MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL RURAL DA AMAZÔNIA
PROGRAMA DE PÓS-GRADUAÇÃO EM AGRONOMIA
DOUTORADO EM AGRONOMIA

QUÉSIA SÁ PAVÃO

ANÁLISE DE TECNOLOGIA VERDE E INTELIGÊNCIA ARTIFICIAL APLICADAS À PREDIÇÃO DE ATRIBUTOS DO SOLO NA AMAZÔNIA BRASILEIRA

GREEN TECH ANALYSIS AND ARTIFICIAL INTELLIGENCE APPLIED FOR SOIL ATTRIBUTES PREDICTION IN THE BRAZILIAN AMAZON

BELÉM
2024

QUÉSIA SÁ PAVÃO

ANÁLISE DE TECNOLOGIA VERDE E INTELIGÊNCIA ARTIFICIAL APLICADAS À PREDIÇÃO DE ATRIBUTOS DO SOLO NA AMAZÔNIA BRASILEIRA

GREEN TECH ANALYSIS AND ARTIFICIAL INTELLIGENCE APPLIED FOR SOIL ATTRIBUTES PREDICTION IN THE BRAZILIAN AMAZON

Tese apresentada ao curso de doutorado do programa de Pós-graduação em Agronomia da Universidade Federal Rural da Amazônia, como requisito para a obtenção do Título de Doutora. Área de Concentração: Manejo e conservação dos recursos ambientais.

Orientador: Prof. Dr. Silvio Junio Ramos
Coorientadora: Drª Paula Godinho Ribeiro

BELÉM
2024

Dados Internacionais de Catalogação na Publicação (CIP)
Bibliotecas da Universidade Federal Rural da Amazônia
Gerada automaticamente mediante os dados fornecidos pelo(a) autor(a)

P337a Pavão, Quésia Sá
Análise de tecnologia verde e inteligência artificial aplicadas à predição de atributos do solo na amazônia brasileira / Quésia Sá Pavão. - 2024.
93 f. : il. color.

Tese (Doutorado) - Programa de Pós-Graduação em Agronomia (PPGA), Campus Universitário de Belém, Universidade Federal Rural Da Amazônia, Belém, 2024.
Orientador: Prof. Dr. Silvio Junio Ramos
Coorientador: Profa. Dra. Paula Godinho Ribeiro.

1. solos amazônicos. 2. predição. 3. sensoriamento próximo. 4. pXRF. 5. Vis-NIR. I. Ramos, Silvio Junio , *orient.* II. Título

CDD 631.4

QUÉSIA SÁ PAVÃO

ANÁLISE DE TECNOLOGIA VERDE E INTELIGÊNCIA ARTIFICIAL APLICADAS À PREDIÇÃO DE ATRIBUTOS DO SOLO NA AMAZÔNIA BRASILEIRA

GREEN TECH ANALYSIS AND ARTIFICIAL INTELLIGENCE APPLIED FOR SOIL ATTRIBUTES PREDICTION IN THE BRAZILIAN AMAZON

Tese apresentada ao curso de doutorado do programa de Pós-graduação em Agronomia da Universidade Federal Rural da Amazônia, como requisito para a obtenção do Título de Doutora. Orientador: Prof. Dr. Silvio Junio Ramos.
Co-orientadora: Dr[a] Paula Godinho Ribeiro.

Data da Aprovação: 30/10/2024

Banca Examinadora:

_____
Dr. Silvio Junio Ramos (Orientador)
Universidade Federal Rural da Amazônia (UFRA)

_____
Dra. Renata Andrade Reis
Universidade Federal de Lavras (UFLA)

_____
Dr. João Fernandes Da Silva Junior
Universidade Federal Rural da Amazônia (UFRA)

_____
Dra. Taciara Zborowski Horst
Universidade Tecnológica Federal do Paraná (UTFPR)

_____
Dr. Sérgio Henrique Godinho Silva
Universidade Federal de Lavras (UFLA)

Dedico este trabalho a Deus, à minha família, em especial aos meus pais por todos momentos de abnegação e incentivo, aos professores que me auxiliaram durante minha jornada acadêmica e aos amigos que me apoiaram.

AGRADECIMENTOS

A Deus por ser quem Ele é, meu refúgio, meu Pai Celestial.

À família Pavão, meu pai João Jorge Silva Pavão, minha mãe Aldenora Maria Sá Pavão, meus irmãos Davi, Filipe, Elias e Levi, pela compreensão em todos os momentos, pela motivação a continuar nas situações difíceis, pela torcida, apoio, pelo carinho e amor a mim dedicados.

À Universidade Federal Rural da Amazônia pela oportunidade de realizar este doutorado, através do Programa de Pós-Graduação em Agronomia, e aos docentes que ao longo destes anos me ensinaram. Ao Instituto Tecnológico Vale - ITV por ceder o espaço e estrutura para a realização de parte dos ensaios da pesquisa.

Ao meu orientador, Dr. Silvio Junio Ramos, pelo acolhimento, dedicação, esclarecimentos e partilha do seu vasto conhecimento. À minha coorientadora, Dra. Paula Godinho Ribeiro, pela atenção, disponibilidade de tempo, pelas sugestões e conhecimento compartilhado, tanto na experimentação quanto nas correções.

Ao Me. Gutierre Pereira Maciel, pelo apoio no desenvolvimento dessa pesquisa. À professora Suzana Romeiro Araújo, pelas contribuições. Aos meus amigos, Aquila Tavares, Dandara Lima, Delany Albuquerque e Odilene Souza, pelo apoio no desenvolvimento em fase de preparo das amostras desta pesquisa.

Ao professor Antônio Rodrigues Fernandes pela parceria, justamente com o Yan Nunes, pela concessão de amostras do Banco de Solos do Laboratório de Elementos Traços no Ambiente (LETAM) da Universidade Federal Rural da Amazônia.

Aos membros da banca avaliadora de qualificação e defesa, pela contribuição.

