., 2001; Wilson et al., 2012).

The darker horizon identified as Ab had lower SOM than the A horizon (5 cm), but higher content than Cgb (50 cm and below) (Figure 4). This suggests a recent deposition episode, where an older A horizon was buried by more recent sediments (hence its classification as Ab). The presence of more gibbsite in the A horizon (Figure 11) indicated the contribution of sediments by surrounding Oxisols containing this mineral (Figure 3) (A and C horizons).

The TE profile was likely formed from a mixture of autochthonous gneissic material and allochthonous alluvial/colluvial sediments. The XRD analysis revealed greater presence of

mica, kaolinite and quartz in the Cgb horizon, and peaks of feldspar and talc in the A horizon (Figure 11), highlighting some of the mineralogical differences between both horizons.

The higher contents of kaolinite in Cgb were also shown by Vis-NIR data by very accentuated absorption features in 1,414 and 2,205 nm, associated with phyllosilicates (Table 2). Absorptions in these bands were higher in Cgb (110 cm) and lower in the C horizon (20 cm). The differences observed in deeper parts of Cgb were corroborated by the HDI, which indicated a transition of soil attributes below 95 cm (Figure 10). Stronger absorption peaks for goethite and hematite in A, C and Ab also emphasizes the above-mentioned mineralogical differences between deposited and buried horizons.

The MIR spectra in band 1870 cm^{-1} , related to kaolinite (Table 3), presented more accentuated absorption in depth 110 cm, and less accentuated features in 5 cm (Figure 7). This corroborated the XRD results, which showed greater content of kaolinite in Cgb compared to A horizon (Figure 11). Accentuated features above $3,500\text{ cm}^{-1}$, also related to 1:1 phyllosilicates (kaolinite), was also confirmed by XRD, and showed high presence of kaolinite in horizons A and Cgb.

Contents of Fe decreased sharply in Cgb (Figures 8 and 9). In anoxic conditions, where gley horizons are formed, Fe(III) is reduced to Fe(II), which is soluble and hence gradually removed from soils (Dorau et al., 2016). Contrastingly, PXRF detected high Fe content in A, C and Ab horizons (Figure 8), supporting the hypothesis of recent deposition of sediments from neighboring Oxisols/Ultisols (Figure 3).

5. Conclusions

Key morphological differences between the deposited soils and buried horizons were detected by all sensors (PXRF, Vis-NIR and MIR). Our results indicated an Fe-rich nature of sediments that compose the A and C horizon of TU and TE, which likely originated from Oxisols and Ultisols in the upper part of the landscape.

The PXRF provided a comprehensive way to evaluate chemical discontinuities whereas Vis-NIR and MIR excelled at detecting the presence of specific minerals, confirmed by XRD analysis, presenting a qualitative way to infer the origin of the elemental contents supplied by PXRF. Sensor combination provided an efficient, low-cost diagnostic of studied soils. The qualitative (instead of purely quantitative) interpretation of spectral curves was key to complement chemical analyses.

The PXRF detected discontinuities in the Hb horizon of TU and mapped the distribution of elements related to variability of sediments that form the gley horizon of TE –invisible

features to naked eye that would be hard to detect in the field. Vis-NIR and MIR data were capable of detecting variations in SOM along the TU's Hb horizon and in the Ab horizon of TD.

Attributes indistinguishable in the field could be inferred, such as variations of key elements, minerals and SOM, helping to elucidate the origin of the soil material. Sensor data coupled with wet-chemistry and mineralogical analyses provided a rich set of quantitative and qualitative data of soil properties, aiding the detailed characterization of the studied profiles.

Acknowledgements

The authors would like to thank the National Council for Scientific and Technological Development (CNPq), Coordination for the Improvement of Higher Education Personnel (CAPES), and Minas Gerais State Agency for Research and Development (FAPEMIG) for the financial support to develop this research.

References

- Alfredsson, H., Hugelius, G., Clymans, W., Stadmark, J., Kuhry, P., Conley, D.J., 2015. Amorphous silica pools in permafrost soils of the central Canadian arctic and the potential impact of climate change. *Biogeochemistry* 124, 441–459.
- Alvares, C.A., Stape, J.L., Sentelhas, P.C., Moraes Gonçalves, J.L. de, Sparovek, G., 2013. Köppen's climate classification map for Brazil. *Meteorologische Zeitschrift* 22, 711–728. <http://doi.org/10.1127/0941-2948/2013/0507>
- Andrade, R., Mancini, M., Teixeira, A.F. dos S., Silva, S.H.G., Weindorf, D.C., Chakraborty, S., Guilherme, L.R.G., Curi, N., 2022. Proximal sensor data fusion and auxiliary information for tropical soil property prediction: Soil texture. *Geoderma* 422, 115936. <https://doi.org/10.1016/j.geoderma.2022.115936>
- ASF DAAC, 2015. Alos palsar radiometric terrain corrected high-res (AP 26358 FBS F6750 RT1). NASA, Includes material from JAXA/METI (2007). <https://doi.org/10.5067/JBYK3J6HFSVF>
- Benedet, L., Faria, W.M., Silva, S.H.G., Mancini, M., Demattê, J.A.M., Guilherme, L.R.G., Curi, N., 2020. Soil texture prediction using portable x-ray fluorescence spectrometry and visible near-infrared diffuse reflectance spectroscopy. *Geoderma* 376, 114553. <https://doi.org/10.1016/j.geoderma.2020.114553>
- Benedet, L., Silva, S.H.G., Mancini, M., Teixeira, A.F.S., Inda, A.V., Demattê, J.A.M., Curi, N., 2022. Variation of properties of two contrasting Oxisols enhanced by pXRF and vis-NIR. *Journal of South American Earth Sciences* 115, 103748. <https://doi.org/10.1016/j.jsames.2022.103748>
- Bockheim, J.G., Hartemink, A.E., Huang, J., 2020. Distribution and properties of sandy soils in the conterminous USA – a conceptual thickness model, and taxonomic analysis. *Catena* 195, 104746. <https://doi.org/10.1016/j.catena.2020.104746>

- BDT – Blender Development Team, 2022. Blender: A 3D modelling and rendering package (version 3.2.0). Blender Foundation, Amsterdam. Available at: <http://www.blender.org>.
- Brindley, G.W., Brown, G., 1980. Crystal structures of clay minerals and their X-ray identification. Mineralogical Society of Great Britain; Ireland. <https://doi.org/10.1180/mono-5>
- Chen, Y., Gao, S., Jones, E.J., Singh, B., 2021. Prediction of soil clay content and cation exchange capacity using visible near-infrared spectroscopy, portable X-ray fluorescence, and X-ray diffraction techniques. *Environ. Sci. Technol.* 55, 4629–4637. <https://doi.org/10.1021/acs.est.0c04130>
- Coblinski, J.A., Inda, A.V., Dematte, J.A.M., Dotto, A.C., Gholizadeh, A., Giasson, E., 2021. Identification of minerals in subtropical soils with different textural classes by VIS-NIR-SWIR reflectance spectroscopy. *Catena* 203. <https://doi.org/10.1016/j.catena.2021.105334>
- Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., Böhner, J., 2015. System for automated geoscientific analyses (SAGA) v. 2.1.4. *Geoscientific Model Development* 8, 1991–2007. <https://doi.org/10.5194/gmd-8-1991-2015>
- Corradini, F., Bartholomeus, H., Lwanga, E.H., Gertsen, H., Geissen, V., 2019. Predicting soil microplastic concentration using Vis-NIR spectroscopy. *Science of The Total Environment* 650, 922–932. <https://doi.org/10.1016/j.scitotenv.2018.09.101>
- Curcio, D., Ciralo, G., D’Asaro, F., Minacapilli, M., 2013. Prediction of soil texture distributions using VNIR-SWIR reflectance spectroscopy. *Procedia Environmental Sciences* 19, 494–503. <https://doi.org/10.1016/j.proenv.2013.06.056>
- Curi, N., Silva, H.G.S., Poggere, G.C., Menezes, M.D., 2017. Mapeamento de solos e magnetismo no campus da UFLA como traçadores ambientais. UFLA, Lavras.
- Curi, N., Lima, J.M., Andrade, H., Gualberto, V., 1990. Geomorfologia, física, química e mineralogia dos principais solos da região de Lavras (MG). *Ciência e Prática* 14, 297–307.
- Demattê, J.A.M., Dotto, A.C., Paiva, A.F.S., Sato, M.V., Dalmolin, R.S.D., Socorro B. de Araújo, M. do, Silva, E.B., Nanni, M.R., ten Caten, A., Noronha, N.C., Lacerda, M.P.C., Araújo Filho, J.C., Rizzo, R., Bellinaso, H., Francelino, M.R., Schaefer, C.E.G.R., Vicente, L.E., Santos, U.J., Sá Barretto Sampaio, E.V., 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., Anjos, L.H.C., Ceddia, M.B., Pinheiro, É.F.M., Grunwald, S., Vasques, G.M., Marques Júnior, J., Silva, A.J., Vasconcelos Barreto, M.C. de, Nóbrega, G.N., Silva, M.Z., Souza, S.F., Valladares, G.S., Viana, J.H.M., Terra, F.S., Horák-Terra, I., Fiorio, P.R., Silva, R.C., Frade Júnior, E.F., Lima, R.H.C., Alba, J.M.F., Souza Junior, V.S., Brefin, M.D.L.M.S., Ruivo, M.D.L.P., Ferreira, T.O., Brait, M.A., Caetano, N.R., Bringhenti, I., Mendes, W.S., Safanelli, J.L., Guimarães, C.C.B., Poppiel, R.R., Souza, A.B., Quesada, C.A., Couto, H.T.Z., 2019. The brazilian soil spectral library (BSSL): A general view, application and challenges. *Geoderma* 354, 113793. <https://doi.org/10.1016/j.geoderma.2019.05.043>
- Dorau, K., Eickmeier, M., Mansfeldt, T., 2016. Comparison of manganese and iron oxide-coated redox bars for characterization of the redox status in wetland soils. *Wetlands* 36, 133–141. <https://doi.org/10.1007/s13157-015-0724-5>
- FAO. World reference base for soil resources/International soil classification system for naming soils and creating legends for soil maps, 2014. Food and Agriculture Organization of the United Nations, Rome, 191 p.

- Fang, Q., Hong, H., Zhao, L., Kukolich, S., Yin, K., Wang, C., 2018. Visible and near-infrared reflectance spectroscopy for investigating soil mineralogy: A review. *Journal of Spectroscopy*. <https://doi.org/10.1155/2018/3168974>
- Faria, A.J.G., Silva, S.H.G., Melo, L.C.A., Andrade, R., Mancini, M., Mesquita, L.F., Teixeira, A.F.S., Guilherme, L.R.G., Curi, N., 2020. Soils of the Brazilian coastal plains biome: Prediction of chemical attributes via portable x-ray fluorescence (pXRF) spectrometry and robust prediction models. *Soil Research* 58, 683–695. <https://doi.org/10.1071/SR20136>
- Gee, G.W., Bauder, J.W., 1986. Particle-size analysis, in: *Methods of Soil Analysis*. John Wiley & Sons, Ltd, pp. 383–411. <https://doi.org/10.2136/sssabookser5.1.2ed.c15>
- Goldich, S.S., 1938. A study in rock-weathering. *The Journal of Geology* 46, 17–58. <https://doi.org/10.1086/624619>
- Gozukara, G., Acar, M., Ozlu, E., Dengiz, O., Hartemink, A.E., Zhang, Y., 2022. A soil quality index using Vis-NIR and pXRF spectra of a soil profile. *Catena* 211, 105954. <https://doi.org/10.1016/j.catena.2021.105954>
- Gozukara, G., Zhang, Y., Hartemink, A.E., 2021. Using vis-NIR and pXRF data to distinguish soil parent materials – An example using 136 pedons from Wisconsin, USA. *Geoderma* 396, 115091. <https://doi.org/10.1016/j.geoderma.2021.115091>
- Grau, O., Peñuelas, J., Ferry, B., Freycon, V., Blanc, L., Desprez, M., Baraloto, C., Chave, J., Descroix, L., Dourdain, A., Guitet, S., Janssens, I.A., Sardans, J., Hérault, B., 2017. Nutrient-cycling mechanisms other than the direct absorption from soil may control forest structure and dynamics in poor Amazonian soils. *Scientific reports* 7, 45017. <https://doi.org/10.1038/srep45017>
- Hartemink, A.E., Minasny, B., 2014. Towards digital soil morphometrics. *Geoderma* 230–231, 305–317. <https://doi.org/10.1016/j.geoderma.2014.03.008>
- Hartemink, A.E., Zhang, Y., Bockheim, J.G., Curi, N., Silva, S.H.G., Grauer-Gray, J., Lowe, D.J., Krasilnikov, P., 2020. Chapter three - soil horizon variation: A review, in: Sparks, D.L. (Ed.), *Advances in Agronomy*. pp. 125–185. <https://doi.org/10.1016/bs.agron.2019.10.003>
- Horta, A., Malone, B., Stockmann, U., Minasny, B., Bishop, T.F.A., McBratney, A.B., Pallasser, R., Pozza, L., 2015. Potential of integrated field spectroscopy and spatial analysis for enhanced assessment of soil contamination: A prospective review. *Geoderma* 241–242, 180–209. <https://doi.org/10.1016/j.geoderma.2014.11.024>
- 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>
- Ibrahim, M.A., Lal, R., 2014. Soil carbon and silicon pools across an un-drained toposequence in central Ohio. *Catena* 120, 57–63. <https://doi.org/10.1016/j.catena.2014.04.006>
- Jenny, H., 1941. *Factors of soil formation: A system of quantitative pedology*. McGraw-Hill, New York.
- Kämpf, N., Marques, J.J., Curi, N., 2012. Mineralogia de solos brasileiros, in: Ker, J.C., Curi, N., Schaefer, C.E.G.R., Vidal-Torrado, P. (Eds.), *Pedologia - Fundamentos*. SBCS, Viçosa, pp. 81–146.
- Kandpal, L.M., Munna, M.A., Cruz, C., Mouazen, A.M., 2022. Spectra fusion of mid-infrared (MIR) and x-ray fluorescence (XRF) spectroscopy for estimation of selected soil fertility attributes. *Sensors* 22. <https://doi.org/10.3390/s22093459>

- Kassambara, A., 2020. Ggpubr: 'ggplot2' based publication ready plots. Available in: <https://cran.r-project.org/web/packages/ggpubr/ggpubr.pdf>. Accessed in: June, 2022.
- Li, F., Xu, L., You, T., Lu, A., 2021. Measurement of potentially toxic elements in the soil through NIR, MIR, and XRF spectral data fusion. *Computers and Electronics in Agriculture* 187, 106257. <https://doi.org/10.1016/j.compag.2021.106257>
- Madeira Netto, J.S., 1996. Spectral reflectance properties of soil. *Photo Interpretation* 34, 59–70.
- Mancini, M., Silva, S.H.G., Hartemink, A.E., Zhang, Y., Faria, Á.J.G., Silva, F.M., Inda, A.V., Demattê, J.A.M., Curi, N., 2021. Formation and variation of a 4.5 m deep Oxisol in southeastern Brazil. *Catena* 206, 105492. <https://doi.org/10.1016/j.catena.2021.105492>
- Mancini, M., Weindorf, D.C., Chakraborty, S., Silva, S.H.G., Santos Teixeira, A.F. dos, Guilherme, L.R.G., Curi, N., 2019a. Tracing tropical soil parent material analysis via portable x-ray fluorescence (pXRF) spectrometry in Brazilian Cerrado. *Geoderma* 337, 718–728. <https://doi.org/10.1016/j.geoderma.2018.10.026>
- Mancini, M., Weindorf, D.C., Silva, S.H.G., Chakraborty, S., Teixeira, A.F.S., Guilherme, L.R.G., Curi, N., 2019b. Parent material distribution mapping from tropical soils data via machine learning and portable x-ray fluorescence (pXRF) spectrometry in Brazil. *Geoderma* 354, 113885. <https://doi.org/10.1016/j.geoderma.2019.113885>
- McLean, E.O., Heddleson, M.R., Bartlett, R.J., Holowaychuk, N., 1958. Aluminum in soils: I. Extraction methods and magnitudes in clays and ohio soils. *Soil Science Society of America Journal* 22, 382–387. <https://doi.org/10.2136/sssaj1958.03615995002200050005x>
- Mehlich, A., 1984. Mehlich 3 soil test extractant: A modification of Mehlich 2 extractant. *Communications in Soil Science and Plant Analysis* 15, 1409–1416. <https://doi.org/10.1080/00103628409367568>
- Mehlich, A., 1934. An examination of the degtjareff method for determining soil organic matter and a proposed modification of the chromic acid titration method. *Soil Science* 37, 29–38. <http://doi.org/10.1097/00010694-193401000-00003>
- Mendes, W.S., Demattê, J.A.M., Rosin, N.A., Silva Terra, F. da, Poppiel, R.R., Urbina-Salazar, D.F., Boechat, C.L., Silva, E.B., Curi, N., Silva, S.H.G., Santos, U.J., Valladares, G.S., 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>
- Mirabella, A., Carnicelli, S., 1992. Iron oxide mineralogy in red and brown soils developed on calcareous rocks in central Italy. *Geoderma* 55, 95–109. [https://doi.org/10.1016/0016-7061\(92\)90007-T](https://doi.org/10.1016/0016-7061(92)90007-T)
- Neuwirth, E., 2014. RColorBrewer: ColorBrewer palettes. Available at: <https://cran.r-project.org/web/packages/RColorBrewer/RColorBrewer.pdf>. Accessed in: June, 2022.
- O'Rourke, S.M., Stockmann, U., Holden, N.M., McBratney, a. B., Minasny, B., 2016. An assessment of model averaging to improve predictive power of portable vis-NIR and XRF for the determination of agronomic soil properties. *Geoderma* 279, 31–44. <https://doi.org/10.1016/j.geoderma.2016.05.005>
- QGIS, 2022. QGIS geographic information system. QGIS Association. Available at: <https://qgis.org/>. Accessed in: June, 2022.
- R Core Team, 2022. R, a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

- Resende, M., Curi, N., Rezende, S.B., Corrêa, G.F., Ker, J.C., 2014. *Pedologia: Base para distinção de ambientes*, 6th ed. UFLA, Lavras.
- Rezende, S.B., Franzmeier, D.P., Resende, M., Mancini, M., Curi, N., 2022. Pedogenic processes in a chronosequence of very deeply weathered soils in southeastern Brazil. *Catena* 215, 106362. <https://doi.org/10.1016/j.catena.2022.106362>
- Rossel, R.A.V., Behrens, T., 2010. Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma* 158, 46–54. <https://doi.org/10.1016/j.geoderma.2009.12.025>
- Rouillon, M., Taylor, M.P., 2016. Can field portable x-ray fluorescence (pXRF) produce high quality data for application in environmental contamination research? *Environmental Pollution* 214, 255–264. <https://doi.org/10.1016/j.envpol.2016.03.055>
- Santos, H.G., Jacomine, P.K.T., Anjos, L.H.C., Oliveira, V.A., Lumbrreras, J.F., Coelho, M.R., Almeida, J.A., Filho, J.C.A., Oliveira, J.B., Cunha, T.J.F., 2018. *Sistema Brasileiro de Classificação de Solos*, 5th ed. Embrapa Solos, Brasília. 353 p.
- Schaetzl, R.J., Anderson, S., 2005. *Soil: Genesis and Geomorphology*, 1st ed. Cambridge University Press, New York.
- Silva, F.M., Weindorf, D.C., Silva, S.H.G., Silva, E.A., Ribeiro, B.T., Guilherme, L.R.G., Curi, N., 2019. Tropical soil toposequence characterization via pXRF spectrometry. *Soil Sci. Soc. Am. J.* 83, 1153–1166. <https://doi.org/10.2136/sssaj2018.12.0498>
- Silva, S.H.G., Hartemink, A.E., Teixeira, A.F.S., Inda, A.V., Guilherme, L.R.G., Curi, N., 2018. Soil weathering analysis using a portable x-ray fluorescence (PXRF) spectrometer in an Inceptisol from the Brazilian Cerrado. *Applied Clay Science* 162, 27–37. <https://doi.org/10.1016/j.clay.2018.05.028>
- Silva, S.H.G., Ribeiro, B.T., Guerra, M.B.B., Carvalho, H.W.P., Lopes, G., Carvalho, G.S., Guilherme, L.R.G., Resende, M., Mancini, M., Curi, N., Rafael, R.B.A., Cardelli, V., Cocco, S., Corti, G., Chakraborty, S., Li, B., Weindorf, D.C., 2021. Chapter one – pXRF in tropical soils: Methodology, applications, achievements and challenges, in: Sparks, D.L. (Ed.), *Advances in Agronomy*. Academic Press, pp. 1–62. <https://doi.org/10.1016/bs.agron.2020.12.001>
- Soil Survey Staff. 2014. *Keys to soil taxonomy*. 12th ed. United States Department of Agriculture Natural Resources Conservation Service.
- Stevens, A., Ramirez-Lopez, L., 2022. An introduction to the prospectr package. Available at: <https://cran.r-project.org/web/packages/prospectr/vignettes/prospectr.html>. Accessed in: June, 2022.
- Stockmann, U., Cattle, S.R., Minasny, B., McBratney, A.B., 2016. Utilizing portable X-ray fluorescence spectrometry for in-field investigation of pedogenesis. *Catena* 139, 220–231. <https://doi.org/10.1016/j.catena.2016.01.007>
- Stolt, M.H., Genthner, M.H., Daniels, W.L., Groover, V.A., 2001. Spatial variability in palustrine wetlands. *Soil Science Society of America Journal* 65, 527–535. <https://doi.org/10.2136/sssaj2001.652527x>
- Struyf, E., Mörtz, C.-M., Humborg, C., Conley, D.J., 2010. An enormous amorphous silica stock in boreal wetlands. *Journal of Geophysical Research: Biogeosciences* 115. <https://doi.org/10.1029/2010JG001324>

- Sun, F., Bakr, N., Dang, T., Pham, V., Weindorf, D.C., Jiang, Z., Li, H., Wang, Q.B., 2020. Enhanced soil profile visualization using portable X-ray fluorescence (PXRF) spectrometry. *Geoderma* 358, 113997. <https://doi.org/10.1016/j.geoderma.2019.113997>
- Terra, F.S., Demattê, J.A.M., Viscarra Rossel, R.A., 2018. Proximal spectral sensing in pedological assessments: Vis–NIR spectra for soil classification based on weathering and pedogenesis. *Geoderma* 318, 123–136. <https://doi.org/10.1016/j.geoderma.2017.10.053>
- Thomas, F., Petzold, R., Landmark, S., Mollenhauer, H., Becker, C., Werban, U., 2022. Estimating forest soil properties for humus assessment - is visNIR the way to go? *Remote Sensing* 14. <https://doi.org/10.3390/rs14061368>
- USDA, 2010. Buried soils and their effect on taxonomic classification. US Department of Agriculture, National Soil Survey Center, Lincoln, NE. Available at: https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/ref/?cid=nrcs142p2_053568. Accessed in: June, 2022.
- Vettori, L., 1969. Métodos de análise de solo. Escritório de Pedologia e Fertilidade de Solos, Rio de Janeiro, Technical report 7.
- Walkley, A., Black, I.A., 1934. An examination of the Degtjareff method for determining soil organic matter and a proposed modification of the chromic acid titration method. *Soil Science* 37, 29–38.
- Weindorf, D.C., Zhu, Y., Haggard, B., Lofton, J., Chakraborty, S., Bakr, N., Zhang, W., Weindorf, W.C., Legoria, M., 2012. Enhanced Pedon Horizonation Using Portable X-ray Fluorescence Spectrometry. *Soil Science Society of America Journal* 76, 522–531. <https://doi.org/10.2136/sssaj2011.0174>
- Weindorf, D.C., Bakr, N., Zhu, Y., 2014. Advances in portable X-ray fluorescence (PXRF) for environmental, pedological, and agronomic applications. *Advances in Agronomy* 128, 1–45. <https://doi.org/10.1016/B978-0-12-802139-2.00001-9>
- Weindorf, D.C., Chakraborty, S., Abdalsatar, A., Aldabaa, A., Paulette, L., Corti, G., Cocco, S., Michéli, E., Wang, D., Li, B., Man, T., Sharma, A. and Person, T., 2015. Lithologic Discontinuity Assessment in Soils via Portable X-ray Fluorescence Spectrometry and Visible Near-Infrared Diffuse Reflectance Spectroscopy. *Soil Science Society of America Journal* 79, 1704–1716. <https://doi.org/10.2136/sssaj2015.04.0160>
- Weindorf, D.C., Chakraborty, S., Herrero, J., Li, B., Castañeda, C., Choudhury, A., 2016. Simultaneous assessment of key properties of arid soil by combined PXRF and Vis-NIR data. *Eur. J. Soil Sci.* 67, 173–183. <https://doi.org/10.1111/ejss.12320>
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R. ain, Grolemond, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T.L., Miller, E., Bache, S.M., Müller, K., Ooms, J., Robinson, D., Seidel, D.P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., Yutani, H., 2019. Welcome to the tidyverse. *Journal of Open Source Software* 4, 1686. <https://doi.org/10.21105/joss.01686>
- Wickham, H., Seidel, D., 2020. Scales: Scale functions for visualization. Available at: <https://cran.r-project.org/web/packages/scales/scales.pdf>. Accessed in: June, 2022.
- Wilson, M.A., Young, A.V., Knapp, B.D., Hoover, D.R., Swenson, H.K., 2012. Geochemistry of alluvial soils composed of metal-enriched sediments, main stem of the Coeur d’Alene river, Idaho. *Soil Science Society of America Journal* 76, 1462–1477. <https://doi.org/10.2136/sssaj2011.0442>

- Yang, M., Xu, D., Chen, S., Li, H., Shi, Z., 2019. Evaluation of machine learning approaches to predict soil organic matter and pH using Vis-NIR spectra. *Sensors* 19. <https://doi.org/10.3390/s19020263>
- Young, Crawford, J.W., Rappoldt, C., 2001. New methods and models for characterizing structural heterogeneity of soil. *Soil and Tillage Research* 61, 33–45. [https://doi.org/10.1016/S0167-1987\(01\)00188-X](https://doi.org/10.1016/S0167-1987(01)00188-X)
- Zhang, Y., Freedman, Z.B., Hartemink, A.E., Whitman, T., Huang, J., 2022. Characterizing soil microbial properties using MIR spectra across 12 ecoclimatic zones (NEON sites). *Geoderma* 409, 115647. <https://doi.org/10.1016/j.geoderma.2021.115647>
- Zhang, Y., Hartemink, A.E., 2020. Data fusion of vis-NIR and PXRF spectra to predict soil physical and chemical properties. *European Journal of Soil Science* 71, 316–333. <https://doi.org/10.1111/ejss.12875>
- Zhang, Y., Hartemink, A.E., 2019. Digital mapping of a soil profile. *European Journal of Soil Science* 70, 27–41.

Chapter II: Pixel-based spatiotemporal statistics from remotely sensed imagery improves spatial soil predictions and sampling strategies

Accepted in Geoderma in 13 May 2024. <https://doi.org/10.1016/j.geoderma.2024.116919>

Marcelo Mancini^{1,2}, Hans Edwin Winzeler³, Joshua Blackstock⁴, Phillip R. Owens^{4,*}, David M. Miller¹, Sérgio H. G. Silva², Amanda J. Ashworth⁵

^aDepartment of Crop, Soil, and Environmental Sciences, Univ. of Arkansas, 465 Agriculture Building, Fayetteville, AR 72701, USA

^bDepartment of Soil Science, Federal Univ. of Lavras, 1001 Doutor Silvio Menicucci Ave., Lavras, Minas Gerais 37200-900, Brazil

^cUniversity of Texas at Arlington, Arlington, TX 76019, USA

^dUSDA-ARS, Dale Bumpers Small Farms Research Center, 6883 AR-23, Booneville, AR 72927, USA

^eUSDA-ARS, Poultry Production and Product Safety Research Unit, Univ. of Arkansas, O-303 Poultry Science Center, Fayetteville, AR 72701, USA

*Corresponding author: phillip.owens@usda.gov

Abbreviations: SOM: soil organic matter; CEC: cation exchange capacity; cLHS: Conditioned Latin Hypercube Sampling; OV: distribution overlapping index.

Abstract

Alluvial plains are vexing landscapes for soil mapping and spatial soil property predictions. Alluvial sediments often exhibit unpredictable spatial variability from both fluvial and anthropogenic disturbance. The determination of optimal number of soil sampling points for capturing soil variability remains a persistent issue for mapping and monitoring soil conditions. Here, soil organic matter (SOM) and cation exchange capacity (CEC) were estimated in an alluvial plain from a dense array of 2145 soil samples over 250 ha. The primary goals were to i) use pixel-based statistics from time-series Sentinel-2 satellite reflectance data to estimate SOM and CEC; ii) evaluate the use of images with vegetation cover versus bare soil images; and, iii) investigate the optimal number of sampling points to map alluvial soils using different sampling strategies. The optimal sample density was 1 sample per 2.5 ha based on the overlapping of prediction distributions ($OV > 0.9$). Conditioned Latin Hypercube Sampling (cLHS) was the most efficient sampling strategy. Random grid sampling provided the least consistent results. The use of cLHS coupled with the mean and standard deviation bands resulted in optimal sampling strategy. Pixel-based statistics from readily available satellite imagery captured persistent soil-reflectance relationships that enabled the use of row crops as proxies to predict soil properties. The combined use of pixel-based statistics as inputs to cLHS can reduce the chances of oversampling and associated costs. Pixel-based statistics provided persistent spatial information that can enhance precision agriculture management, carbon mapping, and is likely applicable to other soil properties and remotely sensed products. Use of pixel-based statistics from Sentinel-2 and similar global-coverage imagery could provide readily-calculable inputs for improved soil mapping and more optimal site sample selections in other agricultural row crop areas, globally.

Keywords: Inceptisols; Vertisols; soil organic carbon; cation exchange capacity; Sentinel-2; machine learning; random forest; recursive feature elimination.

1. Introduction

Soil organic matter (SOM) and cation exchange capacity (CEC) influence soil chemical and physical properties, and plant growth (Brady and Weil, 2002). Soil is the largest carbon reservoir of the terrestrial biosphere, and therefore a crucial factor in global climate change (Lal et al., 2021). Understanding the distribution of soil carbon through space and time is fundamental to tracking changes in soil health for agriculture and Earth's carbon budget (Kindler et al., 2011, Keenor et al., 2021). However, mapping soil carbon variability across Earth's surface remains a challenging task and considerable uncertainties persist (Keskin et al., 2019).

The spatial distribution of soil carbon in landscapes is not uniform, but varies with topography, climate, vegetation, land management, related soil properties, and other factors (Brady and Weil, 2002; Schaetzl and Anderson, 2015). Topographic variation sometimes provides important clues as to soil carbon variability when other soil-forming factors are relatively constant (Schaetzl and Anderson, 2015). Alluvial areas, however, often have minimal topographic variation, making them particularly vexing landscapes in which to map SOM and CEC based on terrain attributes. Further complicating predictability of alluvial soil properties, parent material in alluvial areas can be highly variable due to primary and secondary depositional environments, i.e., the continued reworking of sediments related to fluvial sediment transport and anthropogenic modifications (Saucier, 1994). Hence, a significant number of samples may be needed to constrain the spatial structure of SOM and CEC in alluvial environments. Knowledge on the spatial structure of a soil property of interest can improve the use of Conditioned Latin Hypercube Sampling (cLHS) for sample allocation, selection, and subsequent spatial estimations (Kravchenko, 2003), but information on soil properties may be limited in some areas. The use of remotely sensed imagery, which is globally and continuously available, may help overcome this limitation and provide important information about the spatial structure of properties.

The sampling of soils and laboratory analyses for SOM and CEC assessment can be expensive, with the additional risk of not capturing the spatial structure for accurate spatial estimations depending on the number and location of selected sampling sites. Soil sample analysis requires physical removal of samples, laboratory processing with chemical extractants or, in the case of SOM, high energy ignition (Donkin, 1991; Mehlich, 1984). The resulting data are limited in scope to the specific point location of extraction, thus requiring modeling to scale data from finite point observations to continuous spatial estimations (Yuzugullu et al., 2020). To develop accurate spatial estimation models utilizing covariates, it is imperative that soil

samples capture the spatial structure and include ancillary data that are spatially correlated with soil properties of interest (Hengl et al., 2021; Zeraatpisheh et al., 2022). Common ancillary data for SOM and CEC spatial estimation models include digital elevation models (DEMs) and topographic attributes (Florinsky, 2016; Gessler et al., 1995; Moore et al., 1993; Winzeler et al., 2022; Winzeler et al., 2008), and remotely sensed reflectance spectra (Castaldi et al., 2019a, 2019b, 2019c; Dvorakova et al., 2020; Gholizadeh et al., 2020, 2018; Gomez et al., 2018; Hengl et al., 2021; Silvero et al., 2021; Yuzugullu et al., 2020).

Elevation and topographic attributes can indicate topographic features where water accumulates, which can be correlated to the preservation of SOM. SOM reportedly relates to sentinel reflectance data because soils with increasing levels of SOM absorb increasing amounts of electromagnetic radiation, leading to decreasing reflectance (Castaldi et al., 2019a). Of all the components in soils that contribute to CEC, organic matter contributes the most charge per unit weight, especially at the surface (Brady and Weil, 2002). SOM and CEC can be considered related and their variability in surface soils are closely linked. These relationships, however, depend on the quality of the data. Soil moisture, crop residue, and green vegetation can interfere with direct relationships between reflectance, SOM, and CEC (Dvorakova et al., 2020). Finding cloud-free, bare soil images is often a concern in digital soil mapping (Castaldi et al., 2019a). For models to capture the spatial structure of soil properties and provide accurate predictions, we emphasize three factors of relevance: the correlations between ancillary (reflectance and topographic data) and target variables (SOM and CEC) (Guo et al., 2021a; Shafizadeh-Moghadam et al., 2022), the quality of the data (Luo et al., 2022; Silvero et al., 2021), and the sampling strategy (Castaldi et al., 2019b; Minasny and McBratney, 2006). Thus, the relationships between reflectance signatures, elevation, and measured soil properties at point locations can be transferred as estimates to other locations using machine learning models, such as Random Forest (Andrade et al., 2022; Guo et al., 2021a, 2021b; Lin et al., 2020).

Here, we introduce the calculation of pixel-based statistics of satellite images. Calculating statistics such as mean and standard deviation of time-series of reflectance data can reveal persistent spatial patterns within the data, mitigating interferences by row crop vegetation. However, sample size, pattern, and density play an important role in modeling. The redundant sampling of similar sites and/or an insufficient number of points will fail to explain the spatial structure of soil properties. Some sampling strategies, such as cLHS, account for the distributions of previously known spatial variables related to the target variable to decide the optimal sampling locations (Minasny and McBratney, 2006). Long-term mean and variability of reflectance captured by pixel-based statistics may aid the performance of spatial estimation

models and cLHS. In this study, we explored the use of pixel-based means and standard deviations of time-series Sentinel-2 satellite imagery as explanatory variables. Our goals were to: i, predict SOM and CEC in alluvial soils; ii, investigate the effect of vegetation, in this case row crops, on prediction accuracy; and iii, use a unique, spatially dense dataset (from 0.02 to 8.6 samples per ha) to find the minimum number of samples required to achieve desirable accuracy of SOM and CEC predictions in alluvial soils. We hypothesized that pixel-based statistics would capture the long-term spectral variability related to the spatial distribution of soil properties, thus better informing sample selection algorithms such as cLHS, leading to accurate estimations of SOM and CEC, and more efficient and cost-effective sampling schemes for digital soil mapping.

2. Material and Methods

2.1. Study area

The study was conducted in the National Center for Alluvial Aquifer Research, in Washington County, Mississippi, USA. The site has an area of approximately 250 hectares (2.5 km²) (Figure 1). It is situated 37 m above sea level and has very smooth relief, with elevation varying less than 4 m across the area. Soil orders found in the studied site are: Alfisols, Inceptisols and Vertisols (Table 1). The north boundary of the area is delimited by Deer Creek (Figure 1a), within the Mississippi River alluvial plain, where the geomorphology is driven by fluvial processes (Saucier, 1974; Saucier, 1994). The northern portion of the study area adjacent to Deer Creek represents a natural levee deposit having coarser sand fraction sediments. These coarse sediments overlie a thick clay layer characteristic of the broader floodplain, with the levee deposit thinning out towards the south. This geomorphology resulted in a generally smooth north-south elevation gradient (Fig 1b), as well as in a texture gradient, shown by the gradation of texture from very fine sandy loam soils at the north (Bosket series) to clay at the south (Downling and Sharkey series) (Table 1).

2.2 Soil sampling and laboratory analyses

Soil samples were collected in a dense grid (8.6 samples per ha) at a depth of 0-15 cm, totaling 2145 samples across 250 ha (Figure 1). SOM was assessed by loss-on-ignition (Donkin, 1991). CEC was calculated as the sum of Ca²⁺, Mg²⁺, K⁺, and Na⁺ in soils. These cations were extracted using Mehlich-3 (Mehlich, 1984) and quantified by inductively coupled argon plasma atomic emission spectroscopy.

2.3. Spatial data, statistical analysis, and modeling

Data analysis and modeling procedures shown in the following sections were performed using R (R Core Team, 2022). Utilized packages were: tidyverse (Wickham et al., 2019), caret (Kuhn, 2008), terra (Hijmans, 2023), sp (Bivand et al., 2013; Pebesma and Bivand, 2005), sf (Pebesma, 2018), clhs (Roudier, 2011), ggpointdensity (Kremer, 2019) and sen2r (Ranghetti et al., 2020).

2.3.1. Digital elevation model

We obtained a 1 m resolution digital elevation model made with laser altimetry for the site in Western Mississippi from the United States Geological Survey 3DEP program (U.S. Geological Survey, 2022a). Data were acquired between January 24, 2019 to February 29, 2019 using the Riegl 1560i system capable of acquiring 1.3 million measurements per second mounted on a Piper Navajo twin-piston aircraft (Quantum Spatial, 2020). The average point spacing was 2.04 points m⁻² (Quantum Spatial, 2020). Light-detection and ranging (LiDAR, or air-borne laser altimetry) bare-earth returns were converted to 1-m elevation rasters through a proprietary linear interpolation algorithm applied to a triangulated irregular network with a required vertical accuracy of 0.196 m (Quantum Spatial, 2022). These DEM data were made freely available to the public through the USGS (U.S. Geological Survey, 2022b).

2.3.2. Sentinel-2 data

Sentinel-2 was used as the source of reflectance data. Sentinel-2A and 2B images in the period between December 17, 2018 and February 19, 2023 were used. Only atmospherically corrected images were used (Level 2A products). All images not originally at 10-m resolution were resampled to 10 m using bilinear interpolation.

Cloud coverage was programmatically removed by eliminating images with pixels over the study area with more than 80% probability of being clouds. Sentinel-2's own cloud probability dataset was utilized, which uses the 'sentinel2-cloud-detector' library to calculate cloud probability via gradient boosting models. Remaining images were manually checked for cloud pixels that may have been misclassified.

After selection, a total of 121 images remained out of 310, including a mix of vegetated images with growing crops and bare soil images. Next, images with predominantly bare soil pixels were selected. Bare soil images were selected by identifying low mean values of Normalized Difference Vegetation Index (NDVI) across the study area, aided by visual assessment of each image (Figure 2). NDVI consists of the normalized difference between

chlorophyll absorption in the red edge of the spectrum (band 4) and leaf scattering in the near-infrared range (band 8), calculated as: $(\text{band } 8 - \text{band } 4)/(\text{band } 8 + \text{band } 4)$. A total of 75 bare soil images were available. In summary, two datasets were tested: a mixed dataset, with images containing both crop canopy and bare soil ($n = 121$), and a bare soil dataset, in which images mostly consisting of bare soil pixels were selected ($n=75$).

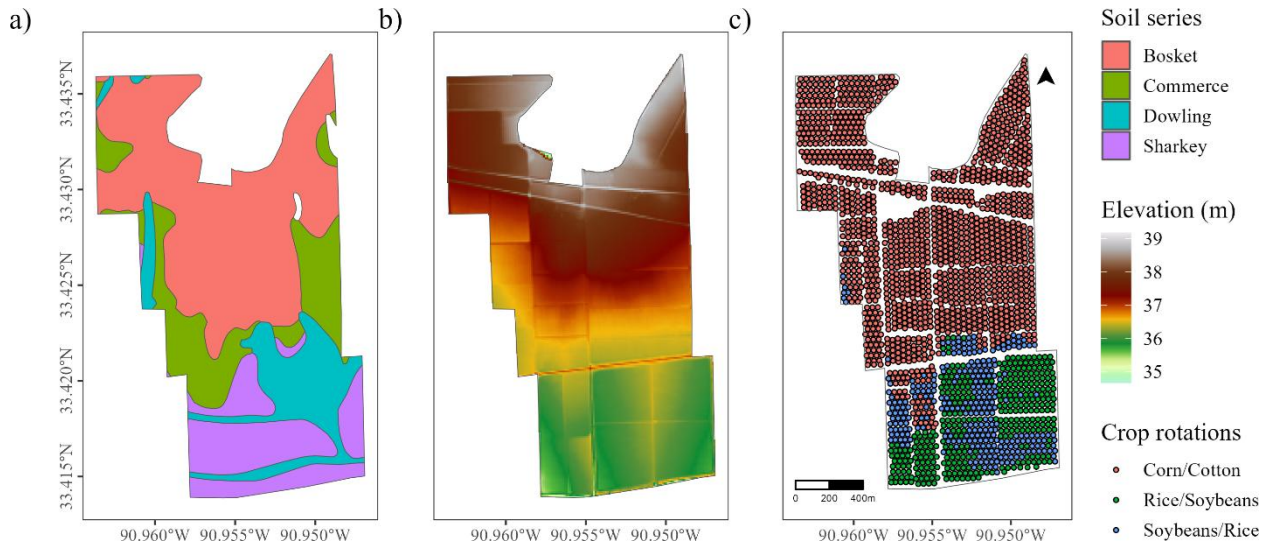


Figure 1. National Center for Alluvial Aquifer Research in Western Washington County, Mississippi, USA. Aerial view of the study area (a, 10 m spatial resolution), digital elevation model (b, 10 m spatial resolution) and soil sampling points (c, $n = 2145$).

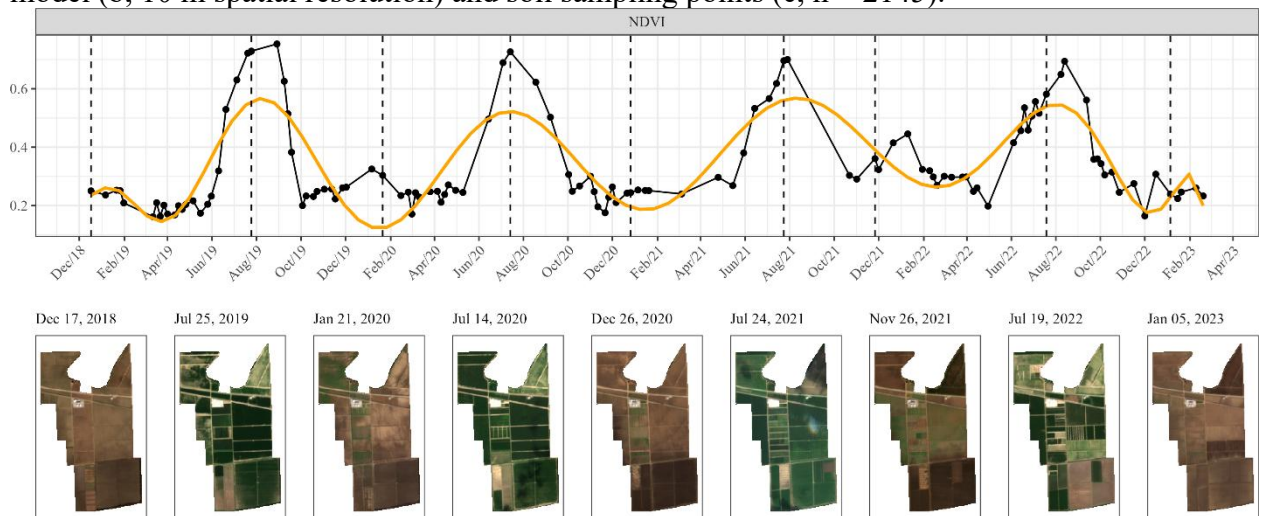


Figure 2. Seasonal variation of mean NDVI between December of 2018 and March of 2023 in Western Mississippi, showing how plant canopy area affects reflectance patterns over time. Dashed lines at the topmost plot indicate the corresponding NDVI values of each image below, in order.

Table 1. Soil classes and their respective texture and drainage classes of series found in the study area, in northeastern Mississippi, USA.

Series	Taxonomic family	Texture class	Drainage class
Bosket	Fine-loamy, mixed, active, thermic Mollic Hapludalfs	Very fine sandy loam	Moderately well drained
Commerce	Fine-silty, mixed, superactive, nonacid, thermic Fluvaquentic Endoaquepts	Silt loam, silty clay loam, or very fine sandy loam	Somewhat poorly drained
Sharkey	Very-fine, smectitic, thermic Chromic Epiaquepts	Clay or silty clay loam	Poorly drained
Dowling	Very-fine, smectitic, nonacid, thermic Vertic Endoaquepts	Clay	Very poorly drained

2.3.3. Pixel-based statistics

Pixel-based statistics were calculated for both mixed (vegetation + bare soils, $n = 121$) and bare soils (only bare soil images, $n = 75$) datasets. This consisted in calculating the mean and standard deviation of each collocated pixel using all the selected images between 2018 and 2023. Point density plots were used in visually assessing potential multi-modal distributions between mean and standard deviation pixel distributions (Figure 3). The plots were made using ggpointdensity R package with the parameter “adjust” set to 0.01. “Adjust” controls the radius in which points should be considered neighbors (see section 3.1).

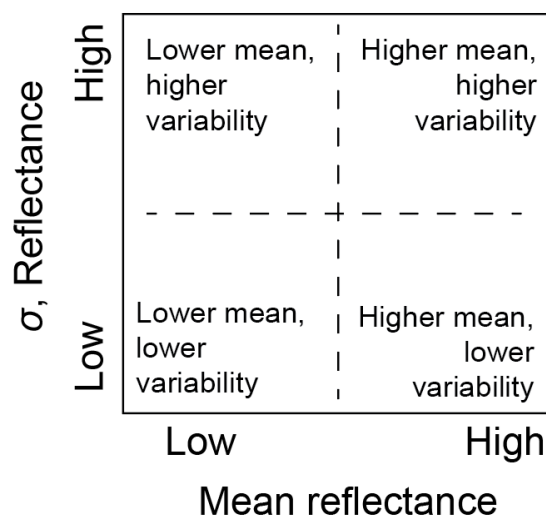


Figure 3. Conceptual two-dimensional histogram space of standard deviation versus mean reflectance and generalized groupings based on relative values.

The resulting rasters were then used as explanatory variables in the modeling process (see section 2.3.4). Spearman’s correlation coefficient was calculated to assess the correlations among SOM, CEC, elevation, and pixel-based statistics for used bands (bands 2, 3, 4, 8, 11 and 12). Comparisons among the distributions of pixel-based statistics calculated from satellite data, SOM and CEC were made using the overlapping index (OV) proposed by Pastore and Calcagni (2019). OV measures the degree of overlap between two distributions. Zero indicates no

overlap; one indicates complete overlap. The data were normalized prior to comparisons (Z-score normalization: $x_{normalized} = (x - \hat{x}) / \sigma_x$).

2.3.4. Spatial estimation of SOM and CEC

The Random Forest algorithm was used for modeling (Liaw and Wiener, 2002). The training phase comprised 5-repeated 10-fold cross-validations (CV). Data were split into a training set (80%) and an external validation set (20%). CV was performed in the training set while varying the mtry parameter. Mtry defines the number of features used in each node of Random Forest, hence each branching node used 5 randomly sampled bands. Goodness of fit was assessed using the coefficient of determination (R^2) and the root mean square error (RMSE). Explanatory variables were selected using the recursive feature elimination (RFE) algorithm (Gregorutti et al., 2017). Since SOM and CEC were found to have similar spatial distributions (see section 3.1.), the union of the seven most important variables selected by the algorithm for both soil attributes were used as explanatory variables. Selected explanatory variables differed for the dataset mixing vegetated and bare soil images, and the set containing only bare soil images ($n = 121$ and 75 , respectively).

Initially, elevation and Sentinel-2 spectral bands 2, 3, 4, 8, 11 and 12 were used as inputs to the RFE algorithm. The sets of variables chosen after RFE were: mixed dataset (vegetation + bare soil) – standard deviation of bands 2, 8, 11 and 12, means of bands 2, 3, 4 and 8, and elevation; bare soil dataset – standard deviation of bands 3, 4, 8, 11 and 12, mean of band 8 and elevation (further discussed in section 3.3.). The cross-validation indicated that the parameter $mtry = 5$ was optimal for both SOM and CEC for all models. It was therefore used in all modeling procedures.

2.3.5. Optimal sampling strategies

We evaluated the optimal sampling number and spatial location of soil sample points which provided reliable predictions of SOM and CEC given a threshold of $OV > 0.9$, i.e., where the distribution of resulting prediction maps overlapped more than 90%. The 0.9 threshold is arbitrary and was used to aid comparisons. Prediction maps obtained using all 2145 samples were considered the ground truth. Optimal sampling tests were done by creating model outputs trained with progressively more samples and comparing them to ground truth maps ($n=2145$). Sample size started with $n = 5$ and increased up to $n = 2100$. Between 5 and 50 samples, sample size increased by 5; between 50 and 100, by 10; between 100 and 500, by 50; and between 500 and 2100, by 200. A total of 31 sample sizes were tested and compared against the ground truth.

Three sampling strategies were tested: regular grid sampling, random sampling and cLHS (Minasny and McBratney, 2006; Roudier, 2011). For each tested sample size, sampling schemes were randomly generated according to each tested sampling strategy and the closest points from the available 2145 samples were selected to represent the sampling schemes (Figure 4). Pixel-based mean and standard deviation of bands 2, 3, 4, 8, 11 and 12 from collected Sentinel-2 images were used as inputs for the cLHS algorithm. The prediction score of each tested sample number is the average score of 1000 simulations. Simulations were performed in SCINet, USDA's high-performance computer clusters.

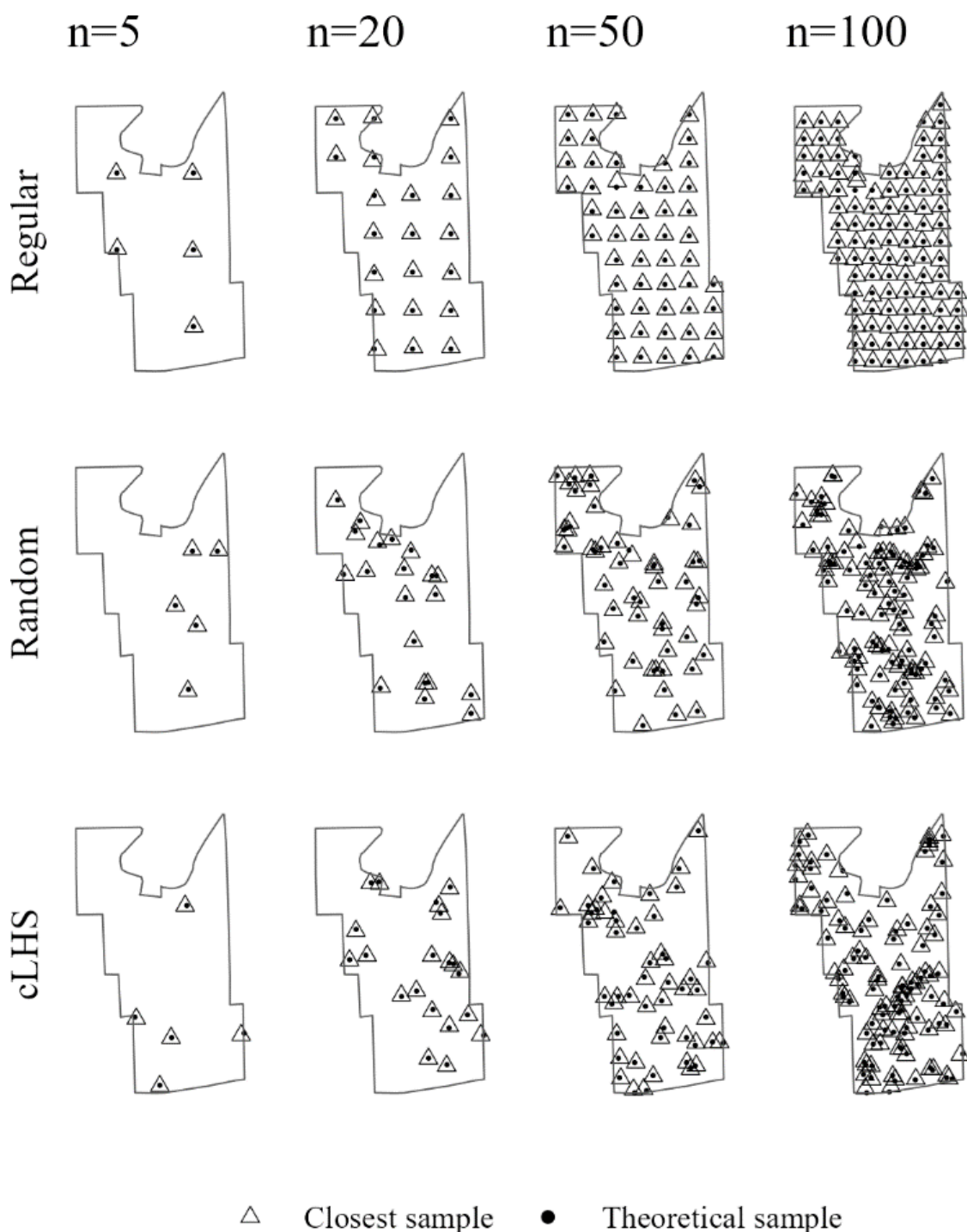


Figure 4. Examples of how each sampling scheme was simulated from available data. A theoretical sampling scheme was created for each sampling strategy (regular grid, random grid, cLHS) and for each tested sample size, from which available samples closest to each theoretical point were selected. The process was randomly repeated 1000 times for each sampling size (from 5 to 2100). cLHS: Conditioned Latin Hypercube Sampling.

The benefits of using pixel-based statistics paired with cLHS were also evaluated. An arbitrarily chosen bare soil image (January 26, 2019) was selected to compare the performance of bands from a single date versus pixel-based statistics bands (i.e., mean and standard deviation bands calculated from several dates) in selecting samples when used as inputs to cLHS. Only bare soil images were used in this comparison. Bands 2, 3, 4, 8, 11 and 12 were used. Similar to the tests among sampling strategies, sample number progressively increased, and models were run 1000 times for each sample density. OV was calculated to compare the distribution of each predicted map to that of the ground truth map (map using all 2145 samples).

3. Results and Discussion

3.1. The relationship between of satellite imagery and soils

Correlations among target (SOM and CEC) and input (pixel-based statistics of satellite bands and elevation) variables were all negative and significant (Figure 5a). Elevation was the attribute most correlated to target variables, followed by bands' means and standard deviation. All band means had higher correlation compared to the standard deviation. The overlapped distributions, however, showed an opposite trend: the standard deviation of bands presented higher OV than the means.

As SOM increases, it lowers the reflectance of soil samples in most wavelengths (Santos et al., 2020), explaining the opposing trends between mean band reflectances and SOM contents. SOM is a heterogeneous mixture of polyfunctional organic molecules which absorb radiation over a significant range of wavelengths (Table 2) (Ben-Dor et al., 2018; Calderón et al., 2011). CEC in soils is controlled mainly by SOM and mineralogy. In addition to wavelengths related to SOM, wavelengths related to 2:1 phyllosilicates are expected to be correlated with CEC. Bands 11 and 12, particularly, are centered close to 1.4 and 1.9 μm , which are two strong absorption features of 2:1 phyllosilicates and water (Fang et al., 2018). Absence of change in reflectance of bands 11 and 12, therefore, indicates spots where water content is less variable, either due to lack of water input, or due to its stronger retention by soils. In irrigated areas, the latter case predominates, wherein the lower standard deviation of bands highlights locations where soil water retention is higher, either due to differences in texture or mineralogy (water OH groups adsorbed in 2:1 phyllosilicates).

Table 2. Summary of Sentinel-2 bands and their physical relationships to soil, water and plant properties.

Band	Central Wavelength (μm)	Related properties	References
Bands 2, 3 and 4 (red, green and blue)	0.492, 0.560 and 0.665, respectively	Electronic transitions of iron cations in oxides, chromophores of organic matter	Ben-Dor et al., (2018); Fang et al., (2019), Thomas et al., (2022)
Band 8 (near-infrared)	0.833	Strong reflectance by plant tissue due to photon scattering in the mesophyll	Castaldi et al., (2019); Frampton et al., (2013), Ustin et al., (2004)
Bands 11 and 12 (short wave infrared)	1.612 and 2.1931	Overtone of organic functional groups, water and 2:1 clay minerals (close to 1.4 and 1.9 μm)	Ben-Dor et al., (2019); Liu et al., (2021), Thomas et al., (2022)

SOM and CEC distributions were almost identical with nearly all exhibiting some degree of bimodality (OV of 0.98, Figure 5b, c). The similarity in distributions suggests that either most of the CEC of soils originates from SOM, or that the variability of texture and mineralogy is closely following that of organic matter accumulation. It is also possible that both are contributing simultaneously to the observed similarity between SOM and CEC (Gallez et al., 1976; Kaiser et al., 2008; Oorts et al., 2003). The negative correlation between both SOM and CEC and elevation suggests a combined effect.

Elevation also decreased as SOM content and CEC increased. Two plausible explanations are possible: the accumulation of water for longer periods at lower elevation areas slows down the decomposition of SOM, favoring its accumulation (less water, more oxygen, and hence more biologic activity); and the segregation of sediments driven by fluvial and geomorphological corollaries from repeated flood events, whereby coarser and finer sediments were deposited in different elevations. Coarser sediments and higher sand fraction contents are located proximal to Deer Creek as natural levee deposits at higher elevations in the study area (Saucier, 1974; 1994). Conversely, fine-textured, more clay-rich sediments were located southward within the flood plain (Figure 1 and Table 1). The development of soil attributes of SOM accumulation, CEC, and associated mean and standard deviation reflectance patterns were controlled by differences in the alluvial parent material as driven by fluvial geomorphology.

The mechanisms by which this north-south texture gradient could affect the spatial distribution of SOM and CEC can be summarized by two main factors: water infiltration capacity and mineralogy. Coarser texture allows for more favorable water infiltration and increased hydraulic conductivity, thereby reducing the occurrence of anoxic conditions that

would favor SOM accumulation (and hence raise CEC). Conversely, finer texture hinders drainage by creating fewer macropores, favoring water accumulation for longer periods, leading to SOM accumulation.

Anoxic conditions hamper the oxidation of SOM and may slow down the weathering of some minerals by reducing the leaching rate of cations from the soil solution. In addition, clay content has a well-known synergistic relationship with organic matter (Brady and Weil, 2002; Chi et al., 2022; Kida and Fujitake, 2020), and the less weathered, 2:1-type phyllosilicates have higher CEC than their more weathered, 1:1-type counterparts (Pincus et al., 2017). Relating back to the conceptual model (Figure 3), greater mean SOM and CEC and lower standard deviation values can be expected in the floodplain deposits, where the relief is flatter and clay content is greater, as opposed to the natural levee deposits, where sediments are comparatively coarser. This relationship justifies the negative correlation (Figure 3) and explains the mean and standard deviation spatial patterns. This combined effect of water retention and mineralogy was also observed in the positions of soil classes in the landscape: Inceptisols with mixed mineralogy northward, and smectitic Vertisols in the southern portions of the landscape (Figure 1a, Table 1).

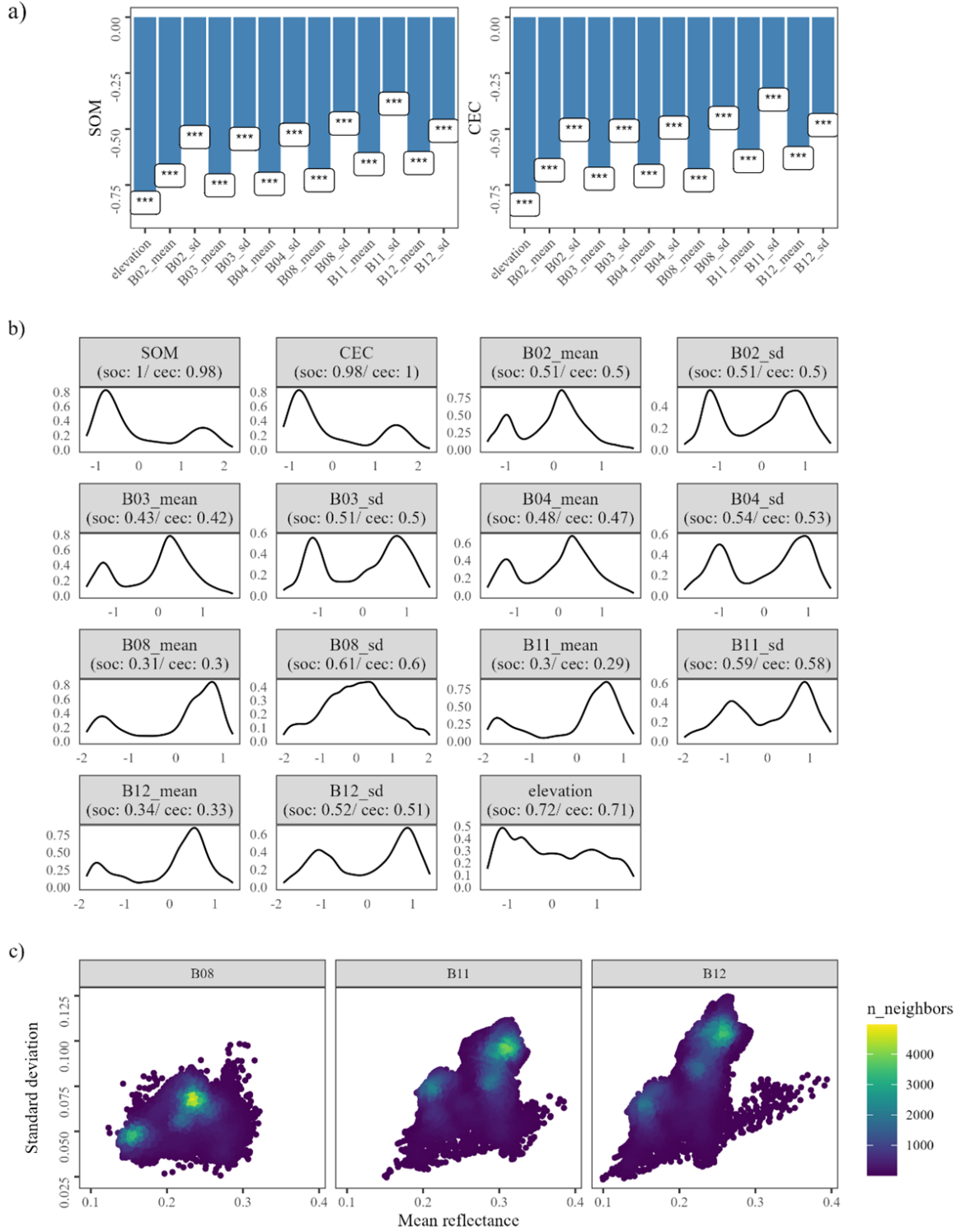


Figure 5. Spearman's correlation coefficients between target variables (SOM and CEC, $n = 2145$) and input variables (pixel-based statistics of Sentinel-2 bands and elevation) (a) and their respective distribution shapes (b). Satellite data used herein correspond to means and standard deviations (sd) calculated for bare soil images selected from December, 2018 to March, 2023 ($n = 75$). Numbers beneath labels (b) indicate the overlapping degree of the distributions of SOM and CEC compared to all others. 1 indicates complete overlapping; 0 indicates none (Pastore and Calcagni, 2019). Point density plots (c) comparing standard deviation versus mean reflectance of bands 8, 11 and 12.

3.2. The persistence of spectral data over time and its effect on vegetation reflectance

Vegetation is known to interfere in reflectance measurements, leading to the recurring attempt to select bare soil images when studying soil-related properties using satellite data (Diek et al., 2016). The effect of vegetation is clear, especially concerning bands 4 and 8 ($0.8 \mu\text{m}$), used to calculate NDVI (Figure 2). Band 8 is intensely reflected by leaf structures due to photon scattering in the spongy mesophyll, while band 4 is absorbed by chlorophyll (Cortez et al., 2020; Ustin et al., 2004). Yet, the challenge of pursuing bare soil images was also evident: even in periods when NDVI was low, the area was partially vegetated (Figure 2).

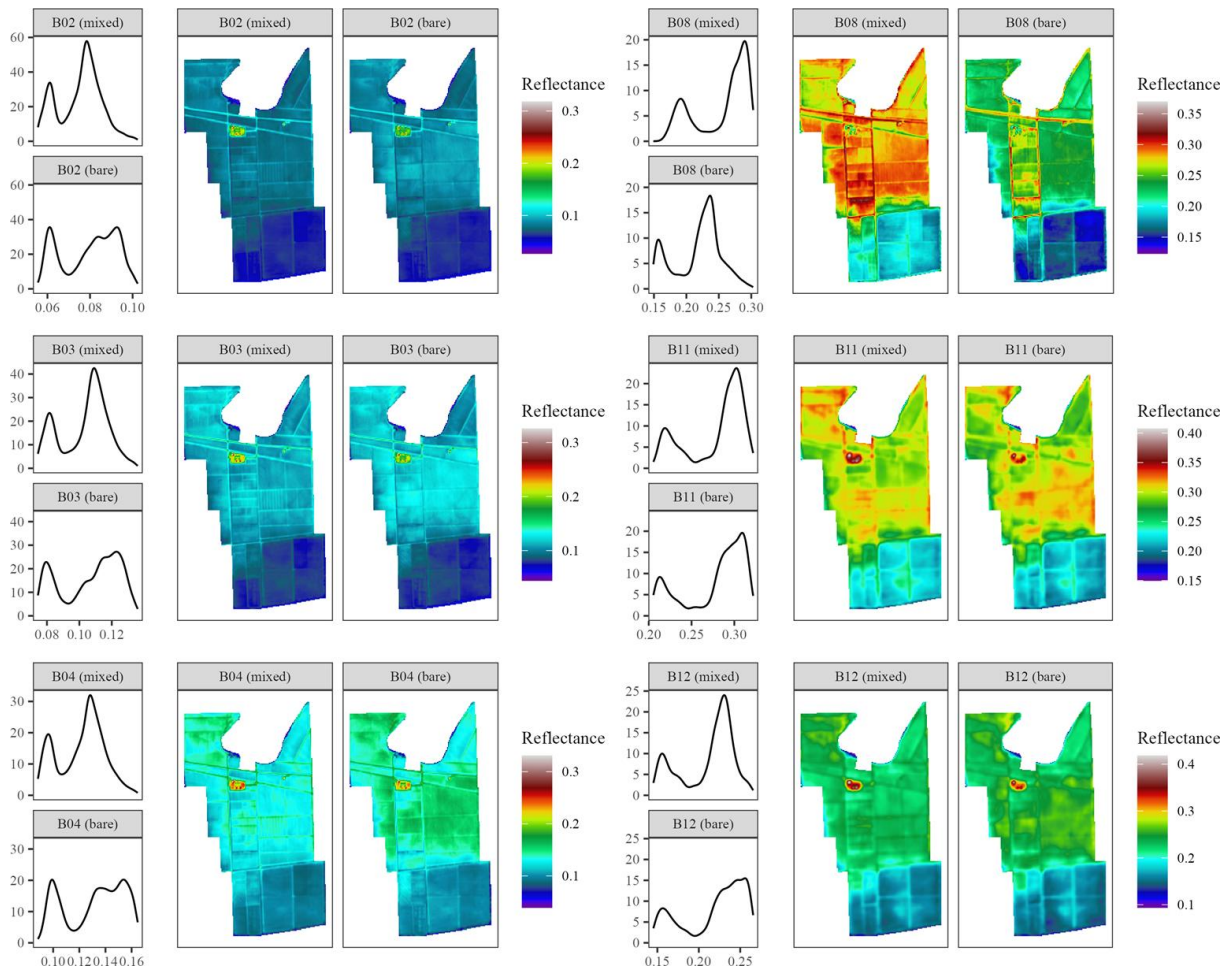


Figure 6. Means of Sentinel-2 images between December of 2018 and March of 2019 comparing bare soil images (n = 75) and a set of mixed images containing both bare soils and vegetated images (n = 121) in Western Mississippi.

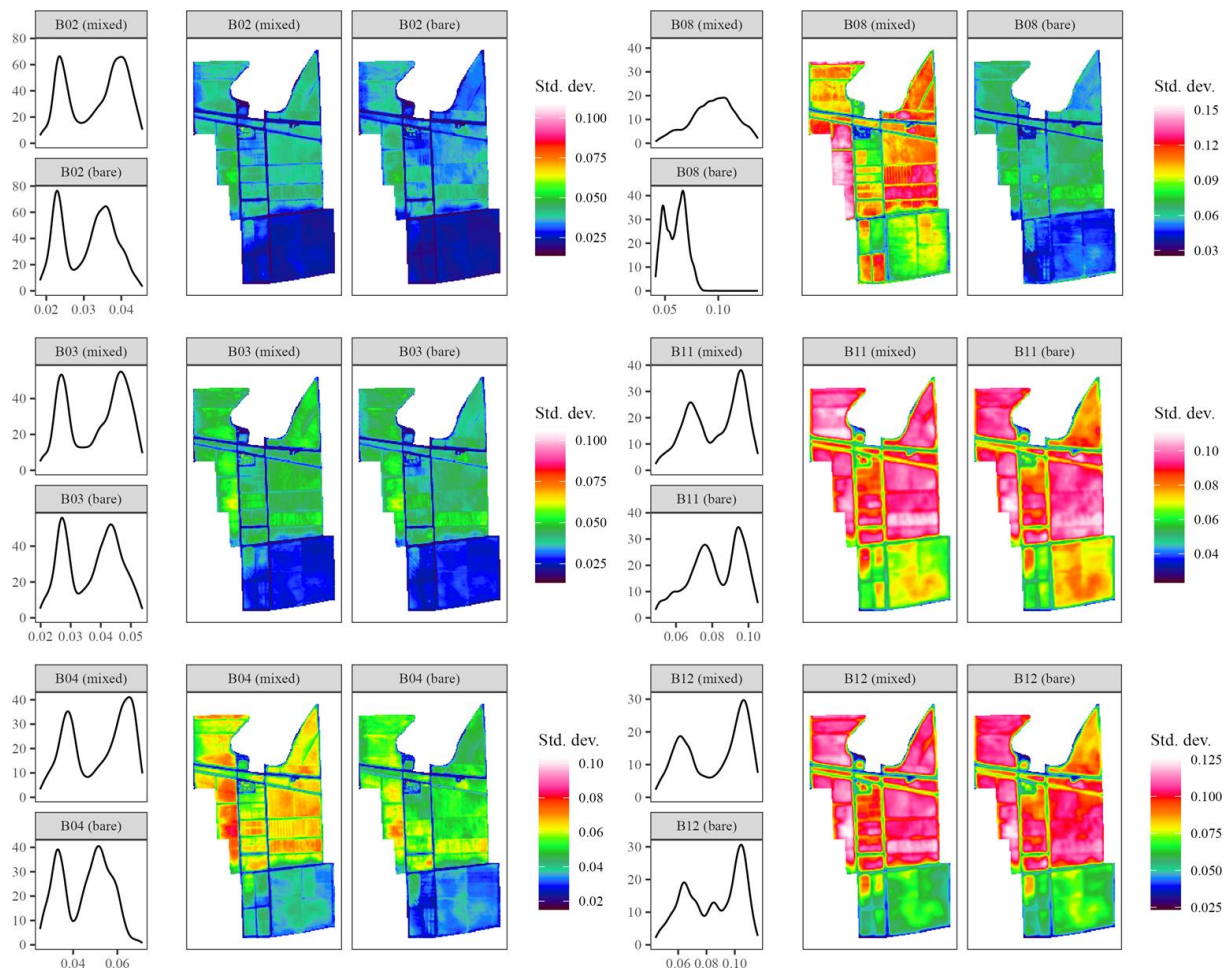


Figure 7. Standard deviation of Sentinel-2 images between December of 2018 and March of 2019 comparing bare soil images ($n = 75$) and a set of mixed images containing both bare soils and vegetated images ($n = 121$) in Western Mississippi. Std. dev. – standard deviation of reflectance values.

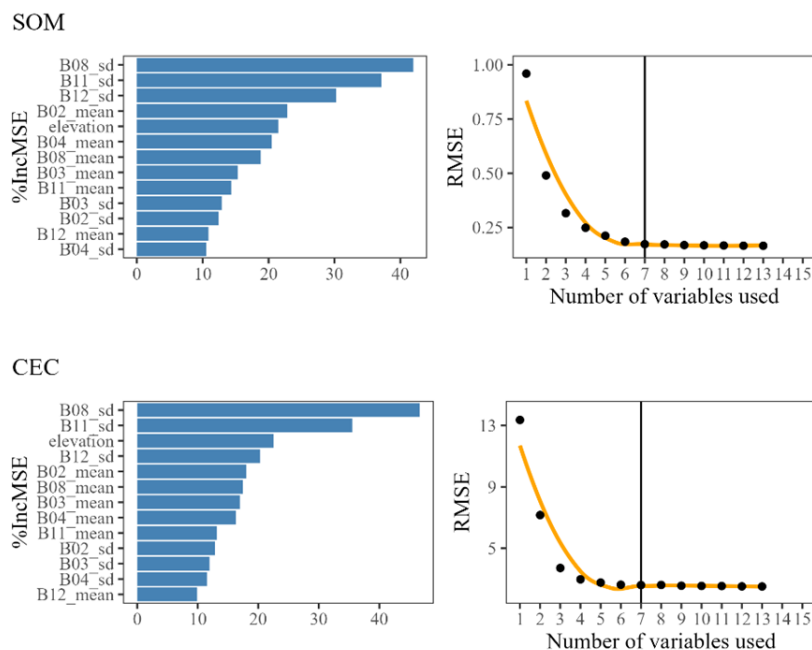
Comparisons between statistics of bare soil images ($n = 75$) and a mixed set of images containing vegetated seasons ($n = 121$) showed some similarities (Figs. 6 and 7). Surprisingly, although band 8 means were slightly shifted to higher values in the mixed dataset, the distribution shapes were similar (Figure 6). Comparing both datasets, all bands presented similar distribution of pixel means, with a few differences in intermediate values between the two modes. Standard deviations of pixels, however, presented more divergence (Figure 7). Although the bimodal distribution remained, shapes were more dissimilar, and band 8 showed contrasting shape and range differences. Nonetheless, the spatial patterns of high and low values between northern and southern areas of the study area were notably similar between datasets in bands 4, 8, 11 and 12, albeit varying in magnitude.

3.3. The importance of Sentinel-2 bands for spatial estimations

The recursive feature elimination algorithm ranked nearly the same variables as most important for SOM and CEC predictions (Figure 8), reaffirming the idea that both SOM and

CEC follow a similar spatial distribution due to the combined effects of soil forming factors and processes (see section 3.1), and related correlation between soil properties and spectral features. Standard deviation of pixels calculated from the respective bands were persistently the most important variables, followed by elevation. Recursive feature elimination using the mixed dataset achieved low RMSE values comparable to those obtained from pure bare soil data (Figure 8).

a)



b)

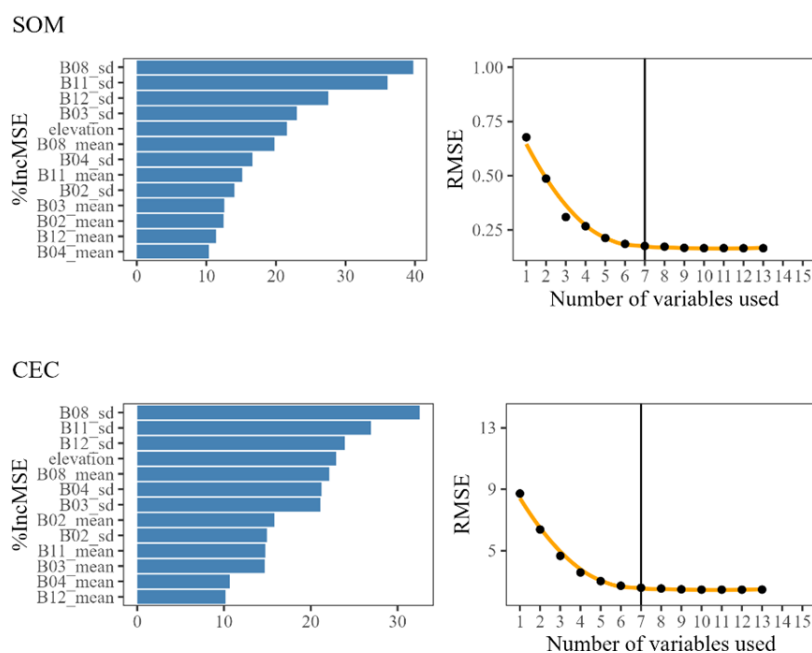


Figure 8. Recursive feature elimination using the Random Forest algorithm to predict soil organic matter (SOM) and cation exchange capacity (CEC) in 2145 samples from alluvial soils

using a set of mixed Sentinel-2 images containing both bare soils and vegetated images (a, $n = 121$), and using a set containing only bare soils (b, $n = 75$) in Western Mississippi. The orange line represents a loess model.

The importance of the visible range (bands 2, 3 and 4) for SOM predictions is easy to explain. The organic compounds that collectively make up SOM contain many structural units, or functional groups, that absorb electromagnetic (EM) radiation in the visible region of the EM spectrum (Viscarra Rossel et al., 2006). The fact that SOM has black color results from the absorption of virtually all wavelengths of visible radiation by these structural units, which are also called chromophores (Ben-Dor et al., 2018). Chromophores of oxides overlap with those of organic matter (Heller Pearlshtien and Ben-Dor, 2020), but this is often not a concern in alluvial plains due to the reductive dissolution of ferric oxides in these landscapes, which are often poorly drained.

It is more challenging, however, to understand the role of row crop canopy in the mixed dataset. This is especially true for band 8, which is highly sensitive to leaf structures. The reflectance of crops may be providing information about areas where plants are consistently greener (and more vigorous), or about different plant species or crop rotations, associated with different degrees of SOM accumulation. By using plants as proxies for soil properties, the temporal trends might interfere less than expected in the explanatory power of satellite images in spectral signals (Sedaghat et al., 2022; Xu and Zhai, 2023).

Bands 11 and 12 are situated close to strong water absorption features (1.4 and $1.9 \mu\text{m}$) (Sedaghat et al., 2022; Xu and Zhai, 2023). Consequently, these bands are often employed in the calculation of water-related spectral indices (Wei et al., 2019; Zhang et al., 2013). Their importance for SOM and CEC predictions is likely related to the effect of water in organic matter accumulation and in the decrease of weathering intensity, as previously discussed (see section 3.1).

Wang et al. (2020) had an interesting insight, suggesting that SOM should be predicted in sufficiently moistened samples, so to avoid the influence of variable water spectral signatures in drier soil samples, thereby arguing that wetting samples is easier than drying. Although their tests were performed in the laboratory, the concept might apply in remotely sensed data as well. The higher moisture content of alluvial soils might minimize the interference of absorption features from other materials by enhancing those related to water, which is closely related to SOM due to its high water retention capacity and due to the well documented tendency for SOM to accumulate under saturated soil conditions.

The high relevance of standard deviation pixels for models suggests that the temporal variation has an important role in explaining the spatial distribution of SOM and CEC. Water

retention trends, lush and healthy vegetation persistence (or not), and variations of color are examples of long-term traits that can be captured by temporal reflectance data. These and all physical and chemical changes undergone by soils and vegetation through time, as detected by remote sensing, are related to soil attributes (Ben-Dor et al., 2009), and seem to have contributed to predictions.

3.4. Random Forest predictions using variable sample density

When using all available soil samples ($n = 2145$), validation indicated models with high accuracy (Figure 9). Models showed no signs of overfitting. All tested models had R^2 values greater than 0.96. SOM models presented RMSE below 0.18%, and CEC models, below 2.56 cmolc kg^{-1} . Prediction performance was similarly good for models using pure bare soil images and those including vegetated images, as shown by the similarity among resulting maps and their distributions (Figure 9c, d).

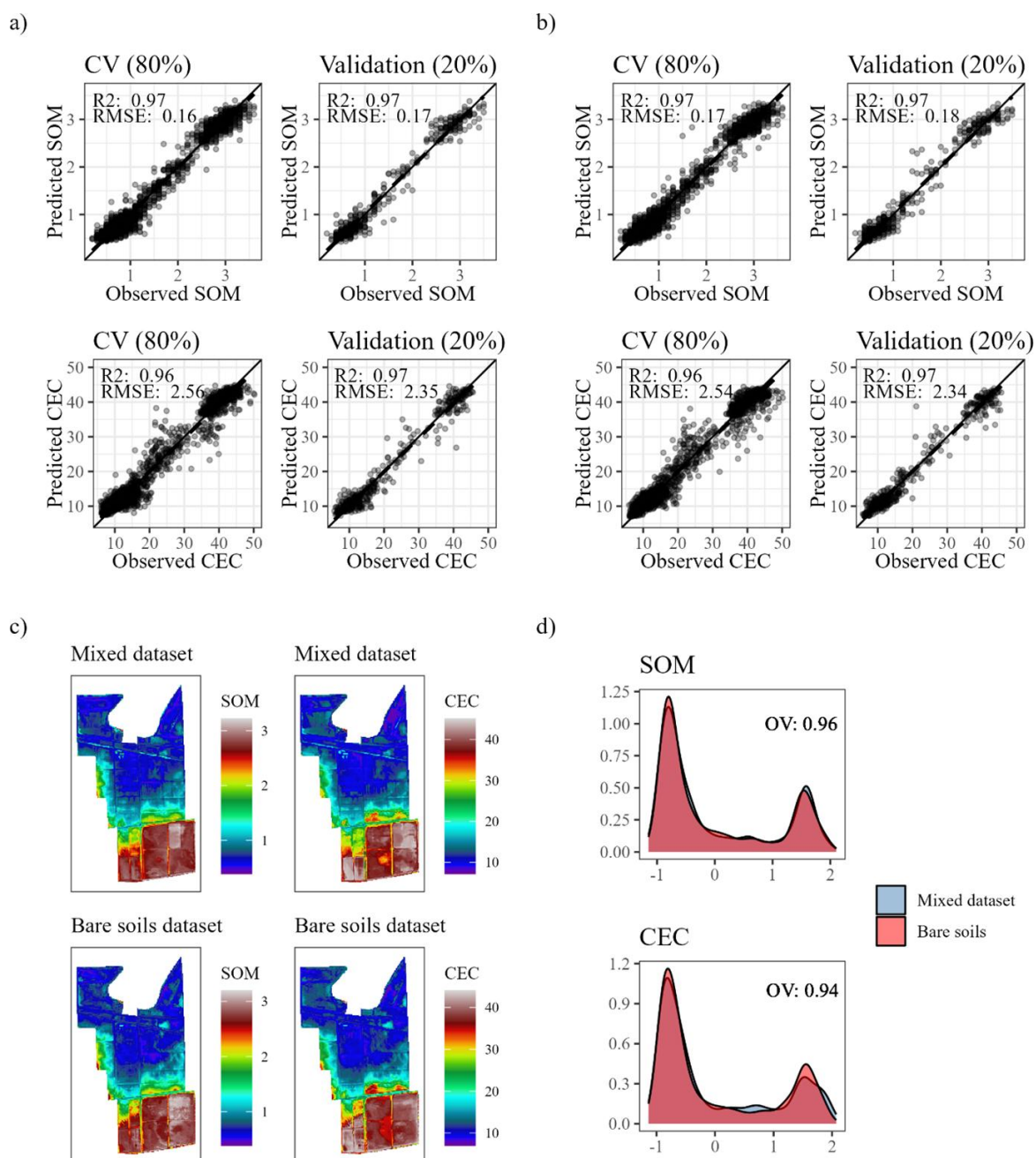


Figure 9. Random Forest predictions of soil organic matter (SOM, %) and cation exchange capacity (CEC, cmolc kg^{-1}) from 2145 samples from alluvial soils of Western Mississippi using a set of mixed Sentinel-2 images containing both bare soils and vegetated images (a, $n = 121$), and images comprising only bare soils (b, $n = 75$). Predicted maps (c) and the overlapping distributions of predictions (d) for both datasets. The continuous line represents the 1:1 line; the dashed line shows a least squares fit. CV – cross-validation; OV – overlapping index.

Sampling size affected prediction accuracy, albeit only prior to a certain number of samples used (Figure 10). When sampling size was greater than 100 samples, the increase of model performance presented asymptotic behavior in all sampling schemes. R^2 and RMSE for different sampling sizes are presented in Supplementary Material (Figs. S1 and S2), as are predicted maps (Figure S3). The decrease of prediction capacity with lower sample size is a

well-known, expected effect (Gomez et al., 2008), as models will possess less information to explain the spatial patterns of the predicted variables. Still, the number of samples sufficient to capture such variability is commonly an unknown parameter. By arbitrarily defining the threshold of $OV > 0.9$, results suggested that 100 samples in 250 ha provided accuracy comparable to that of models trained with over 2000 samples in the studied alluvial plain (Figure 10). The optimal sampling density was hence considered 0.4 samples per hectare, or 1 sample at every 2.5 hectares to obtain an OV of at least 0.9. Samples above this threshold consistently achieved $R^2 > 0.95$ (Figure S1).

Sampling strategy also had a significant impact on the prediction capacity of smaller datasets. Using both the mixed and bare soils dataset, cLHS consistently achieved $OV > 0.9$ with fewer samples (Figure 10 and Table 3). The least consistent strategy was random grid sampling. Results also indicated that using pixel-based statistics bands as inputs to select samples was efficient in improving prediction performance (Figure 11). Although the performance with fewer samples was worse (Figure 11a), after a sample size of 30 samples, models that used pixel-based statistics bands to select samples consistently achieved overlapping indices higher than 0.9. When using single-date bare soil bands, the same performance was only achieved for sample sizes larger than 100 (Figure 11b).

In this study area, the spatial structure of measured values was fairly simple. One group in the bimodal distribution of values was in the north and the second in the south. Given the range of R^2 values observed among sampling methods, in other scenarios where the spatial structure is more complex, a greater minimum number of points would likely be necessary to achieve an OV of 0.9. This study shows that pixel-based statistical layers and their distributions can provide important information for the cLHS selection of sample points that reduced the minimum number of samples required to obtain results with distributions similar to ground truth data ($OV > 0.9$). As such, information from pixel-based statistical layers of reflectance could prove beneficial for SOM and CEC mapping in more complex spatial structures given the correlations between SOM and CEC and spectral reflectance information. These statistical layers provide information on not only persistence, but variability in time and space. While out of the scope of this study, modified cLHS methods, such as spatial cLHS (Gao et al., 2016) and the Principal Components Analysis cost-constrained cLHS sampling method, could directly utilize both the statistical and spatial information from pixel-based statistics to inform sampling locations.

Table 3. The minimum number of samples that achieved an overlapping index (OV) higher than 0.9, and the percentage of tests that surpassed 0.9 for each sampling strategy and dataset. Lowest number of Minimum number of samples, and highest number of Percentage of tests, in each column highlighted in bold.

Sampling strategy	Mixed dataset		Bare soil dataset	
	SOM1	CEC2	SOM	CEC
Minimum number of samples to achieve OV3 of 0.9				
Regular grid	35	70	90	50
Random grid	150	250	50	40
cLHS4	30	45	30	90
Percentage of tests with OV above 0.9 (out of 31 tests)				
Regular grid	61.29	54.84	58.06	61.29
Random grid	41.94	41.94	54.84	54.84
cLHS	58.06	54.84	64.52	58.06

¹Soil organic matter; ²cation exchange capacity; ³Overlapping index; ⁴Conditioned Latin Hypercube Sampling.

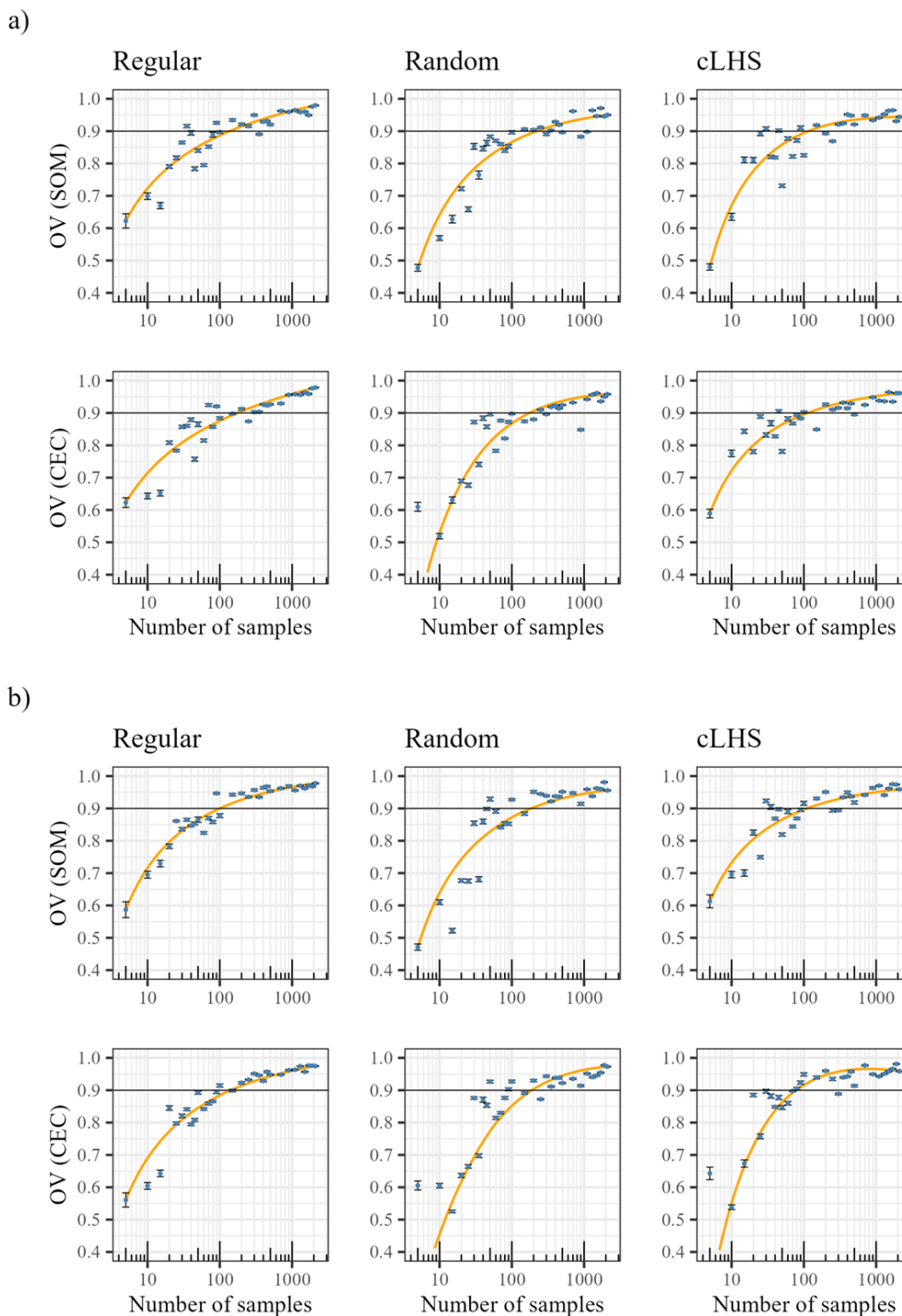


Figure 10. Comparison of the overlapping indices (OV) between Random Forest predictions of soil organic matter (SOM) and cation exchange capacity (CEC) for different sampling strategies. One indicates complete overlap; zero indicates no overlap. The overlapping was calculated comparing the model built with all samples ($n = 2145$) with models built with increasing sample densities. Values shown are the average from 1000 simulations. Error bars indicate the standard deviation from all simulations. Model training used two datasets: one composed of mixed Sentinel-2 images containing both bare soils and vegetated images (a, $n = 121$), and one comprised of only bare soil images (b, $n = 75$). The orange line represents a loess model. The black line indicates the arbitrarily chosen threshold of 0.9 to identify good models. cLHS – Conditioned Latin Hypercube Sampling.

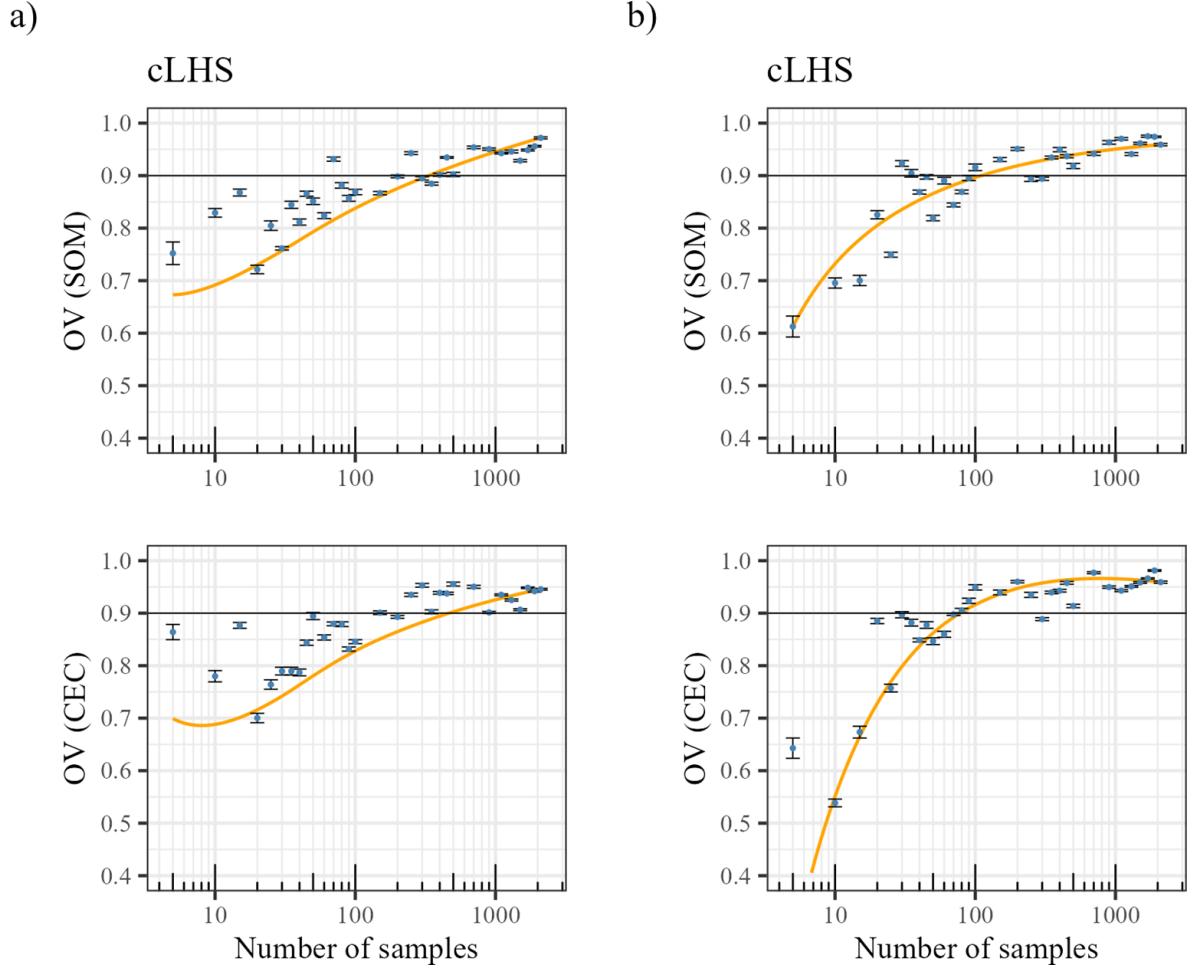


Figure 11. Comparison of overlapping index (OV) between Random Forest predictions of soil organic matter (SOM) and cation exchange capacity (CEC) when using samples selected by single-date bands (a) and pixel-based statistics (b) as inputs to Conditioned Latin Hypercube Sampling (cLHS). The overlapping was calculated by comparing the distributions of results from models built with all samples ($n = 2145$) with those of models built with increasing sample densities (from 5 to 2100 samples per 250 ha). Values shown are the average from 1000 simulations. Error bars indicate the standard deviation from all simulations. Model training used the dataset comprised of only bare soil images ($n = 75$). The orange line represents a loess model. The black line indicates the arbitrarily chosen threshold of 0.9 to identify good models.