Quésia Sá Pavão

SUMÁRIO

1. CONTEXTUALIZAÇÃO

A conservação do solo é uma preocupação global que reflete no 15º Objetivo de Desenvolvimento Sustentável (ODS) das Nações Unidas, pois a vida terrestre é um dos desafios mundiais a serem atingindo até 2030 (Nações Unidas no Brasil, 2024). Os solos representam segurança alimentar, tendo em vista serem a base de sustentação da vida e fonte de nutrientes para as plantas, além de constituírem abrigo para macro e microfauna, por isso que ao conservar o solo é possível atender a outros ODS, como por exemplo os ODS 1 (sem pobreza) e o ODS 2 (fome zero) (ALBALADEJO, DÍAZ-PEREIRA & DE VENTE, 2021).

Para garantir o cumprimento desses objetivos é necessário o conhecimento dos atributos físicos, químicos, mineralógicos e biológicos dos solos para obter a eficiência no manejo e conservação ambiental do solo, assim como o monitoramento das mudanças ocorrente do solo. Essas informações são fundamentais para a tomada de decisão, porém os métodos de análises laboratoriais de solo já consolidados para essa finalidade são complexos, demoradas, exigem grande quantidade de reagentes químicos e produzem resíduos químicos que podem afetar o ambiente essas análises, além de demandar equipamentos sofisticados (BARRA et al., 2021; TSAKIRIDIS et al., 2021).

Nesse sentido, técnicas de sensoriamento próximo são uma alternativa que permitem a medição de propriedades-chave do solo com vantagens de menor tempo de análise e sem geração de resíduos, minimizando os impactos ambientais; apesar de que em condições laboratoriais ainda são necessários alguns procedimentos padrões no preparo da amostra, como a secagem, moagem e peneiramento (GUERRERO, DE NEVE & MOUAZEN, 2021; JAVADI, MUNNAF & MOUAZEN, 2021; VISCARRA ROSSEL et al., 2011). Vários sensores proximais podem avaliar a capacidade do solo de acumular e conduzir carga elétrica, de absorver, refletir e/ou emitir energia eletromagnética, de liberar íons e de resistir à distorção mecânica (VISCARRA ROSSEL et al., 2011).

O espectrômetro portátil de fluorescência de raios-X (pXRF) e o espectrômetro de reflectância difusa na região do visível ao infravermelho próximo (Vis-NIR) são exemplos desses sensores. O pXRF é capaz de medir com precisão a concentração de elementos inorgânicos do solo, enquanto o Vis-NIR tem a capacidade avaliar a reflectância da energia eletromagnética aplicada na amostra na região do visível ao infravermelho próximo e seus dados podem ser utilizados para estimar o componente orgânico e o conjunto de mineralogia

do solo (O'ROURKE et al., 2016). Os dados gerados desses sensores são usados para predição de diversas variáveis, como textura (MANCINI et al., 2024) e atributos da fertilidade dos solos (JAVADI, MUNNAF & MOUAZEN, 2021).

O presente estudo visa a explorar esses sensores proximais (pXRF e Vis-NIR) na predição de atributos dos solos da Amazônia, especificamente dos solos do estado do Pará, para contribuir com o avanço do conhecimento sobre o uso desses sensores em solos tropicais e adotar o seu uso nos laboratórios. Um estado com uma extensão territorial grande, com forte aptidão agrícola e crescente demanda dos agricultores por informações básicas de textura e fertilidade dos solos são motivações para superar os obstáculos de distância geográfica dos laboratórios de análises. À medida que se inicia essa pesquisa com solos amazônicos, pode-se avançar na criação de modelos precisos e aplicáveis a essa região, tornando os sensores uma ferramenta eficiente para o conhecimento dos solos regionais, seja visando a produção agrícola ou para com a preservação destes em áreas protegidas. Nossa hipótese é que dados dos sensores pXRF e Vis-NIR combinados são capazes de predizer a textura e fertilidade de solos da Amazônia com resultados mais acurados que os dados dos sensores isolados.

REFERÊNCIAS

ALBALADEJO, Juan; DÍAZ-PEREIRA, Elvira; DE VENTE, Joris. Eco-holistic soil conservation to support land degradation neutrality and the sustainable development goals. **Catena**, v. 196, p. 104823, 2021. https://doi.org/10.1016/j.catena.2020.104823

BARRA, Issam et al. Soil spectroscopy with the use of chemometrics, machine learning and pre-processing techniques in soil diagnosis: Recent advances–A review. **TrAC Trends in Analytical Chemistry**, v. 135, p. 116166, 2021. https://doi.org/10.1016/j.trac.2020.116166

GUERRERO, A., DE NEVE, S., & MOUAZEN, A. M. (2021). Current sensor technologies for in situ and on-line measurement of soil nitrogen for variable rate fertilization: A review. In **Advances in Agronomy** (Vol. 168). https://doi.org/10.1016/bs.agron.2021.02.001

JAVADI, S. H., MUNNAF, M. A., & MOUAZEN, A. M. (2021). Fusion of Vis-NIR and XRF spectra for estimation of key soil attributes. **Geoderma**, *385*. https://doi.org/10.1016/j.geoderma.2020.114851

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. (2024). Multinational prediction of soil organic carbon and texture via proximal sensors. **Soil Science Society of America Journal**, *88*(1), 8–26. https://doi.org/10.1002/SAJ2.20593

NAÇÕES UNIDAS NO BRASIL. Sobre o nosso trabalho para alcançar os Objetivos de Desenvolvimento Sustentável no Brasil. Disponível em:< https://brasil.un.org/pt-br/sdgs>. Aceeso em 23 de dezembro de 2024.

O'ROURKE, S. M. et al. An assessment of model averaging to improve predictive power of portable vis-NIR and XRF for the determination of agronomic soil properties. **Geoderma**, v. 279, p. 31-44, 2016.

TSAKIRIDIS, Nikolaos L. et al. Improving the predictions of soil properties from VNIR–SWIR spectra in an unlabeled region using semi-supervised and active learning. **Geoderma**, v. 387, p. 114830, 2021. https://doi.org/10.1016/j.geoderma.2020.114830.

VISCARRA ROSSEL, R. A et al. Proximal soil sensing: An effective approach for soil measurements in space and time. **Advances in agronomy**, v. 113, p. 243-291, 2011.

O tópico 2 corresponde ao primeiro artigo resultado da tese, publicado na Geoderma Regional, disponível no link: https://doi.org/10.1016/j.geodrs.2024.e00813

2 TEXTURE PREDICTION OF NATURAL SOILS IN THE BRAZILIAN AMAZON THROUGH PROXIMAL SENSORS

RESUMO

Os sensores próximos fornecem análises rápidas, de baixo custo, ambientalmente amigáveis e confiáveis para a caracterização de solos e outros materiais. Inúmeros estudos têm sido realizados em solos de regiões temperadas, mas há lacunas de conhecimento quanto ao uso desses dispositivos em solos tropicais, especialmente na região amazônica. Nesse sentido, o presente estudo utilizou sensores proximais portáteis de espectroscopia de fluorescência de raios X (pXRF) e espectroscopia de reflectância difusa na região do visível ao infravermelho próximo (Vis-NIR) para predizer a textura de solos naturais em 61 municípios do estado do Pará, região Amazônica, Brasil. Os objetivos foram: i) investigar a precisão da previsão da textura do solo com base em dados de sensores separadamente e na fusão de sensores (dados pXRF e Vis-NIR) utilizando dois algoritmos supervisionados (Random Forest, RF, e Support Vector Machine, SVM) e ii) avaliar o efeito do horizonte do solo (horizontes superficiais e subsuperficiais, e a sua combinação) na predição da textura de solos naturais tropicais. No total, foram coletadas 233 amostras de solo nas profundidades de 0-20 cm e 80-100 cm, equivalentes aos horizontes superficial e subsuperficial, em áreas com cobertura florestal primária ou secundária com pelo menos 20 anos de regeneração natural e aproximadamente 20 ha de área de cobertura. Para a análise da textura do solo foi utilizado o método do hidrômetro. Paralelamente, uma parte das amostras de solo foi analisada por pXRF e Vis-NIR, em triplicata, em condições laboratoriais. Os modelos preditivos com RF foram mais robustos em comparação com os modelos obtidos com SVM, de acordo com a distância interquartil da razão de desempenho (RPIQ), o coeficiente de determinação (R^2), a raiz do erro quadrático médio (RMSE) e o erro absoluto médio (MAE). Os valores de R^2 obtidos por pXRF, Vis-NIR e combinação de dados de sensores foram, respectivamente, 0,89, 0,87 e 0,93 para areia; 0,92, 0,90 e 0,93 para argila; e 0,91, 0,67 e 0,93 para silte. Em geral, os modelos de predição de argila obtiveram valores de R^2 mais elevados em comparação com os modelos de areia e silte. A predição da textura do solo utilizando a fusão de sensores apresentou valores de RMSE mais

baixos e valores de R^2 e RPIQ mais elevados, respectivamente (areia: 7,79, 0,93, 4,69; argila: 5,58, 0,93, 3,86; e silte: 5,72, 0,92, 2,92) em comparação com o sensor com melhor desempenho individual (Vis-NIR). No que diz respeito ao modelo ótimo que utiliza dados de sensores individuais, os modelos Vis-NIR apresentaram um erro reduzido para a predição de argila e areia. O efeito da combinação de horizontes num único e maior conjunto de dados foi minimamente importante para os modelos. Os resultados demonstram confiança no uso de sensores próximos para avaliação da textura do solo em solos naturais da Amazônia, visando reduzir os custos e o tempo necessário para as análises.

Palavras-chave: pXRF; Vis-NIR; Floresta aleatória

ABSTRACT

Proximal sensors provide fast, low-cost, environmentally friendly, and reliable analyses for the characterization of soils and other materials. Numerous studies have been conducted on soils in temperate regions, but there are knowledge gaps regarding the use of these devices in tropical soils, especially in the Amazon region. In this regard, this study utilized portable proximal sensors of X-ray fluorescence spectroscopy (pXRF) and diffuse reflectance spectroscopy in the visible to near-infrared region (Vis-NIR) for predicting the texture of natural soils in 61 municipalities in the state of Pará, Amazon region, Brazil. The objectives were: i) to investigate the accuracy of soil texture prediction based on data from sensors separately and sensor fusion (pXRF and Vis-NIR data) using two supervised algorithms (Random Forest, RF, and Support Vector Machine, SVM) and ii) to assess the effect of soil horizon (superficial and subsuperficial horizons, and their combination) in predicting the texture of tropical natural soils. In total, 233 soil samples were collected in the 0-20 cm and 80-100 cm depths, equivalent to superficial and subsuperficial horizons in areas with primary or secondary forest cover with at least 20 years of natural regeneration and approximately 20 ha of coverage area. The hydrometer method was used for soil texture analysis. In parallel, a portion of the soil samples was analyzed by pXRF and Vis-NIR, in triplicate, under laboratory conditions. The predictive models with RF were more robust compared to the models obtained with SVM, according to ratio performance interquartile distance (RPIQ), coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE). The R^2 values obtained by pXRF, Vis-NIR, and sensor data combination were, respectively, 0.89, 0.87, and 0.93 for sand; 0.92, 0.90, and 0.93 for clay; and 0.91, 0.67, and 0.93 for silt. Overall, clay prediction models achieved higher R^2 values compared to sand and silt models. Soil texture

prediction using sensor fusion showed lower RMSE values and higher R^2 and RPIQ values, respectively (sand: 7.79, 0.93, 4.69; clay: 5.58, 0.93, 3.86; and silt: 5.72, 0.92, 2.92) compared to the best-performing sensor individually (Vis-NIR). With regard to the optimal model utilizing individual sensor data, Vis-NIR models exhibited reduced error for clay and sand prediction. The effect of combining horizons to a single and bigger dataset was minimally important for the models. The results demonstrate confidence in the use of proximal sensors for soil texture assessment in natural Amazon soils, aiming to reduce costs and the time required for analyses.
Keywords: pXRF; Vis-NIR; Random forest

2.1 Introduction

The Amazon region is known for its vast biodiversity (Peres et al., 2010), which includes not only animals and plants, but also its soils. Soil diversity is essential for sustaining life, including nutrient cycling, habitat, and water reservoir, etc. The different types of soil reflect the natural processes taking place over time, such as weathering, and leaching, in addition to human activities, such as wildfires that occur in these environments (Bezdicek et al., 2015; Smith et al., 2016). Given the importance of soils for ecosystem services, it is important to study, preserve, and monitor the soils to maintain a balanced environment for future generations.