4. Conclusions

Pixel-based statistics coupled with elevation data provided accurate SOM and CEC predictions. Pixel-based standard deviations were the most important variables for predictions. This indicates that the variability of soil spectral signatures over time has an important correlation with SOM and CEC. We further speculate that observed persistence and variability derived from reflectance pixel-based statistics could inform other soil chemical and physical properties, e.g., soil texture. The performance of models using mixed vegetated and bare soil data were similar, suggesting that pixel-based statistics modeling can be applied under row crop scenarios if bare soil datasets are not available. Sufficiently large temporal datasets of satellite

images might be useful even with partially vegetated areas when using pixel-based statistics as explanatory variables.

Maps were very accurate ($OV > 0.9$) when at least 100 samples were used, suggesting that a sampling density of 1 sample per 2.5 hectares was optimal. Sampling strategy had a significant impact on the prediction capacity of models. cLHS achieved better prediction results using fewer samples compared to regular grid and random sampling. Sampling strategy performance was ranked: cLHS > regular grid > random grid. The sample selection performance of cLHS was better when using pixel-based statistics bands as inputs compared to using unprocessed bare soil bands. Consequently, the combination of pixel-based statistics and cLHS for soil mapping is advised. A minimum sampling density of 1 sample per 2.5 hectares was able to capture the spatial structure of SOM and CEC at an OV of 0.9 and an R^2 of 0.95 in the studied alluvial plain. Pixel-based statistics provided persistent spatial information that could accurately predict SOM and CEC and enhance sampling efficiency. This persistent spatial structure can be used to improve precision agriculture management, carbon mapping and might be applicable to other soil properties and remotely sensed products. Similar global-coverage imagery could provide readily-calculable inputs for improved soil mapping and more optimal site sample selections in other agricultural row crop areas, globally.

Acknowledgements

Funding was provided by NIFA Award #2022-67013-37319. This research used resources provided by the SCINet project of the USDA Agricultural Research Service, ARS project number 0500-00093-001-00-D. Authors appreciate CNPq, CAPES and FAPEMIG for supporting this research, USDA-ARS Sustainable Water Management Research Unit employees James Dean and Roderick Patterson for their crucial assistance in soil sampling, and Poultry Production and Product Safety Unit employee Taylor Adams for sample laboratory analysis.

Supplementary Material

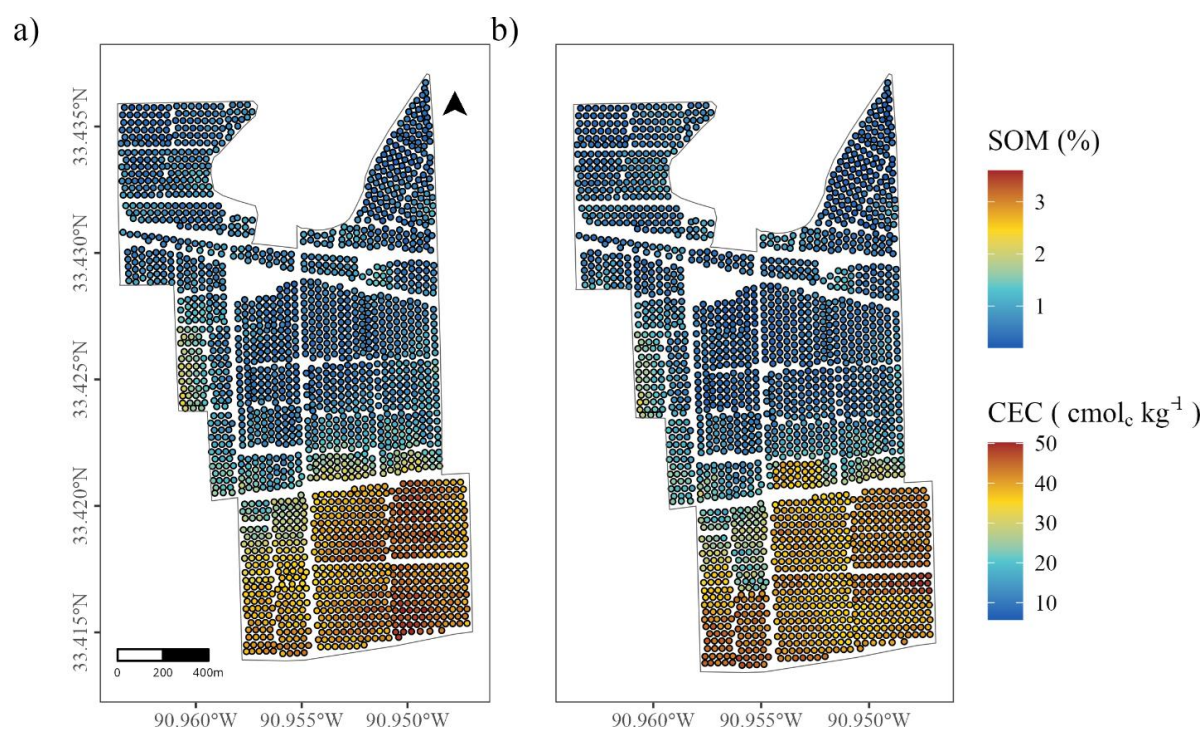


Figure S1. Measured soil organic matter (SOM, a) and cation exchange capacity (CEC, b) from 2145 surficial soil samples collected in the National Center for Alluvial Aquifer Research in Western Washington County, Mississippi, USA.

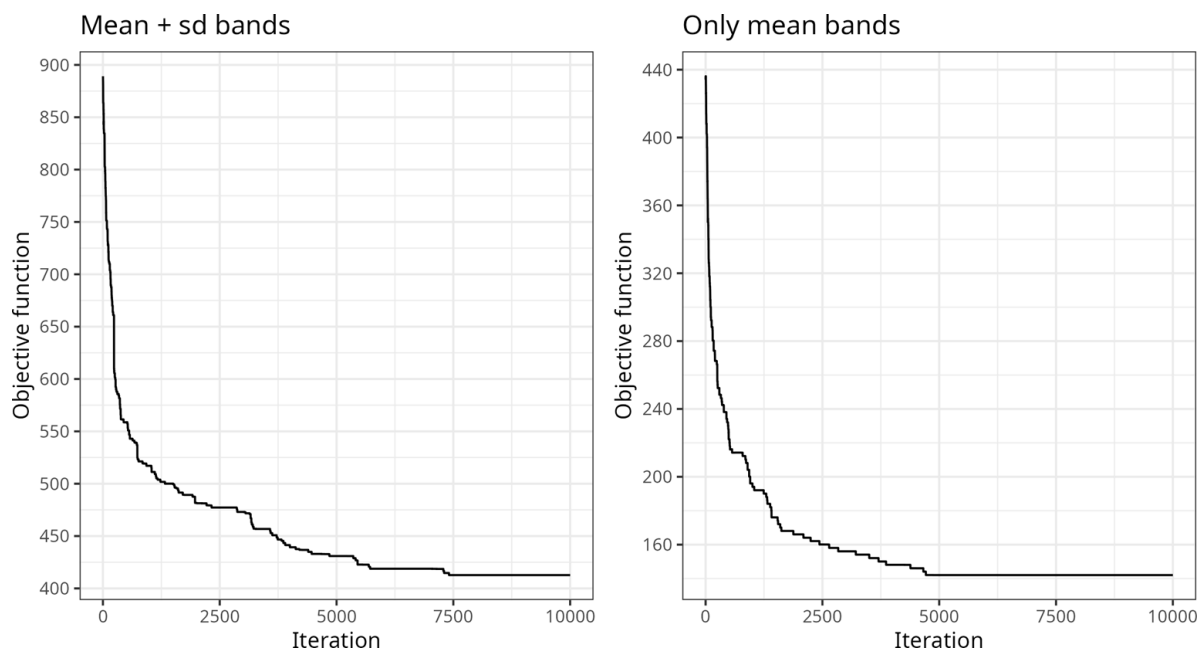


Figure S2. Loss of the objective function of conditioned Latin Hypercube Sampling (cLHS) after 10,000 iterations. This example chooses 100 sampling points for the study area (250 ha) while using as input variables the means and standard deviations (sd) of bands 2, 3, 4, 8, 11, and 12. The intention is to exemplify how increasing the number of variables (bands in this case) increases the loss and the required number of iterations to achieve an optimum selection of samples. When using many pixel-based statistics bands (mean, sd, percentiles, etc.), increasing the number of iterations may be necessary.

References

- Andrade, R., Mancini, M., Teixeira, A.F.S., Silva, S.H.G., Weindorf, D.C., Chakraborty, S., Guilherme, L.R.G., Curi, N., 2022. Proximal sensor data fusion and auxiliary information for tropical soil property prediction: Soil texture. *Geoderma* 422, 115936. <https://doi.org/10.1016/j.geoderma.2022.115936>
- Ben-Dor, E., Chabrillat, S., Demattê, J.A.M., 2018. Characterization of Soil Properties Using Reflectance Spectroscopy, in: Thenkabail, P.S., Lyon, J.G., Huete, A. (Eds.), *Fundamentals, Sensor Systems, Spectral Libraries, and Data Mining for Vegetation*. CRC Press, pp. 187–247. <https://doi.org/10.1201/9781315164151-8>
- 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 Sensing of Environment, Imaging Spectroscopy Special Issue* 113, S38–S55. <https://doi.org/10.1016/j.rse.2008.09.019>
- Bivand, R.S., Pebesma, E., Gomez-Rubio, V., 2013. *Applied spatial data analysis with R*, Second edition. Springer, New York.
- Brady, N.C., Weil, R.R., 2002. *The nature and properties of soils*, 13th ed. Prentice Hall, Upper Saddle River, New Jersey.
- 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, 568–579. <https://doi.org/10.2136/sssaj2009.0375>
- Castaldi, F., Chabrillat, S., Don, A., van Wesemael, B., 2019a. Soil organic carbon mapping using LUCAS topsoil database and Sentinel-2 data: An approach to reduce soil moisture and crop residue effects. *Remote Sensing* 11, 2121. <https://doi.org/10.3390/rs11182121>
- Castaldi, F., Chabrillat, S., van Wesemael, B., 2019b. Sampling strategies for soil property mapping using multispectral Sentinel-2 and hyperspectral EnMAP satellite data. *Remote Sensing* 11, 309. <https://doi.org/10.3390/rs11030309>
- Castaldi, F., Hueni, A., Chabrillat, S., Ward, K., Buttafuoco, G., Bomans, B., Vreys, K., Brell, M., van Wesemael, B., 2019c. Evaluating the capability of the Sentinel 2 data for soil organic carbon prediction in croplands. *ISPRS Journal of Photogrammetry and Remote Sensing* 147, 267–282. <https://doi.org/10.1016/j.isprsjprs.2018.11.026>
- Chi, J., Fan, Y., Wang, L., Putnis, C.V., Zhang, W., 2022. Retention of soil organic matter by occlusion within soil minerals. *Rev Environ Sci Biotechnol* 21, 727–746. <https://doi.org/10.1007/s11157-022-09628-x>
- Cortez, M.L.J., Alves, M.C., Carvalho, G.R., Pozza, E.A., 2020. Relationship between Sentinel-2 orbital data and in situ monitoring of coffee rust. *SN Appl. Sci.* 2, 1460. <https://doi.org/10.1007/s42452-020-03257-1>
- Diek, S., Fornallaz, F., Schaepman, M.E., De Jong, R., 2017. Barest pixel composite for agricultural areas using Landsat time series. *Remote Sensing* 9, 1245. <https://doi.org/10.3390/rs9121245>
- Diek, S., Schaepman, M.E., De Jong, R., 2016. Creating multi-temporal composites of airborne imaging spectroscopy data in support of digital soil mapping. *Remote Sensing* 8, 906. <https://doi.org/10.3390/rs8110906>

- Donkin, M.J., 1991. Loss-on-ignition as an estimator of soil organic carbon in A-horizon forestry soils. *Communications in Soil Science and Plant Analysis* 22, 233–241. <https://doi.org/10.1080/00103629109368411>
- Dvorakova, K., Shi, P., Limbourg, Q., van Wesemael, B., 2020. Soil organic carbon mapping from remote sensing: The effect of crop residues. *Remote Sensing* 12. <https://doi.org/10.3390/rs12121913>
- Fang, Q., Hong, H., Zhao, L., Kukolich, S., Yin, K., Wang, C., 2018. Visible and near-infrared reflectance spectroscopy for investigating soil mineralogy: A review. *Journal of Spectroscopy* 2018, e3168974. <https://doi.org/10.1155/2018/3168974>
- Florinsky, I.V., 2016. *Digital terrain analysis in soil science and geology*, 2nd ed. Elsevier, Amsterdam.
- Frampton, W.J., Dash, J., Watmough, G., Milton, E.J., 2013. Evaluating the capabilities of Sentinel-2 for quantitative estimation of biophysical variables in vegetation. *ISPRS Journal of Photogrammetry and Remote Sensing* 82, 83–92. <https://doi.org/10.1016/j.isprsjprs.2013.04.007>
- Gallez, A., Juo, A.S.R., Herbillon, A.J., 1976. Surface and charge characteristics of selected soils in the tropics. *Soil Science Society of America Journal* 40, 601–608. <https://doi.org/10.2136/sssaj1976.03615995004000040039x>
- Gao, B., Pan, Y., Chen, Z., Wu, F., Ren, X., Hu, M., 2016. A spatial conditioned Latin Hypercube Sampling method for mapping using ancillary data. *Transactions in GIS* 20, 735–754. <https://doi.org/10.1111/tgis.12176>
- Gessler, P.E., Moore, I.D., McKenzie, N.J., Ryan, P.J., 1995. Soil-landscape modelling and spatial prediction of soil attributes. *International journal of geographical information systems* 9, 421–432. <https://doi.org/10.1080/02693799508902047>
- Gholizadeh, A., Saberioon, M., Viscarra Rossel, R.A., Boruvka, L., Klement, A., 2020. Spectroscopic measurements and imaging of soil colour for field scale estimation of soil organic carbon. *Geoderma* 357, 113972. <https://doi.org/10.1016/j.geoderma.2019.113972>
- Gholizadeh, A., Zizala, D., Saberioon, M., Boruvka, L., 2018. Soil organic carbon and texture retrieving and mapping using proximal, airborne and Sentinel-2 spectral imaging. *Remote Sensing of Environment* 218, 89–103. <https://doi.org/10.1016/j.rse.2018.09.015>
- Gomez, C., Adeline, K., Bacha, S., Driessen, B., Gorretta, N., Lagacherie, P., Roger, J.M., Briottet, X., 2018. Sensitivity of clay content prediction to spectral configuration of VNIR/SWIR imaging data, from multispectral to hyperspectral scenarios. *Remote Sensing of Environment* 204, 18–30. <https://doi.org/10.1016/j.rse.2017.10.047>
- Gomez, C., Viscarra Rossel, R.A., McBratney, A.B., 2008. Soil organic carbon prediction by hyperspectral remote sensing and field vis-NIR spectroscopy: An Australian case study. *Geoderma* 146, 403–411. <https://doi.org/10.1016/j.geoderma.2008.06.011>
- Gregorutti, B., Michel, B., Saint-Pierre, P., 2017. Correlation and variable importance in random forests. *Stat Comput* 27, 659–678. <https://doi.org/10.1007/s11222-016-9646-1>
- Guo, L., Fu, P., Shi, T., Chen, Y., Zeng, C., Zhang, H., Wang, S., 2021a. Exploring influence factors in mapping soil organic carbon on low-relief agricultural lands using time series of remote sensing data. *Soil and Tillage Research* 210. <https://doi.org/10.1016/j.still.2021.104982>
- Guo, L., Sun, X., Fu, P., Shi, T., Dang, L., Chen, Y., Linderman, M., Zhang, G., Zhang, Y., Jiang, Q., Zhang, H., Zeng, C., 2021b. Mapping soil organic carbon stock by hyperspectral and

- time-series multispectral remote sensing images in low-relief agricultural areas. *Geoderma* 398. <https://doi.org/10.1016/j.geoderma.2021.115118>
- Heller Pearlshien, D., Ben-Dor, E., 2020. Effect of organic matter content on the spectral signature of iron oxides across the VIS–NIR spectral region in artificial mixtures: An example from a red soil from Israel. *Remote Sensing* 12, 1960. <https://doi.org/10.3390/rs12121960>
- Hengl, T., Miller, M.A.E., Krizan, J., Shepherd, K.D., Sila, A., Kilibarda, M., Antonijevic, O., Glusica, L., Dobermann, A., Haefele, S.M., McGrath, S.P., Acquah, G.E., Collinson, J., Parente, L., Sheykhmousa, M., Saito, K., Johnson, J.-M., Chamberlin, J., Silatsa, F.B.T., Yemefack, M., Wendt, J., MacMillan, R.A., Wheeler, I., Crouch, J., 2021. African soil properties and nutrients mapped at 30 m spatial resolution using two-scale ensemble machine learning. *Scientific Reports* 11. <https://doi.org/10.1038/s41598-021-85639-y>
- Hijmans, R.J., 2023. *Terra: Spatial data analysis*.
- Kaiser, M., Ellerbrock, R.H., Gerke, H.H., 2008. Cation exchange capacity and composition of soluble soil organic matter fractions. *Soil Science Society of America Journal* 72, 1278–1285. <https://doi.org/10.2136/sssaj2007.0340>
- Keenor, S.G., Rodrigues, A.F., Mao, L., Latawiec, A.E., Harwood, A.R., Reid, B.J., 2021. Capturing a soil carbon economy. *Royal Society Open Science* 8, 202305. <https://doi.org/10.1098/rsos.202305>
- Keskin, H., Grunwald, S., Harris, W.G., 2019. Digital mapping of soil carbon fractions with machine learning. *Geoderma* 339, 40–58. <https://doi.org/10.1016/j.geoderma.2018.12.037>
- Kida, M., Fujitake, N., 2020. Organic carbon stabilization mechanisms in mangrove soils: A review. *Forests* 11, 981. <https://doi.org/10.3390/f11090981>
- Kindler, R., Siemens, J., Kaiser, K., Walmsley, D.C., Bernhofer, C., Buchmann, N., Cellier, P., Eugster, W., Gleixner, G., Grünwald, T., Heim, A., Ibrom, A., Jones, S.K., Jones, M., Klumpp, K., Kutsch, W., Larsen, K.S., Lehuger, S., Loubet, B., McKenzie, R., Moors, E., Osborne, B., Pilegaard, K., Rebmann, C., Saunders, M., Schmidt, M.W.I., Schrumpf, M., Seyfferth, J., Skiba, U., Soussana, J.-F., Sutton, M.A., Tefs, C., Vowinkel, B., Zeeman, M.J., Kaupenjohann, M., 2011. Dissolved carbon leaching from soil is a crucial component of the net ecosystem carbon balance. *Global Change Biology* 17, 1167–1185. <https://doi.org/10.1111/j.1365-2486.2010.02282.x>
- Kravchenko, A.N., 2003. Influence of Spatial Structure on Accuracy of Interpolation Methods. *Soil Science Society of America Journal* 67, 1564–1571. <https://doi.org/10.2136/sssaj2003.1564>
- Kremer, L.P.M., 2019. *ggpointdensity: A cross between a 2D density plot and a scatter plot*.
- Kuhn, M., 2008. Building predictive models in R using the caret package. *Journal of Statistical Software* 28, 1–26. <https://doi.org/10.18637/jss.v028.i05>
- Lal, R., Monger, C., Nave, L., Smith, P., 2021. The role of soil in regulation of climate. *Philosophical Transactions of the Royal Society B: Biological Sciences* 376, 20210084. <https://doi.org/10.1098/rstb.2021.0084>
- Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. *R News* 2.
- Lin, C., Zhu, A.-X., Wang, Z., Wang, X., Ma, R., 2020. The refined spatiotemporal representation of soil organic matter based on remote images fusion of Sentinel-2 and Sentinel-3. *International Journal of Applied Earth Observations and and Geoinformation* 89. <https://doi.org/10.1016/j.jag.2020.102094>

- Liu, Y., Qian, J., Yue, H., 2021. Comprehensive evaluation of Sentinel-2 red edge and shortwave-infrared bands to estimate soil moisture. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 14, 7448–7465. <https://doi.org/10.1109/JSTARS.2021.3098513>
- Luo, C., Zhang, X., Wang, Y., Men, Z., Liu, H., 2022. Regional soil organic matter mapping models based on the optimal time window, feature selection algorithm and Google Earth Engine. *Soil and Tillage Research* 219. <https://doi.org/10.1016/j.still.2022.105325>
- Mehlich, A., 1984. Mehlich 3 soil test extractant: A modification of Mehlich 2 extractant. *Communications in Soil Science and Plant Analysis* 15, 1409–1416. <https://doi.org/10.1080/00103628409367568>
- Minasny, B., McBratney, A.B., 2006. A conditioned latin hypercube method for sampling in the presence of ancillary information. *Computers & Geosciences* 32, 1378–1388. <https://doi.org/10.1016/j.cageo.2005.12.009>
- Moore, I.D., Gessler, P.E., Nielsen, G.A., Peterson, G.A., 1993. Soil attribute prediction using terrain analysis. *Soil Science Society of America Journal* 57, 443–452. <https://doi.org/10.2136/sssaj1993.03615995005700020026x>
- Oorts, K., Vanlauwe, B., Merckx, R., 2003. Cation exchange capacities of soil organic matter fractions in a Ferric Lixisol with different organic matter inputs. *Agriculture, Ecosystems & Environment, Balanced Nutrient Management Systems for cropping systems in the tropics: From concept to practice* 100, 161–171. [https://doi.org/10.1016/S0167-8809\(03\)00190-7](https://doi.org/10.1016/S0167-8809(03)00190-7)
- Pastore, M., Calcagni, A., 2019. Measuring Distribution Similarities Between Samples: A Distribution-Free Overlapping Index. *Frontiers in Psychology* 10.
- Pebesma, E., 2018. Simple features for R: Standardized support for spatial vector data. *The R Journal* 10, 439–446. <https://doi.org/10.32614/RJ-2018-009>
- Pebesma, E.J., Bivand, R.S., 2005. Classes and methods for spatial data in R. *R News* 5, 9–13.
- Pincus, L.N., Ryan, P.C., Huertas, F.J., Alvarado, G.E., 2017. The influence of soil age and regional climate on clay mineralogy and cation exchange capacity of moist tropical soils: A case study from Late Quaternary chronosequences in Costa Rica. *Geoderma* 308, 130–148. <https://doi.org/10.1016/j.geoderma.2017.08.033>
- Quantum Spatial, 2022. Mississippi Delta 2018 Work Unit 78037 (metadata).
- Quantum Spatial, 2020. MS_MississippiDelta_2018_D18 Lidar Processing Report.
- R Core Team, 2022. R: A Language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
- Ranghetti, L., Boschetti, M., Nutini, F., Busetto, L., 2020. Sen2r: An R toolbox for automatically downloading and preprocessing Sentinel-2 satellite data. *Computers & Geosciences* 139, 104473. <https://doi.org/10.1016/j.cageo.2020.104473>
- Roudier, P., 2011. cLHS: An R package for conditioned Latin hypercube sampling.
- Santos, U.J., 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.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>
- Saucier, R.T., 1974. Quaternary geology of the Lower Mississippi Valley. *Arkansas Archeological Survey, Research Series* 6.

- Saucier, R.T., 1994. Geomorphology and quaternary geologic history of the Lower Mississippi Valley, Geomorphology and Quaternary Geologic History of the Lower Mississippi Valley. U.S. Army Engineer Waterways Experiment Station.
- Schaetzl, R.J., Anderson, S., 2015. Soils: Genesis and geomorphology, 2nd Revised. ed. Cambridge University Press, New York.
- Sedaghat, A., Shahrestani, M.S., Noroozi, A.A., Fallah Nosratabad, A., Bayat, H., 2022. Developing pedotransfer functions using Sentinel-2 satellite spectral indices and Machine learning for estimating the surface soil moisture. *Journal of Hydrology* 606, 127423. <https://doi.org/10.1016/j.jhydrol.2021.127423>
- Shafizadeh-Moghadam, H., Minaei, F., Talebi-khiyavi, H., Xu, T., Homaei, M., 2022. Synergetic use of multi-temporal Sentinel-1, Sentinel-2, NDVI, and topographic factors for estimating soil organic carbon. *Catena* 212. <https://doi.org/10.1016/j.catena.2022.106077>
- Silvero, N.E.Q., Demattê, J.A.M., Amorim, M.T.A., Santos, N.V. dos, Rizzo, R., Safanelli, J.L., Poppiel, R.R., Mendes, W.S., Bonfatti, B.R., 2021. Soil variability and quantification based on Sentinel-2 and Landsat-8 bare soil images: A comparison. *Remote Sensing of Environment* 252, 112117. <https://doi.org/10.1016/j.rse.2020.112117>
- Thomas, F., Petzold, R., Landmark, S., Mollenhauer, H., Becker, C., Werban, U., 2022. Estimating Forest Soil Properties for Humus Assessment—Is Vis-NIR the Way to Go? *Remote Sensing* 14, 1368. <https://doi.org/10.3390/rs14061368>
- U.S. Geological Survey, 2022b. 3D elevation program 1-meter resolution digital elevation model.
- U.S. Geological Survey, 2022a. USGS 1 Meter 15 X69y371 MS_Central_Delta_2018_D18.
- Ustin, S.L., Ryerson, R.A., Rencz, A.N., 2004. Manual of remote sensing, remote sensing for natural resource management and environmental monitoring, Manual of remote sensing. Wiley.
- Viscarra Rossel, R.A., Minasny, B., Roudier, P., McBratney, A.B., 2006. Colour space models for soil science. *Geoderma* 133, 320–337. <https://doi.org/10.1016/j.geoderma.2005.07.017>
- Wang, Y.-P., Lee, C.-K., Dai, Y.-H., Shen, Y., 2020. Effect of wetting on the determination of soil organic matter content using visible and near-infrared spectrometer. *Geoderma* 376, 114528. <https://doi.org/10.1016/j.geoderma.2020.114528>
- Wei, Z., Meng, Y., Zhang, W., Peng, J., Meng, L., 2019. Downscaling SMAP soil moisture estimation with gradient boosting decision tree regression over the Tibetan Plateau. *Remote Sensing of Environment* 225, 30–44. <https://doi.org/10.1016/j.rse.2019.02.022>
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R., Grolemond, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T.L., Miller, E., Bache, S.M., Müller, K., Ooms, J., Robinson, D., Seidel, D.P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., Yutani, H., 2019. Welcome to the tidyverse. *Journal of Open Source Software* 4, 1686. <https://doi.org/10.21105/joss.01686>
- Winzeler, H.E., Owens, P.R., Joern, B.C., Camberato, J.J., Lee, B.D., Anderson, D.E., Smith, D.R., 2008. Potassium fertility and terrain attributes in a Fragiudalf drainage catena. *Soil Science Society of America Journal* 72, 1311–1320. <https://doi.org/10.2136/sssaj2007.0382>
- Winzeler, H.E., Owens, P.R., Read, Q.D., Libohova, Z., Ashworth, A., Sauer, T., 2022. Topographic wetness index as a proxy for soil moisture in a hillslope catena: Flow algorithms and map generalization. *Land* 11, 2018. <https://doi.org/10.3390/land11112018>

- Xu, X., Zhai, X., 2023. Mapping soil organic matter content during the bare soil period by using satellite data and an improved deep learning network. *Sustainability* 15, 323. <https://doi.org/10.3390/su15010323>
- Yuzugullu, O., Lorenz, F., Froehlich, P., Liebisch, F., 2020. Understanding fields by remote sensing: Soil zoning and property mapping. *Remote Sensing* 12. <https://doi.org/10.3390/rs12071116>
- Zeraatpisheh, M., Garosi, Y., Owliaie, H.R., Ayoubi, S., Taghizadeh-Mehrjardi, R., Scholten, T., Xu, M., 2022. Improving the spatial prediction of soil organic carbon using environmental covariates selection: A comparison of a group of environmental covariates. *Catena* 208. <https://doi.org/10.1016/j.catena.2021.105723>
- Zhang, N., Hong, Y., Qin, Q., Liu, L., 2013. VSDI: A visible and shortwave infrared drought index for monitoring soil and vegetation moisture based on optical remote sensing. *International Journal of Remote Sensing* 34, 4585–4609. <https://doi.org/10.1080/01431161.2013.779046>

Chapter III: Multinational prediction of soil organic carbon and texture via proximal sensors

Accepted in the Soil Science Society of America Journal in 30 August 2023.
<https://doi.org/10.1002/saj2.20593>

Marcelo Mancini^a, Renata Andrade^a, Sérgio Henrique Godinho Silva^a, Rogério Borguete Alves Rafael^b, Swagata Mukhopadhyay^c, Bin Li^d, Somsubhra Chakraborty^c, Luiz Roberto Guimarães Guilherme^a, Autumn Acree^e, David C. Weindorf^f, Nilton Curi^a

^aDepartment of Soil Science, Federal University of Lavras, Lavras, Minas Gerais, Brazil

^bDepartment of Rural Engineering, University Eduardo Mondlane, Maputo, Mozambique

^cAgricultural and Food Engineering Department, Indian Institute of Technology, Kharagpur, India

^dDepartment of Experimental Statistics, Louisiana State University, Baton Rouge, Louisiana, USA

^eEnvironmental Laboratory, US Army Engineer Research and Development Center, Vicksburg, Mississippi, USA

^fDepartment of Earth and Atmospheric Sciences, Central Michigan University, Mount Pleasant, Michigan, USA

Abstract

Novel technologies help to monitor environmental impact of human activities, but tests involving datasets from several countries, encompassing a large variability of soil properties, are still scarce. This study utilized proximal sensors to predict soil organic carbon (OC) and soil texture of samples from Brazil, France, India, Mozambique, and United States of America. 1,749 samples were analyzed by portable X-ray fluorescence spectrometry (pXRF) and visible near-infrared diffuse reflectance spectroscopy (Vis-NIR). Sand ($R^2 = 0.89$), silt (0.87) and clay (0.84) predictions were very accurate, despite contrasting climates, soil parent materials, and weathering degrees. Soil OC predictions were similarly successful (0.74) using samples from five countries. pXRF was the optimal sensor for soil texture predictions. The addition of international data may improve local models. Proximal soil sensing can be successfully used with a multinational soil database offering a clean, rapid, and accurate alternative to estimate soil texture and OC with international datasets.

Keywords: pXRF; Vis-NIR; carbon stocks; machine learning; prediction models.

1. Introduction

Worldwide, the adoption of novel technologies in support of sustainable management has increased; the sustainability of agroecosystems chief among them. One of the most important components of modern agriculture is the precise mapping of soil properties as it is essential for specific soil management (Resende et al., 2014). Special emphasis is given to soil texture (mineral particle size distribution encompassing sand, silt, and clay fractions) and soil organic carbon (OC). Both are crucial properties in soil systems due to their strong relation to associated soil properties (e.g., cation exchange capacity, pH, etc.), and their implications on C sequestration in support of the Sustainable Development Goals of the United Nations (SDG 13 – Climate action; SDG 15 Life on land).

Soil texture is highly influential in water infiltration and retention, soil aeration, plant nutrient availability, cation exchange capacity, and ease of compaction (Tümsavaş et al., 2019). Soil texture also provides insights into pedogenetic processes and factors of soil formation (Phogat et al., 2015). Soil texture has a unique synergy with OC stocks in soils (Schaetzl & Anderson, 2015) in that OC is protected from oxidation by soil aggregates whose stability depends on soil texture and mineralogy (Angst et al., 2018). Adopting management practices that are adequate for different soil texture classes prevents land degradation and helps preserve (and even further increase) soil OC stocks (Yudina et al., 2022). This is relevant because soils play an important role as carbon sinks (Zhang et al., 2018). Soil OC is an important indicator of soil health and fertility, influencing plant growth, soil aggregation, and soils' ability to retain nutrients (cation exchange capacity) (Rahman et al., 2022).

However, obtaining detailed soil texture and OC data over large areas is often not viable. Agricultural lands are extensive, whereby dense sampling grids can result in excessive laboratory analyses. These analyses are expensive, time consuming and use many chemical reagents (Silva et al., 2021; Weindorf et al., 2014). Traditional laboratory analyses commonly increase crop production costs and the time needed for decision making.

Fortunately, newer sensing technologies in tandem with high performance modeling techniques have emerged as a viable alternative to achieve highly detailed spatial information about soils with minimum sample preparation and wet-chemistry analysis (Chatterjee et al., 2021). Proximal sensors have proven successful for predicting multiple soil properties across the world, especially the portable X-ray fluorescence spectrometer (pXRF) and the visible near-infrared spectrometer (Vis-NIR) (Gozukara et al., 2022). pXRF provides multi-elemental quantification from Mg to U in the periodic table simultaneously, features wide dynamic range (e.g., low mg kg⁻¹ to high % elemental abundance), and requires minimal or no sample

preparation (Silva et al., 2021; Weindorf et al., 2014); Vis-NIR determines the spectral signature of the analyzed material from visible to near-infrared regions of the electromagnetic spectrum. Both sensors are field portable and have been increasingly adopted by the soil science community (Benedet et al., 2021; Teixeira et al., 2022).

Both pXRF and Vis-NIR have proven successful over different soil classes developed from varying parent materials, under multiple land uses, and in different parts of the world after the development of local prediction models (Andrade et al., 2022; Mancini et al., 2022; Oliveira et al., 2022; Silvero et al., 2022; Souza et al., 2022; Teixeira et al., 2021; Teixeira et al., 2022). Yet few investigations have tested prediction model applicability across multinational soil datasets, encompassing soils with wide variability in properties. Soils from temperate regions tend to present more silt and sand contents, with clay mineralogy dominated by primary minerals including multiple 2:1 clay minerals. Conversely, soils from tropical regions usually have higher clay content, very low silt contents, lack of 2:1 minerals and feature abundance of Fe and Al oxides (Buol et al., 2011; Resende et al., 2014; Schaetzl & Anderson, 2015; Silva et al., 2021). These differences cause contrasting behavior of soils in terms of water retention, infiltration, structure, cation exchange capacity, natural fertility, etc. Will proximal sensor-based predictions perform well when working with large soil datasets across countries with contrasting conditions?

Given the advanced use of sensors worldwide and focus on pushing forward their practical use under different conditions, this study aimed to predict soil texture and OC in soils from countries located in South America, North America, Europe, Sub-Saharan Africa, and Asia using pXRF and Vis-NIR data to inform predictive machine learning algorithms. Specifically, the goals of the research were to: 1) evaluate the accuracy of predictive modeling using pXRF and Vis-NIR data both in isolation and in combination, and 2) create prediction models using data from each country separately and compare their predictive performance to a combined, multinational model. We hypothesized that the accuracy of a multinational prediction model for soil texture and OC would rival the performance of country-specific prediction models. This is the first step towards testing this approach globally to certify its effectiveness before ultimately proposing its adoption for soils worldwide with deference to local/regional adjustments, if needed.

2. Materials and methods

2.1. Sampling locations and available datasets

Soil samples were obtained from five countries: Brazil, United States of America (USA), France, Mozambique and India (Figure 1). A total of 1,749 surface soil samples (0-20 cm) were evaluated. Mean temperature and annual precipitation of studied soils samples per country were extracted from the WorldClim database (Fick & Hijmans, 2017) and are presented in Table 1. Samples in Brazil were collected under native vegetation ($n = 70$), pasture (118), forest plantation (*Eucalyptus* sp.) (186) and agricultural areas (99). Samples from Mozambique (141), USA (300), and France (149) were collected predominantly in agricultural areas. Of the 686 soil samples from India, 335 were collected in landfill-adjacent agricultural areas; the other 351 were collected in agricultural areas. The high variability of land uses, soil management, soil properties and parent materials will help to ascertain if the predictive model is able to predict soil attributes across different countries.

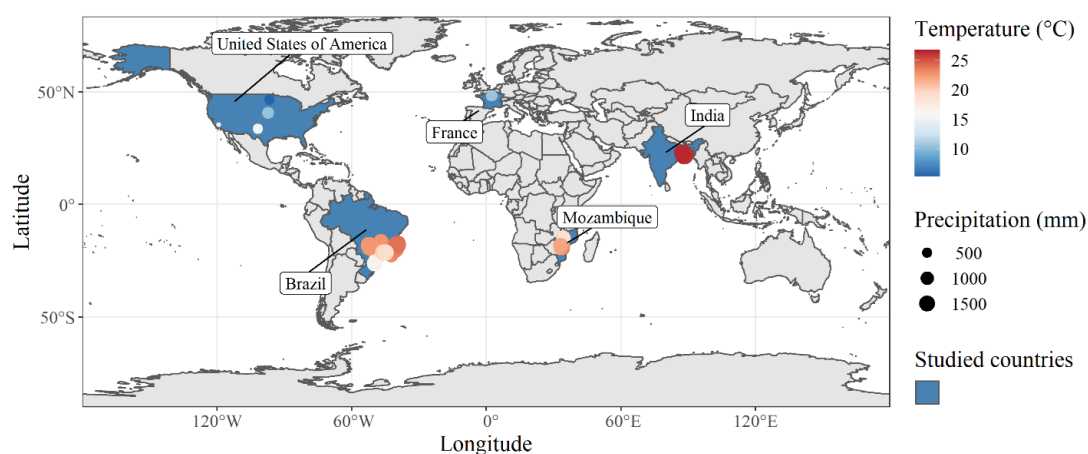


Figure 1. Location of studied soil samples illustrating their respective mean temperatures and mean annual precipitation values.

Table 1. Mean temperature and annual precipitation of studied soils samples per country extracted from the WorldClim database.

Country	Max.	Min.	Mean	SD
-----Precipitation (mm)-----				
Mozambique	1590.00	604.00	1056.77	119.91
Brazil	1610.00	1038.00	1446.63	159.20
France	647.00	642.00	646.21	0.89
India	1786.00	1191.00	1672.62	128.16
US	729.00	170.00	482.00	199.85
-----Temperature (oC)-----				
Mozambique	26.5	18.1	22.1	1.96
Brazil	24.5	16.4	20.9	2.26
France	10.6	10.5	10.6	0.01
India	26.8	26.3	26.5	0.06
US	17.2	5.3	12.2	4.66

Datasets presented the following variables: particle size fractions (soil texture), OC, elemental contents from pXRF and Vis-NIR (350-2500 nm); more details about the analytical procedures follow. Available data for collected soil samples is presented in Table 2.

Table 2. Summary of available data for collected soil samples across five countries.

Country	Organic carbon		Soil texture	
	pXRF	Vis-NIR	pXRF	Vis-NIR
Number of soil samples				
Mozambique	141	0	0	0
Brazil	473	410	426	363
France	149	149	149	149
India	686	686	336	336
US	300	300	300	300

2.2. Laboratory analyses

Samples were air-dried and sieved through a 2 mm mesh (air-dried fine earth, ADFE). Particle size distribution was determined by the pipette method (Brazil, India and France samples) and hydrometer method using a 152H hydrometer (US samples) (Gee & Bauder, 1986). Soil samples were classified per the USDA soil texture triangle (Soil Survey Staff, 2014).

In samples from Brazil and Mozambique, OC was measured using oxidation by dichromate acid and titration with ammonium ferrous sulfate (Walkley & Black, 1934). Data from Brazil were originally stored as soil organic matter. Values were converted to OC by dividing them by the van Bemmelen factor (1.724), commonly used by laboratories in Brazil (Pribyl, 2010). Samples from other countries were further sieved (0.25 mm mesh) and analyzed via high temperature combustion, following Soil Survey Staff (Soil Survey Staff, 2022). The different analysis methods used herein will serve as a test of models' capability of delivering accurate predictions when trained with combined data from different countries, similarly to tests

conducted by Khaledian et al. (2018) for global soil P predictions. This is essential, as different countries frequently utilize different laboratory methods.

For landfill-adjacent agricultural samples from India, three randomly selected samples were investigated for mineralogical composition via X-ray diffraction (XRD) analysis with BRUKER D2 PHASER (Bruker, Billerica, MA, USA). These three samples were further investigated for elemental composition using a ZEISS EVO 60 scanning electron microscope (SEM) (Carl ZEISS SMT, Germany) coupled with an Oxford INCA PentaFET x3 energy dispersive spectroscopy (EDS) detector (Oxford Instrument, UK). The SEM works with tungsten filament with a maximum acceleration voltage of 30 kV.

2.3. Proximal sensors

Most ADFE samples ($n = 1,276$) were analyzed by an Olympus DP-6000 Premium pXRF (Waltham, MA, USA); some samples from Brazil ($n = 473$) and Mozambique ($n=141$) were analyzed by a Bruker S1 Titan (Billerica, MA, USA). Thus, it will be also possible to test whether predictions will be accurate when combining data from different pXRF instruments.

Samples were scanned in Dual Soil Mode for 60 s in triplicate. Mean values from the triplicates were used for all statistical analysis. The methodology is described in detail by Weindorf and Chakraborty (Weindorf & Chakraborty, 2020) and Silva et al. (Silva et al., 2021). The accuracy of the equipment was verified using the certified reference samples 2710a and 2711a from the National Institute of Standards and Technology (NIST).

Next, ADFE samples were analyzed by a Spectral Evolutions PSR-3500 spectroradiometer (Haverhill, MA, USA). The equipment measures spectral data between 350 and 2,500 nm at 1 nm resolution (2,151 bands). Analyses were performed as follows: a probe with a 5 W halogen lamp was placed upon the leveled soil samples for 10 s three times, rotating the sample at every repetition. The results of the triplicates were averaged and the mean value was used for all statistical analysis. After scanning each sample, the equipment was calibrated by a traceable radiance calibration panel made of polytetrafluoroethylene (PTFE) standardized by NIST.

2.4. Data analyses and modeling

Data manipulation and processing were facilitated via R software (R Development Core Team, 2021). Specifically, the following packages were used: tidyverse (Wickham et al., 2019), prospectr (Stevens & Ramirez-Lopez, 2022) resemble (Ramirez-Lopez et al., 2023), rnaturalearthdata (South, 2017), caret (Kuhn, 2008), sp (Bivand et al., 2013), sf (Pebesma,

2018), raster (Hijmans, 2022), ggpubr (Kassambara, 2020), patchwork (Pedersen, 2022) and RColorBrewer (Neuwirth, 2014).

The following pXRF predictor variables were selected: K, Ca, Ti, V, Cr, Mn, Fe, Ni, Cu, Zn, As, Rb, Sr, Zr, Ba, Pb. For each different dataset (Table 3), recursive feature elimination was used to select the optimal pXRF variables. This algorithm fits the model using all variables and ranks their importance for the predictions. Next, it tests different subsets of the most important variables until it finds the best performing combination. The caret package was used for this procedure (Kuhn, 2008). Variables importance was ranked using the Random Forest algorithm (further discussed below) (Liaw & Wiener, 2002).

Table 3. Number of samples in each dataset used for training and validation of soil organic carbon (OC) and texture prediction models using pXRF and Vis-NIR data from five countries.

Carbon (CC) and texture prediction models using pXRF and Vis-NIR data from five countries.							
Country	OC						Total samples
	pXRF		Vis-NIR		pXRF +Vis-NIR		
	n train.1	n valid.2	n train.	n valid.	n train.	n valid.	
All countries*	1401	348	-	-	-	-	1749
All countries (except Mozambique)**	1237	308	1237	308	1237	308	1545
Mozambique	114	27	114	27	114	27	141
Brazil	330	80	330	80	330	80	410
France	121	28	121	28	121	28	149
India	551	135	551	135	551	135	686
US	240	60	240	60	240	60	300
Texture							
	pXRF		Vis-NIR		pXRF +Vis-NIR		
	n train.	n valid.	n train.	n valid.	n train.	n valid.	
All countries	920	228	920	228	920	228	1148
Mozambique	-	-	-	-	-	-	-
Brazil	291	72	291	72	291	72	363
France	121	28	121	28	121	28	149
India	271	65	271	65	271	65	336
US	240	60	240	60	240	60	300

¹Number of samples used for training.

²Number of samples used for validation.

*Included samples from Mozambique and Brazil that did not have Vis-NIR data. These samples were not used in other models so that they shared the same dataset and results were comparable.

**Mozambique samples do not have Vis-NIR data.

A dimensionality reduction technique was used to reduce the number of spectral variables while retaining the most relevant information from the 2,151 bands. Principal component analysis with non-linear iterative partial least squares was used (PCA-NIPALS, Wold, 1975). The optimal number of principal components was selected using Optimal Component Selection (Ramirez-Lopez et al., 2013). For visualization purposes, the continuum removal was applied to spectral data from each country. This method consists in creating a

convex hull that serves as a new baseline for the spectrum. The hull baseline receives value 1. The resulting preprocessed spectrum is calculated based on its distance from the new reference.

Predictor variables from pXRF selected by recursive feature selection and principal components from Vis-NIR were used to train machine learning models to predict soil OC and texture (clay, silt, and sand contents). The utilized algorithm was Random Forest. Models used the following parameters: nodesize = 5, number of trees = 500 and mtry = 1/3 of the number of predictor variables following Liaw and Wiener (Liaw & Wiener, 2002). Datasets were randomly split into training (80%) and validation sets (20%). Validation was performed by calculating the coefficient of determination (R^2), the root mean square error (RMSE), and the concordance correlation coefficient (ρ_c) (Lin, 1989) (Eqs. 1, 2 and 3).

$$R^2 = 1 - \frac{\sum_{i=1}^n (m_i - y_i)^2}{\sum_{i=1}^n (m_i - \bar{m})^2} \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - m_i)^2} \quad (2)$$

$$\rho_c = \frac{2\sigma_{12}}{(\bar{m}_1 - \bar{m}_2)^2 + \sigma_1^2 + \sigma_2^2} \quad (3)$$

where y_i is the predicted value, m_i is the observed value, and $\bar{m}$ is the mean of observations. In Eq. 3, 1 and 2 are the compared distributions (in this case, predicted and observed values), σ is the standard deviation, and σ_{12} is their covariance.

There were no Vis-NIR data available for samples from Mozambique and for some samples from Brazil (Table 2). Modeling was therefore done in two parts: first, in order to compare the performance of pXRF and Vis-NIR, modeling was done using only the same samples that had data from both sensors available (excluding Mozambique and some samples from Brazil). This was done for each country and with data from all countries combined ($n = 1,545$). Lastly, one OC model was created using only pXRF data. Soil OC analyses and pXRF data were available for all samples ($n = 1,749$), comprising all five countries (Brazil, USA, France, India, and Mozambique). A summary showing how datasets were organized for modeling is presented at Table 3.

3. Results

3.1. Multinational descriptive statistics for soil texture and soil organic carbon

A wider range of clay content was observed in Brazil (Figure 1 and Supplemental Figure S1). In India and Brazil, clay content was higher than in the other countries. Sand contents were high for some samples from all countries, while silt contents were higher (>50%) in most

samples from France, USA, and India. Silt content was higher in the USA and France samples. Higher soil OC contents were observed for samples from India, which were mostly collected from landfill areas (Figure 2). Native forest and pastures presented higher soil OC content compared to samples collected under agricultural use.

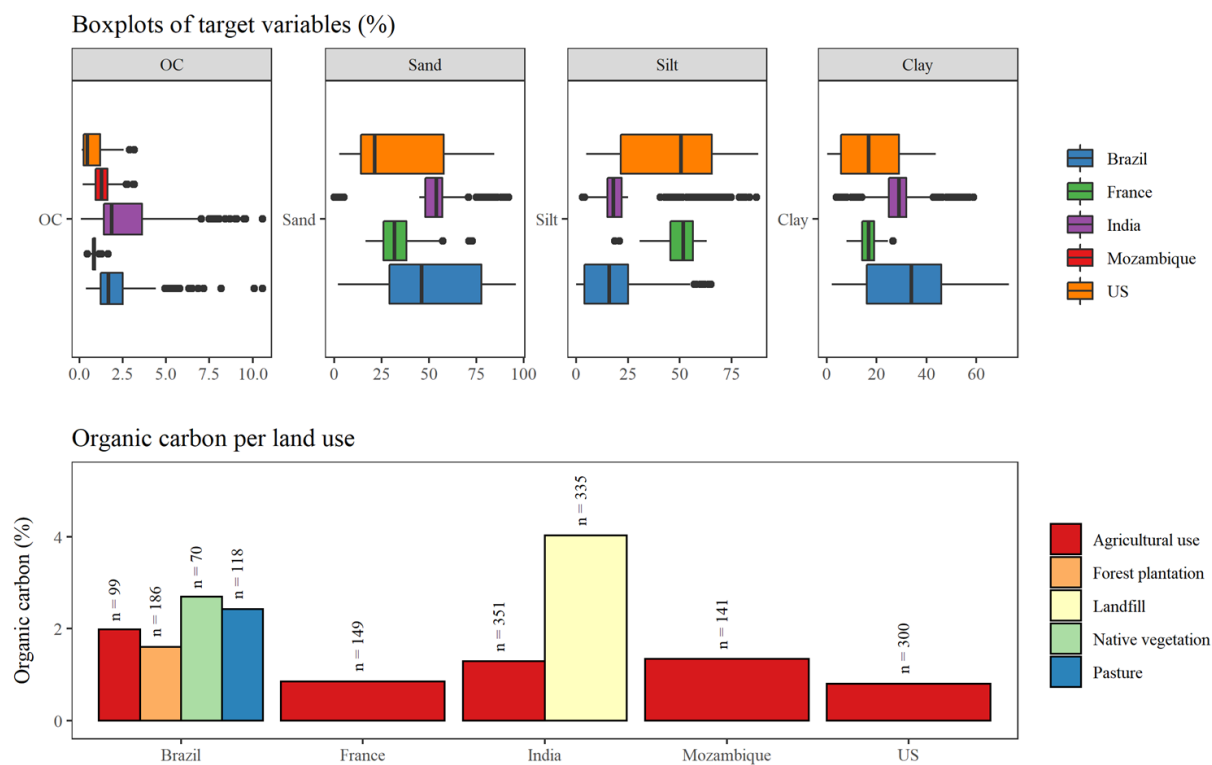


Figure 2. Boxplots of soil organic carbon (OC) and texture by country and the distribution of OC by land use from 1,749 soil samples collected in five countries. n = number of samples.

Examining the entire dataset, samples covered all soil texture classes (Figure 2 and Supplemental Figure S1). Most samples were classified under sandy classes (leftmost area of the texture triangle). Accordingly, sand was the predominant fraction in soils (43.62%), followed by silt and clay, (29.53 and 26.34%, respectively), on average (Table 4). Samples from India were mostly sandy clay loam; samples from France, were mostly loam and silt loam. Samples from Brazil and USA presented a wider variety of texture classes (Figure 2). Samples within the clay textural class were mostly from Brazil (only three from the USA); very clayey textures were exclusively from Brazil.

Table 4. Descriptive statistics of soil texture, organic carbon (OC) and portable X-ray fluorescence (pXRF) analyses of soil samples collected in 5 countries: Mozambique, Brazil, France, India and the United States of America.

	Max.	Min.	Mean	Median	SD	CV%	IQR
Target variables (%)							
Clay	74.00	0.17	26.34	26.00	15.20	57.70	19.58
Silt	87.93	0.00	29.53	21.00	22.33	75.61	35.20
Sand	96.00	0.00	43.62	44.12	25.58	58.63	38.00
OC	10.58	0.08	1.90	1.52	1.56	0.82	1.57
pXRF predictor variables (mg kg ⁻¹)							
As	94.14	0.67	9.26	8.20	6.36	68.74	4.97
Ba	8132.00	97.67	773.69	599.00	827.15	106.91	356.00
Ca	63881.00	37.37	7379.34	3724.58	7883.14	106.83	10373.74
Cr	5910.00	10.33	203.54	102.62	408.28	200.59	163.95
Cu	108636.00	5.00	237.03	34.33	3651.74	1540.62	50.95
Fe	324366.00	466.33	42155.69	34435.01	36865.22	87.45	32204.69
K	92172.00	345.00	17296.09	12088.00	15382.54	88.94	11665.05
Mn	5364.00	13.50	592.37	437.00	571.32	96.45	375.55
Ni	2138.94	2.00	54.48	43.00	92.73	170.23	31.53
Pb	1003.24	4.80	102.22	28.41	160.70	157.20	48.67
Rb	853.00	1.67	112.98	96.87	88.06	77.95	86.31
Sr	4610.00	2.00	128.52	77.69	245.89	191.32	135.52
Ti	46367.33	16.00	6686.50	5295.41	5056.08	75.62	3900.00
V	1341.00	10.00	146.21	108.36	165.26	113.03	103.13
Zn	2316.42	2.33	203.60	62.15	326.32	160.27	74.76
Zr	4638.00	21.00	412.77	290.00	348.61	84.46	295.00

Max.: maximum value; Min.: minimum value; SD: standard deviation; CV: coefficient of variation; IQR: interquartile range.

3.2. Worldwide description of predictor variables

The elements Fe, K, Ca, and Ti were the most abundant on average (Table 4). Fe and K are likely related to iron oxides (hematite and goethite), and muscovite and feldspars, respectively. Ca content can be ascribed to soils under lower weathering intensity in temperate regions (mainly the USA and France), likely from plagioclase and calcite. Ti forms some of the most resistant minerals on Earth (ilmenite, rutile and anatase) and thereby tends to be very immobile in soils. Dispersion, as measured by the coefficient of variation (CV, standard deviation/mean), was consistently below 200%, with the exception of Cu.

Contents of Fe were higher in samples from Brazil, India, and Mozambique (Figure 3). Higher K contents were observed for Mozambique and India, and in some samples in Brazil. In India, high contents of Pb and Zn were detected, attributed to samples from landfill-adjacent agricultural areas. The presence of these heavy metals was further substantiated by the SEM/EDS analyses (Supplemental Figure S2). Further, the elemental results from XRD

analysis showed the presence of As_2O_3 and PbCl_2 , indicating both As and Pb enrichment. Many chemical elements (As, Ba, Fe, K, Mn, Rb, Ti, Sr, V and Zr) were detected in higher concentrations in samples from Mozambique compared to soils from other countries.

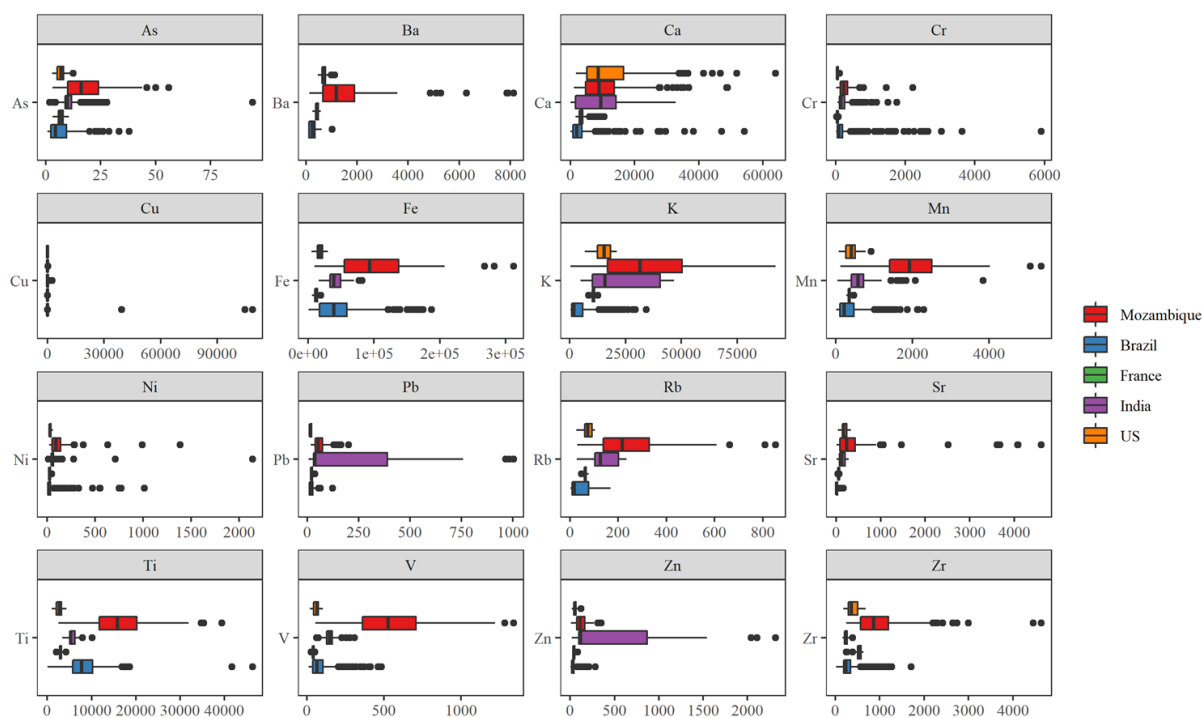


Figure 3. Boxplots of portable X-ray fluorescence (pXRF) analyses of 1,749 soil samples collected across five countries. All X axes indicate elemental concentrations in mg kg^{-1} .

Spectra from India and Brazil, countries with higher precipitation and temperature, showed flatter features compared to US and France, countries with temperate climate (Figure 4). Clear absorption peaks at 420 and 480 nm could be observed for India, the USA, and France, which are associated with the presence of goethite (Viscarra Rossel and Beherens, 2010, Zheng et al., 2019). A depression at 650 nm was present in Brazil's spectra, related to hematite (Zheng et al., 2019). The concavity between 800 and 1,000 nm can also indicate highly crystalline Fe minerals (Fang et al., 2018), which suggests the abundance of crystalline hematite in Brazil. Absorption features at 1,400 and 1,900 nm were more distinct in the spectra of USA and France and are associated with O–H groups and water, and indicate the presence of 2:1 minerals, more common in younger soils, formed under climatic conditions of lower weathering intensity. Absorption features at 2,200 nm are indicative of kaolinite, and were more accentuated for Brazil and the US.

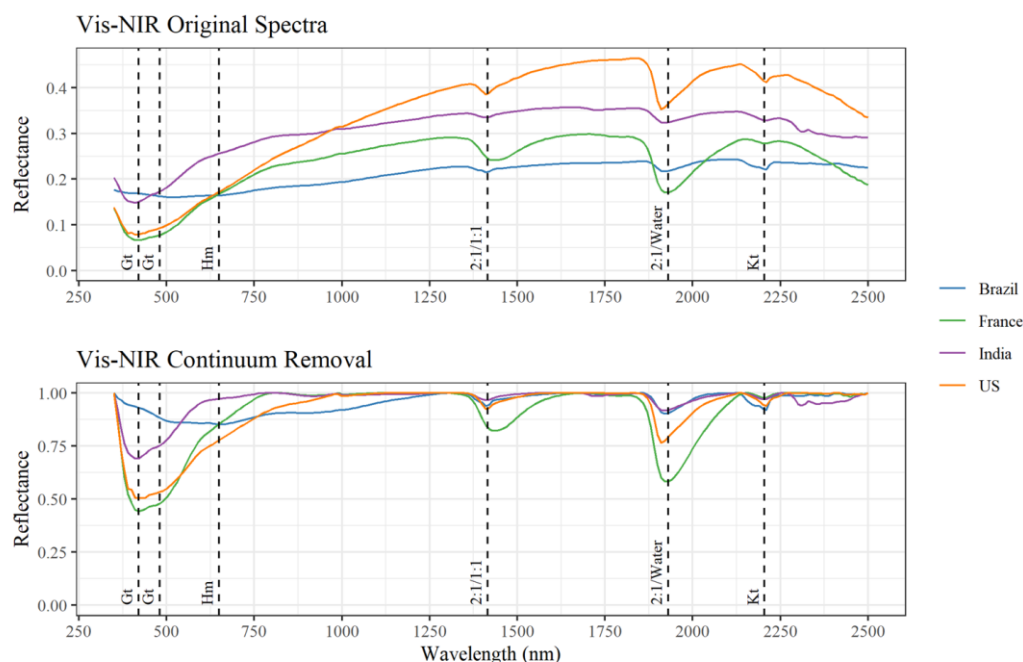


Figure 4. Vis-NIR spectra of 1,545 soil samples collected from four countries. Out of the 1,749 available data, Vis-NIR data was not available for Mozambique samples ($n = 141$) and for some samples from Brazil ($n = 63$). Gt = goethite; Hm = hematite; 2:1 = 2:1 phyllosilicates; 1:1 = 1:1 phyllosilicates; Kt = kaolinite.

As Brazil and India presented higher clay content and smoother absorption shapes between 350 and 500 nm, positive correlations for clay content were higher in this spectral range (Figure 5). Great sand content usually confers lighter color to soil, hence positive correlations were observed throughout the spectrum, especially in Brazil, where sand content was higher (Figure 2). Soil OC presented higher correlations in lower wavelengths (350-500 nm). Positive correlations were observed in countries where soil OC content was higher (Brazil and India), and negative correlations were evident where soil OC content was lower (US and France) (Figs. 2 and 6). This is the visible range of the spectrum related to chromophores (Yang et al., 2019), which consists of compounds that confer color to soils. Absorption features in this region are commonly related to soil OC (Thomas et al., 2022).

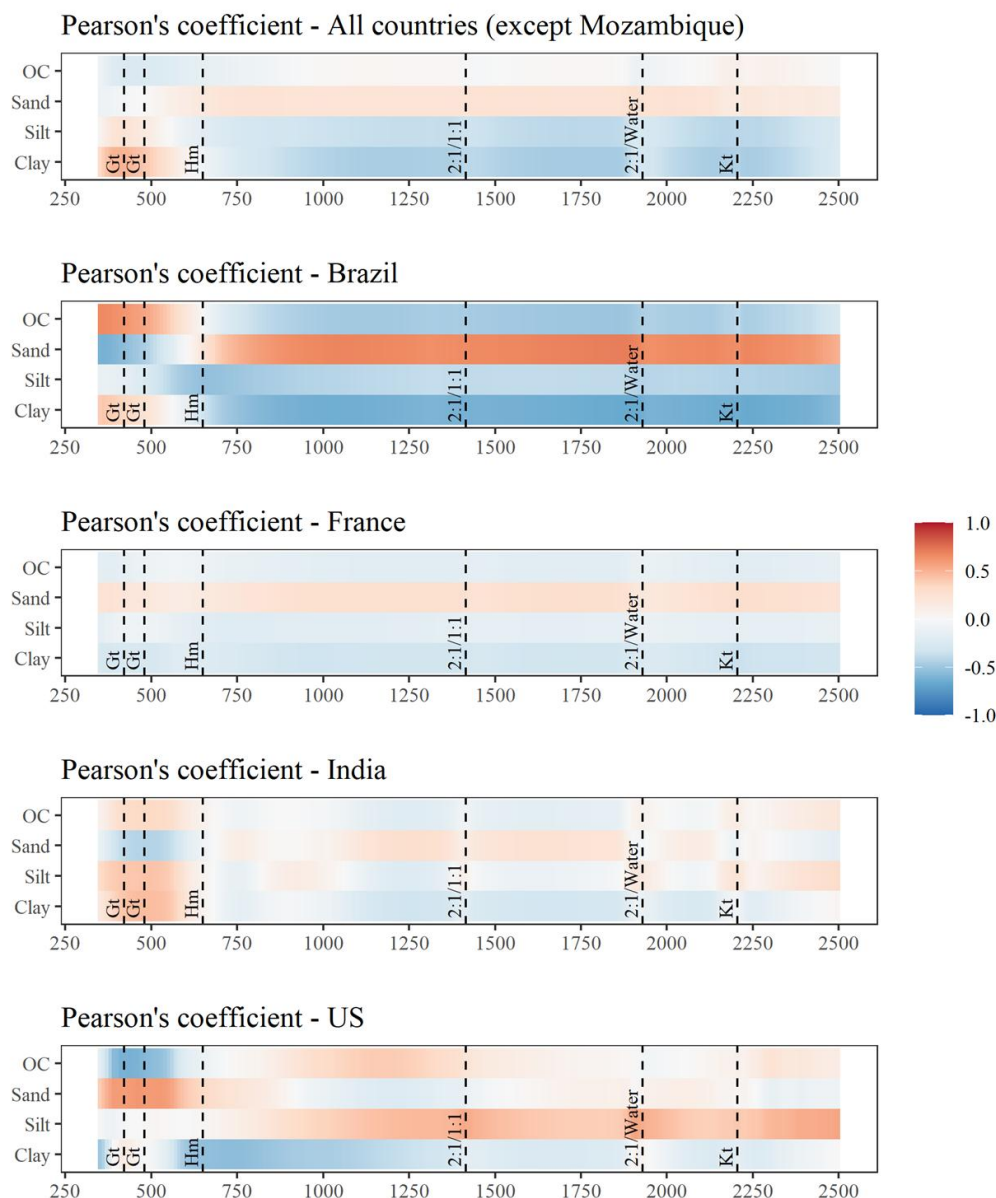


Figure 5. Pearson's correlation coefficient from unprocessed Vis-NIR spectra of 1,545 soil samples collected over four countries. Out of the 1749 available data, Vis-NIR data was not available for Mozambique samples ($n = 141$) and for some samples from Brazil ($n = 63$). Gt = goethite; Hm = hematite; 2:1 = 2:1 phyllosilicates; 1:1 = 1:1 phyllosilicates; Kt = kaolinite. All X axes indicate wavelength in nm.

3.3. Multinational prediction models of organic carbon and soil texture

The best prediction results for soil OC were achieved mostly using the combination of pXRF and Vis-NIR data (Figure 6; Table 5). The best scores were obtained for the US model (0.93) in isolation; the second best, for four countries combined (0.76). The worst performing models were built with data from France and Mozambique (the latter only with pXRF data). The soil OC prediction model trained using all available pXRF data, including Mozambique's samples, presented no significant improvement over the other models (Table 5).

Table 5. Hold-out validation (Valid., 20% of datasets) and 10-fold cross validation (CV, 80% of the datasets) results of prediction models of soil organic carbon (OC) and texture for soil samples from five countries. The best models were selected based on the lowest root mean squared error (RMSE) value of the hold-out validation.

Country	OC																	
	pXRF						Vis-NIR						pXRF+Vis-NIR					
	RMSE		R ²		ρ_c		RMSE		R ²		ρ_c		RMSE		R ²		ρ_c	
	CV	Valid.	CV	Valid.	CV	Valid.	CV	Valid.	CV	Valid.	CV	Valid.	CV	Valid.	CV	Valid.	CV	Valid.
All countries	0.90	0.84	0.66	0.74	0.78	0.84	-	-	-	-	-	-	-	-	-	-	-	-
Mozambique	0.54	0.64	0.18	0.02	0.18	0.08	-	-	-	-	-	-	-	-	-	-	-	-
All countries (except Mozambique)	0.94	0.81	0.67	0.75	0.80	0.84	0.97	0.80	0.65	0.75	0.78	0.85	0.91	0.79	0.69	0.76	0.81	0.85
Brazil	0.99	0.74	0.49	0.70	0.57	0.75	1.00	0.86	0.46	0.52	0.57	0.69	0.90	0.72	0.59	0.72	0.65	0.76
France	0.11	0.09	0.42	0.45	0.48	0.65	0.12	0.11	0.30	0.17	0.25	0.31	0.11	0.08	0.49	0.48	0.49	0.66
India	1.00	1.41	0.68	0.52	0.80	0.67	1.02	1.36	0.67	0.57	0.79	0.68	1.01	1.36	0.68	0.56	0.80	0.69
US	0.16	0.21	0.95	0.92	0.97	0.96	0.28	0.29	0.89	0.86	0.90	0.91	0.16	0.20	0.96	0.93	0.97	0.96
Sand																		
All countries (except Mozambique)	9.17	10.70	0.87	0.82	0.92	0.90	11.72	11.77	0.80	0.79	0.87	0.87	8.36	8.65	0.90	0.89	0.94	0.93
Brazil	8.36	8.82	0.90	0.89	0.94	0.94	13.89	12.97	0.73	0.77	0.82	0.85	8.43	9.09	0.90	0.88	0.94	0.93
France	4.34	4.70	0.81	0.86	0.85	0.89	6.94	10.08	0.48	0.22	0.49	0.30	4.56	5.85	0.80	0.85	0.82	0.80
India	13.01	13.30	0.71	0.75	0.82	0.84	11.31	11.18	0.83	0.88	0.86	0.87	10.47	11.79	0.83	0.82	0.88	0.87
US	4.65	5.55	0.97	0.96	0.98	0.98	7.20	6.39	0.94	0.95	0.95	0.97	4.59	5.01	0.97	0.96	0.98	0.98
Silt																		
All countries (except Mozambique)	8.38	8.61	0.86	0.87	0.92	0.92	9.26	9.70	0.84	0.83	0.90	0.90	7.26	8.38	0.90	0.87	0.94	0.92
Brazil	6.50	7.19	0.79	0.70	0.87	0.82	8.42	9.01	0.65	0.54	0.75	0.68	6.50	7.08	0.79	0.72	0.87	0.82
France	3.91	4.36	0.75	0.75	0.81	0.86	5.98	7.33	0.41	0.36	0.44	0.39	3.83	4.37	0.78	0.77	0.81	0.84
India	11.10	11.45	0.71	0.70	0.79	0.80	10.37	10.46	0.79	0.79	0.81	0.82	9.64	10.22	0.80	0.77	0.84	0.84
US	5.38	4.43	0.95	0.97	0.97	0.98	6.86	5.79	0.92	0.95	0.95	0.97	5.36	4.15	0.95	0.97	0.97	0.99
Clay																		

All countries (except Mozambique)	5.39	6.04	0.86	0.84	0.91	0.91	7.07	7.33	0.75	0.77	0.84	0.85	5.52	6.28	0.85	0.83	0.91	0.90
Brazil	6.97	6.32	0.85	0.87	0.91	0.92	10.60	10.47	0.65	0.63	0.75	0.75	7.38	6.87	0.83	0.84	0.89	0.90
France	2.39	2.27	0.56	0.59	0.69	0.72	2.79	2.63	0.45	0.52	0.48	0.54	2.36	2.29	0.57	0.60	0.69	0.71
India	5.94	3.66	0.60	0.79	0.69	0.87	6.00	3.91	0.63	0.78	0.66	0.83	5.69	3.31	0.64	0.83	0.71	0.90
US	2.89	1.75	0.95	0.98	0.97	0.99	4.67	3.80	0.90	0.93	0.92	0.94	2.92	1.68	0.95	0.98	0.97	0.99

*Mozambique data was modeled separately so that models shared the same dataset in order to assure that sensor performances were comparable.

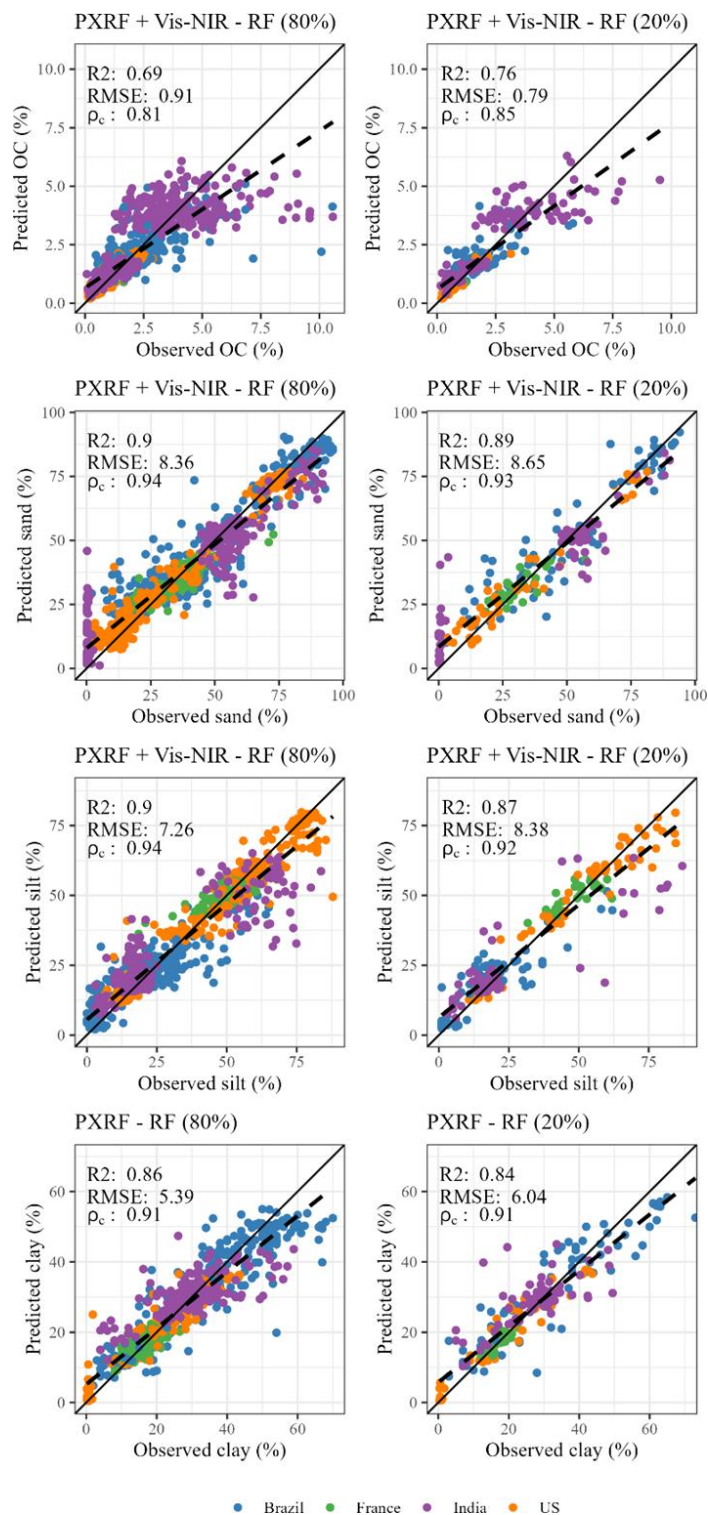


Figure 6. Hold-out validation (right, 20% of datasets) and 10-fold cross-validation (left, 80% of datasets) results of the best prediction models of soil organic carbon (OC), sand, silt, and clay using pXRF and Vis-NIR data from four countries (1,545 samples, Mozambique data not included).

Soil texture predictions were very accurate for all datasets (Table 5). The best performances were predominantly obtained using pXRF data only (Table 5). The only exception was the silt prediction model built with data from Brazil, which presented optimal

results using pXRF + Vis-NIR data. The best performances were achieved by models using USA data ($R^2 > 0.95$) while the weakest performances for sand and silt were observed for models using India data ($R^2 = 0.70$ and 0.75 , respectively). For clay predictions, France data presented the poorest results (0.59). Poor results could be related to the low average contents of textural fractions in some countries: low silt contents in Brazil and low clay contents in France (Figure 2). Combining data from all countries, the best R^2 scores for sand, silt, and clay were 0.82 , 0.87 and 0.84 , respectively (Figure 6; Table 5).

Prediction biplots did not show any significant sign of overfitting (e.g., large differences between cross-validation and hold-out validation) (Figure 6). Biplots made clear that most errors in soil OC predictions came from India samples, especially those collected in landfill areas (samples with higher soil OC content in the topmost plots of Figure 6).

3.4. Predictions outside the training scope

To test how models perform outside of their training scope, models were applied to regions they were not trained to predict. The generalization power of models was overall poor (Figure 7). The highest concordance scores were: models trained with the data from Brazil, which presented ρ_c of 0.55 when predicting OC from India samples using Vis-NIR data; models trained with France data achieved ρ_c of 0.40 predicting clay from US samples using PXRF data; and models trained with the India dataset when predicting sand contents from Brazil samples using Vis-NIR data which achieve ρ_c of 0.65 .

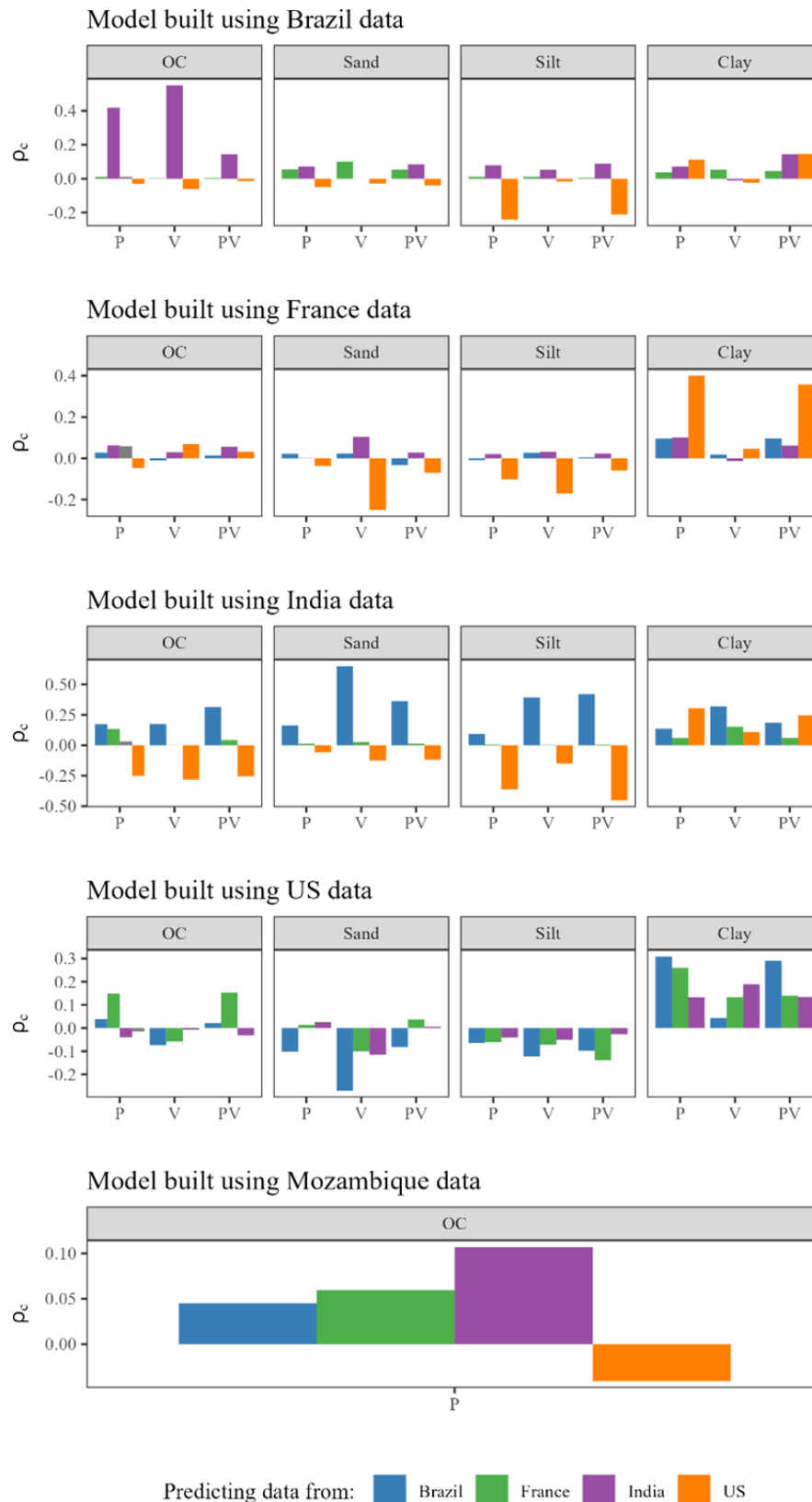


Figure 7. Concordance correlation coefficient (ρ_c) of models trained with regional data predicting samples from other countries. Each row consists of the performance of a model built using a dataset from a specific country applied to predict available samples from all other countries.

3.5. Local applications of the multinational model

The validation set of the multinational model was split per country to test its local performance. The performance of models was similar to that of the local ones in general (Figure 8). Bar labels show the RMSE difference between the multinational model and local ones. Positive values indicate error increase compared to local models, negative values, error decrease.

In some cases, such as for clay predictions in US, error was higher compared to local model predictions. In other cases, error was lower, as was observed for almost all OC predictions from US and India samples, and for silt predictions from France samples. Other noteworthy trends were texture predictions for India, which were worse in almost all cases, and clay predictions from France samples, which improved compared to local models.

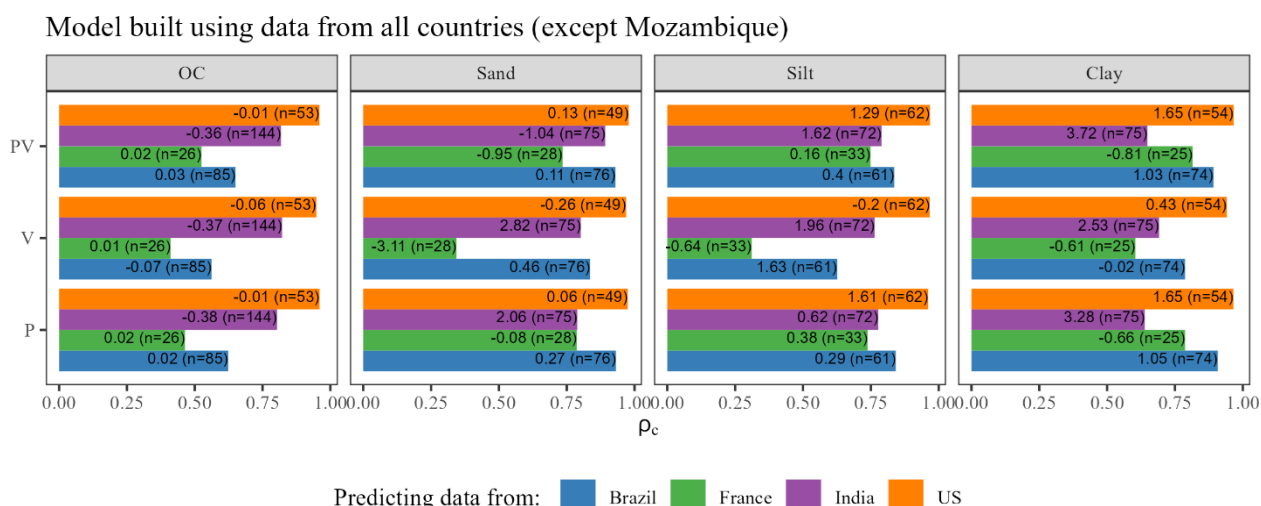


Figure 8. Concordance correlation coefficient (ρ_c) of the multinational models predicting regional datasets (Mozambique data not included). Only unseen data was used for validation. The labels indicate the RMSE difference between the multinational model performance and that of the model trained with local data from the respective countries. Positive values indicate an increase in error, negative values indicate a decrease in error.

3.6. Importance of variables in organic carbon and soil texture predictions

Figure 9 presents the pXRF variables that were selected as important most often by recursive feature elimination. For soil OC prediction models, Ca, Fe, Rb, V, Zn, and Sr were the variables that were selected most frequently (selected by 7 models, including those using Mozambique data). For soil texture models, Fe and Ti were the most used variables, followed by Cu, K, and Mn. The most important Vis-NIR bands for prediction models were consistently the first few principal components (Figure 9), reaffirming that they explain most of the variability in the datasets.

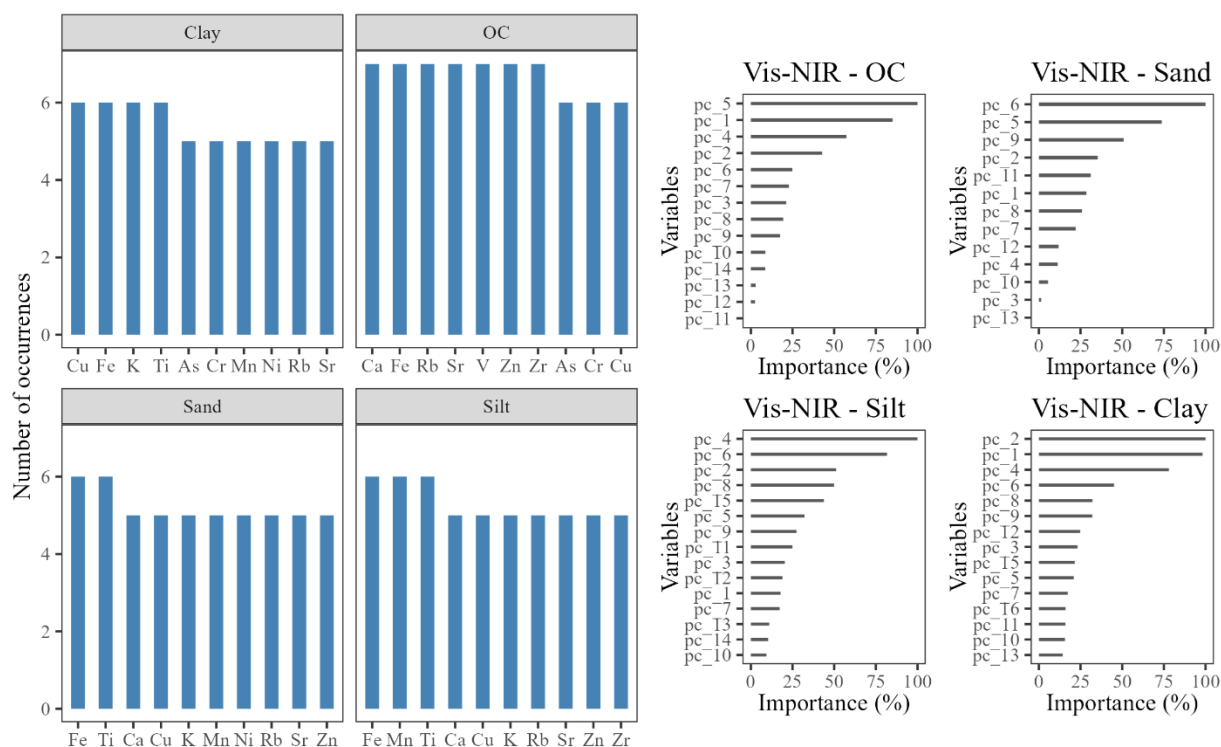


Figure 9. Number of times that pXRF variables were selected by the recursive feature elimination algorithm (left, only pXRF data were used) and the most important Vis-NIR variables (right, only Vis-NIR data were used) for prediction models of soil organic carbon (OC) and soil texture. pXRF models include 1,749 soil samples, Vis-NIR models include 1,545 soil samples (Vis-NIR data were not available for Mozambique and some samples of Brazil).

4. Discussion

The studied soils reflected strongly contrasting climatic conditions. Differences in precipitation and temperature highlighted these differences (Figure 1; Table 1). Naturally, among the five soil forming factors (relief, parent material, climate, organisms and time), climate had a significant impact in soil formation (Resende et al., 2014; Schaetzl & Anderson, 2015). Weathering is more intense under high temperature and high precipitation. One key weathering process involves the hydrolysis of primary minerals, which are commonly phyllosilicates (2:1 minerals), followed by the precipitation of secondary minerals (smaller particles). Higher clay contents in soils from India, Brazil, and Mozambique can therefore be ascribed to tropical climate, in general, where weathering is more intense (Figure 2). Moreover, some of these had long periods without glaciation. Reaffirming the connection between climate and mineralogy, Vis-NIR results presented a clear absorption feature for hematite in spectra from Brazil (Figure 4), and less accentuated features at 1,415 and 1,930 nm, indicating a lower presence of 2:1 minerals.

Conversely, more silt content was observed in soils from the USA and France, where climate is colder and drier (Figs. 2 and 3). The spectra for these two countries presented the most distinct absorption features for 2:1 minerals (1,415 and 1,930 nm) (Figure 4), in accordance with the assumption that these soils tend to be less weathered. Kaolinite absorption features were more accentuated in Brazil and USA data (Figure 4). In Brazil, kaolinite is the most abundant clay mineral in soils, but it is also commonly found in the silt fraction of less developed soils, such as in silty soils from the USA (Figs. 2 and Supplemental Figure S1) (Kämpf et al., 2012). In Brazil, Araujo et al. (2014) found that kaolinite is the dominant mineral in the clay fraction of most soils derived from eight parent materials commonly found in the country, except for some Oxisols. Similarly, in Mozambique, most soils have demonstrated kaolinitic mineralogy (Maria & Yost, 2006). Sand content distribution was higher in soils from Brazil and India, but also in some soils from the USA (Figure 2). In the USA, this can be attributed to very young soils, especially from arid regions (western US), in which pedogenesis is slow, and larger particles are preserved (Buol et al., 2011). In soils from tropical countries, the residual accumulation of quartz and a smaller amount of muscovite in the sand fraction is a very common process (Kämpf et al., 2012).

The high soil OC contents observed in India were specifically related to landfill areas (Figure 2). The second highest values were found in soils from Mozambique and Brazil. This likely reflects the protective effect clays exert on soil organic matter. Also, higher precipitation in tropical regions commonly results in more biomass production. Native vegetation and pasture areas presented comparatively higher mean soil OC contents than those of agricultural areas and planted forests (Yudina et al., 2022). A long-term study (Rothamsted Research, United Kingdom) tested how soils responded to the conversion of grasslands into arable land (Gregory et al., 2016). Results showed that > 60% of soil OC and nitrogen were lost over more than 50 years. Planted forest soils presented lower OC content compared to those of agricultural areas in Brazil, but higher contents compared to agricultural areas of other countries. This tendency indicates more accumulation of OC where there is lower manipulation of soils. While specific management practices were not evaluated herein, they are known to influence soil OC contents (Menon et al., 2020).

Studies using Vis-NIR and pXRF together to predict soil OC have shown their applicability regardless of soil texture variability (Weindorf & Chakraborty, 2018), but testing using multinational datasets have not yet been completed. Zhao et al. (2022) successfully applied Vis-NIR to assess soil OC in 639 samples in four sugarcane districts in Australia ($R^2 > 0.7$). Samples included five soil reference groups (Spaargaren & Deckers, 1998). Similar

research was done with pXRF. Faria et al. (2022) have shown that accurate OC predictions are possible (0.64) by using pXRF in 705 Brazilian soil samples including eight different soil orders (Soil Survey Staff, 2014). Here, results indicated the applicability of this methodology for multinational soil datasets (0.74, including all five countries) (Table 5).

Ca was the most important variable for pXRF models (Figure 9). This is likely related to the retention of Ca by soil organic matter (Kalinichev & Kirkpatrick, 2007; Rowley et al., 2021). Ca is highly mobile in soils and almost absent in the mineral matrix of very weathered soils (Resende et al., 2014). In such soils, cations are mainly retained almost exclusively by organic matter. Variables such as Fe and Ti, also selected by many models (Figure 9), reflect mineralogy (mafic/felsic mineralogical suite). Soils with greater Fe contents relative to other elements (e.g., K, Ca, Mg, Si, etc.) tend to be more weathered and thus have greater clay contents (Rezende et al., 2022). The clay fraction has a positive correlation with the preservation of OC in soils (Schweizer et al., 2021). Ti forms stable minerals and is used to evaluate lithologic discontinuities as a parent material fingerprint (Chapman & Horn, 1968; Schaetzl & Anderson, 2015). It helps differentiate the coarser and finer, or mafic and felsic nature of parent materials, which influences the texture of resulting soils, also depending on the degree of weathering.

Vis-NIR had poorer prediction performance compared to pXRF when predicting soil texture (Table 5). Still, its combination with pXRF improved performance for many datasets. For OC predictions, however, Vis-NIR and pXRF validation scores were similar, and their combination provided the best predictions for all countries. This suggests higher relevance of Vis-NIR data in predicting OC-related variables compared to soil texture. This can be attributed to Vis-NIR's capacity of detecting functional groups related to organic matter. pXRF depends solely on indirect relationships when predicting soil OC, as it cannot detect carbon or nitrogen in isolation, important constituents of organic matter (Silva et al., 2021; Weindorf et al., 2014). Predictions of soil OC were less accurate for samples collected from a landfill in India. The different chemical composition of these soils, indicated by abnormally higher contents of Pb and Zn (Figure 2), possibly affected prediction performances negatively. Soil OC predictions were not possible using data from Mozambique.

Soil texture prediction accuracy was outstanding (Table 5, Figure 6), despite the mixture of soil data from contrasting environmental conditions. Studies using Vis-NIR and pXRF to predict soil texture are less common than those focused on assessing OC, but recent studies have reported results similar to those found herein, even though those evaluations were conducted in less variable datasets. For instance, Zhang and Hartemink (2020) used both

sensors to predict soil texture in 25 pedons in Wisconsin. Models that included pXRF data (alone or combined) presented the best scores ($R^2 > 0.80$). Likewise, Andrade et al. (2022) tested a similar approach with 464 samples in Brazil and also achieved R^2 scores > 0.80 for all texture fractions using models including pXRF data alone or combined with Vis-NIR data. pXRF was more effective in predicting soil texture (Table 5). Nevertheless, models were still accurate even when combining data from different countries, with $RMSE \leq 10\%$.

Here, Vis-NIR data did not contribute to improve results when data were combined, and presented less accurate results, similar to what was reported by Andrade et al. (2022). Vis-NIR models using the dataset from France showed the lowest accuracies. Data from France had less clay content and a smaller amplitude of high silt values (Figure 2). Considering that most models found bands related to clay minerals (especially Fe oxides) more important, texture predictions in soils with proportionally more silt and sand contents were likely harder to be modeled with Vis-NIR data. Nonetheless, Vis-NIR data had excellent performance using US soil data ($R^2 \sim 0.80$ and above).

When locally trained models were used to predict samples from other countries, predictions were not accurate, but interesting relationships were revealed. The model trained with Brazil data achieved comparatively better prediction scores when predicting OC from India samples (Figure 7). The same happened for the model trained with India data when predicting sand from Brazil. A similar effect was observed between France and US for clay predictions. The model trained with data from four countries (Figure 8) presented cases where error compared to local models were lower, and some where error increased. This indicates that, although extrapolation may not be feasible in all cases, some datasets share a complementary relationship. In practice, data for similar regions elsewhere can potentially improve local models.

The accurate mapping of soil texture and carbon in high spatial resolution is key for soil conservation and specific management. Soil conservation directly connects to the maintenance and increase of carbon stocks in soils. The detailed mapping of these two variables is critical to support the growing intensity of human activities in a sustainable manner (Henderson et al., 2022; Lorenz et al., 2019). Devising precise management strategies is only possible with high resolution spatial datasoil maps, which is unfeasible for many countries by using traditional methods alone due to the financial expenses required and the amount of chemical waste generatedwithout copious soil data. The use of global datasets is therefore inevitable. The world is in growing need of novel technologies that enable accurate assessment of soil attributes using variate data sources. The studied proximal sensors (pXRF; Vis-NIR) have proven to be efficient

tools for this end using a multinational and highly variable dataset. The results herein provide an important step regarding greater adoption of such technologies towards detailed soil characterization worldwide. Evaluating the effect of several terrain and soil attributes on soil OC stocks using a 30-year- time series in Brazil, Tayebi et al. (Tayebi et al., 2021) found that clay was the most important variable. Comparing carbon stocks in soils derived from basalt, loess, and Tertiary sand, Angst et al. (Angst et al., 2018) found that soil OC content decreased following clay content (soil derived from basalt > loess > sand).