Soil diversity is linked to the interaction between different types of soil formation factors and processes, rendering soil a multifactorial component of the landscape (Schaetzl and Anderson, 2005). Climate-related factors, such as temperature and precipitation, for example, accelerate weathering and leaching, accounting for physical, chemical, and mineralogical differences between soils from temperate and tropical regions (Silva et al., 2021). Besides, soils in temperate regions tend to be younger, while those found in tropical rainforests are commonly of early Quaternary to late Tertiary age (> 1,500,000 years) and contain low amounts of weatherable minerals, low cation exchange capacity (CEC), and silt contents (Carpenter et al., 2014; dos Santos et al., 2024). As a consequence, temperate forest soils tend to present an abundance of 2:1 and 1:1 clay minerals, while the mineralogy of tropical rainforest soils is predominantly kaolinitic (1:1 clay mineral) and rich in Fe and Al oxides (hematite, goethite, gibbsite) in the clay fraction (Carpenter et al., 2014). These minerals in the clay fraction, mainly gibbsite, contribute to the formation of a granular structure in the B horizon of highly-weathered

tropical soils, which completely modifies the porous system of such soils and, hence, increasing water infiltration rate and reduces water holding capacity (Resende et al., 2014).

In addition to exhibiting higher apparent density and lower water content, tropical soils are distinguished by their low silt content and high proportions of either sand or clay. This contrasts with the composition of soils in temperate regions (Mancini et al., 2024; Minasny and Hartemink, 2011; Silva et al., 2021). These unique properties of tropical soils and the lack of detailed information on tropical soil variability promote challenges and opportunities which is why there is a need for more precise and tailored analytical approaches to rapidly determine their mineralogical, physical, and chemical attributes.

One of the key physical attributes of tropical soils is texture, given its strong influence on CEC, hydrothermal regime, plant growth, etc. (Coblinski et al., 2020; Mancini et al., 2024). Additionally, it plays an important role in the accessibility of nutrients to plants, the diversity of living organisms, and the overall soil quality (Coblinski et al., 2020). Texture analysis provides fundamental information for decision-making regarding sustainable soil use (Benedet et al., 2020a), as it is related to several physical, chemical, biological, hydrological, and mechanical soil processes (Coblinski et al., 2020; Hossain et al., 2018). However, conventional laboratory analyses for determining this soil attribute are time-consuming, expensive, and require the use of chemical reagents (e.g., sodium hexametaphosphate) that generate residues (Andrade et al., 2020; Parent et al., 2021).

One of the recent alternatives becoming widely used in soil science encompasses proximal sensors, such as visible and near-infrared reflectance spectrometry (Vis-NIR) and portable X-ray fluorescence spectrometry (pXRF) for soil texture prediction (Zhang and Hartemink, 2019). These sensors promote non-destructive soil analyses, are less costly, do not require the use of reagents, and deliver results in a few seconds (Coblinski et al., 2020; Silva et al., 2020). Data from these proximal sensors have been successfully applied in predicting texture, both by using single sensor and combination of data from multiple sensors (Andrade et al., 2020; Hobley and Prater, 2019; Mancini et al., 2024; Zhang and Hartemink, 2020), applying methods such as the partial least-squares regression (PLSR) (Naimi et al., 2022), cubist regression algorithm (Coblinski et al., 2020), random forest (RF) (Hobley and Prater, 2019; Silva et al., 2020), support vector machine (SVM) (Silva et al., 2020) and multiple linear regression (Zhu et al., 2011). These methods have shown better texture prediction results than other multivariate calibrations, such as the stepwise multiple linear regression (Silva et al.,

2020). However, this approach based on proximal sensors and machine learning has not been investigated in the Amazon region and its peculiar soils, which motivates this study.

Research on Brazilian soils has shown that RF and data fusion techniques can effectively predict soil texture (Andrade et al., 2022). However, studies in diverse environmental conditions are necessary to confirm these techniques' effectiveness across different scenarios. This is particularly true for tropical environments like the Amazon region, where accurate and current data on soil attributes are in demand (Benedet et al., 2020a; Minasny and Hartemink, 2011). In addition, it is known that the soil texture varies spatially across the landscape and along a soil profile. This motivates studies to investigate the impact of soil texture at different layers on the accuracy of predictive models.

Given that studies combining pXRF and Vis-NIR data along with machine learning algorithms for predicting soil attributes are practically nonexistent in the Amazon region, keeping in mind the global environmental importance of this region and, hence, the necessity of characterizing its soils in detail, our objectives were to: i) investigate the accuracy of texture estimation of Amazon soils based on proximal sensor data separately (pXRF or Vis-NIR) and sensor fusion (pXRF and Vis-NIR) by using two supervised algorithms (RF and SVM), and ii) to compare the effect of surface, subsurface horizons, and the combined horizons (superficial and subsuperficial) datasets on texture prediction. We hypothesize that through these two supervised machine learning algorithms, predicted Amazon soil texture using proximal sensor data will be analogous to actual values obtained by conventional laboratory analysis, and that sensor fusion will provide greater prediction accuracy of soil texture, in the data set containing surface and subsurface horizons combined and individually.

2.2 Material and Methods

2.2.1 Study Area and Soil Sampling

Samples of rainforest soils from the Amazon region were collected in 61 municipalities in the state of Pará, northern Brazil (Figure 1). The Köppen climate classification of the state is tropical, with 66% of the territory being Am (monsoon), 28% Af (no dry season), and 4% Aw (with dry winter). The average annual temperature is 25 ºC, and the average annual precipitation is 1,900 mm (Alvares et al., 2013). The state of Pará has 144 municipalities and is the second largest state in Brazil, with a territorial extension of 1,245,870.704 km² (IBGE, 2022). Pará is

the largest state in the eastern Amazon, and it is precisely located within the ecological zone known as the Humid Tropics, in the Amazon biome (Brasil et al., 2020; Gonçalves et al., 2022). Soils from this state are influenced by the parent materials, climatic zones, and variability of geomorphologic positions. This territory is marked by the Amazonian craton and the Amazonian sedimentary basin (Gonçalves et al., 2022). It presents a great diversity of soils and holds intensive farming and mining activities (de Souza et al., 2015). According to the Brazilian Soil Classification System (Santos et al., 2018), the main soil classes are Argissolo Vermelho-Amarelo (39.81%), Latossolo Amarelo (21.08%), and Latossolo Vermelho (19.57%) (Santos et al., 2018; Brasil et al., 2020), corresponding to Ultisols, and Oxisols of the USA Soil Taxonomy (Soil Survey Staff, 2022). These two soils are usually developed from sediments of the Barriers Formation (Plio-Pleistocene) and the Alter do Chão Formation (Cretaceous) (Birani et al., 2015).