5. Conclusions

Modeling data from different countries improved local predictions in some cases, indicating that soil data from similar regions may have a complementary effect. pXRF was the best-performing sensor and yielded excellent prediction results for sand, silt, and clay contents ($R^2 = 0.82, 0.87, \text{ and } 0.84$, respectively). Based upon these results, development of a custom prediction mode for soil texture using contemporary pXRF units is possible. Comparatively, Vis-NIR data less was relevant for soil texture predictions.

Soil OC content predictions were likewise accurate (0.74) using samples from five countries combined (USA, India, Brazil, France, and Mozambique). Vis-NIR data were more relevant for soil OC predictions compared to soil texture. Soil OC predictions using only Mozambique data presented poor results. Landfill samples from India presented higher residuals, indicating that mixing data from landfill and agricultural soils was not an optimal strategy. The most relevant pXRF variables were Fe and Ti, related to pedogenesis (Fe, residual accumulation of Fe oxides) and parent material (Ti, parent material fingerprint).

Soil texture and organic matter have a synergetic relationship, and their accurate mapping is of utmost importance for the conservation of soils and their associated carbon stocks. We advise the application of proximal sensors as an efficient method of accurate soil texture and OC mapping, aiming to provide efficient strategies of soil conservation and monitoring of carbon stocks at lower costs and time needed for soil characterization.

Supplemental Material

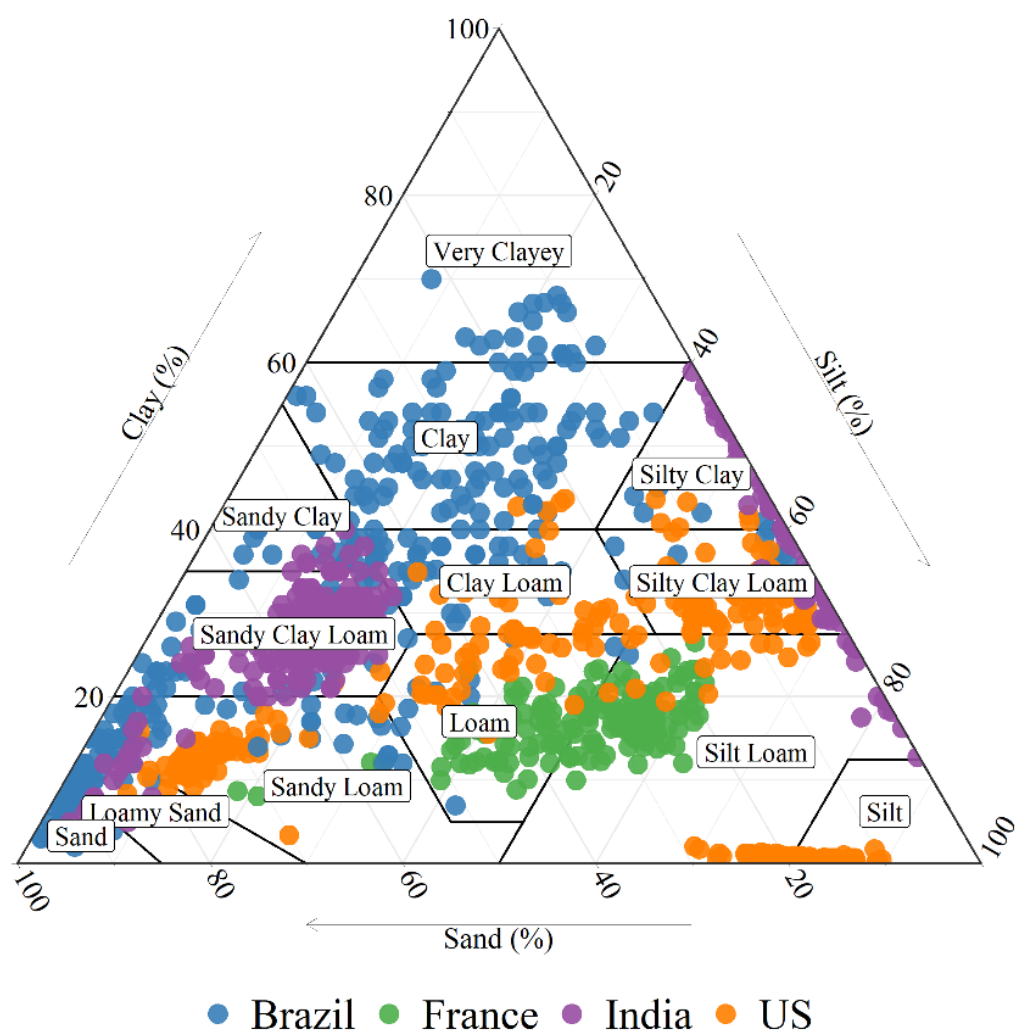


Figure S1. Texture classes of 1,749 soil samples collected across four countries.

References

- Andrade, R., Mancini, M., Teixeira, A. F. dos S., Silva, S. H. G., Weindorf, D. C., Chakraborty, S., Guilherme, L. R. G., & Curi, N. (2022). Proximal sensor data fusion and auxiliary information for tropical soil property prediction: Soil texture. *Geoderma*, 422, 115936. <https://doi.org/10.1016/j.geoderma.2022.115936>
- Angst, G., Messinger, J., Greiner, M., Häusler, W., Hertel, D., Kirfel, K., Kögel-Knabner, I., Leuschner, C., Rethemeyer, J., & Mueller, C. W. (2018). Soil organic carbon stocks in topsoil and subsoil controlled by parent material, carbon input in the rhizosphere, and microbial-derived compounds. *Soil Biology and Biochemistry*, 122, 19–30. <https://doi.org/10.1016/j.soilbio.2018.03.026>
- Araujo, M. A., Pedroso, A. V., Amaral, D. C., & Zinn, Y. L. (2014). Paragênese mineral de solos desenvolvidos de diferentes litologias na região sul de Minas Gerais. *Revista Brasileira de Ciência Do Solo*, 38(1), 11–25. <https://doi.org/10.1590/S0100-06832014000100002>
- Benedet, L., Acuña-Guzman, S. F., Faria, W. M., Silva, S. H. G., Mancini, M., Teixeira, A. F. dos S., Pierangeli, L. M. P., Acerbi Júnior, F. W., Gomide, L. R., Pádua Júnior, A. L., Souza, I. A. de, Menezes, M. D., Marques, J. J., Guilherme, L. R. G., & Curi, N. (2021). Rapid soil fertility prediction using X-ray fluorescence data and machine learning algorithms. *Catena*, 197, 105003. <https://doi.org/10.1016/j.catena.2020.105003>
- Bivand, R. S., Pebesma, E., & Gomez-Rubio, V. (2013). *Applied spatial data analysis with R* (2nd ed.). Springer.
- Buol, S. W., Southard, R. J., Graham, R. C., & McDaniel, P. A. (2011). *Soil Genesis and Classification* (1st ed.). Wiley. <https://doi.org/10.1002/9780470960622>
- Chapman, S. L., & Horn, M. E. (1968). Parent material uniformity and origin of silty soils in northwest Arkansas based on Zirconium-Titanium contents. *Soil Science Society of America Journal*, 32(2), 265–271. <https://doi.org/10.2136/sssaj1968.03615995003200020030x>
- Chatterjee, S., Hartemink, A. E., Triantafilis, J., Desai, A. R., Soldat, D., Zhu, J., Townsend, P. A., Zhang, Y., & Huang, J. (2021). Characterization of field-scale soil variation using a stepwise multi-sensor fusion approach and a cost-benefit analysis. *Catena*, 201, 105190. <https://doi.org/10.1016/j.catena.2021.105190>
- Fang, Q., Hong, H., Zhao, L., Kukulich, S., Yin, K., & Wang, C. (2018). Visible and near-infrared reflectance spectroscopy for investigating soil mineralogy: A review. *Journal of Spectroscopy*, 2018, 1–14. <https://doi.org/10.1155/2018/3168974>
- Faria, A. J. G., Silva, S. H. G., Andrade, R., Mancini, M., Melo, L. C. A., Weindorf, D. C., Guilherme, L. R. G., & Curi, N. (2022). Prediction of soil organic matter content by combining data from Nix Pro™ color sensor and portable X-ray fluorescence spectrometry in tropical soils. *Geoderma Regional*, 28, e00461. <https://doi.org/10.1016/j.geodrs.2021.e00461>
- Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, 37(12), 4302–4315. <https://doi.org/10.1002/joc.5086>
- Gee, G. W., & Bauder, J. W. (1986). Particle-size analysis. In *Methods of soil analysis: Part 1—Physical and mineralogical methods* (2nd ed., Vol. 5, pp. 383–411). Soil Science Society of America, American Society of Agronomy. <https://doi.org/10.2136/sssabookser5.1.2ed.c15>

- Gozukara, G., Acar, M., Ozlu, E., Dengiz, O., Hartemink, A. E., & Zhang, Y. (2022). A soil quality index using Vis-NIR and pXRF spectra of a soil profile. *Catena*, 211, 105954. <https://doi.org/10.1016/j.catena.2021.105954>
- Gregory, A. S., Dungait, J. A. J., Watts, C. W., Bol, R., Dixon, E. R., White, R. P., & Whitmore, A. P. (2016). Long-term management changes topsoil and subsoil organic carbon and nitrogen dynamics in a temperate agricultural system: Long-term management, soil carbon and nitrogen. *European Journal of Soil Science*, 67(4), 421–430. <https://doi.org/10.1111/ejss.12359>
- Henderson, B., Lankoski, J., Flynn, E., Sykes, A., Payen, F., & MacLeod, M. (2022). Soil carbon sequestration by agriculture: Policy options. *OECD - Food Agriculture and Fisheries*.
- Hijmans, R. (2022). raster: Geographic data analysis and modeling, version 3.5-21 (<https://rpkgs.datanovia.com/factoextra/index.html>). R Package.
- Kalinichev, A. G., & Kirkpatrick, R. J. (2007). Molecular dynamics simulation of cationic complexation with natural organic matter. *European Journal of Soil Science*, 58(4), 909–917. <https://doi.org/10.1111/j.1365-2389.2007.00929.x>
- Kämpf, N., Marques, J. J., & Curi, N. (2012). Mineralogia de solos brasileiros. In J. C. Ker, N. Curi, C. E. Schaefer, & P. V. Torrado, *Pedologia: Fundamentos* (pp. 81–145). SBCS.
- Kassambara, A. (2020). Ggpubr: ‘Ggplot2’ based publication ready plots (<https://rpkgs.datanovia.com/factoextra/index.html>). R Package.
- Khaledian, Y., Quinton, J. N., Brevik, E. C., Pereira, P., & Zeraatpisheh, M. (2018). Developing global pedotransfer functions to estimate available soil phosphorus. *Science of The Total Environment*, 644, 1110–1116. <https://doi.org/10.1016/j.scitotenv.2018.06.394>
- Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, 28(5), Article 5. <https://doi.org/10.18637/jss.v028.i05>
- Liaw, A., & Wiener, M. (2002). Classification and regression by random forest. *R News*, 2, 18–22.
- Lin, L. I. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45(1), 255–268.
- Lorenz, K., Lal, R., & Ehlers, K. (2019). Soil organic carbon stock as an indicator for monitoring land and soil degradation in relation to United Nations’ Sustainable Development Goals. *Land Degradation and Development*, 30(7), 824–838. <https://doi.org/10.1002/ldr.3270>
- Mancini, M., Andrade, R., Teixeira, A. F. dos S., Silva, S. H. G., Weindorf, D. C., Chakraborty, S., Guilherme, L. R. G., & Curi, N. (2022). Proximal sensor data fusion for Brazilian soil properties prediction: Exchangeable/available macronutrients, aluminum, and potential acidity. *Geoderma Regional*, 30, e00573. <https://doi.org/10.1016/j.geodrs.2022.e00573>
- Maria, R. M., & Yost, R. (2006). A survey of soil fertility of four agroecological zones of Mozambique. *Soil Science*, 171(11), 902–914. <https://doi.org/10.1097/01.ss.0000228058.38581.de>
- Menon, M., Mawodza, T., Rabbani, A., Bland, A., Lair, G. J., Babaei, M., Kercheva, M., Rousseva, S., & Banwart, S. (2020). Pore system characteristics of soil aggregates and their relevance to aggregate stability. *Geoderma*, 366, 114259. <https://doi.org/10.1016/j.geoderma.2020.114259>

- Neuwirth, E. (2014). RColorBrewer: ColorBrewer palettes (<https://rpkgs.datanovia.com/factoextra/index.html>). R Package.
- Oliveira, J. F., Brossard, M., Corazza, E. J., de Fátima Guimarães, M., & Marchão, R. L. (2022). Field-scale spatial correlation between soil and Vis-NIR spectra in the Cerrado biome of Central Brazil. *Geoderma Regional*, 30, e00532. <https://doi.org/10.1016/j.geodrs.2022.e00532>
- Pebesma, E. (2018). Simple features for R: Standardized support for spatial vector data. *The R Journal*, 10(1), 439. <https://doi.org/10.32614/RJ-2018-009>
- Pedersen, T. (2022). Patchwork: The composer of plots, version 1.1.2 (<https://rpkgs.datanovia.com/factoextra/index.html>). R Package.
- Phogat, V. K., Tomar, V. S., & Dahyia, R. (2015). Soil Physical Properties. In R. K. Rattan, J. C. Katyal, B. S. Dwivedi, A. K. Sarkar, T. Bhattachatyia, J. C. Tarafdar, & S. S. Kukal, *Soil Science: An Introduction* (First, pp. 135–171). Indian Society of Soil Science.
- Pribyl, D. W. (2010). A critical review of the conventional SOC to SOM conversion factor. *Geoderma*, 156(3–4), 75–83. <https://doi.org/10.1016/j.geoderma.2010.02.003>
- R Development Core Team. (2021). R: A language and environmental for statistical computing. R Foundation for Statistical Computing.
- Rahman, M. M., Kamal, M. Z. U., Ranamukhaarachchi, S., Alam, M. S., Alam, M. K., Khan, M. A. R., Islam, M. M., Alam, M. A., Jiban, S. I., Mamun, M. A. A., Abdullah, H. M., Biswas, J. C., Akhter, S., Naher, U. A., Maniruzzaman, Md., Haque, M. M., & Ahmed, F. (2022). Effects of organic amendments on soil aggregate stability, carbon sequestration, and energy use efficiency in wetland paddy cultivation. *Sustainability*, 14(8), 4475. <https://doi.org/10.3390/su14084475>
- Ramirez-Lopez, L., Behrens, T., Schmidt, K., Rossel, R. A. V., Demattê, J. A. M., & Scholten, T. (2013). Distance and similarity-search metrics for use with soil vis-NIR spectra. *Geoderma*, 199, 43–53. <https://doi.org/10.1016/j.geoderma.2012.08.035>
- Ramirez-Lopez, L., Stevens, A., Orellano, C., Viscarra Rossel, R., Shen, Z., Wadoux, A., & Breure, T. (2023). resemble: Regression and similarity evaluation for memory-based learning in spectral chemometrics. [Manual]. <https://CRAN.R-project.org/package=resemble>
- Resende, M., Curi, N., Rezende, S. B., Corrêa, G. F., & Ker, J. C. (2014). *Pedologia: Base para distinção de ambientes* (6.ed.). Editora UFPA.
- Rezende, S. B. de, Franzmeier, D. P., Resende, M., Mancini, M., & Curi, N. (2022). Pedogenic processes in a chronosequence of very deeply weathered soils in southeastern Brazil. *Catena*, 215, 106362. <https://doi.org/10.1016/j.catena.2022.106362>
- Rowley, M. C., Grand, S., Spangenberg, J. E., & Verrecchia, E. P. (2021). Evidence linking calcium to increased organo-mineral association in soils. *Biogeochemistry*, 153(3), 223–241. <https://doi.org/10.1007/s10533-021-00779-7>
- Schaetzl, R. J., & Anderson, S. (2015). *Soils: Genesis and geomorphology* (2nd Revised). Cambridge University Press.
- Schweizer, S. A., Mueller, C. W., Höschen, C., Ivanov, P., & Kögel-Knabner, I. (2021). The role of clay content and mineral surface area for soil organic carbon storage in an arable toposequence. *Biogeochemistry*, 156(3), 401–420. <https://doi.org/10.1007/s10533-021-00850-3>

- Silva, S. H. G., Ribeiro, B. T., Guerra, M. B. B., de Carvalho, H. W. P., Lopes, G., Carvalho, G. S., Guilherme, L. R. G., Resende, M., Mancini, M., Curi, N., Rafael, R. B. A., Cardelli, V., Cocco, S., Corti, G., Chakraborty, S., Li, B., & Weindorf, D. C. (2021). PXRF in tropical soils: Methodology, applications, achievements and challenges. In *Advances in Agronomy* (Vol. 167, pp. 1–62). Elsevier. <https://doi.org/10.1016/bs.agron.2020.12.001>
- Silvero, N. E. Q., Demattê, J. A. M., Minasny, B., Rosin, N. A., Nascimento, J. G., Rodríguez Albarracín, H. S., Bellinaso, H., & Gómez, A. M. R. (2022). Sensing technologies for characterizing and monitoring soil functions: A review. In *Advances in Agronomy* (p. S006521132200092X). Elsevier. <https://doi.org/10.1016/bs.agron.2022.08.002>
- Soil Survey Staff. (2014). *Keys to Soil Taxonomy* (12th ed.). USDA.
- Soil Survey Staff. (2022). *Kellogg soil survey laboratory methods manual—Soil survey investigations report no. 42* (1–6). Department of Agriculture, Natural Resources Conservation Service.
- South, A. (2017). *rnaturalearthdata: World vector map data from natural earth used in “rnaturalearth”, version 0.1.0.* (<https://rpkgs.datanovia.com/factoextra/index.html>). R Package.
- Souza, A. B., Di Raimo, L. A. D. L., César de Mello, D., Guimarães, C. C. B., Urbina-Salazar, D. F., Silva, S. H. G., Curi, N., & Demattê, J. A. M. (2022). Surface reflectance and pXRF for assessing soil weathering indexes. *Journal of South American Earth Sciences*, 115, 103747. <https://doi.org/10.1016/j.jsames.2022.103747>
- Spaargaren, O. C., & Deckers, J. (1998). The World Reference Base for Soil Resources. In A. Schulte & D. Ruyat (Eds.), *Soils of Tropical Forest Ecosystems* (pp. 21–28). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-03649-5_2
- Stevens, A., & Ramirez-Lopez, L. (2022). An introduction to the prospectr package. R package Vignette R package version (2), 1. (<https://rpkgs.datanovia.com/factoextra/index.html>). R Package.
- Teixeira, A. F. dos S., Silva, S. H. G., Soares de Carvalho, T., Silva, A. O., Azarias Guimarães, A., & de Souza Moreira, F. M. (2021). Soil physicochemical properties and terrain information predict soil enzymes activity in phytophysiognomies of the Quadrilátero Ferrífero region in Brazil. *Catena*, 199, 105083. <https://doi.org/10.1016/j.catena.2020.105083>
- Teixeira, A. F. S., Andrade, R., Mancini, M., Silva, S. H. G., Weindorf, D. C., Chakraborty, S., Guilherme, L. R. G., & Curi, N. (2022). Proximal sensor data fusion for tropical soil property prediction: Soil fertility properties. *Journal of South American Earth Sciences*, 103873. <https://doi.org/10.1016/j.jsames.2022.103873>
- Thomas, F., Petzold, R., Landmark, S., Mollenhauer, H., Becker, C., & Werban, U. (2022). Estimating forest soil properties for humus assessment—Is Vis-NIR the way to go? *Remote Sensing*, 14(6), 1368. <https://doi.org/10.3390/rs14061368>
- Tümsavaş, Z., Tekin, Y., Ulusoy, Y., & Mouazen, A. M. (2019). Prediction and mapping of soil clay and sand contents using visible and near-infrared spectroscopy. *Biosystems Engineering*, 177, 90–100. <https://doi.org/10.1016/j.biosystemseng.2018.06.008>
- Walkley, A., & Black, I. A. (1934). An examination of the Degtjareff method for determining soil organic matter, and a proposed modification of the chromic acid titration method. *Soil Science*, 37(1), Article 1. <https://doi.org/10.1097/00010694-193401000-00003>

- Weindorf, D. C., Bakr, N., & Zhu, Y. (2014). Advances in portable X-ray fluorescence (PXRF) for environmental, pedological, and agronomic applications. In *Advances in Agronomy* (Vol. 128, pp. 1–45). Elsevier. <https://doi.org/10.1016/B978-0-12-802139-2.00001-9>
- Weindorf, D. C., & Chakraborty, S. (2018). Portable apparatus for soil chemical characterization (Patent US10107770B2).
- Weindorf, D. C., & Chakraborty, S. (2020). Portable X-ray fluorescence spectrometry analysis of soils. *Soil Science Society of America Journal*, 84(5), 1384–1392. <https://doi.org/10.1002/saj2.20151>
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Golemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., ... Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, 4(43), 1686. <https://doi.org/10.21105/joss.01686>
- Wold, H. (1975). Soft Modelling by Latent Variables: The Non-Linear Iterative Partial Least Squares (NIPALS) Approach. *Journal of Applied Probability*, 12(S1), 117–142. <https://doi.org/10.1017/S0021900200047604>
- Yudina, A. V., Klyueva, V. V., Romanenko, K. A., & Fomin, D. S. (2022). Micro- within macro: How micro-aggregation shapes the soil pore space and water-stability. *Geoderma*, 415, 115771. <https://doi.org/10.1016/j.geoderma.2022.115771>
- Zhang, S., Wang, L., Wei, W., Hu, J., Mei, S., Zhao, Q., & Tsang, Y. F. (2018). Enhanced roles of biochar and organic fertilizer in microalgae for soil carbon sink. *Biodegradation*, 29(4), 313–321. <https://doi.org/10.1007/s10532-017-9790-0>
- Zhang, Y., & Hartemink, A. E. (2020). Data fusion of vis-NIR and PXRF spectra to predict soil physical and chemical properties. *European Journal of Soil Science*, 71(3), 316–333. <https://doi.org/10.1111/ejss.12875>
- Zhao, X., Zhao, D., Wang, J., & Triantafilis, J. (2022). Soil organic carbon (SOC) prediction in Australian sugarcane fields using Vis–NIR spectroscopy with different model setting approaches. *Geoderma Regional*, 30, e00566. <https://doi.org/10.1016/j.geodrs.2022.e00566>

Chapter IV: Interpreting machine-learning models trained to estimate soil texture and organic carbon in soils of Southeast Brazil

Marcelo Mancini^a, Sérgio Henrique Godinho Silva^a, Zamir Libohova^b, Nilton Curi^a

^aDepartment of Soil Science, Federal University of Lavras, Lavras, Minas Gerais, Brazil

^bUSDA-ARS, Dale Bumpers Small Farms Research Center, 6883 AR-23, Booneville, AR 72927, USA

Abstract

Machine-learning algorithms are capable of modeling complex soil phenomena using many environmental covariates, but bear the drawback of being hard to interpret, commonly known as “black boxes”. Here we use model interpretation methods to make machine-learning models more transparent and explainable, aiming to learn from relationships found by models between environmental covariates and soil properties. The chemical composition of 292 soil samples were measured using portable X-ray fluorescence (pXRF) and combined with topographic attributes to train machine-learning models to estimate sand, clay, and soil organic carbon contents (SOC). The utilized interpretation methods include partial dependence plots (PDP), accumulated local effects (ALE), and Shapley values. Sand and clay content models were accurate, and their estimations relied mostly on Si, Al, Fe, and K contents. Models implicitly identified the parent materials of samples, associating high Fe contents with clayey samples derived from gabbro, and high Si and K contents with sandier soil samples formed from gneiss. SOC models were less accurate than soil texture models due to relying excessively on the total Ca contents of soils, which may vary based on land use and management practices. Soil texture models may be used in areas where Si and Fe contents are within the ranges of soils studied herein, but SOC models are not transferable due to their reliance on Ca content, which may be site-specific. Machine-learning interpretation methods can make information produced by models more accessible to the scientific community, increasing the reliability and trust in models, turning models into sources of pedologic knowledge instead of utilitarian prediction tools.

Keywords: Oxisols, Ultisols, pedometrics, sand fraction, clay fraction, pXRF, explainable AI, interpretable machine-learning.

1. Introduction

Machine-learning has been widely used in soil science research (Padarian et al., 2020). Its quick dissemination was incentivized by the search for easier modeling methods, and not because traditional statistical models were inadequate, but because they can be harder to use, especially when problems involve complex data distributions, spatial dependence, and large numbers of covariates. Machine-learning models offer easier pathways to work with non-linearity and non-stationarity (Wydiatmanti et al., 2025) and make no strong assumptions about the distributions of the data. However, the advent of machine-learning models brought new problems (Arrouays et al., 2020). Such models are incredibly flexible and capable of fitting to diverse data and may produce seemingly accurate answers even when posed with absurd problems (Fourcade et al., 2023). But although complex, machine-learning models are not truly black boxes, and methods exist through which their internal workings can be understood (Molnar, 2024). In this work we use these methods to examine the role of X-ray fluorescence spectrometry data (contents of chemical elements in soil samples) in machine-learning models trained to predict soil properties.

Poor understanding of the internal decision processes inside models creates problems beyond inaccurate predictions. For environmental scientists, the inability to understand their models inhibits gaining insights about possible connections between covariates (i.e., predictors) and their variables of interest (i.e., predicted). Without comprehending a model's internal functioning, its capability of generalization can only be assessed by data used by the modeler, which often is not representative of all situations in which the model may be used. This can obfuscate limitations of models and lead to their misuse or misconceptions about their adequate scope of application. Wadoux et al. (2020a) highlighted these issues by showing that seemingly accurate predictions of soil organic carbon (SOC) can be achieved even when models are trained with pictures of soil scientists, i.e., pseudo-covariates or predictors that have no relation to SOC. In a similar work, Fourcade et al. (2023) used pictures of famous paintings to predict the distributions of species across Europe and showed that the resulting evaluation metrics of models trained with pseudo-covariates (paintings) can be comparable to, or even better than, those trained with real environmental covariates. Both works are valuable documentations of how machine-learning models can find ways of fitting to any data and are not to be fully trusted without proper knowledge of their inner workings.

These pitfalls have long been known, and many methods have been developed in attempt to explain the predictions made by complex machine-learning models (Holzinger et al., 2022), although they remain widely unexplored in soil science. These methods can be local or

global, depending on whether their objective is to analyze individual predictions (local) or the behavior of the model as a whole (global). This work will explore mainly the global methods named permutation feature importance (PFI), partial dependence plots (PDP), and accumulated local effects (ALE), which provide an assessment of the relevance and effect of each covariate in predictions. In addition, Shapley values, a local method used to evaluate the contribution of covariates in individual predictions, was also applied. The work by Wadoux and Molnar (2022) does a thorough exposition of the mathematical background behind these model interpretation methods and provides several examples of their application to soil properties.

Many environmental covariates relate to soil formation processes and can be used to train machine-learning models capable of estimating soil properties (Silva et al., 2025; Wadoux et al., 2020b). Commonly utilized covariates include topographic attributes calculated from elevation data, remotely-sensed spectral data from satellites and unmanned aerial vehicles, and data acquired by probing samples at close range using proximal sensors. This study focuses on studying measurements obtained through portable X-ray fluorescence spectrometry (pXRF), a proximal sensor used to estimate the total chemical composition of samples by measuring the fluorescence emitted by atoms when irradiated with X-rays (Silva et al., 2021a; Weindorf et al., 2014). PXRF data has been used extensively in the modeling of soil properties and has shown encouraging results when used to estimate properties related to the parent material of soils, such as texture (Andrade et al., 2020a; Gozukara et al., 2022). Understanding how machine-learning models utilize pXRF data can shed light on the relationship between the chemical composition of soils and their properties, besides helping to further document the limitations and strengths of these models.

The goal of this study was to investigate the relationships between pXRF covariates, soil texture and SOC learned by machine-learning models from samples collected in the campus of the Federal University of Lavras, in Brazil. For several years, soil data collected throughout the campus was used to feed machine-learning models applied extensively for the sole purpose of providing predictions of soil properties. By applying model interpretation methods, we hope to expose the strengths and weaknesses of these models, and to document and share knowledge about the role of pXRF covariates in the prediction of clay, sand, and SOC contents. Machine-learning models were trained using pXRF data as covariates to predict clay, sand, and SOC contents from 292 soil samples with PFI, PDP, ALE, and Shapley values used to assess and pedologically interpret their performance. Specifically, the objectives were to i) understand the importance and investigate the effect of pXRF covariates in predictions; ii) use Shapley values to expose the limitations of models by investigating the contribution of covariates to prediction

errors. We believe that machine-learning interpretation methods can serve to better document and share data about models and provide knowledge on possible causal connections between soil properties and the chemical composition of soils as provided by pXRF.

2. Material and Methods

2.1. Study area and soil sampling

The study was conducted on the campus of the Federal University of Lavras, Brazil (Figure 1). A total of 292 soil samples were collected from 128 profiles across varying depths, which were grouped into three depth classes: 0-20, 20-60, and 60-100 cm, based on the classification of genetic horizons. Intermediate depths represent B horizons of younger soils or transitional horizons of deep, weathered soils. The climate in the study area is classified as Cwa according to the Köppen classification system, with dry winters and hot summers. Average temperature in the region is 20.4°C, with mean rainfall of 1460 mm, occurring between November and March (Alvares et al., 2013).

Sampled soil profiles included four soil orders: Oxisols, Ultisols, Inceptisols, and Entisols. Detailed classifications of the studied soil profiles according to the Soil Taxonomy (Soil Survey Staff, 2014) and the Brazilian Soil Classification System (SiBCS) (Santos et al., 2018) are presented in Table 1. Most soils in the study area were formed from two parent materials: orthogneiss, a banded metamorphic rock with predominance of felsic minerals (Si-rich minerals such as quartz, feldspars, and muscovite); and gabbro, a phaneritic igneous rock composed predominantly of mafic minerals (Si-poor minerals such as pyroxene). The parent material map of the study area is presented in Mancini et al. (2019). The 128 soil profiles were under the following land uses: pasture (51), native forest (42), Eucalyptus and Pinus plantations (18), coffee plantations (10), urban areas (4), and areas used for agronomic research (3).

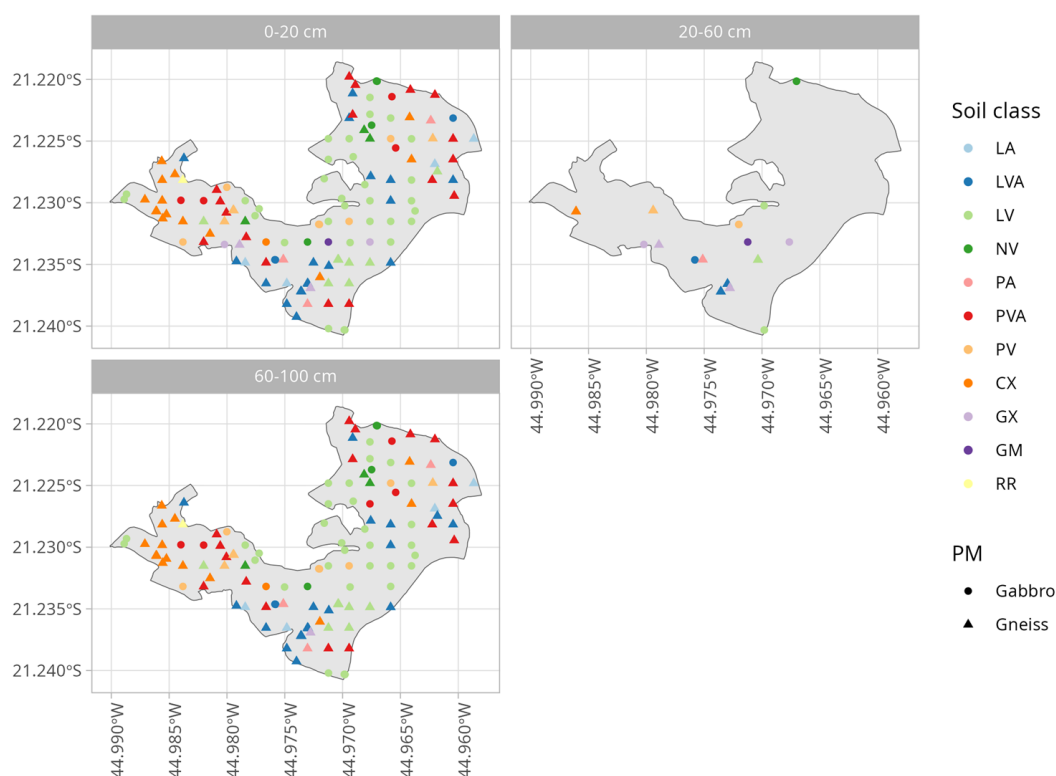


Figure 1. Locations of 292 soil samples collected in the Federal University of Lavras, Brazil. Sampled soils were found within 5 different land uses and were formed from two different parent materials (PM). PM data adapted from Mancini et al. (2019) and soil classifications were adapted from Curi et al. (2017). Map datum: WGS 84. Soil classes are described in Table 1.

Table 1. Soil classes of the 128 studied profiles located within the campus of the Federal University of Lavras, Brazil. Classifications are provided according to criteria posed by the Brazilian Soil Classification System (SiBCS) (Santos et al., 2018), along with its correlations with the Soil Taxonomy (Soil Survey Staff, 2014).

Symbol	n	Soil Taxonomy	SiBCS
LA	8	Xanthic Hapludox	Latossolo Amarelo Distrófico típico. Latossolo Amarelo Eutrófico típico
LVA	51	Typic Hapludox	Latossolo Vermelho-Amarelo Distrófico típico, Latossolo Vermelho-Amarelo Distroférrico típico, Latossolo Vermelho-Amarelo Eutrófico típico
LV	17	Anionic Acrudox	Latossolo Vermelho Distroférrico típico
LV	70	Rhodic Hapludox	Latossolo Vermelho Distrófico típico. Latossolo Vermelho Eutrófico típico, Latossolo Vermelho Eutroférrico típico
NV	17	Rhodic Kandiudult	Nitossolo Vermelho Distrófico típico, Nitossolo Vermelho Eutrófico típico
PA	11	Inceptic Hapludult	Argissolo Amarelo Distrófico típico. Argissolo Amarelo Eutrófico típico
PVA	43	Typic Hapludult	Argissolo Vermelho-Amarelo Distrófico típico, Argissolo Vermelho-Amarelo Eutrófico típico
PV	26	Typic Rhodudult	Argissolo Vermelho Distrófico típico,

			Argissolo Vermelho Eutrófico típico, Argissolo Vermelho Eutroférico típico
CX	33	Typic Dystrustept	Cambissolo Tb Distrófico típico, Cambissolo Tb Eutrófico típico
GX	12	Typic Endoaquent	Gleissolo Háplico Tb Distrófico típico, Gleissolo Háplico Tb Eutrófico típico
GM	2	Mollic Endoaquent	Gleissolo Melânico Tb Distrófico típico
RR	2	Typic Udorthent	Neossolo Regolítico Distrófico típico

2.2. Laboratory analyses

Samples were air-dried and passed through a 2-mm sieve prior to undergoing the following analyses. SOC was determined through oxidation with potassium dichromate followed by titration with ammonium ferrous sulfate (Walkley and Black, 1934). Sand and clay contents were determined using the pipette method (Gee and Bauder, 1986). Samples were dispersed in an aqueous solution with 1 mol L⁻¹ of NaOH and passed through a sieve with 0.053 mm openings to separate sand-sized particles. The portion of the sample retained in the sieve was used to determine sand content by weight. The remaining solution was put to rest until silt-sized particles settled enough to extract an aliquot of dispersed clay from the top of the solution. After the sedimentation of silt, an aliquot of dispersed clay was extracted from the top of the solution, dried and weighted in order to determine the clay content.

The pXRF analysis was conducted by placing approximately 10g of air-dried and sieved samples in cylindrical holders covered with Prolene® films, following Weindorf and Chakraborty (2020). The utilized device was a Bruker S1 Titan LE Handheld XRF Analyzer (Billerica, MA, USA), equipped with a Rh X-ray tube that operates at 15-50 kV and 5-100 µA. Each sample was scanned for 60 s in triplicates, while rotating the sample between repetitions. The mean value of the triplicates was used for subsequent data analysis. To assess the accuracy of the pXRF device, three reference samples were used: 2710a (Montana soil I) and 2711b (Montana soil II), certified by the National Institute of Standards and Technology, and one reference sample provided by the manufacturer of the device.

2.3. Modeling procedures

2.3.1. Variable selection

Machine-learning problems involve large sets of features; therefore, the relevance of the features is often unknown beforehand. For this reason, it is a common practice to use feature selection algorithms to filter the most relevant variables and to simplify the resulting models. However, a choice was made to use a fixed set of variables selected based on pedologic

knowledge for all modeled attributes instead of using feature selection strategies. The reasons for this choice were: i) the focus of this work is to investigate the internal workings of models, which would be much more complex if different sets of variables were tested; ii) the sample sizes of the studied datasets are not adequate for a reliable feature selection assessment (Kuncheva et al., 2020); iii) there is previous knowledge available about which variables are expected to be relevant, both from tacit knowledge of local pedologists and from previous works conducted with using pXRF covariates in soils from the studied region (Andrade et al., 2020b; Pelegrino et al., 2021; Silva et al., 2021).

The pXRF variables used for modeling were selected based on their relevance for pedogenetic processes, sorption complex dynamics, and parent material identification. The selected variables were: Si and Al as indicators of soil weathering and parent material; Ca, K, and P, as indicators of the sorption capacity of soils controlled by clay or SOC content; Fe, Cr, Cu, Sr, Ti, Zn, Zr, V, and Rb, since the abundance of these elements are largely influenced by the chemical composition of the parent material. Additionally, sampling depth and variables that represent geomorphology were used as complementary variables. Sampling depth was selected due to its relevance in SOC estimations, since organic matter is deposited at the surface and its content tends to decrease with depth. Geomorphology data was selected to represent the relationship between relief and the soil formation processes that drive changes in soil texture and SOC, such as melanization and argilluviation (Bockheim and Gennadiyev, 2000).

Geomorphological variables were obtained from altimetric data available for the study area. First, contour lines with 1-m resolution were used to create a digital elevation model of the study area with spatial resolution of 5 m. Next, SAGA GIS (Conrad et al., 2015) was used to calculate the following terrain attributes from the elevation model: slope, valley depth, and SAGA's topographic wetness index (TWI). Data from the elevation model itself was also used as a covariate. Values of covariate rasters were extracted to each of the sampled points (Figure 1) and used in the modeling process. The raster-to-point extraction and the handling of spatial data were done using QGIS 3.44 (QGIS, 2025).

In summary, the following covariates were chosen as inputs to train the machine-learning models: sampling depth, Si, Al, Ca, K, P, Fe, Cr, Cu, Ti, Rb, Sr, V, Zn, Zr, elevation, slope, valley depth, and TWI.

2.4. Machine-learning algorithms

Models were trained to predict SOC, sand, and clay contents using the following machine-learning algorithms: K-nearest neighbors (KNN), random forest (RF), gradient

boosting machine (GBM), and support vector machine with linear kernel (SVM). Prior to modeling, data was split into two sets, the first set used to train the models (70% of the data, 203 samples), and a second set (30%, 89 samples) to evaluate the performance the of models. The split was done based on soil profiles to assure that the test set contained only samples from profiles that were not used during the training phase to avoid data leakage and overoptimistic evaluation scores. The tuning of hyperparameters was done using 10-fold cross-validation with adaptive resampling (Kuhn, 2014), which was repeated 10 times. Similar to the train/test split, cross-validations were done based on soil profiles, assuring that samples from the same profile would not be included in more than one fold.

The models' performance was evaluated using the coefficient of determination (R^2), Lin's concordance correlation coefficient (ρ_c), and the root mean square error (RMSE), combined with the analysis of residuals. Models that presented low RMSE, no patterns in residuals, and approximately normal residual distributions were considered more reliable. The R packages 'dplyr' and 'caret' were used for data handling and all modeling procedures.

2.5. Model interpretation methods

Feature permutation importance (FPI) was used to assess the global effect of covariates in prediction errors. FPI consists of shuffling the values of a covariate and testing whether prediction errors increase. If errors increase substantially after shuffling, the covariate is assumed to contribute to the accuracy of predictions; otherwise its removal has little effect in the outcomes of the model. In other words, the increase or decrease of RMSE after permutations is used to assess the FPI. The same process is then repeated for all covariates. A total of 1000 permutations were performed for each covariate during this test to assess their relevance. Additionally, interactions between covariates were calculated using Friedman's H statistic (Fiedman and Popescu; Molnar, 2024), as implemented by the R package 'iml' (Molnar et al., 2018).

Partial dependence plots (PDP), individual conditional expectations (ICE) and accumulated local effects (ALE) were used to assess the effect of covariates on target variables (Molnar, 2024). PDPs and ICEs evaluate how the outcomes of a model change when a single covariate varies (Goldstein et al., 2015; Friedman, 2001). Their main difference is that PDPs express the global effect of covariates, while ICEs present their effect in each individual prediction. However, PDPs and ICEs use marginal distributions and can therefore be misleading when covariates are correlated, because it might consider absurd combinations between covariates (e.g., very high SOC at deeper horizons). ALEs solve that problem by evaluating the

global effect of a covariate using conditional distributions and calculating the effects over small data intervals (Apley and Zhu, 2020). ALEs suffer less from the problems caused by correlated covariates but can be harder to interpret due to being calculated per intervals. Generally, when PDPs/ICEs and ALEs result in similar curves, both can be used but the former can provide more intuitive interpretations; and when they differ, ALEs will offer a more accurate representation of the effects of covariates (Molnar, 2024). All three methods were applied as they can be complementary to each other.

Lastly, Shapley values were calculated to shed light on the contribution of covariates in individual soil samples (Molnar, 2024; Shapley, 1952). If we assume that the estimation task is a game, and each covariate is a player, the prize of the game is the difference between the mean value of predictions and the specific prediction being evaluated. The Shapley value evaluates how to divide the prize (the deviation from the mean prediction) to each player (covariate) (Molnar, 2024). In other words, the Shapley value measures the contribution of each covariate in deviating one specific prediction from the mean. This method can be utilized to understand the reasons why a sample is being incorrectly predicted, or to learn which covariates influenced an accurate prediction the most.

All interpretations methods used in this study were calculated with the R package ‘iml’ (Molnar et al., 2018) and plotted using ‘ggplot2’ (Wickham, 2016).

3. Results

3.1. Descriptive statistics of soil texture and OC

Soil samples presented clayey texture, especially those from Oxisols derived from gabbro (Table 2, Figure 2). The average values of clay and sand contents were 50 and 32% (Table 2), respectively, depicting the predominantly clayey nature of the studied soils. Only a few samples were classified as loamy (sandy clay loam, clay loam or loam), and silt contents were overall low (mostly < 15%) given the highly weathered status of these tropical soils. SOC content was low compared to soils of temperate regions, with an average of 1.78% (Table 2), which is expected due to the hot and humid climate that causes high oxidation rates of organic matter, hindering its accumulation and lowering its persistence (Resende et al., 2021). A few samples presented SOC contents unusually high for tropical climate (>5%). These high values were observed in native forests, pastures, and depositional parts of the landscape, and were attributed to the sheltering provided by plant canopy, grasses biomass, and the deposition of plant litter in lowlands. The descriptive statistics of the train and test sets were similar, showing that the test set is a representative split of the overall data (Table 2).

Table 2. Descriptive statistics training (n = 203) and test (n = 89) datasets of soil organic carbon (SOC), sand, and clay contents target variables used to train and evaluate machine-learning models. Samples were collected from the campus of the Federal University of Lavras, Brazil. SD: standard deviation; Max: maximum value; Min: minimum value; Q3 and 1: third and first quantiles; IQR: interquantile range.

Dataset	Statistic	Clay (%)	Sand (%)	SOC (%)
Training (n = 203)				
	Mean	50.11	32.05	1.78
	SD	11.69	11.71	1.48
	Max	76.00	60.00	6.84
	Q3	59.00	42.00	2.67
	Median	50.00	32.00	1.16
	Q1	41.00	23.00	0.70
	Min.	17.00	9.00	0.12
	IQR	18.00	19.00	1.89
Test (n = 89)				
	Mean	50.72	32.45	1.71
	SD	13.80	12.88	1.39
	Max	75.00	63.00	6.32
	Q3	60.00	44.00	2.49
	Median	51.00	32.00	1.39
	Q1	40.00	22.00	0.64
	Min	17.00	7.00	0.06
	IQR	20.00	22.00	1.86

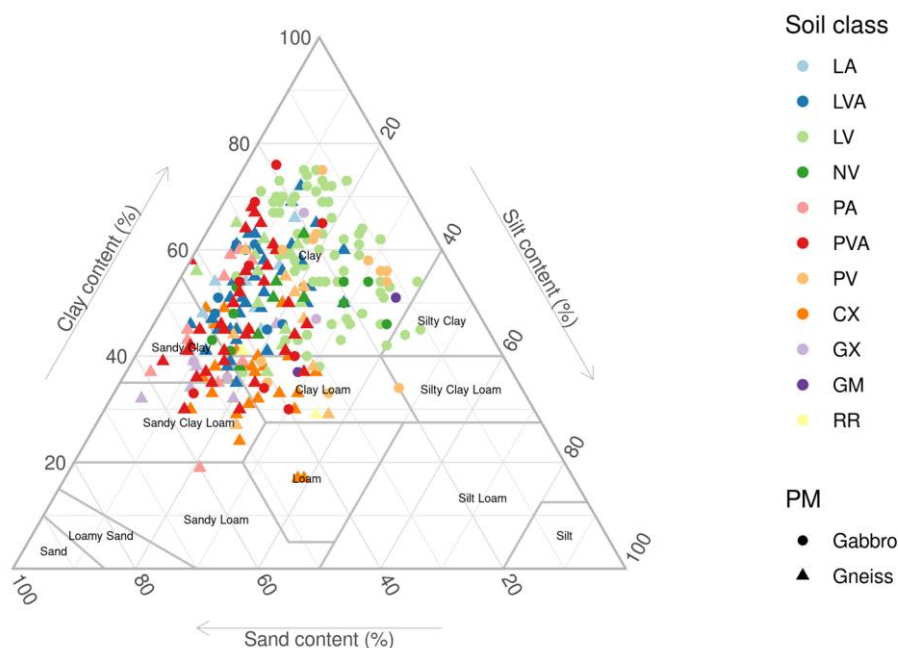


Figure 2. Texture classes of 292 soil samples collected in the Federal University of Lavras, Brazil. Sampled soils were found within 5 different land uses and were formed from two different parent materials (PM). The abbreviations of soil classes are described in Table 1. PM data adapted from Mancini et al. (2019) and soil classifications were adapted from Curi et al. (2017).

3.2. Model performance and the importance of covariates

GBM presented the lowest RMSE for SOC and clay content, while RF achieved the lowest RMSE for sand content estimations (Table 3). Although achieving the highest

performances, the tree-based models (RF and GBM) presented the highest performance differences between the train and test sets, raising concerns about overfitting. The plots comparing observed and predicted data, along with QQ plots showing the normality of the data confirmed that sand and clay content models were indeed very accurate and presented approximately normally distributed residuals (Figure 3), although clay models showed a slight underestimation tendency for samples above 50% clay. SOC models, however, had difficulty in predicting higher contents of SOC, presenting a heavier right tail. RF was considered the best performing model for clay, GBM for sand contents, and SVM for SOC due to its more reliable predictions of low SOC values. All algorithms had poor performance when predicting SOC above 2% (Figure 3). The complete set of observed vs. predicted plots and QQ plots are presented in the Supplementary Material (Figures S1, S2, and S3).