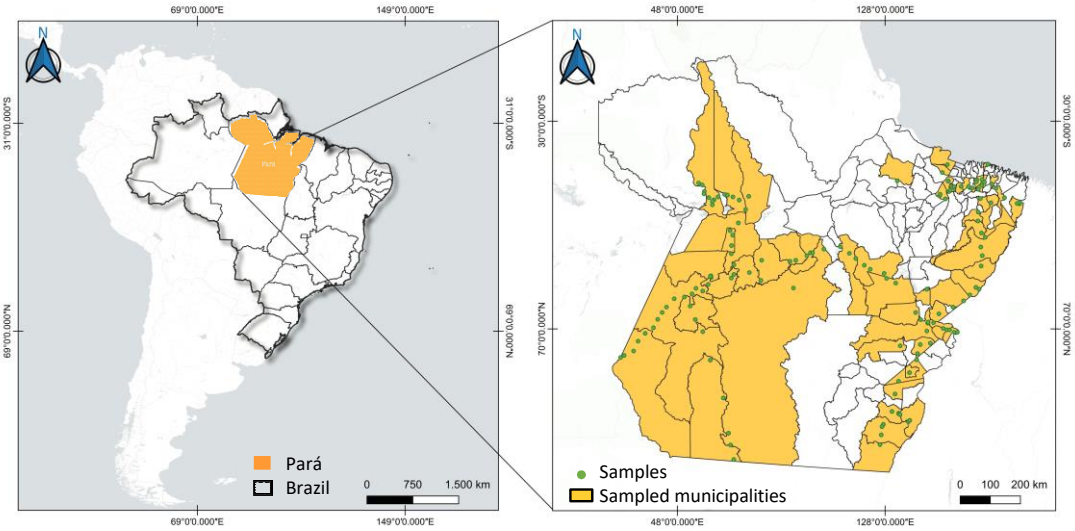

Figure 1 - Study area and soil sample collection sites in the state of Pará, Brazil.

A total of 233 soil samples were collected, with 140 taken from the superficial layer (0-20 cm) and 93 from the subsuperficial layer (80-100 cm). The sampling areas were selected in locations with either primary or secondary forest cover, which had undergone natural regeneration for at least 20 years, and had a forest coverage of approximately 20 ha.

2.2.2 Laboratory analyses and spectral data acquisition

Before the analyses, soil samples were air-dried, disaggregated, and sieved through a 2 mm mesh. Soil texture was determined using the hydrometer method. In this method, the clay fraction is determined by sedimentation, the sand fraction by sieving, and the silt fraction is obtained by the difference between the other fractions, as described by Gee and Bauder (1986).

A portion of each soil sample was placed in a polypropylene bag for total content analysis using the portable X-ray fluorescence spectrometer (pXRF) S1 TITAN 800 (Bruker, Billerica, MA, USA). This model is equipped with a high-performance graphene window silicon drift detector (SDD), providing the widest application range and best light element performance for Mg, Al, and Si. Before analyzing the samples, the recovery values of each element content were calculated using certified reference material from the manufacturer (Check Sample - CS), and from the National Institute of Standards & Technology - NIST (Standard Reference Material® 2710a and Standard Reference Material® 2711a). The recovery values are given by the following formula: Recovery (in %) = 100 x (obtained content / reference sample content).

After that, the soil samples were analyzed by the device in triplicate, with each reading lasting 60 s. Given the high recovery, the following 13 elements were used for prediction models: magnesium (Mg), aluminum (Al), silicon (Si), phosphorus (P), calcium (Ca), titanium (Ti), vanadium (V), manganese (Mn), iron (Fe), copper (Cu), zinc (Zn), strontium (Sr), and zirconium (Zr). Their recovery values were: (CS/2710a/2711a), Mg (0/113/123), Al (0/94/93), Si (0/94/96), P (97/34/56), Ca (102/88/99), Ti (0/93/98), V (0/0/114), Mn (99/94/104), Fe (100/98/100), Cu (102/97/98), Zn (108/101/105), Sr (101/101/97), and Zr (0/0/0). A value of 0 indicates that either the equipment could not detect the element or that the element has no certified value.

For spectral data acquisition, soil samples were dried at 45 °C, ground, and sieved through a 2 mm mesh, then distributed on Petri dishes. The Fieldspec® 3 spectroradiometer (Analytical Spectral Devices, Boulder, CO, USA) was used, with a spectral range from visible to shortwave infrared (350–2500 nm) and spectral resolution of 1 nm from 350 to 700 nm, 3 nm from 700 to 1400 nm, and 10 nm from 1400 to 2500 nm (Demattê et al., 2019). The data output sampling interval is 1 nm, with 2151 reported channels. During analysis, the spectral sensor (Vis-NIR) was used in a dark room (eliminating the entry of external light during spectrum acquisition) with controlled temperature and humidity and no reflective interference from other

objects. It was positioned 8 cm from the surface of the sample, scanning an area of approximately 2 cm². A standard Spectralon white plate was scanned every 20 minutes during the scans. Two replicates were obtained (one involving a 180° rotation of the Petri dish) for each sample. Each spectrum was generated by averaging 100 readings over 10 s. The average value of the two replicates was used for each sample. The soil spectra were measured following the protocol proposed by Ben Dor et al. (2015) (Demattê et al., 2019).

2.2.3 Statistical analyses

All statistical analyses were made in R software (R Core Team, 2023). For a better understanding of the soils variability, the data were previously subjected to descriptive analyses (minimum, maximum, standard deviation, and coefficient of variation values), textural classification, and creation of boxplots and spectral reflectance curves based on Vis-NIR results. For that, the spectral reflectance was transformed for continuum removal (CR) (Clark and Roush, 1984). This preprocessing eliminates the continuous features of the spectra and is often used to isolate specific absorption features, capable of providing calibration models with high precision (Demattê et al., 2019). This process was applied to the raw spectra only to highlight the absorption characteristics of the soil spectrum.

The whole dataset was divided into three parts: I) data from the superficial horizon only, II) data from the subsuperficial horizon only, and III) data from the combined horizons. This classification was used in the analysis of pXRF, Vis-NIR, and sensor fusion (pXRF and Vis-NIR).

According to the contents of clay, silt and sand, each sample was represented within the corresponding textural class of the textural triangle (Schoeneberger et al., 2012), following the Brazilian Soil Classification System - SiBCS (Santos et al., 2018), using the "dplyr" (Wickham et al., 2023) and "soiltexture" packages (Moeys, 2018). Principal component analysis (PCA) was applied to visually explore the total element contents delivered by pXRF at two soil depths (superficial and subsuperficial horizons), using the "FactoMineR" (Lê et al., 2008), "factoextra" (Kassambara and Mundt, 2020), and "ggplot2" packages (Wickham, 2016). The Vis-NIR spectra of soil samples were generated using the "prospectr" (Stevens and Ramirez Lopez, 2014) package, considering the spectra of all samples and the maximum, minimum, and median wavelength values for each horizon and the dataset with both horizons combined.

2.2.4 Modeling

To model and predict soil texture, the predictor variables for pXRF included the elements: Mg, Al, Si, P, Ca, Ti, V, Mn, Fe, Cu, Zn, Sr, and Zr. Meanwhile, the Vis-NIR variables included the reflectance of the samples at wavelengths in the visible to near-infrared range (350 to 2500 nm). We created prediction models using three different approaches - raw pXRF data, raw Vis-NIR data, and a fusion strategy that involved concatenating pXRF and Vis-NIR data. We chose to use low-level fusion, which involved using raw data as an input for the fusion procedure. This approach is the most direct and conceptually simple method to merge data (Azcarate et al., 2021). This process was done for samples from superficial, subsuperficial, and combined horizons.

To obtain a parsimonious model, a data dimensionality reduction method was used, called Boruta. The variable selection process for both sensors was carried out using the 'Boruta' package (Kursa and Rudnicki, 2010), which categorizes variables as confirmed, tentative, or rejected. Only confirmed-labeled variables were considered for predictions. Prediction of tropical soil texture was developed using two machine learning algorithms: RF and SVM with Radial Basis Function Kernel , with the assistance of the 'caret' package (Kuhn, 2008).

RF is based on ensemble learning (Breiman, 2001), combining multiple decision trees where several independent bootstrap samples and a random selection of sample points with replacement are used in the construction of each tree (Shahriari et al., 2019). Training is done by splitting the dataset according to a series of conditional rules (if-then), with the most important predictors placed at the top of the trees to create nodes and leaves of greater homogeneity. The result is the average prediction of all trees. The main hyperparameter adjusted to improve model performance was "mtry," which is the number of predictors used in each tree (Siqueira et al., 2023). The "ntree" parameter is the number of trees in the forest.

SVM performs training by seeking a hyperplane capable of separating the data with the largest possible margin (Cortes and Vapnik, 1995). The margin is constructed as the space between the decision boundary and the first sample points on each side. Radial SVM uses a radial basis function kernel to transform the data into a higher-dimensional space, thus producing flexible decision boundaries when the data is not easily separable. The algorithm's performance depends on a good balance between the "sigma" and "cost" hyperparameters.

"Sigma" controls the degree of non-linearity of the model, while "cost" controls the trade-off between margin and training errors (Siqueira et al., 2023).

For RF, the “radomForest” package (Liaw and Wiener, 2002) was used with mtry parameters for pXRF data ranging from 2 to 8; for Vis-NIR data between 2 and 5, and for sensor fusion from 2 to 13, with ntree = 500. For SVM, the "svmRadial" method was used with the “kernlab” package (Karatzoglou et al., 2023), for pXRF data, with cost parameters between 0.25 and 2 and sigma between 0.14 and 0.74; for Vis-NIR data, cost between 0.25 and 2 and sigma between 0.10 and 2.17; for sensor fusion, cost between 0.25 and 2 and sigma between 0.06 and 0.20. The models underwent cross-validation (*leave-one-out* – LOOCV), where one sample from the dataset (N) is removed, and the model is calibrated for N-sample in the dataset. The excluded sample is then predicted by the model. In the end, all samples are predicted by an independent empirical statistical model and are subsequently used to evaluate prediction accuracy (Richter et al., 2009).

Model evaluation was based on the following metrics: root mean square error – RMSE (Eq.1), mean absolute error – MAE (Eq.2), coefficient of determination – R^2 (Eq.3), and ratio of performance to interquartile distance - RPIQ (Eq.4). RPIQ index accounts much better for the spread of the population for soil-sample sets often present a highly skewed distribution (Bellon-Maurel et al., 2010). The R^2 equal to 1 means an optimal value, that is, all the variance of the response variable (clay, silt, and sand contents) was explained by the model. The RMSE closer to zero, the better the prediction power of the model. We adopted the best model to be the one with the highest values of R^2 and RPIQ combined with lower values of RMSE (Coblinski et al., 2020; Nawar et al., 2016).

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

$$MAE = \frac{1}{n}\sum_{i=1}^{n} \left|y_i - \hat{y}\right| \quad \text{(Eq. 2)}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad \text{(Eq. 3)}$$

$$RPIQ = \frac{Q3 - Q1}{RMSE} \quad \text{(Eq. 4)}$$

Where, n: number of observations; y_i: value measured by laboratory analysis; $\hat{y}_i$: value estimated by the model; $\bar{y}$: mean of all values measured by chemical analysis; Q3: third quartile; Q1: first quartile

2.3 Results and Discussion

2.3.1 Exploratory analyses

The soil samples from natural areas of the Amazon covered all five textural classes (Figure 2). The texture classification revealed that 42.9% of the samples are of loamy, clayey (30%), very fine clayey (13.7%), sandy (10.7%), and silty (2.6%).