Si, Fe, and Ca were the most relevant covariates for clay, sand, and SOC content predictions, respectively (Figure 4). K and Al for texture, and Depth for SOC were noteworthy due to their impact in estimation error. Stronger interactions between covariates were observed in KNN and RF models, showing that these algorithms found more complex relationships between covariates compared to the other models. Comparatively, GBM relied less on interactions between covariates, and SVM showed negligible interactions. Si and Ca presented the strongest interactions in soil texture models, while Ca and Depth interacted the most with other covariates in SOC models (Figure 4).

Table 3. Descriptive performance statistics for the training and test datasets of soil organic carbon (SOC), sand, and clay content models. Samples were collected from the campus of the Federal University of Lavras, Brazil. SD: standard deviation; Max: maximum value; Min: minimum value; Q3 and 1: third and first quantiles; IQR: interquantile range.

Target variable	Machine learning models	ρ_c		R^2		RMSE		MAE		RPIQ	
		traini ng	ρ_c test	traini ng	R^2 test	training	test	traini ng	MAE test	traini ng	RPIQ test
Clay (%)	KNN	0.84	0.79	0.77	0.73	5.87	7.77	4.62	5.79	3.07	2.57
	RF†	0.96	0.84	0.96	0.83	2.93	6.63	2.22	5.24	6.14	3.01
	GBM	0.96	0.87	0.92	0.80	3.29	6.34	2.46	4.97	5.47	3.16
	SVM	0.82	0.83	0.69	0.73	6.49	7.35	4.59	5.48	2.78	2.72
Sand (%)	KNN	0.86	0.80	0.79	0.74	5.52	6.94	4.45	5.20	3.44	3.17
	RF	0.97	0.88	0.96	0.84	2.56	5.53	1.96	4.16	7.43	3.98
	GBM†	0.95	0.87	0.90	0.79	3.70	6.04	2.86	4.72	5.13	3.65
	SVM	0.88	0.85	0.79	0.78	5.41	6.22	3.94	5.08	3.51	3.54
SOC (%)	KNN	0.85	0.72	0.80	0.53	0.71	0.97	0.52	0.62	2.65	1.91
	RF	0.97	0.85	0.95	0.73	0.34	0.72	0.24	0.51	5.53	2.59
	GBM†	0.90	0.80	0.83	0.65	0.62	0.83	0.45	0.56	3.06	2.25
	SVM	0.86	0.79	0.77	0.63	0.72	0.87	0.51	0.57	2.64	2.13

†Models elected as best performing from evaluation metrics and residual analysis presented in Figure 3 and in the Supplementary Material (Figures S1, S2, and S3).

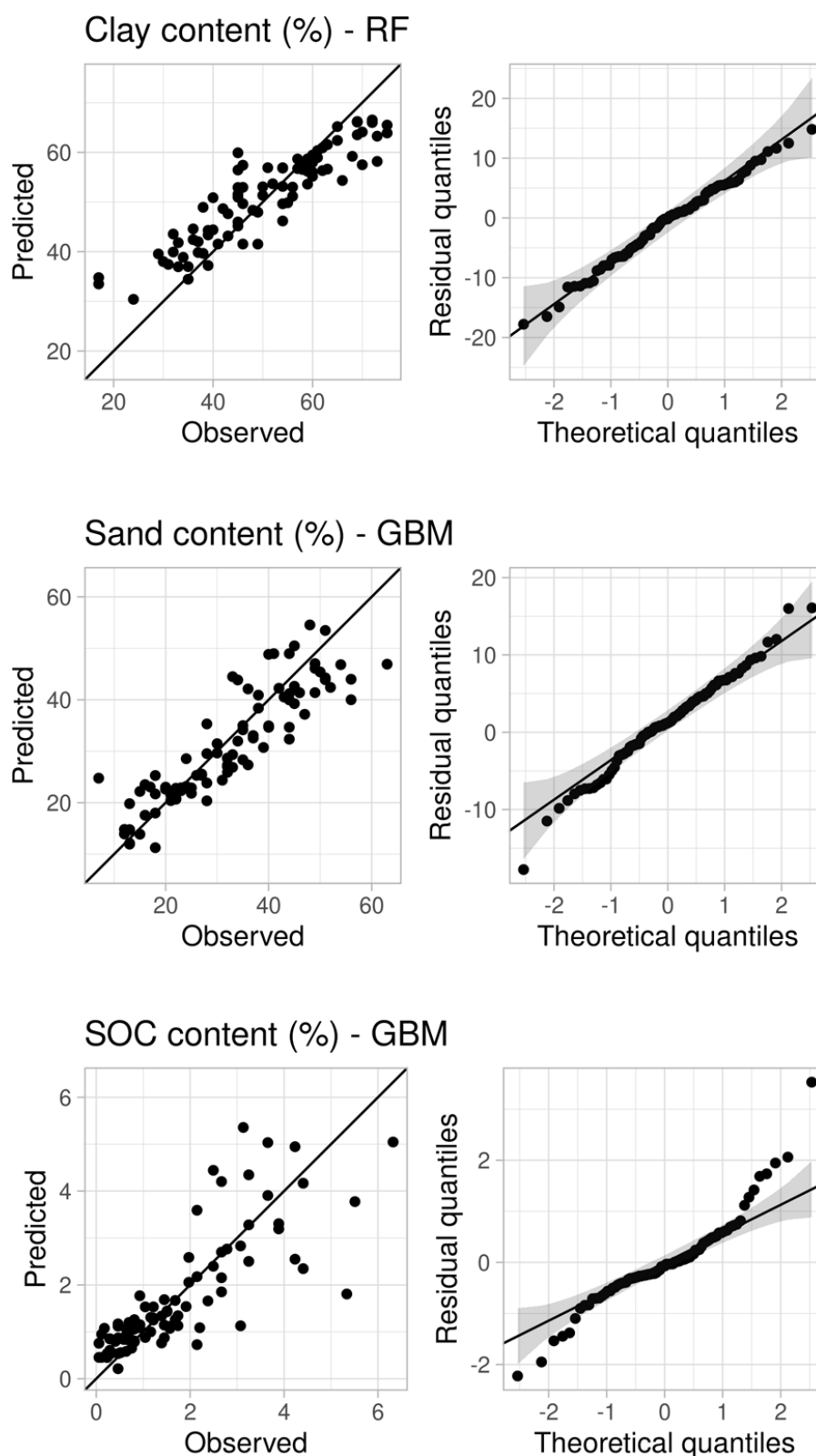
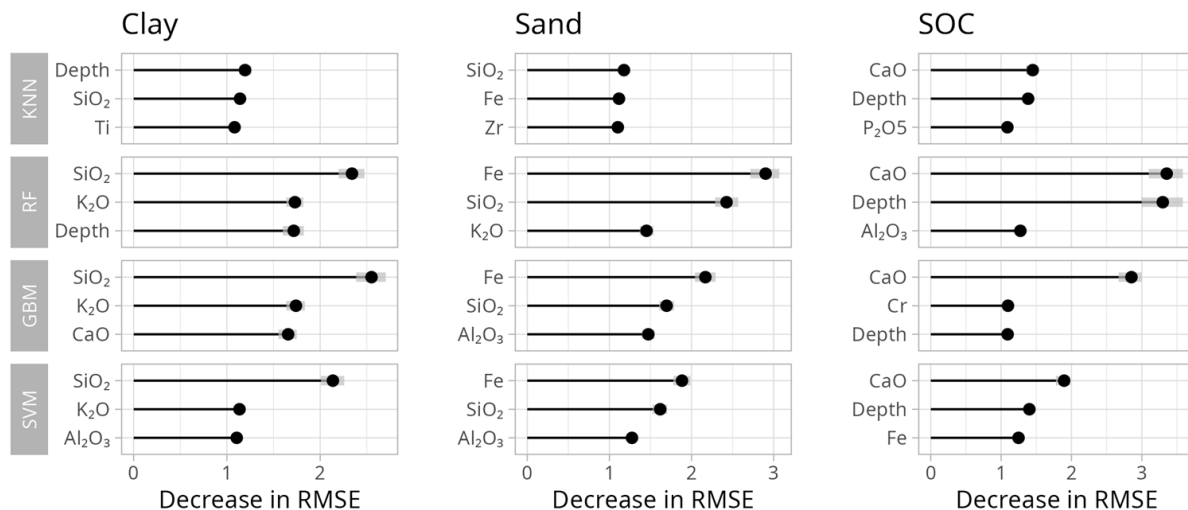


Figure 3. Observed vs. predicted plots (left) and QQ plots (right) of models trained with portable fluorescence spectrometry (pXRF) data to estimate clay, sand, and soil organic carbon contents. Models were trained with 203 data points, but results presented refer to the test dataset, comprising of 89 data points unseen by models during training. This figure contains only the best performing models, consisting of random forests (RF) and gradient boosting machines (GBM). The complete set of plots is presented in the Supplementary Material (Figures S1, S2, and S3).

Covariate importances



Interactions between covariates

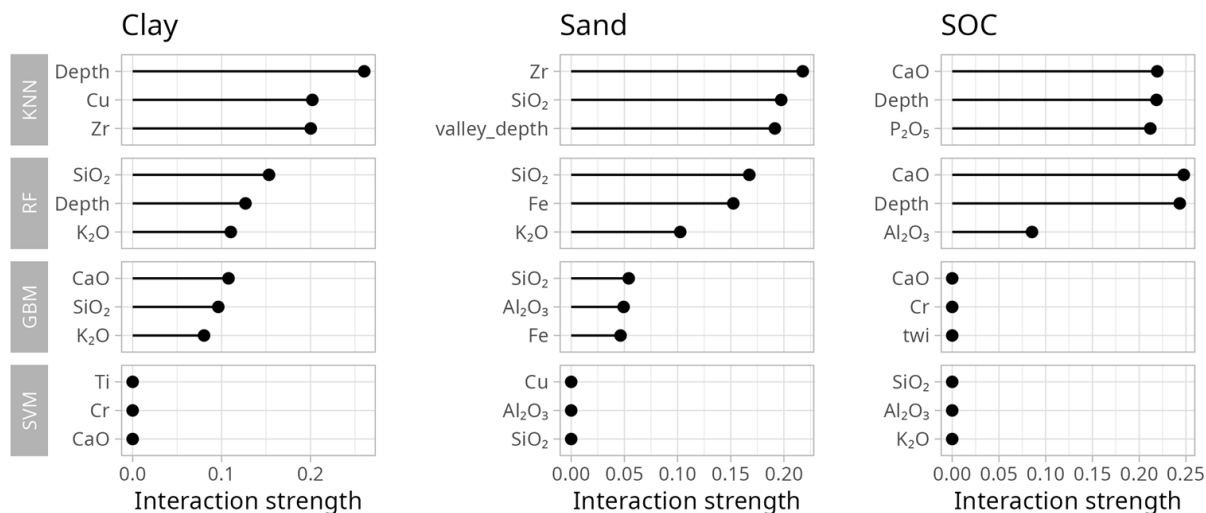


Figure 4. The top three covariates that impacted prediction error the most and that presented the strongest global interactions within the trained machine-learning models. The relevance of covariates was measured by Feature Permutation Importance (FPI) using 1000 repetitions. The models were trained to estimate clay, sand, and organic carbon (SOC) contents in 292 soil samples.

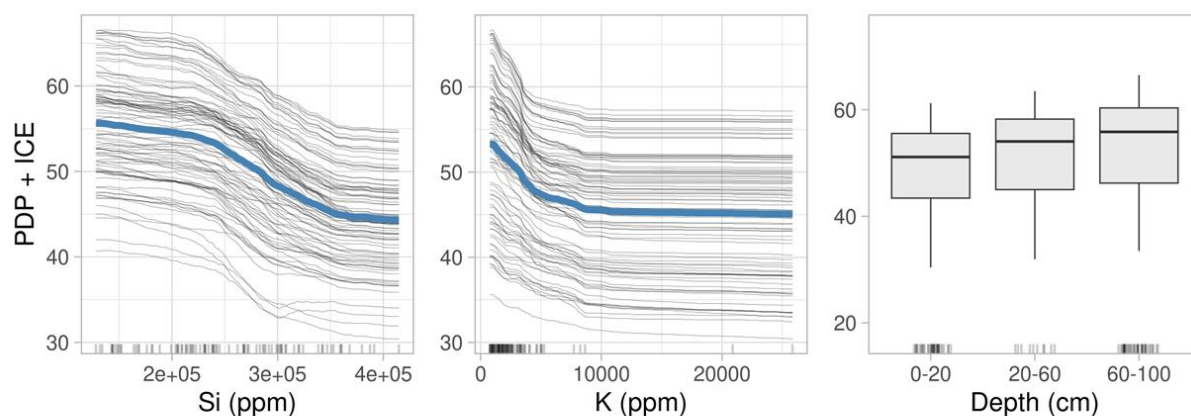
3.3. Clay prediction models

Samples low in Si and K contents were assumed by models to be clay-rich (Figure 5), since these elements are constituents of quartz (SiO_2) and mica, mainly muscovite ($\text{KAl}_2(\text{AlSi}_3\text{O}_{10})(\text{OH},\text{F})_2$), and these minerals are seldom found in the clay fraction of highly weathered soils. Increasing depth also had a positive effect on clay content predictions, which can be explained by the translocation of clay downward in soil profiles during the weathering process, forming the Argissolos (Ultisols) of the area.

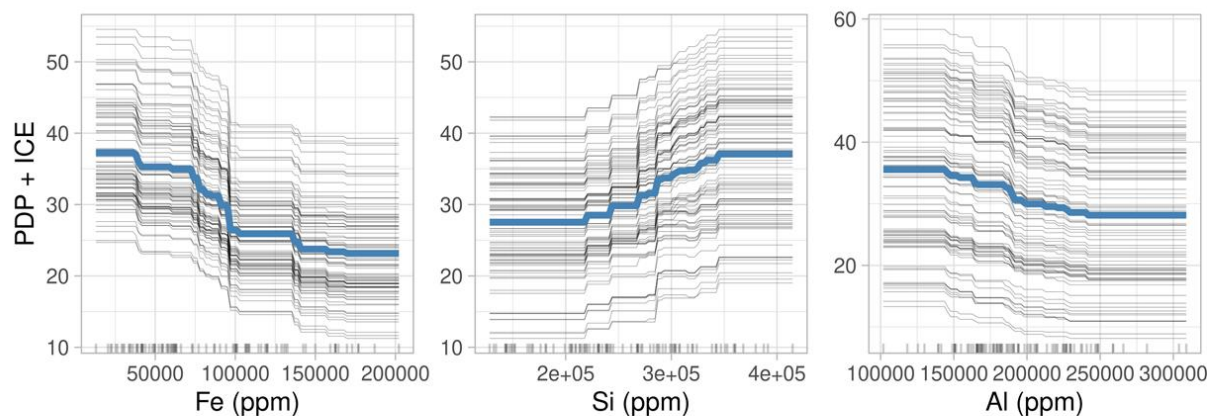
An interesting interaction between Si and K was observed wherein the simultaneous occurrence of low Si and K contents further increased the estimated values of clay content

beyond the main effects (Figure 6). This interaction seems to indicate that models use K contents to separate soils derived from gabbro and gneiss, which could be clearly observed when plotting the parent materials of samples on top of the interactions (Figure 6). Gabbro does not contain many K-bearing minerals that are resistant to weathering, such as muscovite, whereas muscovite is commonly found in felsic bands of gneiss and tends to remain in small quantities in the sand and silt fractions even after long weathering periods (Wilson, 2004). Models estimated that samples low in Si and K (gabbro) tend to have higher clay content, while some combinations of comparatively higher Si and K contents (gneiss) have their clay predictions decrease (Figure 6).

Covariate effects on clay content predictions - (RF)



Covariate effects on sand content predictions - (GBM)



Covariate effects on SOC content predictions - (GBM)

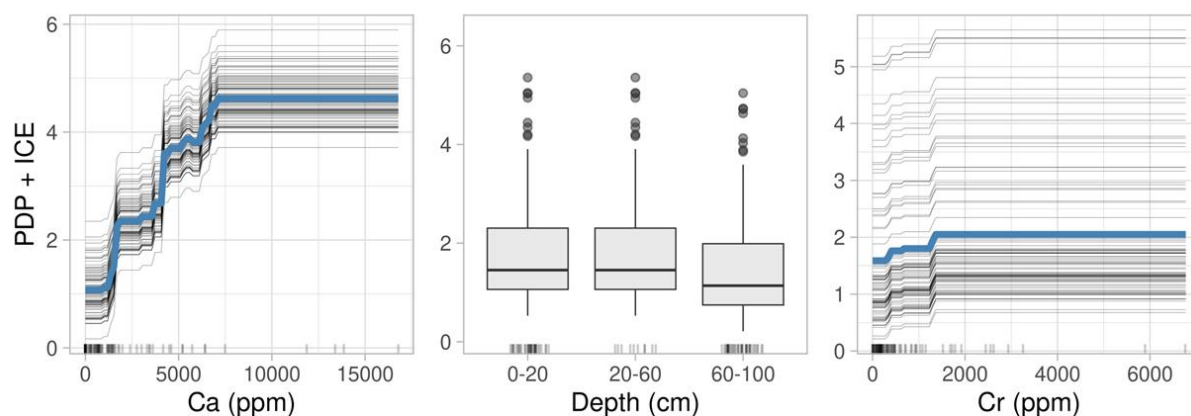


Figure 5. The effects of the top three most relevant covariates in the estimation of clay, sand, and soil organic carbon (SOC) contents (from left to right, from most to least relevant). The effects were measured by partial difference plots (PDP), which describe the global effects of covariates shown as blue lines, and by individual accumulated expectation (ICE), representing the effects of covariates on individual samples, shown as black lines. Random forest (RF) and gradient boosting machine models (GBM) were used. Rugged lines in the x axis indicate the density of samples. The models were trained with 203 soil samples, but the effects were calculated using 89 samples of the test set, unseen during the training phase.

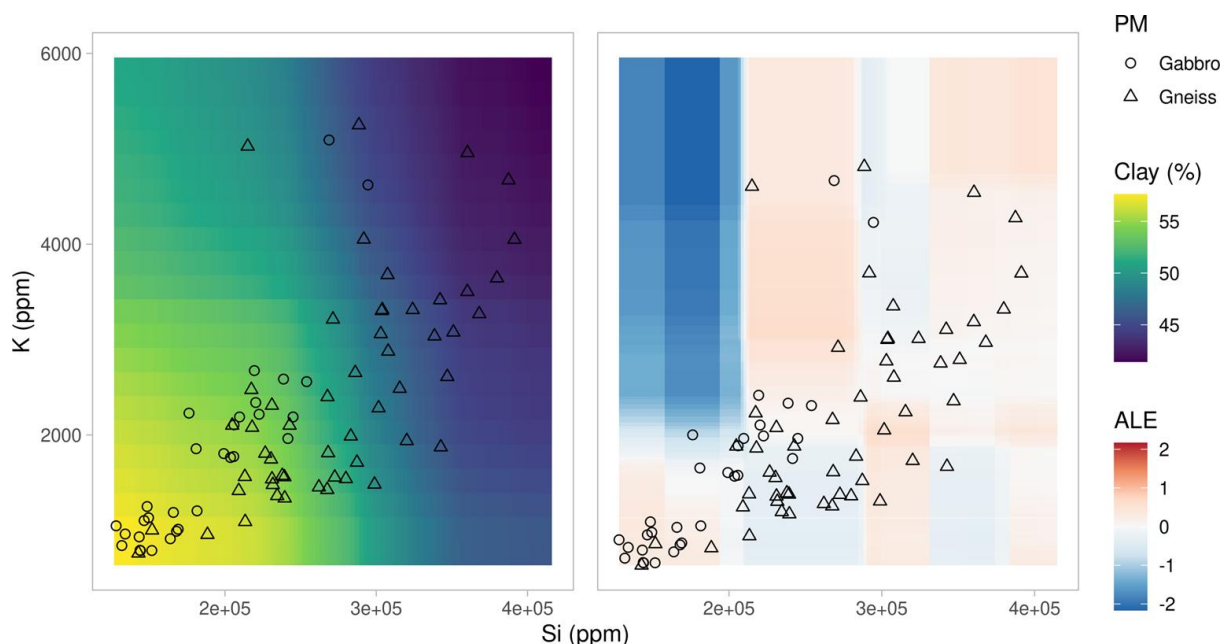


Figure 6. The combined main effects of K and Si contents (left) and the interaction between these two variables (right) in a random forest (RF) model trained to predict the content of clay in soil samples. The main effects were extracted from partial difference plots (PDP), whereas interactions represent the accumulated local effects (ALE), describing only the effect of different combinations of both covariates while discounting the main effect. The model was trained with 203 soil samples, but the effects were calculated using 89 samples of the test dataset, unseen during training. Five samples with high K contents were clipped from the plots for better visualization, so only 84 samples are shown.

3.4. Sand prediction models

Models inferred that sand contents decrease with the increase of Fe, predicting sand to be lower than the mean (32%) when Fe is above roughly 82,000 ppm (Figure 7A). The opposite trend was found for Si, with sand content increasing with the increase of sand (Figure 5). The GBM model was considered the best sand model (Table 3), but its pairwise interactions were weak. The RF model presented the strongest interactions, good evaluation metrics (Figure 4 and Table 3), and PDPs and ICEs similar to the GBM model, being therefore used as a reference to investigate interactions between covariates (Figure 8). The strongest pairwise interactions were found between Si and the elements Al, Cr, and Cu. The negative effect of low Si on sand content was more pronounced in samples with lower Al contents (Figure 8). This is an interesting insight showing that sand content models also created interactions based on parent materials, as samples from gabbro tend to be low in both Si and Al contents (Figure 8). The interaction between the effect of Si and the contents of Cr and Cu were similar.

Following that high sand content estimations were associated with Fe-poor and Si-rich soil samples, the largest residuals occurred due to the combination of Fe and Si contents in samples not corresponding to the contents of sand expected by models. On one hand,

overestimations happened when Fe contents were not sufficiently high and Si contents not sufficiently low to match the observed low sand contents (sample 196, Figure 7C). On the other hand, underestimations happened when samples were not sufficiently low in Fe and high in Si to increase sand estimations enough to match the observed sand contents (sample 66, Figure 7D). The strong effect of Fe and Si and the lesser impact of other covariates on estimations is an indication that this model is only fit for weathered soils wherein the sand fraction is predominantly composed of Si (quartz-rich sand) and bears low contents of Fe (goethite- and hematite-bearing clay minerals). Hence, these models are likely not transferable to soils in which the sand fraction contains a wider variety of minerals.

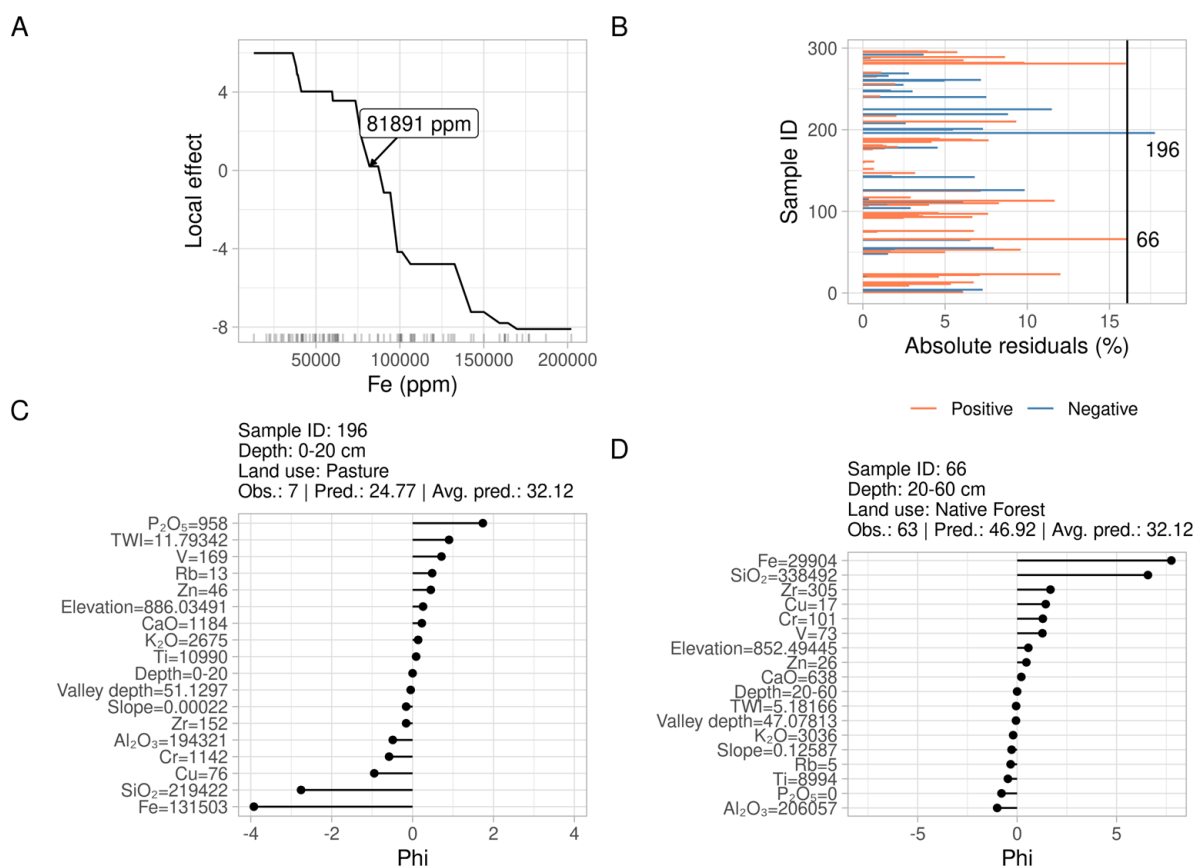


Figure 7. Accumulated local effects (ALE) of the most important covariate in sand content predictions (A), absolute residuals of predictions made using the test dataset ($n = 89$) (B), and Shapley values calculated for the two largest residuals (C and D). All panels reflect results from the gradient boosting model (GBM), considered the best performing model for sand content predictions (Table 3). Obs.: observed value; Pred.: value predicted by the model; Avg. pred.: average prediction of the model.

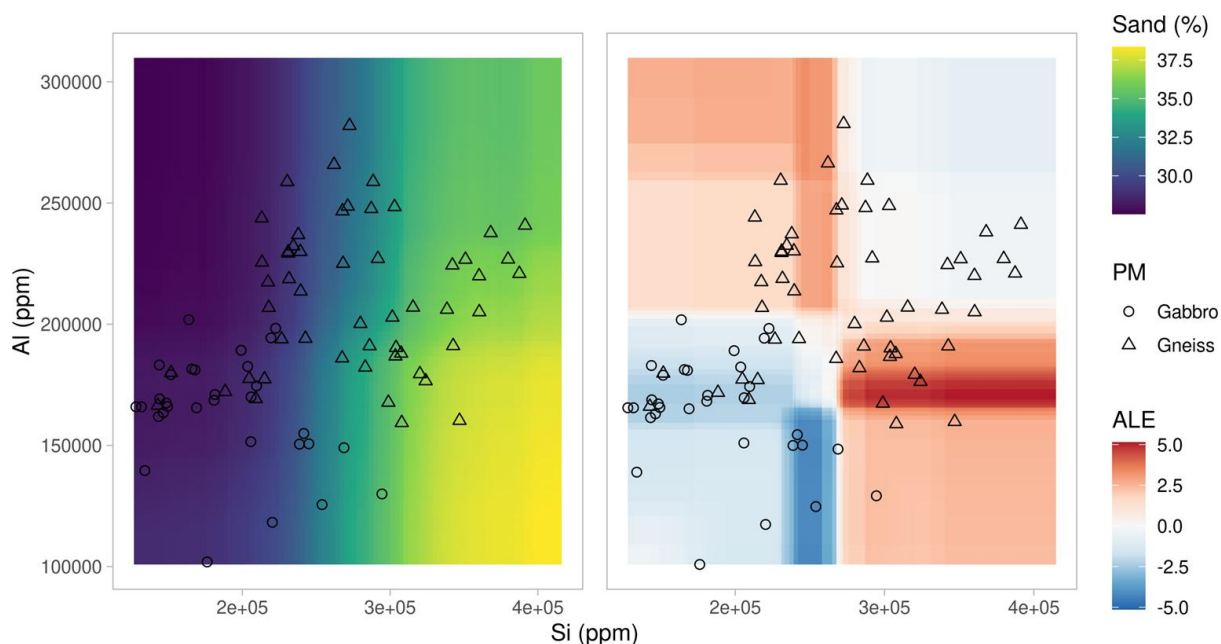


Figure 8. The combined main effects of Al and Si contents (left) and the interaction between these two variables (right) in a random forest (RF) model trained to predict the content of clay in soil samples. The main effects were extracted from partial difference plots (PDP), whereas interactions represent the accumulated local effects (ALE), describing only the effect of different combinations of both covariates while discounting the main effect. The model was trained with 203 soil samples, but the effects were calculated using 89 samples of the test dataset, unseen during training.

3.5. SOC prediction models

The effect of Ca in the tree-based models (RF and GBM) was similar and sharply positive until 7,500 ppm, after which Ca contents had no substantial effect on SOC predictions (Figure 9A). Residual analysis showed that the weakness of SOC models was the underestimation of SOC contents higher than 2%, observed in both training and test datasets (Figure 3). The largest overestimation happened in a coffee plantation, where the anthropic input of Ca substantially contributed to increase the predicted value (sample 288, Figure 9C). On the contrary, the highest underestimation happened under a eucalyptus plantation, where SOC content was high, but the Ca content was low, inducing a prediction close to the mean due to the absence of the effect of Ca (sample 240, Figure 9D). These two cases depict the overreliance of the model on Ca content, a liability that is aggravated by the scarcity of samples with high SOC content, as only 25% of the data has SOC contents above 2.5%. The most important interaction for SOC models was that with depth, as organic matter is deposited and accumulates mainly in the soil surface, causing most of SOC to be concentrated in the topmost horizons.

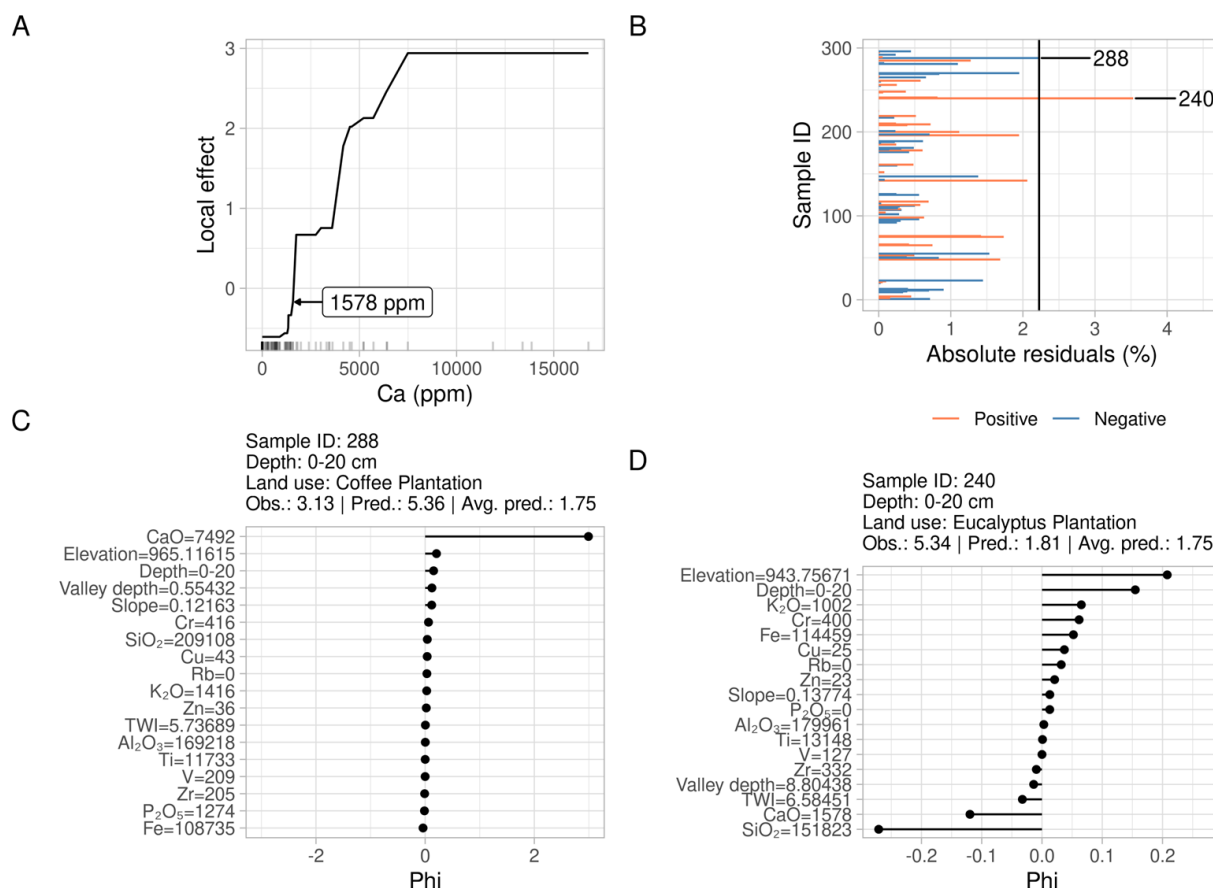


Figure 9. Accumulated local effects (ALE) of the most important covariate in soil organic carbon (SOC) content predictions (A), absolute residuals of predictions made using the test set ($n = 89$) (B), and Shapley values calculated for the two largest residuals (C and D). All panels reflect results from the gradient boosting model (GBM), considered the best performing model for SOC content predictions (Table 3). Obs.: observed value; Pred.: value predicted by the model; Avg. pred.: average prediction of the model.

4. Discussion

The role of total Si and Fe contents in explaining soil texture stems from the presence of quartz (SiO₂) in the sand fraction, and Fe oxides in the clay fraction of tropical soils (Rezende et al., 2022; Schaefer et al., 2008). As soils undergo weathering, relatively more soluble primary minerals dissolve, while sand-sized grains of quartz, which are very insoluble, remain preserved. Elements released from the dissolution of primary minerals precipitate, forming smaller and weathering-resistant clay minerals rich in Al and Fe (e.g., kaolinite, goethite, hematite, and gibbsite). Therefore, it is common for total Si to be higher in the sand fraction due to the abundance of quartz (He et al., 2022; Rezende et al., 2022), and for Fe and Al to increase in finer fractions as many clay minerals bear Al or Fe in their composition, especially in tropical environments where weathering is particularly intense (Journet et al., 2014; Schaefer et al., 2008). The models correctly identified these pedogenetic processes by associating low

Fe and Al contents with higher concentrations of sand, and low Si with higher contents of clay (Figure 5).

Parent material was also a key driver of soil texture. The negative effect of K in clay content predictions (Figure 5) was associated with the presence of sand-sized muscovite, a K-bearing phyllosilicate that occurs in the felsic bands of gneiss. Similar to quartz, muscovite is very resistant to weathering, and there are evidences showing that it becomes less resistant after it is broken down to clay-sized grains (Wilson, 2004). This is supported by Mancini et al. (2023) and Silva et al. (2021b), who studied soils in the same region and observed higher presence of muscovite in the sand and silt fractions compared to the clay fraction. The pairwise interactions between Si and K (Figure 6) show that the association between sand-sized muscovite and sand content of soils derived from gneiss was implicitly learned by models. Both main effects and interactions are driven by differences between the chemical composition of gabbro, composed of Fe-bearing minerals (e.g., pyroxenes), and gneiss, in which Si-rich and K-bearing minerals (e.g., quartz and muscovite) are present in felsic bands.

The reliance of SOC models on total Ca content comes from the high cation exchange capacity (CEC) of organic matter. The mineralogical composition of most tropical soils is dominated by kaolinite and gibbsite, along with Fe oxides, minerals whose charges are pH-dependent (for Fe and Al oxides, charges tend to be positive, forming anion exchange capacity instead of cation exchange capacity) and much lower than that of 2:1 phyllosilicates that are more abundant in soils of temperate climate, such as smectites and vermiculites. For this reason, the capacity of weathered-leached tropical soils to retain cations originates mostly from organic matter, making available cation contents good indicators of the presence of SOC in soils. Ca is particularly useful as an indicator for two main reasons. First, Ca has a smaller hydrated radius compared to other common cations in the soil solution (e.g. Mg^{2+} , K^{+} , and NH_4^{+}), being more strongly held by adsorption sites (Sparks et al., 2024). And second, in sufficiently weathered A horizons, Ca is not present in the crystalline structure of minerals, since Ca-bearing minerals are more soluble than feldspars and quartz, therefore the total Ca content detected by pXRF can be almost entirely attributed to available Ca adsorbed on organic matter. This latter argument is supported by Silva et al. (2019), who found a strong correlation between plant-available Ca contents and total Ca measured by a pXRF device in tropical soils.

Although Ca is a useful indicator of SOC, the dependence of models on Ca content is a liability. Ca can vary with liming practices and with time due to the presence of plant roots that can uptake Ca. Another problem arises when Ca is present in the structure of minerals or when is adsorbed by reactive 2:1 phyllosilicates (although such cases are rare in tropical soils). In

such cases, total Ca contents detected by pXRF would no longer represent only the available Ca present on SOC adsorption sites, but also the Ca present in the structure of minerals or held by their surfaces. It is desirable to find models that rely on variables that do not change depending on management practices and whose effects do not substantially change depending on the mineralogical composition of soils. These limitations are likely connected to the heavier right tails of residuals (Figure 5), which suggest that models cannot properly fit samples with higher SOC content. Shapley values are also evidence of the issues caused by overreliance on Ca, showing that models are unable to output accurate predictions when SOC is not purely explained by Ca content. This was observed when Ca was high due to the liming in cultivated areas (sample 288, Figure 9C), or in forest plantations where soil fertility management is not frequently conducted, resulting in lower Ca contents even in samples that are high in SOC (sample 240, Figure 9D). Improving SOC models will require obtaining more samples to represent the right tail of SOC's distribution and finding auxiliary covariates capable of explaining SOC that are less prone to anthropic influence and preferably whose effects are transferable to younger soils with more reactive 2:1 clay minerals.

4.1. The importance of model interpretation

Reporting simply that a variable is important for predictions may not seem like a novel discovery, and in fact it is a common practice in machine-learning research. Yet it is worthwhile emphasizing the relevance of this information and their often overlooked implications. The reliance of SOC predictions on Ca content is an example of important pitfalls of machine-learning models that are often neglected. With the current SOC model, soil samples that have high SOC content but bear low amounts of Ca in their exchange sites would invariably be underestimated.

Similarly, the asymptotic behaviors in the PDPs and ALEs of tree-based models (RF and GBM) show a commonly overlooked limitation: their inability to extrapolate beyond the range of values of covariates used to train them. This happens because tree-based models create splitting rules, dividing samples depending on their covariates' values. Naturally, rules can only be created within the ranges of data used to train the models; beyond that, there are no rules created to differentiate samples further. Consequently, when tree-based models are used to predict data with covariate values beyond those used for training, they will tend to output the same prediction for every sample. This is particularly obvious in the PDPs of the SOC models (Figure 5), which are incapable of differentiating samples with more than 7,500 ppm of Ca.

Some of the highlights found in this study include i) the dependence of SOC models on Ca; ii) the stratification of parent materials learned by models based on Si, Fe, and K contents; iii) and the ability to confirm that models are learning physically accurate relationships between covariates and soil properties, but they need a pedologist's interpretation for benefiting from their full potentials. This is key information that is often ignored when creating and shipping machine-learning models, but that could be made explicit by the diagnostic techniques explored in this study. Scientific research is a collective effort and, if machine-learning models are to become ubiquitous in science, it will require that scientists find effective and transparent ways to communicate what models are learning, and what can be learned from them.

5. Conclusions

This study presented insights on what machine-learning models could learn from pXRF data when trained to predict soil texture (sand and clay contents) and SOC in tropical soils. Total contents of Si, Al, Fe, and K were the main drivers of soil texture and differed substantially between soils derived from gabbro and gneiss, attesting the influence of parent material even after long and intense weathering. Models implicitly learned that low contents of Si, K, and Al identified soils derived from gabbro and were associated with higher clay and lower sand contents. Sand and clay models were reliable and can likely be generalized to tropical soils where Si and Fe are associated with the sand and clay fractions, respectively, and whose concentrations are found within the ranges of soils studied herein. Total Ca contents explained most of SOC variability, but this relationship cannot be easily generalized due to Ca being highly influenced by land use. In the tropical soils studied, overreliance on Ca contents was found to be a limitation of SOC models trained with pXRF data. In similar environments, SOC models will require covariates that are less influenced by land use and management practices for appropriate modeling. Model interpretation methods help in sharing the pedologic knowledge learned by machine-learning models and reveals models' strengths and liabilities, contributing to advance the knowledge by explicitly showing the relations found between target and predictor variables.

Supplementary Material

Table S1. Descriptive statistics of the train and test sets of covariates used to train and evaluate machine-learning models built to predict soil organic carbon (SOC), sand, and clay contents in the campus of the Federal University of Lavras, Brazil. Total elemental contents were determined via portable X-ray fluorescence and topographic attributes were calculated from a 5-m resolution digital elevation model. SD: standard deviation; Max: maximum value; Min: minimum value; Q3 and 1: third and first quantiles; IQR: interquartile range; rad: radians.

Training dataset								
Covariate	Mean	SD	Max	Q3	Median	Q1	Min	IQR
Al (ppm)	185411.21	35070.22	293321.00	213858.00	179672.00	160966.00	103111.00	52398.00
Si (ppm)	253092.23	69405.60	411333.00	303677.00	260897.00	204898.00	101184.00	98544.50
Ca (ppm)	2090.87	3040.70	20453.00	2849.00	873.00	302.00	0.00	2394.50
K (ppm)	2975.38	3442.97	34101.00	3597.00	2315.00	1606.00	705.00	1974.00
P (ppm)	364.75	444.25	2671.00	553.00	241.00	0.00	0.00	549.50
Rb (ppm)	10.16	19.16	221.00	12.00	8.00	3.00	0.00	9.00
Fe (ppm)	84206.45	50292.12	239446.00	113576.00	70374.00	45233.00	11792.00	65624.00
Cu (ppm)	39.04	31.29	183.00	50.00	29.00	18.00	8.00	31.50
Cr (ppm)	770.68	1080.15	5471.00	1018.00	260.00	130.00	0.00	864.00
Ti (ppm)	8988.68	3447.19	17665.00	11381.00	9074.00	6309.00	2412.00	5015.50
V (ppm)	100.85	131.89	624.00	159.00	40.00	0.00	0.00	155.50
Zn (ppm)	37.95	19.06	123.00	44.00	33.00	25.00	14.00	19.00
Zr (ppm)	207.77	47.31	354.00	239.00	205.00	171.00	124.00	68.00
Elevation	897.00	28.77	976.80	914.80	892.55	873.06	845.78	39.62
Slope (rad.)	0.19	0.08	0.44	0.25	0.17	0.13	0.00	0.11
Valley depth (m)	25.08	15.93	60.51	37.60	24.12	11.29	0.00	25.28
TWI	6.36	1.58	16.44	7.21	6.20	5.46	3.17	1.72
Test dataset								
Covariate	Mean	SD	Max	Q3	Median	Q1	Min	IQR
Al (ppm)	194558.20	39924.15	308821.00	225022.00	187946.00	167458.00	101877.00	57564.00
Si (ppm)	250591.93	71729.80	414752.00	303005.00	239510.00	204779.00	128053.00	98226.00
Ca (ppm)	1917.52	3143.91	16782.00	1750.00	758.00	210.00	0.00	1540.00
K (ppm)	2881.18	3505.80	25849.00	3077.00	2079.00	1427.00	761.00	1650.00
P (ppm)	317.31	395.05	1443.00	523.00	207.00	0.00	0.00	523.00
Rb (ppm)	8.89	14.08	93.00	10.00	6.00	2.00	0.00	8.00
Fe (ppm)	85192.67	47350.96	202181.00	118627.00	72937.00	46273.00	12970.00	72354.00
Cu (ppm)	35.07	21.12	118.00	44.00	34.00	17.00	10.00	27.00
Cr (ppm)	663.46	1113.16	6775.00	616.00	248.00	132.00	0.00	484.00
Ti (ppm)	9903.91	3946.11	18600.00	12981.00	9490.00	6615.00	3241.00	6366.00
V (ppm)	117.17	115.01	488.00	206.00	87.00	16.00	0.00	190.00
Zn (ppm)	34.67	16.93	107.00	40.00	30.00	24.00	14.00	16.00
Zr (ppm)	233.75	54.27	358.00	262.00	229.00	194.00	98.00	68.00
Elevation	897.94	34.32	973.29	923.77	891.89	866.65	849.69	57.12
Slope (rad.)	0.18	0.10	0.48	0.23	0.15	0.13	0.00	0.10
Valley depth (m)	23.99	16.51	51.30	34.83	26.19	8.80	0.33	26.03
TWI	6.48	1.39	11.79	6.73	6.41	5.70	4.13	1.04

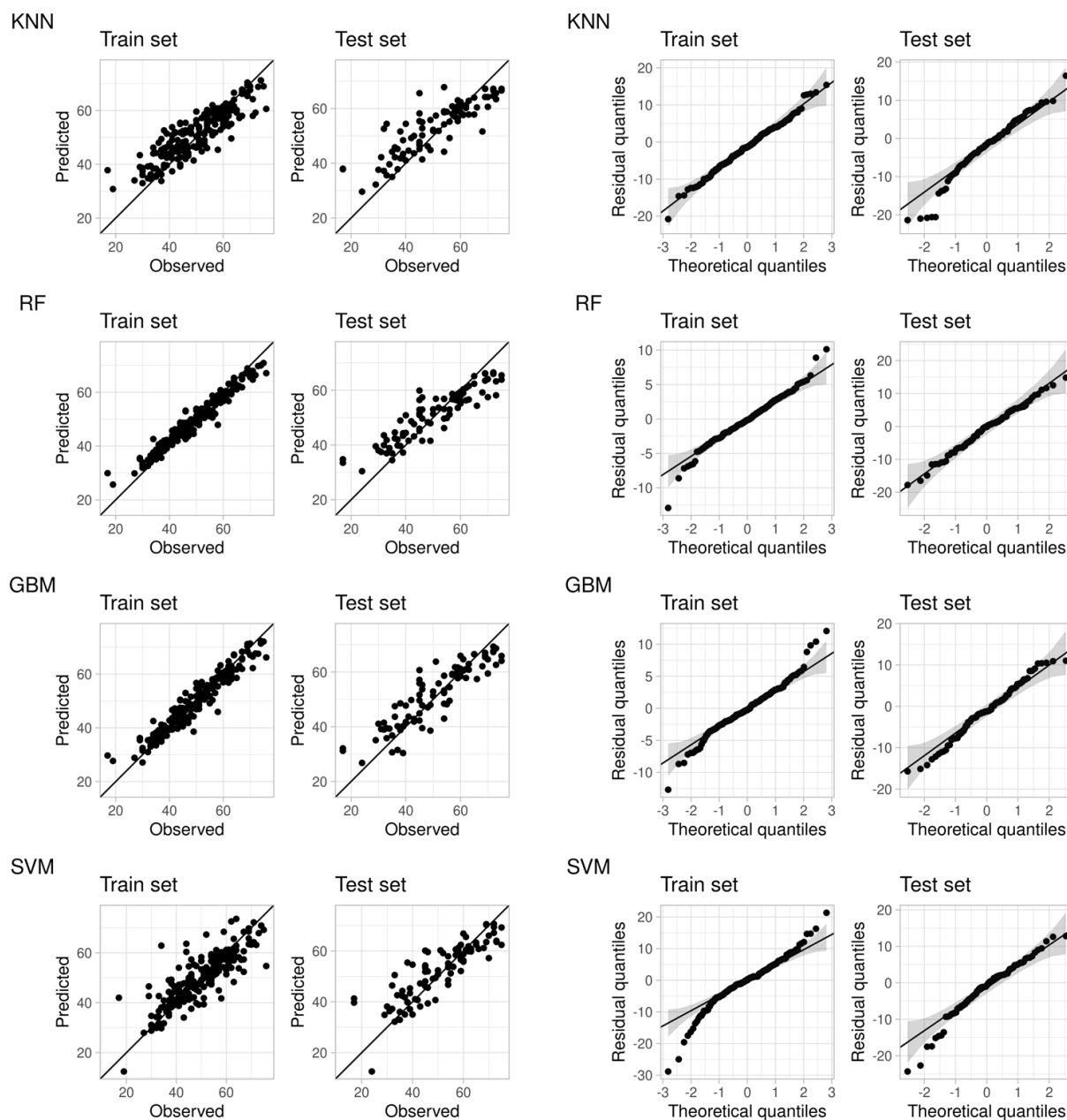


Figure S1. Performance of clay content (%) prediction models. Comparison between observed and predicted values (left) and normality of residuals (right) for the train (n = 203) and test (n = 89) sets. KNN: K-nearest neighbors; RF: random forest; GBM: gradient boosting machines; SVM: support vector machines.