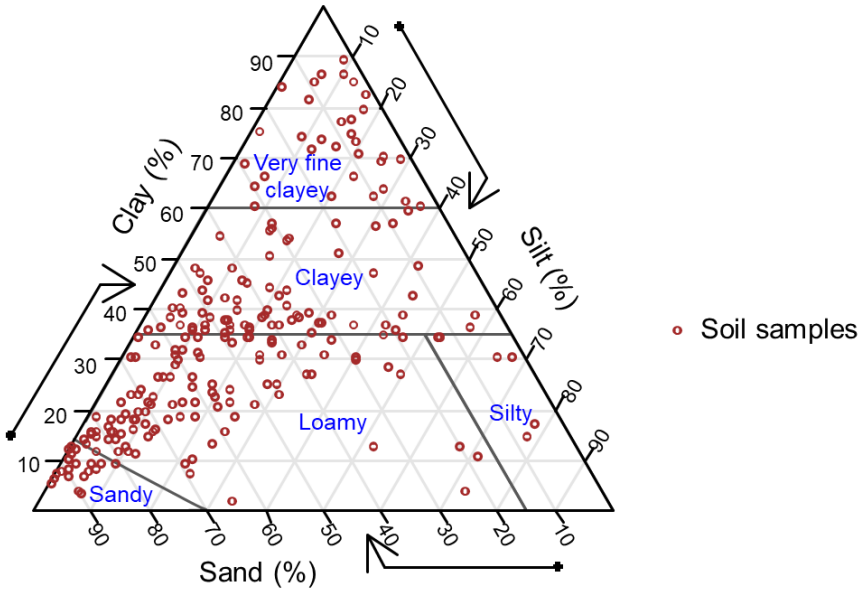

Figure 2 - Distribution of soil samples collected in Brazilian Amazon region in the Textural Triangle by the Brazilian soil texture classification

The high clay content, found in most soil samples, can be attributed to geological, pedological, and climatic factors. The region experiences high rainfall, which coupled with the occurrence of various types of fine-textured rocks, such as gabbro, basalt, shale, and clayey sediments, contributes to the formation of clayey soils such as Oxisols and Ultisols. These soil types are predominant in the state of Pará (Delarmelinda et al., 2017; Quesada et al., 2010). The low silt content is common due to the intense weathering of most Brazilian soils, except for C horizons, which are less weathered (Silva et al., 2020). Compared to temperate soils, tropical

soils have low silt content (except when weathering is limited) and high sand or clay content, which reflects the parent material and degree of weathering (Andrade et al., 2020; Silva et al., 2021).

The table 1 shows that the sand, silt, and clay contents varied between 1% and 94%, 0% and 78%, and 2% and 89%, respectively, for both horizons. The subsuperficial horizon had higher average silt (19%) and clay (37%) fractions compared to the superficial horizon. On the other hand, the superficial horizon had a higher average sand fraction (51%) than the subsuperficial horizon. As a result of strong chemical weathering in the Amazon, clay minerals such as kaolinite, hematite, and goethite are formed. However, the clay fraction tends to increase in the subsurface at various depths, which is evidence of clay translocation (eluviation) from the topsoil to deeper soil layers. This phenomenon is particularly intense in the region (Delarmelinda et al., 2017; Fritsch et al., 2002).

Some of the soil samples were taken from areas that had previously been covered by secondary forests. These areas may have been affected by natural factors like burning, or by human activity like soil tillage for agriculture or pasture. This disturbance to the soil may have caused a decrease in the amount of clay in the superficial horizon. When vegetation is removed, it can trigger various processes such as loss of organic matter and water erosion, which can result in the dispersion and redistribution of clay in the soil profile or across the landscape (Chaves et al., 2020; Pereira et al., 2020).

Table 1 - Descriptive statistics of soil texture, in %, in the 0-20 cm (superficial) and 80-100 cm (subsuperficial) layers (n= 233).

Soil layer	Soil Fraction	Min[1]	Max[2]	Mean	Median	SD[3]	CV[4]
Superficial	Sand	2	94	51	50	25	49
	Silt	1	78	17	15	13	78
	Clay	3	86	31	32	18	59
Subsuperficial	Sand	1	89	42	44	25	61
	Silt	0	78	19	13	17	90
	Clay	2	89	37	36	21	57

[1]Minimum values, [2]Maximum values, [3]Standard deviation, and [4]Coefficient of variation

A wide variety of soil textures was observed in both the superficial and subsuperficial layers. The sand fraction had the lowest variability, while the silt and clay fractions had higher variability. Coefficients of variation ranged from 49% to 61% for the sand fraction, 78% to 90% for the silt fraction, and 57% to 59% for the clay fraction, as shown in Table 1. The texture diversity of tropical soil samples at different depths is better illustrated in the boxplots (Figure 3). Soil depth is often overlooked in studies, but it is crucial to consider as soil attributes can vary significantly along the profile, which can impact prediction model performance (Coblinski et al., 2020).

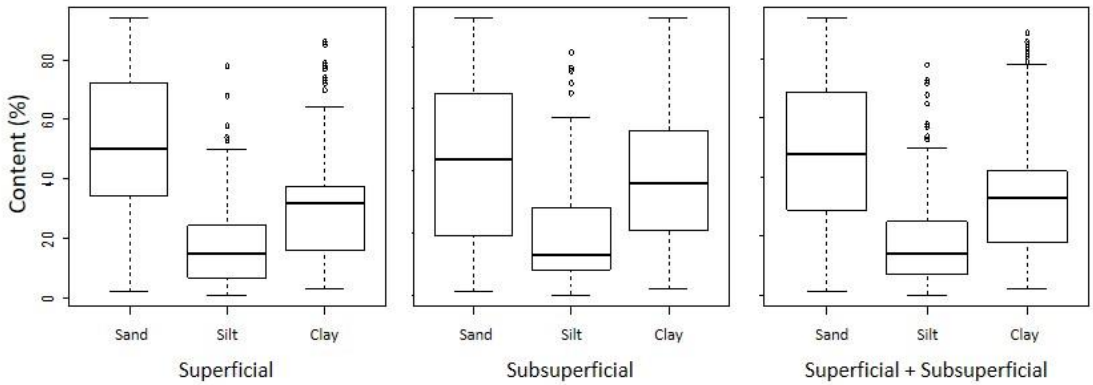

Figure 3 – Boxplots of soil texture in superficial horizons (0 – 20 cm), subsuperficial horizons (80 – 100 cm) and combined.