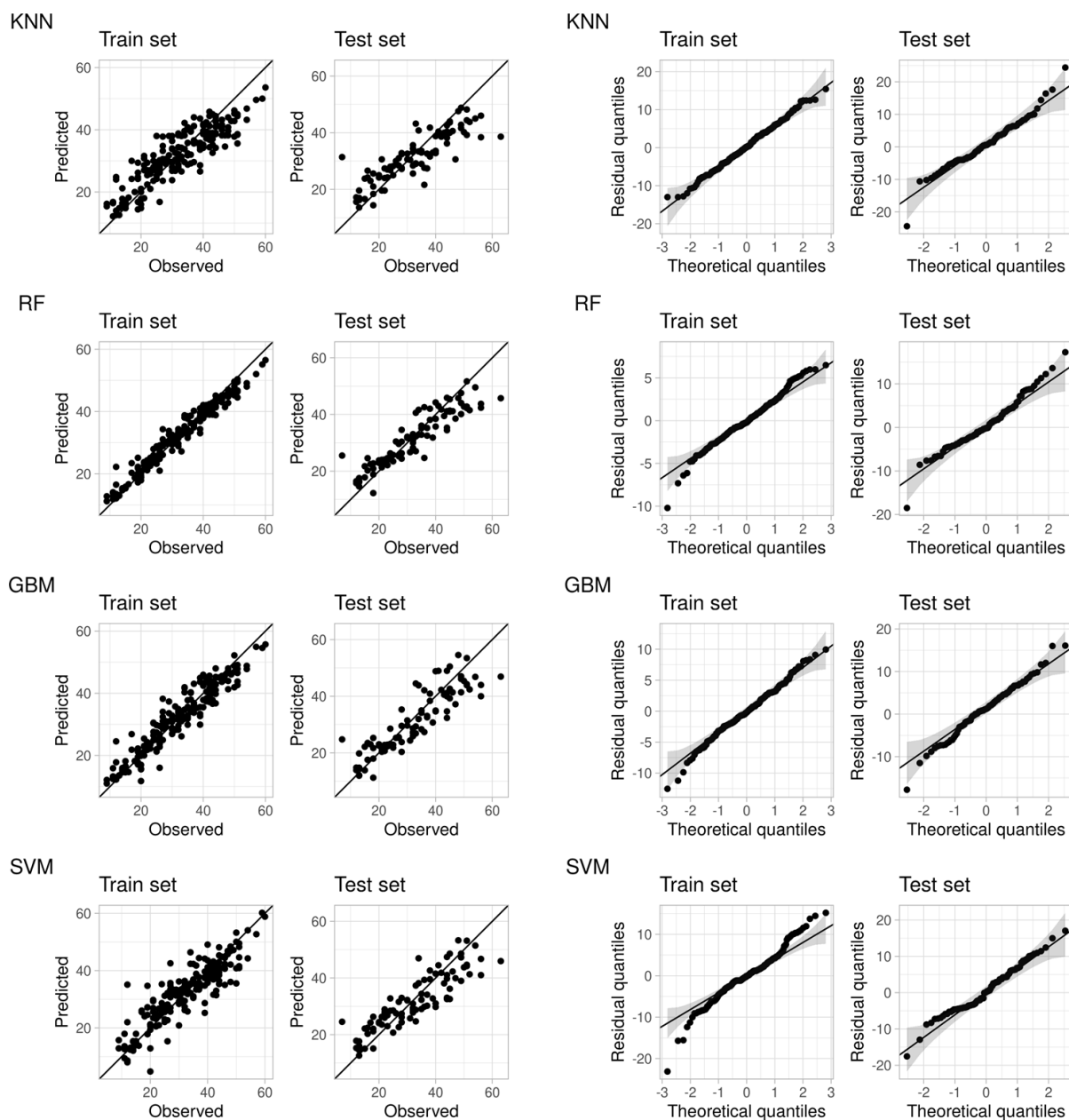


Figure S2. Performance of sand content (%) prediction models. Comparison between observed and predicted values (left) and normality of residuals (right) for the train (n = 203) and test (n = 89) sets. KNN: K-nearest neighbors; RF: random forest; GBM: gradient boosting machines; SVM: support vector machines.

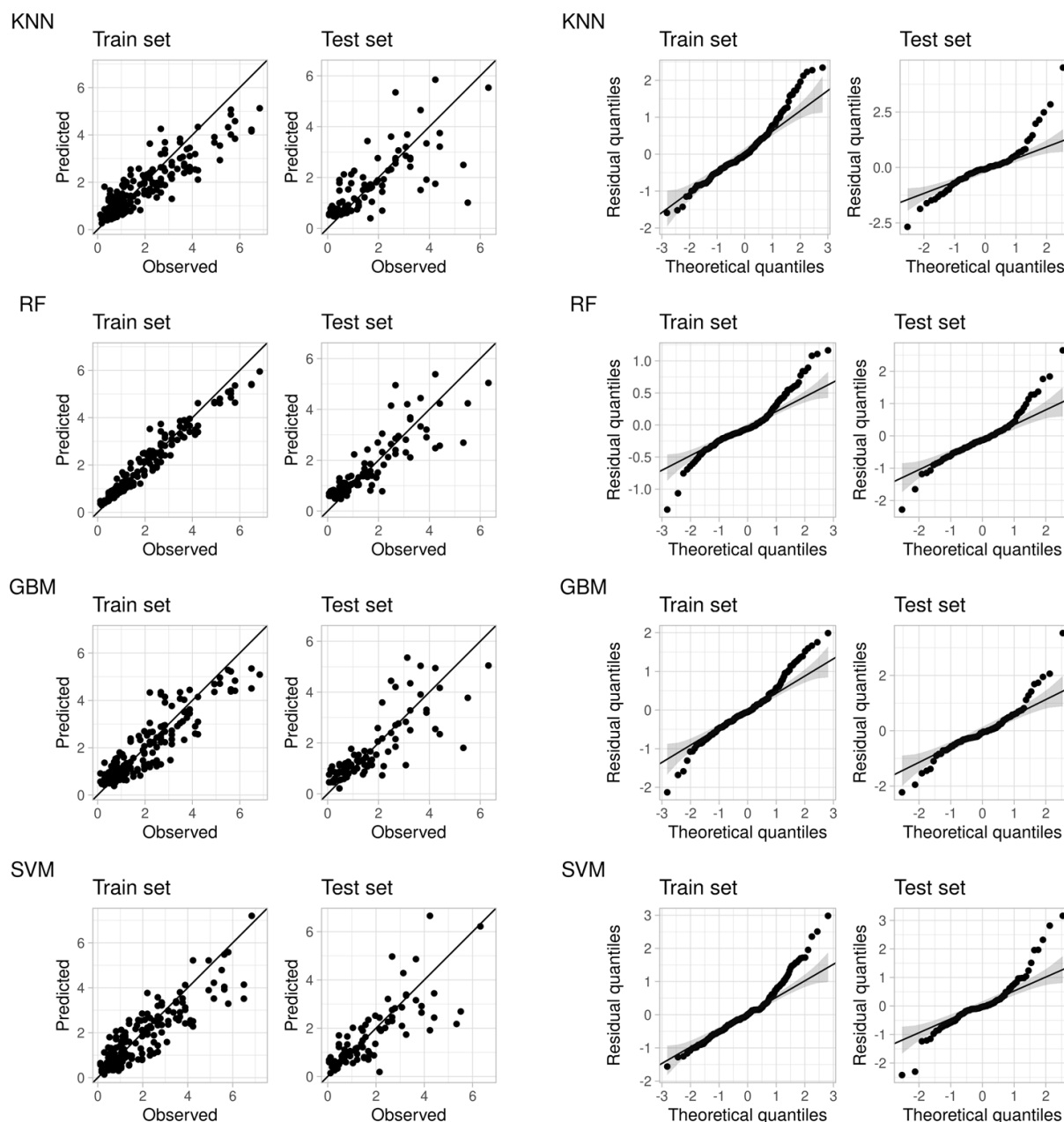


Figure S3. Performance of soil organic carbon (SOC) content (%) prediction models. Comparison between observed and predicted values (left) and normality of residuals (right) for the train (n = 203) and test (n = 89) sets. KNN: K-nearest neighbors; RF: random forest; GBM: gradient boosting machines; SVM: support vector machines.

References

- Andrade, R., Faria, W.M., Silva, S.H.G., Chakraborty, S., Weindorf, D.C., Mesquita, L.F., Guilherme, L.R.G., Curi, N., 2020a. Prediction of soil fertility via portable X-ray fluorescence (pXRF) spectrometry and soil texture in the Brazilian Coastal Plains. *Geoderma* 357, 113960. <https://doi.org/10.1016/j.geoderma.2019.113960>
- Andrade, R., Silva, S.H.G., Faria, W.M., Poggere, G.C., Barbosa, J.Z., Guilherme, L.R.G., Curi, N., 2020b. Proximal sensing applied to soil texture prediction and mapping in Brazil. *Geoderma Reg.* 23, e00321. <https://doi.org/10.1016/j.geodrs.2020.e00321>
- Alvares, C.A., Stape, J.L., Sentelhas, P.C., Moraes Gonçalves, J.L., Sparovek, G., 2013. Köppen's climate classification map for Brazil. *Meteorol. Zeitschrift* 22, 711–728. <https://doi.org/10.1127/0941-2948/2013/0507>
- Apley, D.W., Zhu, J., 2020. Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 82, 1059–1086. <https://doi.org/10.1111/rssb.12377>
- Arrouays, D., McBratney, A., Bouma, J., Libohova, Z., Richer-de-Forges, A.C., Morgan, C.L.S., Roudier, P., Poggio, L., Mulder, V.L., 2020. Impressions of digital soil maps: The good, the not so good, and making them ever better. *Geoderma Regional* 20, e00255. <https://doi.org/10.1016/j.geodrs.2020.e00255>
- Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., Böhner, J., 2015. System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. *Geoscientific Model Development* 8, 1991–2007. <https://doi.org/10.5194/gmd-8-1991-2015>
- Curi, N., Silva, S.H.G., Poggere, G.C., Menezes, M.D., 2017. Mapeamento de solos e magnetismo no campus da UFPA como traçadores ambientais. UFPA, Lavras, MG, Brazil.
- Bockheim, J.G., Gennadiyev, A.N., 2000. The role of soil-forming processes in the definition of taxa in Soil Taxonomy and the World Soil Reference Base. *Geoderma* 95, 53–72. [https://doi.org/10.1016/S0016-7061\(99\)00083-X](https://doi.org/10.1016/S0016-7061(99)00083-X)
- Fourcade, Y., Besnard, A.G., Secondi, J., 2018. Paintings predict the distribution of species, or the challenge of selecting environmental predictors and evaluation statistics. *Global Ecology and Biogeography* 27, 245–256. <https://doi.org/10.1111/geb.12684>
- Friedman, J.H., 2001. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* 29, 1189–1232. <https://doi.org/10.1214/aos/1013203451>
- Friedman, J.H., Popescu, B.E., 2008. Predictive Learning via Rule Ensembles. *The Annals of Applied Statistics* 2, 916–954.
- Gee, G.W., Bauder, J.W., 1986. Particle-size Analysis, in: *Methods of Soil Analysis*. John Wiley & Sons, Ltd, pp. 383–411. <https://doi.org/10.2136/sssabookser5.1.2ed.c15>
- Goldstein, A., Kapelner, A., Bleich, J., Pitkin, E., 2015. Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. *Journal of Computational and Graphical Statistics* 24, 44–65. <https://doi.org/10.1080/10618600.2014.907095>
- Gozukara, G., Zhang, Y., Hartemink, A.E., 2022. Using pXRF and vis-NIR spectra for predicting properties of soils developed in loess. *Pedosphere* 32, 602–615. [https://doi.org/10.1016/S1002-0160\(21\)60092-9](https://doi.org/10.1016/S1002-0160(21)60092-9)

- He, J., Garzanti, E., Jiang, T., Barbarano, M., Resentini, A., Liu, E., Chen, S., Shi, G., Wang, H., 2022. Mineralogy and geochemistry of modern Red River sediments (North Vietnam): Provenance and weathering implications. *Journal of Sedimentary Research* 92, 1169–1185. <https://doi.org/10.2110/jsr.2022.045>
- Holzinger, A., Saranti, A., Molnar, C., Biecek, P., Samek, W., 2022. Explainable AI Methods - A Brief Overview, in: Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, K.-R., Samek, W. (Eds.), *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*. Springer International Publishing, Cham, pp. 13–38. https://doi.org/10.1007/978-3-031-04083-2_2
- Journet, E., Balkanski, Y., Harrison, S.P., 2014. A new data set of soil mineralogy for dust-cycle modeling. *Atmospheric Chemistry and Physics* 14, 3801–3816. <https://doi.org/10.5194/acp-14-3801-2014>
- Kuhn, M., 2014. Futility Analysis in the Cross-Validation of Machine Learning Models. <https://doi.org/10.48550/arXiv.1405.6974>
- Kuncheva, L.I., Matthews, C.E., Arnaiz-González, Á., Rodríguez, J.J., 2020. Feature Selection from High-Dimensional Data with Very Low Sample Size: A Cautionary Tale. <https://doi.org/10.48550/arXiv.2008.12025>
- Mancini, M., Weindorf, D.C., Silva, S.H.G., Chakraborty, S., Teixeira, A.F.S., Guilherme, L.R.G., Curi, N., 2019. Parent material distribution mapping from tropical soils data via machine learning and portable X-ray fluorescence (pXRF) spectrometry in Brazil. *Geoderma* 354, 113885. <https://doi.org/10.1016/j.geoderma.2019.113885>
- Molnar, C., 2024. Interpretable Machine Learning.
- Padarian, J., Minasny, B., McBratney, A.B., 2020. Machine learning and soil sciences: a review aided by machine learning tools. *SOIL* 6, 35–52. <https://doi.org/10.5194/soil-6-35-2020>
- Pelegrino, M.H.P., Silva, S.H.G., Faria, A.J.G., Mancini, M., Teixeira, A.F.S., Chakraborty, S., Weindorf, D.C., Guilherme, L.R.G., Curi, N., 2021. Prediction of soil nutrient content via pXRF spectrometry and its spatial variation in a highly variable tropical area. *Prec Agric*, 23, 18–34.
- QGIS, 2025. QGIS Geographic Information System 3.44. QGIS Association. Available at: <https://qgis.org/>. Accessed in: September, 2025.
- Resende, M., Curi, N., Poggere, G.C., Barbosa, J.Z., Pozza, A.A.A., Teixeira, A.F. dos S., 2021. *Pedologia, fertilidade, água e planta: Inter-relações e aplicações*, 2nd ed. Editora UFLA, Lavras.
- Rezende, S.B. de, Franzmeier, D.P., Resende, M., Mancini, M., Curi, N., 2022. Pedogenic processes in a chronosequence of very deeply weathered soils in southeastern Brazil. *Catena* 215, 106362. <https://doi.org/10.1016/j.catena.2022.106362>
- Santos, H.G., Jacomine, P.K.T., Anjos, L.H.C., Oliveira, V.Á., Lumberras, J.F., Coelho, M.R., Almeida, J.A., Araújo Filho, J.C., Lima, H.N., Marques, F.A., Oliveira, J.B., Cunha, T.J.F., 2025. *Sistema brasileiro de classificação de solos*, 6th ed. Embrapa, Brasília, DF, Brazil.
- Schaefer, C.E.G.R., Fabris, J.D., Ker, J.C., 2008. Minerals in the clay fraction of Brazilian Latosols (Oxisols): a review. *Clay miner.* 43, 137–154. <https://doi.org/10.1180/claymin.2008.043.1.11>
- Shapley, L.S., 1952. A value for n-person games. RAND Corporation, Santa Monica, CA. <https://doi.org/10.7249/P0295>

- Silva, E.A., Weindorf, D.C., Silva, S.H.G., Ribeiro, B.T., Poggere, G.C., Carvalho, T.S., Gonçalves, M.G.M., Guilherme, L.R.G., Curi, N., 2019. Advances in Tropical Soil Characterization via Portable X-Ray Fluorescence Spectrometry. *Pedosphere* 29, 468–482. [https://doi.org/10.1016/S1002-0160\(19\)60815-5](https://doi.org/10.1016/S1002-0160(19)60815-5)
- Silva, F.M., Silva, S.H.G., Acuña-Guzman, S.F., Silva, E.A., Ribeiro, B.T., Fruett, T., Inda, A.V., Teixeira, A.F. dos S., Mancini, M., Guilherme, L.R.G., Curi, N., 2021b. Chemical and mineralogical changes in the textural fractions of quartzite-derived tropical soils, along weathering, assessed by portable X-ray fluorescence spectrometry and X-ray diffraction. *J. South Am. Earth Sci.* 112, 103634. <https://doi.org/10.1016/j.jsames.2021.103634>
- Silva, S.H.G., Mancini, M., Teixeira, A.F. dos S., Andrade, R., Curi, N., 2025. Chapter 5 - Applications of machine learning and artificial intelligence in soil science, in: Weindorf, D.C., Chakraborty, S., Li, B. (Eds.), *Unlocking the Secrets of Soil*. Elsevier, pp. 155–179. <https://doi.org/10.1016/B978-0-443-29879-0.00004-6>
- Silva, S.H.G., Ribeiro, B.T., Guerra, M.B.B., de Carvalho, H.W.P., Lopes, G., Carvalho, G.S., Guilherme, L.R.G., Resende, M., Mancini, M., Curi, N., Rafael, R.B.A., Cardelli, V., Cocco, S., Corti, G., Chakraborty, S., Li, B., Weindorf, D.C., 2021a. Chapter One - pXRF in tropical soils: Methodology, applications, achievements and challenges, in: Sparks, D.L. (Ed.), *Advances in Agronomy*. Academic Press, pp. 1–62. <https://doi.org/10.1016/bs.agron.2020.12.001>
- Silva, S.H.G., Weindorf, D.C., Faria, W.M., Pinto, L.C., Menezes, M.D., Guilherme, L.R.G., Curi, N., 2021. Teixeira, A.F. dos S., Andrade, R., Curi, N., 2025. Proximal sensor-enhanced soil mapping in complex soil-landscape areas of Brazil. *Pedosphere*, 31, 615–626.
- Soil Survey Staff, 2014. *Keys to soil taxonomy*, 12th ed. USDA-NRCS.
- Sparks, D.L., Singh, B., Siebecker, M.G., 2024. Chapter 6 - Ion Exchange Processes, in: Sparks, D.L., Singh, B., Siebecker, M.G. (Eds.), *Environmental Soil Chemistry (Third Edition)*. Academic Press, Boston, pp. 283–304. <https://doi.org/10.1016/B978-0-443-14034-1.00006-X>
- Wadoux, A.M.J.-C., Samuel-Rosa, A., Poggio, L., Mulder, V.L., 2020a. A note on knowledge discovery and machine learning in digital soil mapping. *European Journal of Soil Science* 71, 133–136. <https://doi.org/10.1111/ejss.1290>
- Wadoux, A.M.J.-C., Minasny, B., McBratney, A.B., 2020b. Machine learning for digital soil mapping: Applications, challenges and suggested solutions. *Earth-Science Reviews* 210, 103359. <https://doi.org/10.1016/j.earscirev.2020.103359>
- Wadoux, A.M.J.-C., Molnar, C., 2022. Beyond prediction: methods for interpreting complex models of soil variation. *Geoderma* 422, 115953. <https://doi.org/10.1016/j.geoderma.2022.115953>
- Walkley, A., Black, I.A., 1934. An examination of the Degtjareff method for determining soil organic matter, and a proposed modification of the chromic acid titration method. *Soil Sci.* 37, 29. <https://doi.org/10.1097/00010694-193401000-00003>
- Weindorf, D.C., Bakr, N., Zhu, Y., 2014. Advances in portable X-ray fluorescence (PXRF) for environmental, pedological, and agronomic applications, in: *Advances in Agronomy*. Elsevier, pp. 1–45. <https://doi.org/10.1016/B978-0-12-802139-2.00001-9>
- Weindorf, D.C., Chakraborty, S., 2020. Portable X-ray fluorescence spectrometry analysis of soils. *Soil Science Society of America Journal* 84, 1384–1392. <https://doi.org/10.1002/saj2.20151>

- Wickham, H., 2016. *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York.
- Wilson, M.J., 2004. Weathering of the primary rock-forming minerals: processes, products and rates. *Clay miner.* 39, 233–266. <https://doi.org/10.1180/0009855043930133>
- Widyatmanti, W., Rossiter, D.G., Cahyana, D., 2025. The definition of soil science from a geographical perspective. *Soil Security* 20, 100199. <https://doi.org/10.1016/j.soisec.2025.100199>

Chapter V: Machine-learning indicates which chemical elements are most influential in explaining soil texture in Brazil

Marcelo Mancini^a, Sérgio Henrique Godinho Silva^a, Zamir Libohova^b, Nilton Curi^a

^aDepartment of Soil Science, Federal University of Lavras, Lavras, Minas Gerais, Brazil

^bUSDA-ARS, Dale Bumpers Small Farms Research Center, 6883 AR-23, Booneville, AR 72927, USA

Abstract

The chemical elements concentrations in soil samples measured by portable X-ray fluorescence (pXRF) spectrometry have proven to be powerful prediction tool for soil texture. Despite the accuracy of the models, not much is known about the effect of each chemical element on estimations of sand, silt, and clay contents. We applied machine-learning interpretation methods to investigate what soil texture models are learning from the chemistry of soils. Soil texture prediction models were trained using as input the chemical composition of soils measured by pXRF. A total of 2267 soil samples from six study areas located in different Brazilian biomes were analyzed. Permutation feature importance, partial dependence, and individual conditional expectation curves were utilized to characterize the effect of each chemical element on texture predictions. Sand and clay models were very accurate ($\rho_c > 0.9$, except for soil samples from the Cerrado biome); however, satisfactory silt predictions were not possible, indicating lower correlation between silt contents and soil chemistry in tropical soils. Fe, Si and Al were the most important elements for sand and clay predictions. Fe and Al showed a positive effect on clay contents, suggesting that models identified the residual accumulations of Fe and Al caused by the intense pedogenetic processes in the tropics. Si concentrations had a positive effect on sand contents, likely associated with sand-sized quartz. In soils low in Fe content, the role of Si and Al increased, and elements such as Rb and V were relevant clay predictors, because they can substitute other elements in clay minerals or be adsorbed on their surfaces. The chemical composition of soils are outstanding estimators of texture in tropical soils and their effects agree with our current conceptual models of soil genesis. Although Fe, Si, and Al are the prevailing predictors of soil texture, trace elements like Rb and V should not be neglected.

Keywords: Oxisols; Ultisols; pXRF; soil parent material; pedogenesis; soil chemistry; soil mineralogy

1. Introduction

Soil texture is of utmost importance to understand the role of soils in the environment, as it drives water retention and infiltration (Wang et al., 2023), nutrient adsorption (Roy and Jílková, 2025; Ye et al., 2024), soil biological communities (Fernandez-Illescas et al., 2001; Zhu et al., 2021), and correlates strongly with soil structure and other physical properties (Hamza et al., 2025). Texture describes the distribution of fine and coarse particles in soils, separating soil particles into three size classes: sand (>2 mm), silt (between 2 and 0.05 mm), and clay (<0.002 mm). There are many methods for assessment of soil texture, but the most utilized methods rely on the sieving and sedimentation of coarser particles as a way to separate them from finer, clay-sized particles. To facilitate these often time-consuming methods, texture estimation techniques have been developed using proximal and remote sensing. The success of these methods indicates a correlation between sensor data and pedologic processes responsible for the formation of the current texture observed in soils. Most studies, however, focus on estimation accuracy and do not further explore the sensor-soil relationship further. In this study, we investigate the role of soil chemical composition measured by portable X-ray fluorescence (pXRF) sensors in estimating soil texture.

X-ray fluorescence is an analytical method used to determine the chemical composition of solid and liquid materials (Silva et al., 2021). When X-rays interact with a sample, part of the energy is absorbed by atoms, causing electrons to be ejected from their orbitals. To fill the created gaps, other electrons must lose energy, and hence they emit energy in the form of fluorescence. Each element emits its own characteristic fluorescence and the number of emitted energy pulses can be used to estimate the concentration of chemical elements in a sample. It is a fast and non-destructive method that can characterize the chemical composition of samples in a few seconds. The portable version of this method (pXRF) has been used in soil research to estimate a plethora of soil properties, including but not limited to: the available/exchangeable content of plant nutrients in soils (Tavares et al., 2023; Teixeira et al., 2022), soil organic matter (Faria et al., 2022), soil pollution (Li et al., 2021; Tighe et al., 2018), and soil texture (Andrade et al., 2022). Above all other soil properties, pXRF has shown outstanding capability of estimating soil texture (Nunes et al., 2025; Zhang and Hartemink, 2020). This is attributed to the influence of soil parent material (i.e., the rocks or sediments from which soils are formed), degree of weathering and the dominance of specific minerals in each particle size. For instance, quartz and Fe oxides are commonly found in the sand and clay fractions, respectively, making Si and Fe potential fingerprints of soil texture (Mancini et al., 2019; Silva et al., 2020).

To connect the chemical composition of a soil to its texture, a calibrated prediction model is needed. From the many available alternatives, machine-learning algorithms are currently the most successful methods applied to build soil texture models using sensor data as input variables (Hengl et al., 2018; Mancini et al., 2023; Padarian et al., 2020). Machine-learning models excel in finding complex relationships among environmental covariates and have the advantage of facilitating the use of non-linear and skewed data, as they do not make assumptions regarding the distribution or the structure of the data (Wadoux et al., 2020a). The downside is that the structures created by these models are complex and not easily interpretable by humans, potentially leading to blind trust in defective models (Arrouays et al., 2020; Rossiter, 2018; Wadoux et al., 2020b). But methods exist through which the inner workings of these models can be inspected (Holzinger et al., 2022; Molnar, 2024). Sometimes referred to as “explainable artificial intelligence (AI)”, these statistical methods help understand how models are using each covariate, and whether they match the pedologic knowledge at hand (Wadoux and Molnar, 2022).

As an attempt to connect machine learning modeling with our current pedological knowledge, our general goal was to better understand the contribution of the chemical composition of soils measured by pXRF in explaining soil texture. More specifically, the objectives were to: i) train sand, silt and clay models using the chemical composition of soils from six distinct Brazilian biomes, thus creating a diverse dataset of soil properties across Brazil; ii) apply machine learning interpretation methods to identify and document key elements for soil texture estimations and characterize their effects on sand, silt, and clay contents; iii) relate these findings with pedogenic processes as an effort to link the relationships between environmental variables and soil properties learned by machine learning algorithms with the current pedological knowledge. We hypothesize that machine learning interpretation methods will improve the current understanding of the effects of pXRF on soil texture predictions and confirm whether relationships found by models find background on the current pedologic knowledge.

2. Material and methods

2.1. Study areas

The study was conducted in six different regions of Brazil. Each area was named according to its biome, namely: Amazon, transition between Atlantic Forest and Cerrado, transition between Atlantic Forest and Caatinga, Cerrado, Coastal Plains, and Pampa. Soil classes are presented in the text per the Soil Taxonomy (Soil Survey Staff, 2014). Table 1

presents their corresponding classes in the Brazilian Soil Classification System (SiBCS; Santos et al., 2025) and the World Reference Base for Soil Resources (WRB; IUSS Working Group WRB, 2022). It is important to emphasize that the biomes are being used here as labels to identify each study area and function as a summary of the climatic conditions therein, but these samples do not represent the entire variety of soils in each biome.

2.1.1. Amazon Forest

This study area is located in the Iranduba municipality, in the state of Amazonas, Brazil. Data was collected from 52 soil profiles from 5 depths (0-20, 20-40, 40-60, 60-80, and 80-100 cm), totaling 254 samples. Climate in this area is Af following the Köppen Geiger classification, consisting of tropical humid climate with no dry season. The lithology in this region is a transition between sandstone with intercalated claystone and alluvial deposits. Common soil classes in this area include Oxisols (Xanthic Ferralsols) and plinthic soils (Plithosols).

2.1.2. Atlantic Forest

Atlantic Forest data was collected on the campus of the Federal University of Lavras, located in southern Minas Gerais, in the municipality of Lavras. This dataset is composed of soil samples from 129 locations collected in 4 depths (0-20, 20-40, 40-60 and 80-100 cm, except when soils were shallower). Climate in this region is humid subtropical (Cwa), with mean annual temperature of 20.3 °C and mean annual precipitation of 1380 mm (INMET et al., 2022). Spanning across 315 ha, this study area has a rich soil class variability, containing: Oxisols, Inceptisols, Entisols, and Histosols (Curi et al., 2017). The parent material forming most soils are granite-gneiss and gabbro (Mancini et al., 2019), but soils formed from alluvial and colluvial sediments are also found in lowlands.

2.1.3. Caatinga

The Caatinga dataset was collected in the physiographic region of Jequié, Bahia. The study areas are farms located in the municipalities of Maracás and Planaltino, across 4276 ha. Soil samples were collected in 89 locations; 52 in the proximity of Maracás, and 37 near Planaltino. Three depths were sampled in each location when possible (0-20, 40-70 and 100-150 cm), totaling 267 samples. Maracás has subtropical highland climate (Cfb) with mean annual precipitation of 850 mm, without sharp differences in mean precipitation between seasons. Planaltino has tropical humid climate (As), characterized by dry summers and rainy winters, with mean annual precipitation of 700 mm. In the sampled areas the predominant soil

classes are Oxisols (57% of the area) and Ultisols (47%) formed from nutrient-poor sediments of varied particle sizes.

2.1.4. Coastal Plains

This dataset is composed of samples collected during the soil survey of farms located in the east of Bahia, comprising an area of 6774 ha. Soil samples were collected from 93 points at three depths, when possible (0-20, 40-70 and 100-150 cm), totaling 267 samples. The region has tropical humid climate (Af), with well-distributed rainfall across the year. The predominant soil classes are Ultisols (45% of the area), followed by Entisols (predominantly sandy soils, i.e. Quartzpsamments, 18% of the area). Plinthic soils (Plinthaquoxes) also occur in the region (7%). Soils in the region are influenced by sediments from the Barreiras Formation, leading formed soils to be predominantly kaolinitic and hard-setting (clayey, with very dense and hard B horizons when dry).

2.1.5. Cerrado

This dataset is composed of soil samples collected in 121 locations in two farms located near the municipality of Três Lagoas, in Mato Grosso do Sul state. Soil samples were collected in three depths when possible (0-20, 40-70 and 100-150 cm), totaling 321 samples collected in 121 locations. Climate in this region is tropical humid (Aw), characterized by rainy summers and dry winters, with mean annual precipitation of 1340 mm. The geology of the region is characterized by sandstones overlying basalt or dacite. Soils present high sand content due to the influence of sandstone (sedimentary rock rich in sand-sized quartz), but may be influenced by basalt (mafic rock rich in iron oxides) when formed where the sandstone has been carved deep by erosional processes, presenting higher clay contents in such cases. The predominant soil class is Oxisol.

2.1.6. Pampa

The Pampa dataset is the most heterogeneous among study areas. It consists of 806 samples collected across multiple silvicultural lands in the Rio Grande do Sul State, in southern Brazil. Soils were formed from a variety of parent materials, including coarse and fine texture sandstone (sometimes intercalated with clayey deposits), volcano-sedimentary deposits overlaying sandstone, and granite. Soils derived from sandstones can vary from sandy Entisols (Quartzpsamments or Arenosols) to Ultisols. When formed on crystalline rocks such as granite, resulting soils are usually Ultisols. If formed upon more reactive clay mineralogy, or an abrupt

transition from sandy to clayey parent material is present, soils can be Kandic Ultisols or Albaquults (Luvisols or Planosols), respectively. The climate in all study areas within the Pampa biome is subtropical humid with no dry season and hot summers (Cfa).

Table 1. Correspondences between classes of the Brazilian Soil Classification System (SiBCS; Santos et al., 2025), Soil Taxonomy (ST; Soil Survey Staff, 2014), and the World Base Reference for Soil Resources (WRB; IUSS Working Group WRB, 2022).

SiBCS	ST	WRB	Study areas that contain samples from this soil class
Latossolos	Oxisols	Ferralsols	All study areas
Nitossolos	Kandic Alfisols/Ultisols (with pressure faces)	Nitisols	Atlantic Forest
Plintossolos	Plinthaquox and Ultisols (can include other orders)	Plinthosols	Coastal Plains
Argissolos	Ultisols	Acrisols/Lixisols	All study areas
Luvissolos	Alfisols	Luvisols	Pampa
Cambissolos	Inseptisols	Cambisols	Atlantic Forest
Gleissolos	Inseptisols (Aquepts)	Gleysols	Atlantic Forest; Coastal Plains
Planossolos	Alfisols/Ultisols (i.e. Albaqualfs/Albaquults)	Planosols	Pampa
Neossolos	Entisols	Regosol	Atlantic Forest; Coastal Plains
Neossolos Quartzarênicos	Entisols (Quartzpsamments)	Arenosols	Pampa; Coastal Tablelands
Organossolos	Histosols	Histosols	Atlantic Forest



Figure 1. Sampled sites (n=2,267) and the respective biomes of each study area across Brazil. Datum: SIRGAS2000.

2.2. Laboratory analysis

Soil texture analysis was done after drying and sieving samples to separate particles with diameter smaller than 2 mm (air-dried fine earth, ADFE). Texture was determined by the pipette method. An aliquot of 10 g of each soil sample was added to a solution of 100 ml of distilled water and 10 ml of NaOH 1 mol L⁻¹. The suspension was agitated for 16 hours at 30 rpm. Next, the suspension was sieved (0.053 mm) and added to a 500 ml test tube; the volume was completed with distilled water and the material retained in the sieve was dried at 110 °C. The sedimentation time of silt particles was calculated by the Stokes method. After waiting for the settling of silt-sized particles, an aliquot of 10 ml was extracted from the suspension with a pipette and dried at 110 °C. The sand content was determined by weighing the dried material retained in the 0.053 mm sieve, the clay content was determined by weighing the dried extracted 10 ml aliquot, and silt content was calculated as the difference between sand and clay contents.

The total content of chemical elements in soil samples was determined using a Tracer 5G pXRF model from Bruker. Tracer 5G has a Rh X-ray tube and can operate at up to 50kV and 100 mA. The dried and sieved soil samples (ADFE) were scanned in Soil mode for 60 s each in triplicates, rotating the samples after each repetition, following procedures described by Weindorf and Chakraborty (2020). Results represent the mean of triplicates.

2.3. Statistical analysis

The relationship between covariates were assessed using Spearman's correlation coefficient and dendrograms. Dendrograms were built using Spearman's correlation as the distance metric and the nearest neighbor algorithm as the linkage criterion.

The chemical composition of soil samples measured by pXRF was used as input to train the soil texture models. The same covariates (12) were chosen for all models to allow for the comparison between effects observed in each region; namely: Al, Ca, Fe, K, P, Rb, Si, Sr, Ti, V, Zn, and Zr. The choice of covariates was based on data availability (sufficient values above the limit of detection for all samples) and pedologic knowledge. Si and Al are key indicators of pedogenic processes (weathering) and parent material, while other chemical elements (Rb, Sr, Ti, V, Zn, Zr) are representative of parent material composition. Ca and P can indicate management differences (agricultural inputs).

Random forest was used to train soil texture prediction models using soil's chemical composition (pXRF data) as input. Datasets were split into a training set (80%) used to tune model parameters, and a held-out validation set (20%) used to evaluate model performance. Model tuning consisted of 10-fold cross-validated random search. The 10-fold cross validation was repeated 100 times for each model while randomly changing parameters. The parameters of the best performing models were selected. After tuning, models were evaluated using the held-out set. Model evaluation metrics included the coefficient of determination (R^2), Lin's concordance correlation coefficient (ρ_c), root mean square error (RMSE), and mean absolute error (MAE).

The effect of soil's chemical composition on soil texture predictions was assessed using partial dependence plots (PDP), individual conditional expectation (ICE) (Molnar, 2024; Wadoux and Molnar, 2022). PDPs provide the global effect of each input covariate on the predicted outcome. ICEs are similar, but present the effect of each individual sample instead of the global effect. All analyses were conducted using the validation sets.

3. Results and discussion

3.1. Texture and chemical composition of studied soils

The distribution of texture in samples mostly reflected soil parent material and degree of weathering (Table 2). The highest average sand contents were observed in the studied areas under the biomes: Cerrado, Coastal Plains, and Pampa. Studied soils within these biomes were either formed on sandstones (Cerrado and Pampa), or on sandy sediments that occur erratically amidst the very clayey kaolinitic sediments of the Barreiras Formation in the Coastal Plains (Andrade et al., 2020). On the other hand, the highest average clay contents occurred in areas where soils are formed from igneous and metamorphic rocks (Atlantic Forest, where granite-gneiss and gabbro are the predominant parent materials), or clayey sediments (Amazon and Caatinga). The highest silt contents were observed in the Atlantic Forest area, likely due to the presence of Inceptisols (Cambisols) in this dataset which were formed upon fresh granite-gneiss parent material.

The chemical composition of samples denoted parent material differences, while also indicating pedogenetic processes and chemical affinity between elements (Figure 2). Fe and Si were notably clustered together in most study areas and presented strong negative correlation with each other in all datasets. This relationship represents: i) the compositional difference between soils formed from felsic (Si-rich) and mafic (Fe-rich) parent materials, and ii) the latolization process, whereby Si is removed by intense weathering. Latolization causes the residual accumulation of clay minerals resistant to weathering, such as Fe- and Al-rich minerals (e.g., hematite, goethite, kaolinite and gibbsite), a common process in tropical soils (Fanning and Fanning, 1989; Hartemink et al., 2020; Rezende et al., 2022). Other noteworthy pairings, like Ti and V, Rb and K, and Ca and P can be attributed to their occurrence in minerals or their capacity of substituting each other inside phyllosilicates. Ti and V can occur in vanadiferous titanomagnetites (Simandi and Paradis, 2022); Rb can be found together with K in phyllosilicates such as lepidolite and zinnwaldite (Zhang et al., 2026), and can substitute K in phyllosilicates; and Ca and P tend to combine and form a series of insoluble calcium phosphates.

Table 2. Descriptive statistics of training and test sets used to train and evaluate models. The studied soil data (n=2,267) were collected from six study areas, each falling under a different Brazilian biome, in the states of Amazonas, Bahia, Mato Grosso do Sul, Minas Gerais, and Rio Grande do Sul.

Variable	Statistics	Amazon		Caatinga		Atlantic Forest		Cerrado		Coastal Plains		Pampa	
		Test	Training	Test	Training	Test	Training	Test	Training	Test	Training	Test	Training
Sand	n	55	199	54	213	66	228	66	255	70	272	195	594
	Mean	42.08	41.05	47.98	50.75	35.47	31.16	78.03	77.63	70.00	66.08	77.53	76.78
	SD	10.54	11.85	12.31	12.19	11.06	12.15	7.70	9.49	19.00	22.20	15.05	15.55
	Min.	24.00	14.30	28.00	24.00	13.00	7.00	60.00	37.00	7.40	4.95	30.00	32.00
	Q25	33.35	32.10	39.00	43.00	28.00	22.00	75.25	73.00	61.00	49.00	71.00	67.00
	Q50	40.50	38.80	45.50	50.00	37.50	31.00	80.00	80.00	75.50	70.87	83.00	82.00
	Q75	49.55	50.85	56.75	60.00	44.00	41.00	83.00	84.00	81.88	84.23	86.00	87.00
	Max.	63.30	66.90	78.00	81.00	56.00	63.00	91.00	92.00	94.00	98.42	96.00	99.00
Silt	n	55	199	54	213	66	228	66	255	70	272	195	594
	Mean	10.43	11.53	3.59	3.75	16.58	18.00	3.88	3.57	6.75	6.37	8.17	8.61
	SD	3.96	4.35	2.01	2.12	8.42	9.11	2.30	2.55	6.87	7.07	6.05	6.62
	Min.	1.70	0.10	1.00	1.00	5.00	0.00	1.00	1.00	1.00	0.00	1.00	1.00
	Q25	8.73	8.83	2.00	2.00	11.00	11.00	2.00	2.00	2.00	2.00	4.40	4.00
	Q50	10.25	10.70	3.00	3.00	15.00	16.00	3.00	3.00	5.00	4.04	6.20	6.30
	Q75	11.58	13.68	5.00	5.00	20.75	24.00	5.00	4.00	9.00	9.00	11.00	13.00
	Max.	33.30	29.10	8.00	14.00	39.00	46.00	13.00	17.00	42.00	68.00	40.00	44.00
Clay	n	55	199	54	213	66	228	66	255	70	272	195	594
	Mean	47.48	47.42	48.43	45.50	47.95	50.85	18.09	18.83	23.26	27.55	14.33	14.66
	SD	12.07	13.41	12.29	11.88	13.86	11.87	7.55	8.84	15.63	19.57	10.78	10.74
	Min.	25.10	21.15	19.00	17.00	17.00	17.00	8.00	7.00	5.00	0.00	2.00	0.00
	Q25	38.83	35.45	41.25	37.00	38.00	42.00	14.00	13.00	12.31	11.00	9.00	7.50
	Q50	49.15	48.20	51.00	47.00	45.50	51.00	16.00	16.00	17.50	22.00	10.00	12.50
	Q75	58.05	57.73	58.00	54.00	58.75	60.00	19.75	22.00	30.38	43.25	17.25	19.00
	Max.	68.95	77.40	69.00	71.00	74.00	76.00	36.00	55.00	66.67	80.69	58.00	56.00

n: number of samples; SD: standard deviation; Q: quantile; Min.: minimum value; Max.: maximum value.

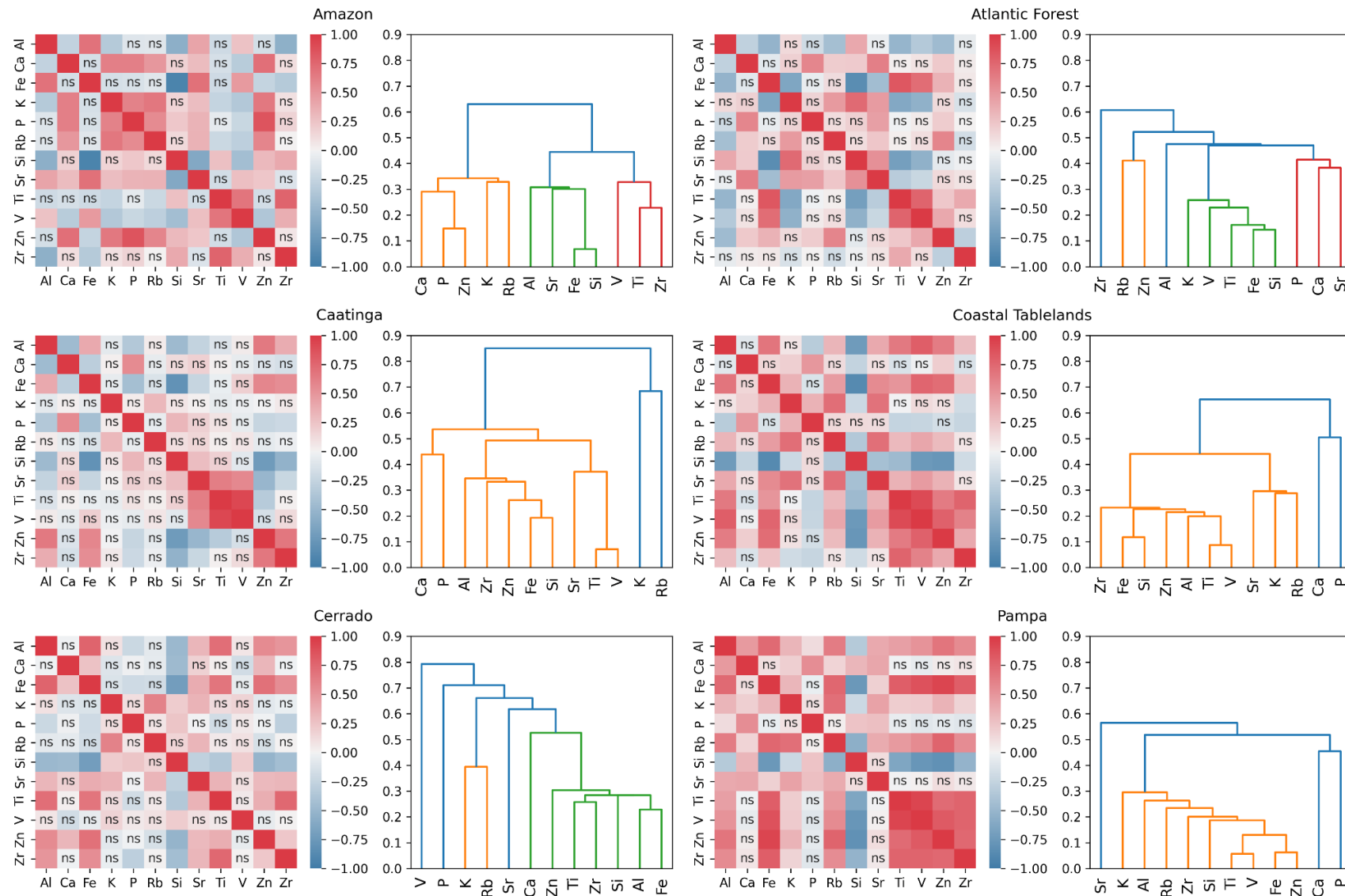


Figure 2. Spearman's correlation matrices and dendrograms of the chemical composition of soil samples measured by portable X-ray fluorescence. Soils were sampled from six study areas under different Brazilian biomes. The Spearman's correlation coefficient was used as the distance metric in the dendrograms.

3.2. Performance of ML models

The contents of sand and clay were accurately predicted in all study areas (Table 3). The exceptions were the models built for the Cerrado study area, which did not generalize as well as others; i.e., the validation scores were much lower for the validation set compared to the training set (Figures 3 and 4). As opposed to sand and clay models, silt predictions were overall inaccurate, presenting satisfactory results only for the Pampa dataset (Figure 5). This indicates that, although the chemistry of samples correlated well with the contents of sand- and clay-sized particles, the chemical composition of tropical soils was not a good predictor of silt content, probably due to the presence of minerals in the silt fraction that are also present either in sand or clay fraction (Silva et al., 2018; Silva et al., 2022).

XRF data is known to excel at predicting soil texture (Zhu et al., 2012), surpassing proximal and remote spectral sensors (Andrade et al., 2022; Swain et al., 2021). Empirical data suggests that this capability is a consequence of certain chemical elements found in abundance in specific particle sizes; particularly Fe in clay-sized minerals (hematite and goethite), and Si in sand-sized quartz particles. The usefulness of XRF in estimating sand and clay is not exclusive to weathered-leached soils of the tropics, being observed in soils derived from a myriad of parent materials across the globe (Mancini et al., 2020; Silva et al., 2020; Andrade et al., 2022, Mancini et al., 2023). In silt loam Alfisols (Luvisols) of Germany formed from loess, good correlations between XRF data and silt were found (Greenberg et al., 2023), wherein Fe, Ni, Al, Si, Mg, Cr, K, and Mn were considered important covariates to estimate silt contents, in this order. Very accurate soil texture predictions were also achieved in mangrove soils (Nunes et al., 2025), where Rb content had the biggest effect on predictions.

But as opposed to sand and clay, silt contents seem not to be accurately estimated in every case. Similar to results found herein, Benedet et al. (2020) and Silva et al. (2020) have also observed lower prediction accuracy for silt contents in tropical soils compared to sand and clay. Croffie et al. (2022) studied Irish soils formed from limestone and siliceous stone and concluded that silt contents were harder to estimate compared to other size fractions.

Table 3. Evaluation metrics of models trained to predict soil texture in study areas within six Brazilian biomes. Training results represent the means of 10-fold cross-validations, while test results were calculated using the test set (20% of the original set, see Table 2).

	Sand					
	Amazon	Caatinga	Atlantic Forest	Cerrado	Coastal Plains	Pampa
Train ρ_c	0.99	0.98	0.98	0.90	0.96	0.98
Test ρ_c	0.91	0.88	0.84	0.61	0.87	0.94
Train R^2	0.98	0.96	0.96	0.84	0.93	0.96
Test R^2	0.83	0.79	0.72	0.40	0.77	0.90
Train RMSE	3.96	6.43	6.11	5.94	9.56	5.89
Test RMSE	4.33	5.53	5.86	5.92	8.97	4.78
Train MAE	1.17	1.95	1.87	2.61	4.34	2.07
Test MAE	3.09	4.38	4.61	4.01	5.91	3.48
	Silt					
Train ρ_c	0.69	0.63	0.94	0.29	0.79	0.90
Test ρ_c	0.27	0.09	0.44	0.11	0.55	0.82
Train R^2	0.58	0.56	0.90	0.24	0.70	0.83
Test R^2	0.12	0.02	0.26	0.07	0.36	0.70
Train RMSE	3.35	1.99	6.60	2.30	4.71	4.01
Test RMSE	3.69	1.98	7.21	2.21	5.44	3.29
Train MAE	1.82	1.09	2.34	1.59	2.12	1.71
Test MAE	2.16	1.51	5.69	1.52	3.21	2.04
	Clay					
Train ρ_c	0.99	0.98	0.97	0.84	0.98	0.98
Test ρ_c	0.90	0.89	0.83	0.64	0.81	0.92
Train R^2	0.99	0.96	0.95	0.74	0.97	0.96
Test R^2	0.81	0.81	0.76	0.43	0.67	0.85
Train RMSE	4.15	5.91	6.78	5.27	8.05	4.29
Test RMSE	5.19	5.33	6.78	5.65	8.92	4.11
Train MAE	1.17	1.81	2.15	3.08	2.36	1.50
Test MAE	3.21	4.43	5.32	3.58	5.18	2.75

ρ_c : Lin's concordance correlation coefficient; RMSE: root mean squared error; MAE: mean absolute error.

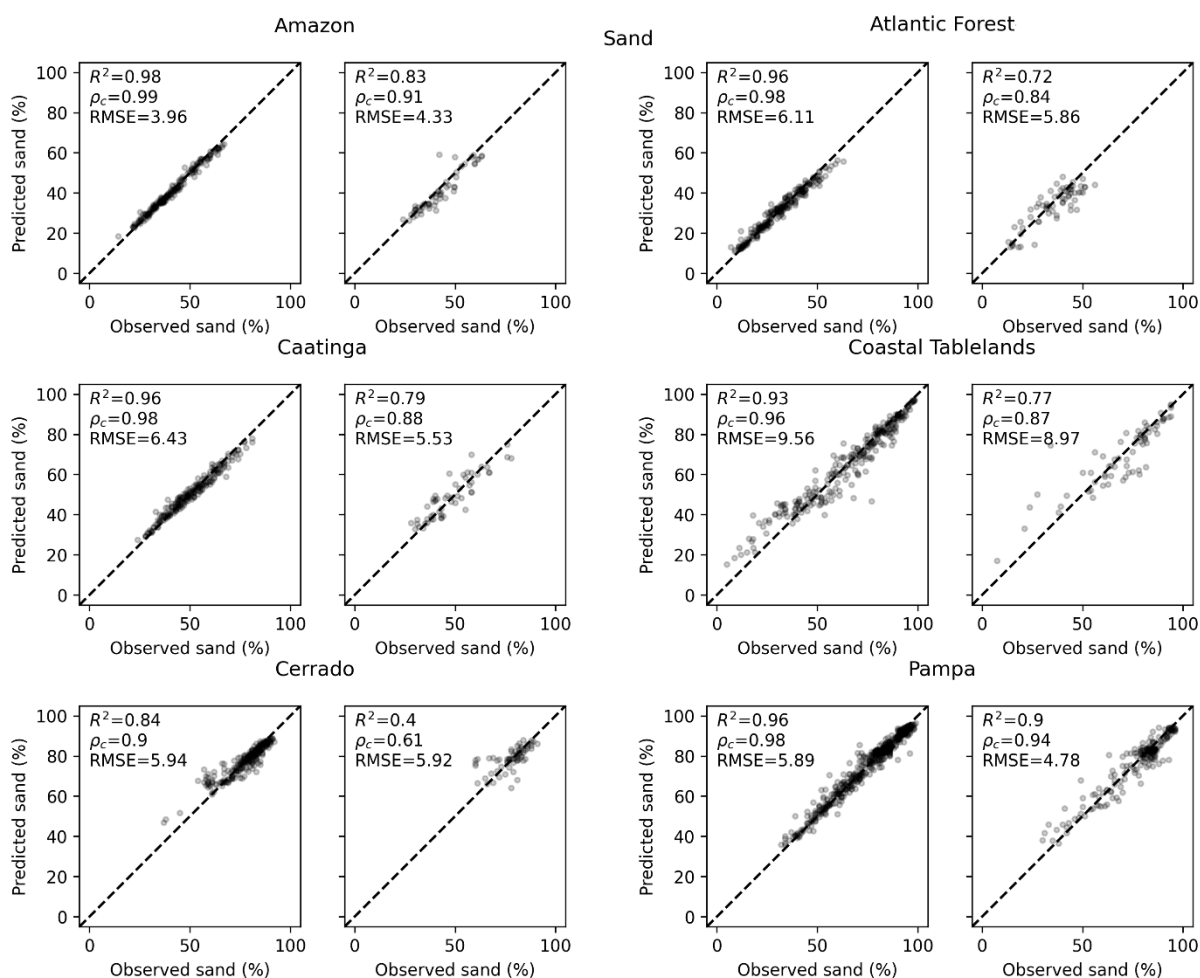


Figure 3. Observed and predicted sand contents estimated from the chemical composition of soil samples using random forest models. Soils were collected from six study areas across Brazil under different biomes. Names separated by a dash indicate the transition between two biomes. ρ_c : Lin's concordance correlation coefficient; RMSE: root mean squared error.

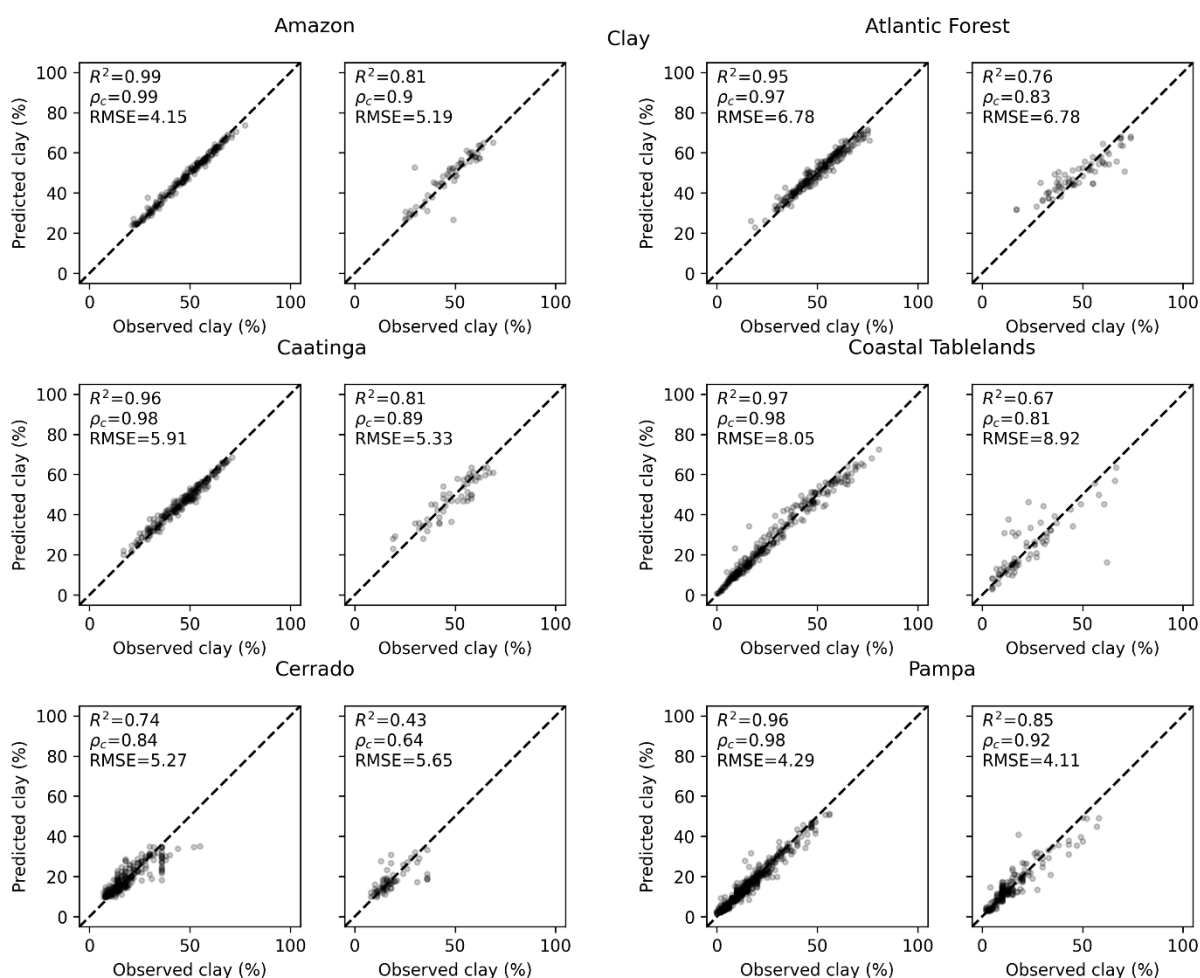


Figure 4. Observed and predicted clay contents estimated from the chemical composition of soil samples using random forest models. Soils were collected from six study areas across Brazil under different biomes. Names separated by a dash indicate the transition between two biomes. ρ_c : Lin's concordance correlation coefficient; RMSE: root mean squared error.

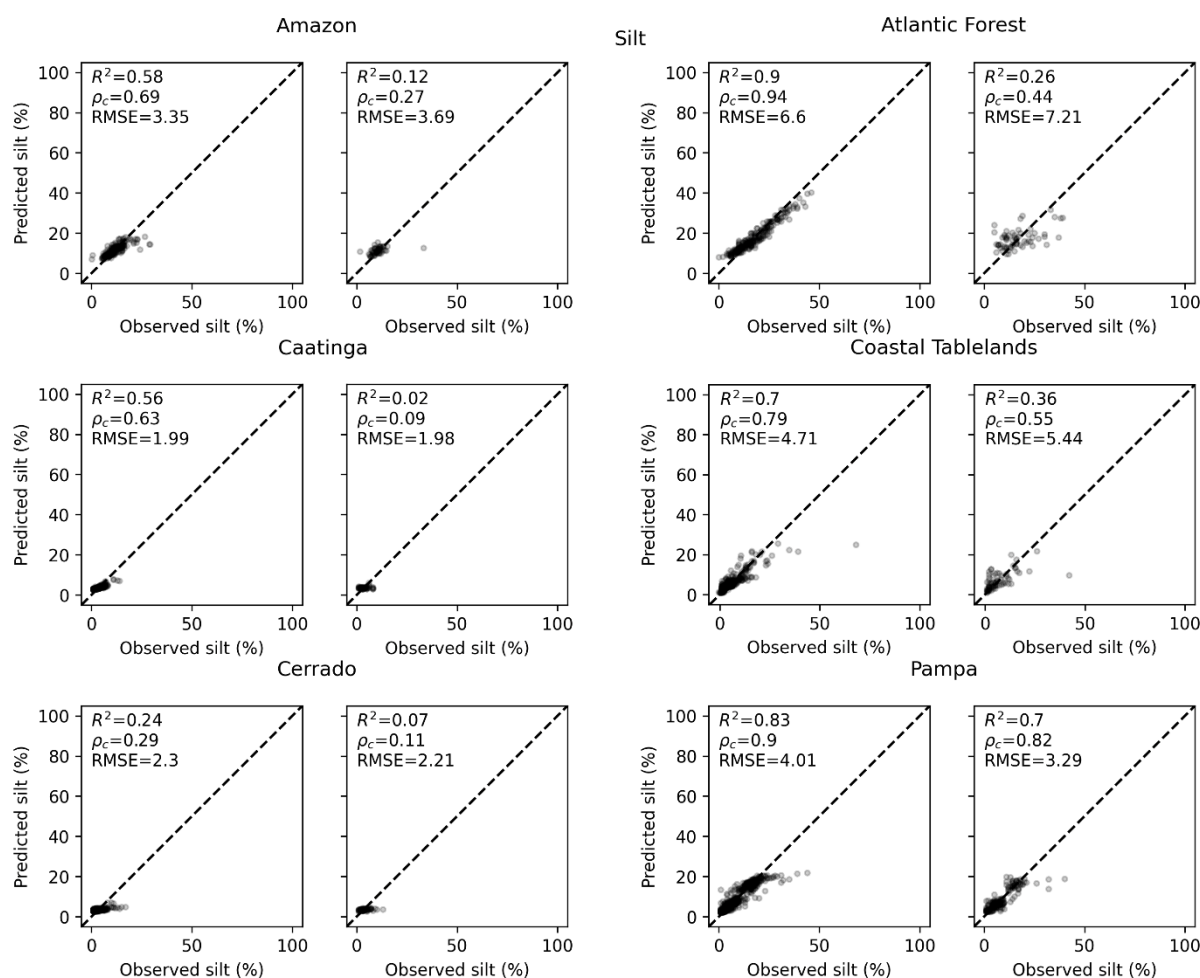


Figure 5. Observed and predicted silt contents estimated from the chemical composition of soil samples using random forest models. Soils were collected from six study areas across Brazil under different biomes. Names separated by a dash indicate the transition between two biomes. ρ_c : Lin's concordance correlation coefficient; RMSE: root mean squared error.

3.3. The effect of soil chemical composition on texture prediction

Fe and Si, followed by Al, were the most relevant elements to explain soil texture (Figure 6). The impacts of Fe and Si on sand and clay prediction errors were almost identical for the Amazon and Atlantic Forest study areas, where parent materials are alluvial deposits and granite-gneiss or gabbro bedrock, respectively. In the Caatinga area, where soils were formed from reworked colluvium, Si and Al had the strongest impact on prediction errors of both sand and clay predictions. Fe and Al were relevant variables in the Cerrado and Pampa areas, with Rb having notable importance in sand and silt predictions in the Pampa study area. The chemical composition of soils had little impact on silt prediction errors, indicating that it lacks the capacity of explaining silt content in the studied areas. The only exception was the model built with samples from the Pampa area, for which silt contents were sensitive to changes in Rb and Ca.

The areas under the Amazon and Atlantic Forest biomes presented very similar covariate effects (Figure 7). The effects of Si and Fe on sand predictions were almost identical, showing that high Si and low Fe are strongly associated with sandier soils. The stronger effect of K on clay predictions in Atlantic Forest samples were attributed to the differences between granite-gneiss and gabbro, the two predominant parent materials of this study area. Old and weathered tropical soils formed from gabbro tend to have higher clay content and absence of K-bearing secondary minerals, whereas granite and gneiss can include muscovite in its mineralogical suite, a weathering-resistant phyllosilicate that persists in the sand fraction even in weathered-leached Oxisols and Ultisols. Hence, soils higher in K are more likely to bear more sand-sized muscovite. The high average K contents in Atlantic Forest samples compared to those of other study areas supports this hypothesis (Figure 8).

In the Caatinga study area, Si and Al had the strongest effects on predictions (Figure 7), followed by Fe. The lesser effect of Fe may be due to its lower content in soils of this region compared to those of the Amazon and Atlantic Forest (Figure 8). The studied soils of this area are mostly yellow Oxisols and Ultisols (Xanthic Ferralsols and Acrisols, according to the WRB), characterized as having less Fe content compared to their hematitic counterparts, which are redder due to the coloring effect of hematite (Rhodic soils). The negative effects of Al and Fe on sand predictions (and the opposite on clay) were attributed to the presence of kaolinite ($\text{Al}_2\text{Si}_2\text{O}_5(\text{OH})_4$) and Fe oxides in the clay fraction. This agrees with the definition of the latolization process, consisting in the residual accumulation of Al and Fe caused by the removal of Si along with other mobile elements, and the precipitation of weathering-resistant Al- and Fe-rich clay minerals (kaolinite, gibbsite, hematite, and goethite).

In the Cerrado study area, the soils derived from sandstones, a rock composed predominantly of quartz, and were therefore rich in sand-sized quartz (SiO_2), since large quartz grains are very insoluble and resist intense weathering. Clay contents in these soils result from the weathering of cementing agents found between quartz grains (e.g., silica, carbonates, and Fe oxides) into clay minerals such as kaolinite, hematite, and goethite. The average chemical composition of these samples indicated these soils are comparatively higher in Si content (likely from quartz and kaolinite), and poorer in Fe (hematite and goethite) (Figure 8). The lower sensitivity observed in models of this area to covariates like Si, Fe, and Al (Figure 7) were attributed to the lower content of Fe in these soils (Figure 8), which would otherwise help in separating clay from sand contents. Covariate effects, although weaker than those observed in the other studied areas, indicated that models relied on Fe and Al for sand and clay predictions,

likely associating higher Fe and Al contents with the presence of clay minerals (hematite/goethite and kaolinite, respectively).

Rb and V, trace elements in soils, had a notable effect on Coastal Plains and Pampas models (Figure 7). Rb substitutes K in phyllosilicates and can also be adsorbed on the surfaces of clay minerals (Croffie et al., 2022), which may explain its negative effect on sand content predictions in Pampa samples. The importance of V for Coastal Plains models was not fully understood, but is hypothesized to follow a mechanism similar to Rb, since V can be present in minerals such as magnetite and rutile, and can also be adsorbed on clay minerals (Widyastuti et al., 2026). However, soils from the Coastal Plains are formed from the Barreiras Formation, a kaolinitic sedimentary material poor in mafic minerals, therefore the origin of the V content is unknown. There is an important V mine (Menchen Mine) west of the Coastal Plains study area in Maracás (Simandi and Paradis et al., 2022), which may be an indication that other deposits of V-rich ores could have existed in that past which could have influenced the sediments of the Barreiras Formation. Another hypothesis would be the presence of mafic rocks under the Barreiras Formation sediments influencing the formation of soil profiles. Regardless of its origin, the reliance of models on V shows that trace elements trapped in clay minerals can function as important chemical proxies for the estimation of soil texture in Fe-poor parent materials.

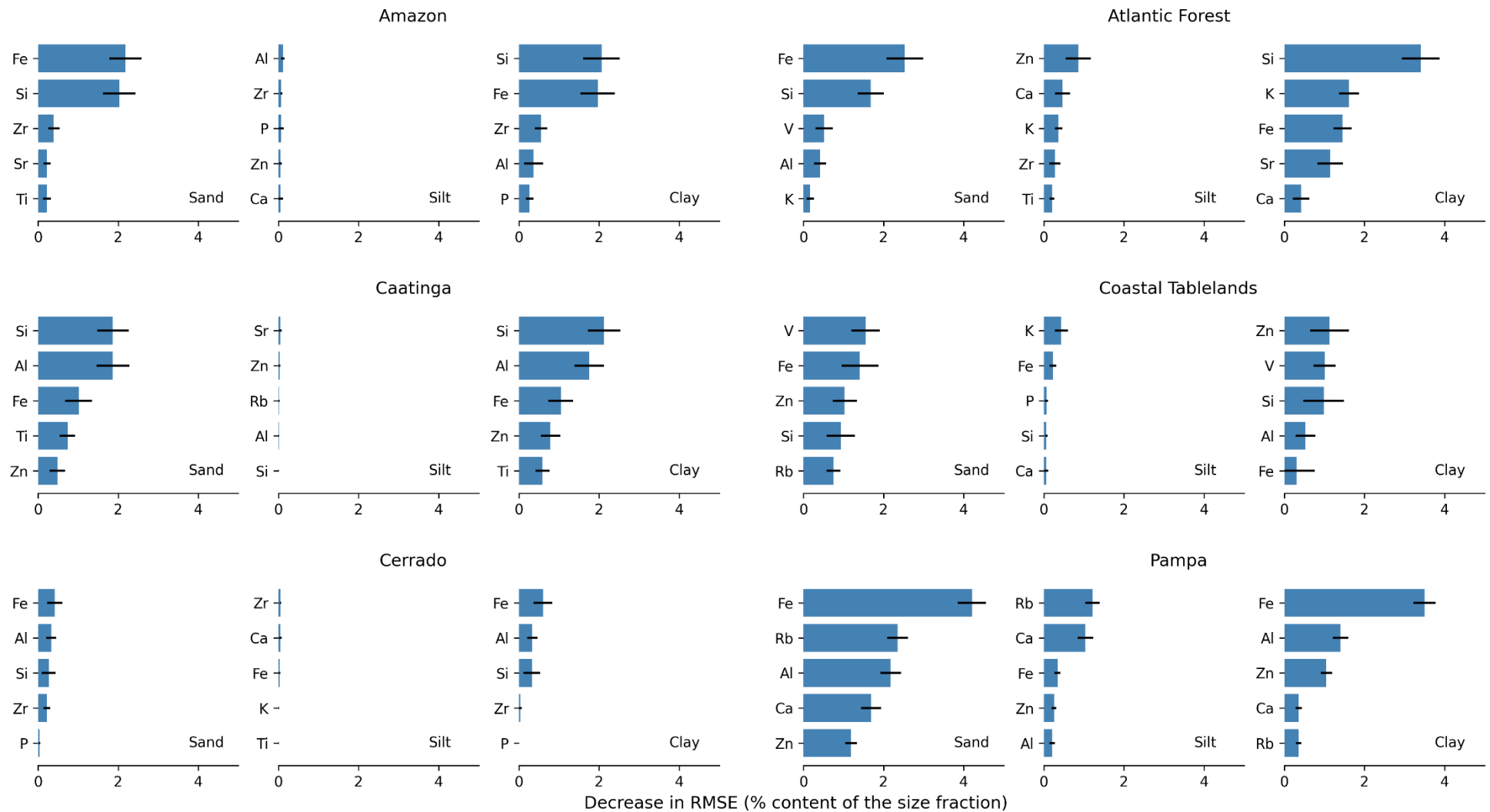


Figure 6. Permutation feature importance of soil texture models trained with portable X-ray fluorescence (pXRF) data. Soils were sampled from six study areas under six different Brazilian biomes. The chosen metric of importance was the root mean squared error (RMSE) decrease after the permutation of each variable. Permutations were repeated 1000 times each. The gray bars indicate the 95% interval of repetitions. Only the 5 most relevant covariates are shown.

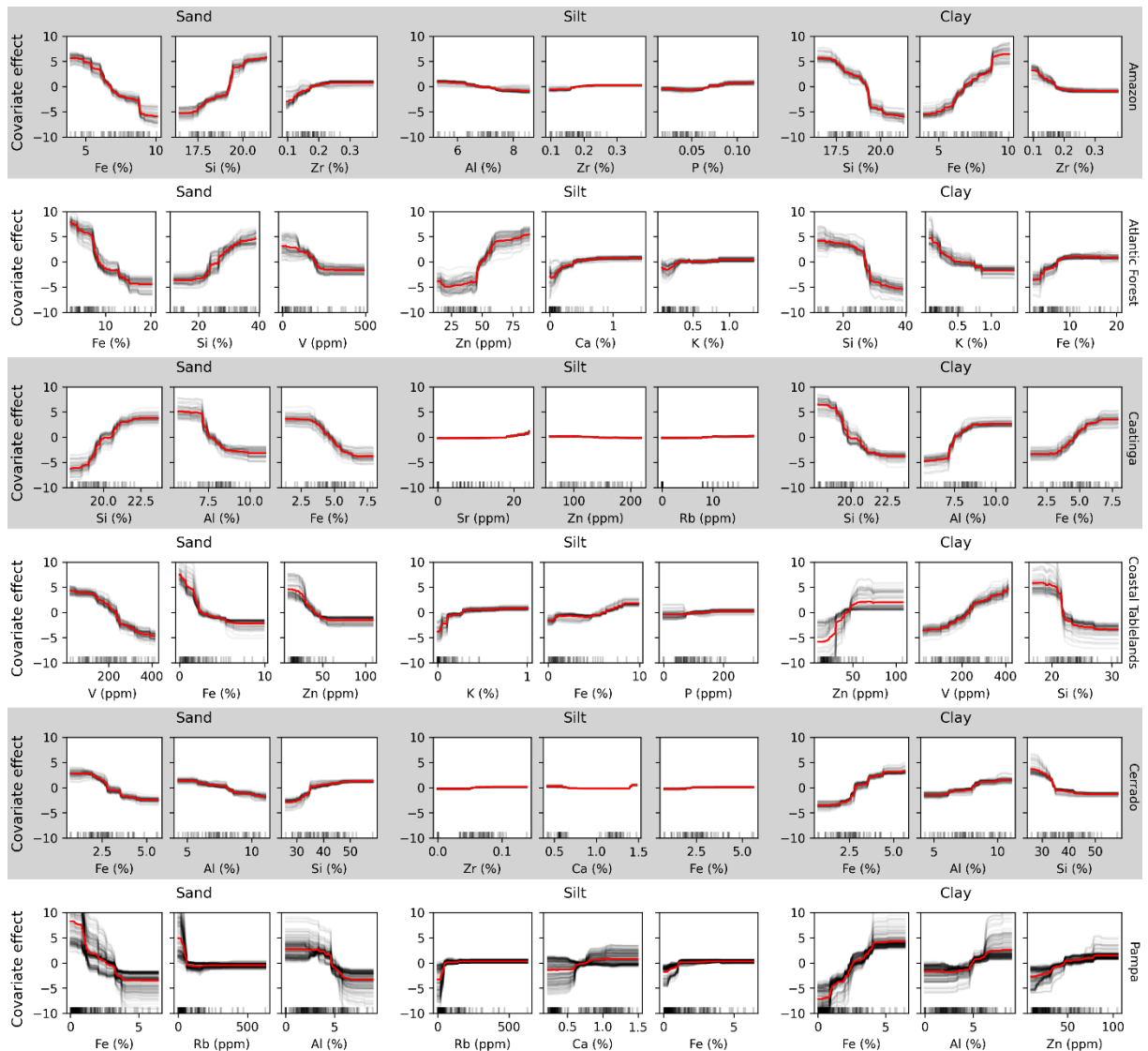


Figure 7. Partial dependence (PDPs, colored red) and individual conditional expectation curves (ICEs, colored black) for soil texture models trained with portable X-ray fluorescence (pXRF) data. Soils were sampled from six study area under six different Brazilian biomes. Presented data was calculated using the test datasets (see Table 2). Rugged lines indicate the density of available soil samples.

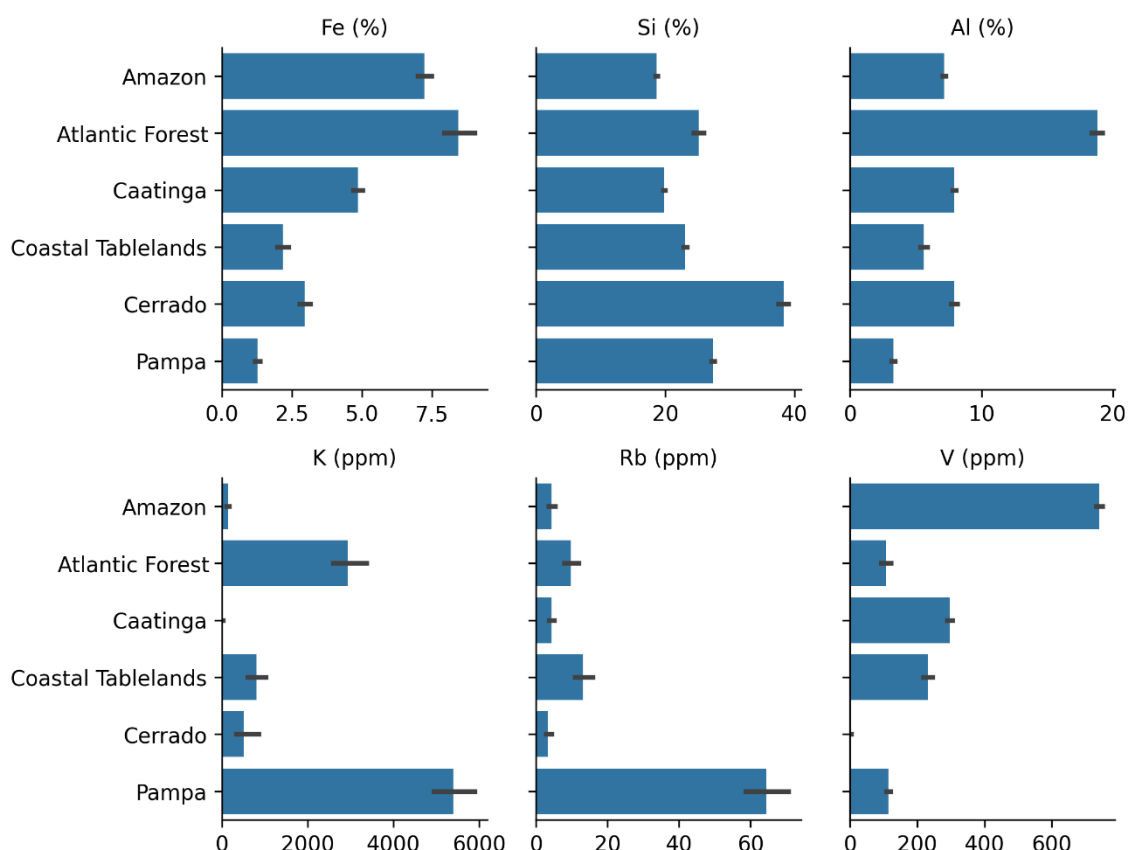


Figure 8. Mean elemental concentrations per study area of covariates considered relevant by soil texture models trained with portable X-ray fluorescence (pXRF) data (see Figures 6 and 7). Soils were sampled from six study area under six different Brazilian biomes. The gray bars indicate the 95% interval.

4. Conclusions

The concentrations of Si and Fe are responsible for the high accuracy of soil texture estimations using pXRF data in tropical soils from different biomes. In Fe-poor soils, Al and other trace elements associated with clay minerals such as Rb and V can have a strong effect on texture estimations. Models learned that higher concentrations of Fe and Al were associated with higher clay contents due to the presence of these elements in clay minerals (hematite, goethite, gibbsite, and kaolinite). High Si concentrations were associated with soils richer in sand due to the predominance of sand-sized quartz particles in the sand fraction of the studied soils. The relationships found by models between pXRF covariates and soil texture in such diversified soil conditions were in agreement with the current scientific knowledge about soil formation. PXRF data can and should be used reliably to estimate soil texture in tropical soils.

References

- Andrade, R., Mancini, M., Teixeira, A.F. dos S., Silva, S.H.G., Weindorf, D.C., Chakraborty, S., Guilherme, L.R.G., Curi, N., 2022. Proximal sensor data fusion and auxiliary information for tropical soil property prediction: Soil texture. *Geoderma* 422, 115936. <https://doi.org/10.1016/j.geoderma.2022.115936>
- Andrade, R., Faria, W.M., Silva, S.H.G., Chakraborty, S., Weindorf, D.C., Mesquita, L.F., Guilherme, L.R.G., Curi, N., 2020. Prediction of soil fertility via portable X-ray fluorescence (pXRF) spectrometry and soil texture in the Brazilian Coastal Plains. *Geoderma* 357, 113960. <https://doi.org/10.1016/j.geoderma.2019.113960>
- Arrouays, D., McBratney, A., Bouma, J., Libohova, Z., Richer-de-Forges, A.C., Morgan, C.L.S., Roudier, P., Poggio, L., Mulder, V.L., 2020. Impressions of digital soil maps: The good, the not so good, and making them ever better. *Geoderma Regional* 20, e00255. <https://doi.org/10.1016/j.geodrs.2020.e00255>
- Benedet, L., Faria, W.M., Silva, S.H.G., Mancini, M., Demattê, J.A.M., Guilherme, L.R.G., Curi, N., 2020. Soil texture prediction using portable X-ray fluorescence spectrometry and visible near-infrared diffuse reflectance spectroscopy. *Geoderma* 376, 114553. <https://doi.org/10.1016/j.geoderma.2020.114553>
- Croffie, M.E.T., Williams, P.N., Fenton, O., Fenelon, A., Daly, K., 2022. Rubidium measured by XRF as a predictor of soil particle size in limestone and siliceous parent materials. *J Soils Sediments* 22, 818–830. <https://doi.org/10.1007/s11368-021-03109-4>
- Curi, N., Silva, S.H.G., Poggere, G.C., Menezes, M.D., 2017. Mapeamento de solos e magnetismo no campus da UFPA como traçadores ambientais. UFPA, Lavras, MG, Brazil.
- Fanning, D.S., Fanning, M.C.B., 1989. Soil morphology, genesis and classification. John Wiley and Sons Inc., New York.
- Faria, A.J.G. de, Silva, S.H.G., Andrade, R., Mancini, M., Melo, L.C.A., Weindorf, D.C., Guilherme, L.R.G., Curi, N., 2022. Prediction of soil organic matter content by combining data from Nix Pro™ color sensor and portable X-ray fluorescence spectrometry in tropical soils. *Geoderma Regional* 28, e00461. <https://doi.org/10.1016/j.geodrs.2021.e00461>
- Fernandez-Illescas, C.P., Porporato, A., Laio, F., Rodriguez-Iturbe, I., 2001. The ecohydrological role of soil texture in a water-limited ecosystem. *Water Resources Research* 37, 2863–2872. <https://doi.org/10.1029/2000WR000121>
- Gholizadeh, A., Zizala, D., Saberioon, M., Boruvka, L., 2018. Soil organic carbon and texture retrieving and mapping using proximal, airborne and Sentinel-2 spectral imaging. *Remote Sensing of Environment* 218, 89–103. <https://doi.org/10.1016/j.rse.2018.09.015>
- Greenberg, I., Vohland, M., Seidel, M., Hutengs, C., Bezard, R., Ludwig, B., 2023. Evaluation of Mid-Infrared and X-ray Fluorescence Data Fusion Approaches for Prediction of Soil Properties at the Field Scale. *Sensors (Basel)* 23, 662. <https://doi.org/10.3390/s23020662>
- Hamza, A., Karčauskienė, D., Mockevičienė, I., Repšienė, R., Tahir, M.A., Manzoor, M.Z., Kousar, S., Lodhi, S.S., Rasool, N., Ullah, I., 2025. Soil Aggregate Dynamics and Stability: Natural and Anthropogenic Drivers. *Agriculture* 15, 2500. <https://doi.org/10.3390/agriculture15232500>
- Hengl, T., Nussbaum, M., Wright, M.N., Heuvelink, G.B.M., Gräler, B., 2018. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ* 6, e5518. <https://doi.org/10.7717/peerj.5518>

- Hartemink, A.E., Zhang, Y., Bockheim, J.G., Curi, N., Silva, S.H.G., Grauer-Gray, J., Lowe, D.J., Krasilnikov, P., 2020. Soil horizon variation: A review, in: *Advances in Agronomy*. Elsevier, pp. 125–185. <https://doi.org/10.1016/bs.agron.2019.10.003>
- Holzinger, A., Saranti, A., Molnar, C., Biecek, P., Samek, W., 2022. Explainable AI Methods - A Brief Overview, in: Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, K.-R., Samek, W. (Eds.), *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*. Springer International Publishing, Cham, pp. 13–38. https://doi.org/10.1007/978-3-031-04083-2_2
- INMET, 2022. Normas climatológicas do Brasil 1991–2020. Available in: <https://clima.inmet.gov.br/NormaisClimatologicas>. Accessed in: June, 2026.
- IUSS Working Group WRB, 2022. World Reference Base for Soil Resources. International soil classification system for naming soils and creating legends for soil maps, 4th ed. International Union of Soil Sciences (IUSS), Vienna, Austria.
- Li, F., Xu, L., You, T., Lu, A., 2021. Measurement of potentially toxic elements in the soil through NIR, MIR, and XRF spectral data fusion. *Computers and Electronics in Agriculture* 187, 106257. <https://doi.org/10.1016/j.compag.2021.106257>
- Mancini, M., Andrade, R., Silva, S.H.G., Rafael, R.B.A., Mukhopadhyay, S., Li, B., Chakraborty, S., Guilherme, L.R.G., Acree, A., Weindorf, D.C., Curi, N., 2023. Multinational prediction of soil organic carbon and texture via proximal sensors. *Soil Science Soc of Amer J* 88, saj2.20593. <https://doi.org/10.1002/saj2.20593>
- Mancini, M., Weindorf, D.C., Silva, S.H.G., Chakraborty, S., Teixeira, A.F.S., Guilherme, L.R.G., Curi, N., 2019. Parent material distribution mapping from tropical soils data via machine learning and portable X-ray fluorescence (pXRF) spectrometry in Brazil. *Geoderma* 354, 113885. <https://doi.org/10.1016/j.geoderma.2019.113885>
- Mancini, M., Silva, S.H.G., Teixeira, A.F.S., Guilherme, L.R.G., Curi, N., 2020. Soil parent material prediction for Brazil via proximal soil sensing. *Geoderma Reg.* 22, e00310. <https://doi.org/10.1016/j.geodrs.2020.e00310>
- Molnar, C., 2024. Interpretable Machine Learning.
- Nunes, L.M., Bomfim, M.R., Conceição, J.N., Costa, O.A.V., Gloaguen, T.V., Almeida, M.C., Cruz, M.J.M., Júnior, A.B.S.R., Santos, J.A.G., 2025. Integrating Portable X-Ray Fluorescence and Machine Learning for Soil Texture Prediction in Mangrove Ecosystems. *Earth Syst Environ.* <https://doi.org/10.1007/s41748-025-00812-x>
- Padarian, J., Minasny, B., McBratney, A.B., 2020. Machine learning and soil sciences: a review aided by machine learning tools. *Soil* 6, 35–52. <https://doi.org/10.5194/soil-6-35-2020>
- Rezende, S.B. de, Franzmeier, D.P., Resende, M., Mancini, M., Curi, N., 2022. Pedogenic processes in a chronosequence of very deeply weathered soils in southeastern Brazil. *Catena* 215, 106362. <https://doi.org/10.1016/j.catena.2022.106362>
- Rossiter, D.G., 2018. Past, present & future of information technology in pedometrics. *Geoderma* 324, 131–137. <https://doi.org/10.1016/j.geoderma.2018.03.009>
- Roy, A., Jílková, V., 2025. Laboratory Optimization of Poultry Feather Hydrolysate Application for Soil Fertility: Effects of Soil Texture and Hydrolysate State and Addition Rate. *Agronomy* 15, 1638. <https://doi.org/10.3390/agronomy15071638>
- Santos, H.G. dos, Jacomine, P.K.T., Anjos, L.H.C. dos, Oliveira, V.Á. de, Lumbreras, J.F., Coelho, M.R., Almeida, J.A. de, Araújo Filho, J.C. de, Lima, H.N., Marques, F.A., Oliveira,

- J.B. de, Cunha, T.J.F., 2025. Sistema brasileiro de classificação de solos, 6th ed. Embrapa, Brasília, DF.
- Simandl, G.J., Paradis, S., 2022. Vanadium as a critical material: economic geology with emphasis on market and the main deposit types. *Applied Earth Science* 131, 218–236. <https://doi.org/10.1080/25726838.2022.2102883>
- Silva, F.M., Silva, S.H.G., Teixeira, A.F. dos S., Inda, A.V., Fruett, T., Weindorf, D.C., Guilherme, L.R.G., Curi, N., 2022. Using proximal sensors to assess pedogenetic development of Inceptisols and Oxisols in Brazil. *Geoderma Reg.* 28, 1–10. <https://doi.org/10.1016/j.geodrs.2021.e00465>
- Silva, S.H.G., Hartemink, A.E., Teixeira, A.F. dos S., Inda, A.V., Guilherme, L.R.G., Curi, N., 2018. Soil weathering analysis using a portable X-ray fluorescence (PXRF) spectrometer in an Inceptisol from the Brazilian Cerrado. *Appl. Clay Sci.* 162, 27–37. <https://doi.org/10.1016/j.clay.2018.05.028>
- Silva, S.H.G., Weindorf, D.C., Pinto, L.C., Faria, W.M., Acerbi Junior, F.W., Gomide, L.R., de Mello, J.M., de Pádua Junior, A.L., de Souza, I.A., Teixeira, A.F. dos S., Guilherme, L.R.G., Curi, N., 2020. Soil texture prediction in tropical soils: A portable X-ray fluorescence spectrometry approach. *Geoderma* 362, 114136. <https://doi.org/10.1016/j.geoderma.2019.114136>
- Soil Survey Staff, 2014. Keys to soil taxonomy, 12th ed. USDA-NRCS.
- Swain, S.R., Chakraborty, P., Panigrahi, N., Vasava, H.B., Reddy, N.N., Roy, S., Majeed, I., Das, B.S., 2021. Estimation of soil texture using Sentinel-2 multispectral imaging data: An ensemble modeling approach. *Soil and Tillage Research* 213, 105134. <https://doi.org/10.1016/j.still.2021.105134>
- Tavares, T.R., Minasny, B., McBratney, A., Cherubin, M.R., Marques, G.T., Ragagnin, M.M., Alves, E.E.N., Padarian, J., Lavres, J., De Carvalho, H.W.P., 2023. Estimating plant-available nutrients with XRF sensors: Towards a versatile analysis tool for soil condition assessment. *Geoderma* 439, 116701. <https://doi.org/10.1016/j.geoderma.2023.116701>
- Teixeira, A.F. S., Andrade, R., Mancini, M., Silva, S.H.G., Weindorf, D.C., Chakraborty, S., Guilherme, L.R.G., Curi, N., 2022. Proximal sensor data fusion for tropical soil property prediction: Soil fertility properties. *Journal of South American Earth Sciences* 116, 103873. <https://doi.org/10.1016/j.jsames.2022.103873>
- Tighe, M., Rogan, G., Wilson, S.C., Grave, P., Kealhofer, L., Yukongdi, P., 2018. The potential for portable X-ray fluorescence determination of soil copper at ancient metallurgy sites, and considerations beyond measurements of total concentrations. *Journal of Environmental Management* 206, 373–382. <https://doi.org/10.1016/j.jenvman.2017.10.052>
- Wadoux, A.M.J.-C., Molnar, C., 2022. Beyond prediction: methods for interpreting complex models of soil variation. *Geoderma* 422, 115953. <https://doi.org/10.1016/j.geoderma.2022.115953>
- Wadoux, A.M.J.-C., Minasny, B., McBratney, A.B., 2020a. Machine learning for digital soil mapping: Applications, challenges and suggested solutions. *Earth-Science Reviews* 210, 103359. <https://doi.org/10.1016/j.earscirev.2020.103359>
- Wadoux, A.M.J.-C., Samuel-Rosa, A., Poggio, L., Mulder, V.L., 2020b. A note on knowledge discovery and machine learning in digital soil mapping. *European Journal of Soil Science* 71, 133–136. <https://doi.org/10.1111/ejss.12909>

- Wang, P., Su, X., Zhou, Z., Wang, N., Liu, J., Zhu, B., 2023. Differential effects of soil texture and root traits on the spatial variability of soil infiltrability under natural revegetation in the Loess Plateau of China. *Catena* 220, 106693. <https://doi.org/10.1016/j.catena.2022.106693>
- Widyastuti, M.T., Minasny, B., Ng, W., de Caritat, P., Padarian, J., McBratney, A., 2025. Digital Mapping of Soil Vanadium Across Australia. *European Journal of Soil Science* 76, e70237. <https://doi.org/10.1111/ejss.70237>
- Ye, C., Zheng, G., Tao, Y., Xu, Y., Chu, G., Xu, C., Chen, S., Liu, Y., Zhang, X., Wang, D., 2024. Effect of Soil Texture on Soil Nutrient Status and Rice Nutrient Absorption in Paddy Soils. *Agronomy* 14, 1339. <https://doi.org/10.3390/agronomy14061339>
- Zhang, Z., Li, W., Zhang, W., Wang, C., Zhao, Z., Choi, Y., Zhang, H., Zhao, T., Sadri, F., 2026. A review of rubidium: Resources, technologies, and applications. *Desalination* 620, 119612. <https://doi.org/10.1016/j.desal.2025.119612>
- Zhang, Y., Hartemink, A.E., 2020. Data fusion of vis–NIR and PXRF spectra to predict soil physical and chemical properties. *European Journal of Soil Science* 71, 316–333. <https://doi.org/10.1111/ejss.12875>
- Zhu, Y., Weindorf, D.C., Zhang, W., 2011. Characterizing soils using a portable X-ray fluorescence spectrometer: 1. Soil texture. *Geoderma* 167–168, 167–177. <https://doi.org/10.1016/j.geoderma.2011.08.010>
- Zhu, Y., Guo, B., Liu, C., Lin, Y., Fu, Q., Li, N., Li, H., 2021. Soil fertility, enzyme activity, and microbial community structure diversity among different soil textures under different land use types in coastal saline soil. *J Soils Sediments* 21, 2240–2252. <https://doi.org/10.1007/s11368-021-02916-z>

Part III

Closing Remarks

Classifying and mapping soils is an extremely difficult task. Soils are the interface between Earth's crust and the atmosphere, and hence suffer the action of all weathering agents. The conceptual five factors of soil formation –namely climate, relief, parent material, organisms, and time– are a vague summary of almost all environmental processes that occur on Earth's surface and attest the nearly immeasurable variability of soils. In the past, pedologists have undertaken the task of understanding soils with very simple tools such as their own eyes and hands, coupled with knives, magnets, and augers, supported by their environmental knowledge at the time. Nowadays, much has improved, and a plethora of data is available: previous pedologic and geologic maps, meteorological data, satellite imagery, digital elevation models, maps of topographic attributes, proximal sensing data, and many others. There is much more data than there are scientists to digest it all. This thesis is a small contribution to this important collective mission of understanding soil processes and providing reliable information to potential users.

Perfectly characterizing soil variability often requires more samples than is allowed by financial and time constraints. The results presented herein showed that sensors can expand soil characterization beyond what is visible by the human eye through spectral and chemical sensors capable of producing data which tend to correlate with key minerals and organic compounds that compose soils. Additionally, sensor data can also be used to optimize sampling by investigating the temporal and spatial trends of spectral data from satellite imagery, indicating changes in moisture contents of soils and vegetation influenced by soil properties; such spatial stratification guarantees a better representation of soils in space using fewer samples, saving time and resources, making soil surveys viable and more accurate. Results also confirm that machine-learning models are fully capable of learning soil patterns from sensor data and that the chemical composition of soils is a powerful predictor of soil texture, allowing for a fast, accurate, and inexpensive texture estimation method. Lastly, machine-learning models have proven to be physically accurate, finding expected relationships between particle size fractions and the chemical elements Si, Fe, and Al. Trace elements such as Rb and V can also be relevant and help unveil new knowledge about the geomorphic processes and soil formation.

Soils compose the interface between Earth and atmosphere and the substrate in which humans produce their food and build their homes. Knowledge about soils is essential and must be accurate to properly guide agricultural management, environmental protection, and proper urban planning. The combination of sensors and machine-learning is recommended as a reliable method of mapping, characterizing, and learning about soils.