The two first principal components of PCA accounted for 48.7% of the total variance explained for different horizons (Figure 4). There was no clear segregation between soil horizons. Samples that are located near each other suggest a connection, but some samples are more spread out in the superficial horizon. This finding is consistent with what was observed in the soil spectra (Figure 5), which displayed corresponding spectral behavior but with differences in reflectance intensity. These similar reflectance patterns at superficial and subsuperficial horizons reinforce that the spectra of a *pedon* can be similar if the soil is uniform at depth (Zhang and Hartemink, 2020). In this respect, Figure 6 illustrates an upward trend in the reflectance of the raw spectrum between 350 and 1855 nm. However, there are exceptions, as evidenced by the pattern observed in bands A, B, and C, which correspond to absorption characteristics at wavelengths close to 1425, 1855, and 2200 nm. These bands were more pronounced in the subsuperficial horizon. These bands are associated with 2:1 clay minerals, kaolinite, and gibbsite. The raw spectra exhibited a practically increasing trend at wavelengths

from 350 to 780 nm, but upon the removal of the continuum, absorption features became evident in this range, which are associated with iron oxides, in addition to the other specified bands. Mancini et al. (2024) showed that the highest clay content in Brazilian soils presented smoother forms between 350 and 500 nm, due to the positive correlations of clay content in this spectral range. Likewise, soil organic carbon showed greater correlations at lower wavelengths (350–500 nm). Thus, pre-processing these spectra is important, precisely to highlight weak bands, as in addition to obvious spectral characteristics observed in the raw spectra, there may be a lot of imperceptible information about functional groups in the spectral curve (Song et al., 2024).

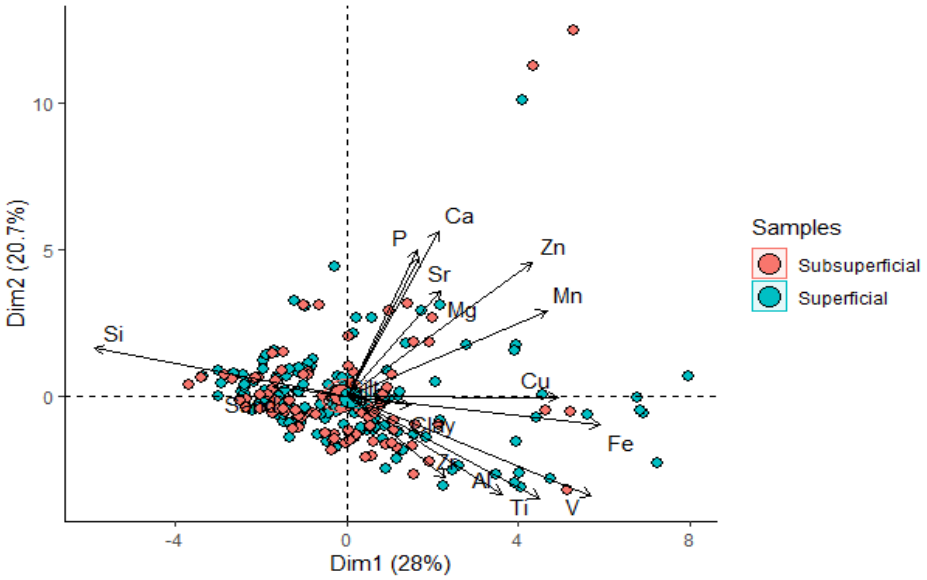

Figure 4 - Principal Component Analysis (PCA) of portable X-ray fluorescence spectrometry (pXRF) data from samples of soil superficial and subsuperficial horizons.

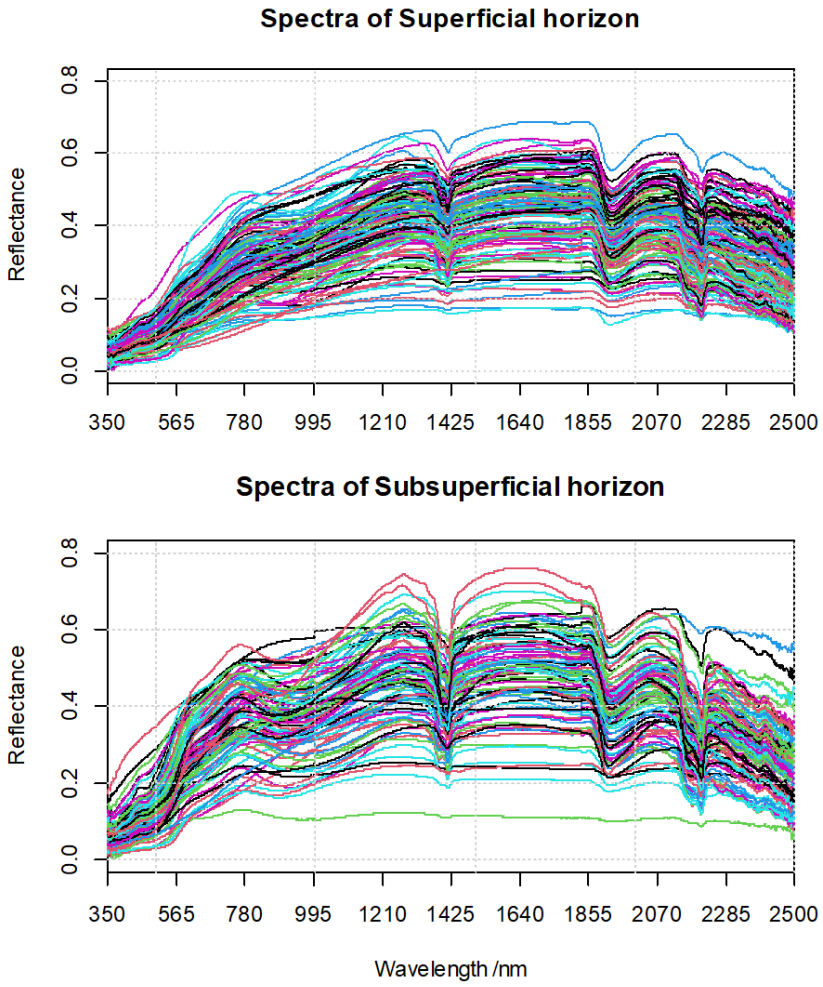

Figure 5 - Visible-Near Infrared (Vis-NIR) spectra of all samples from superficial and subsuperficial horizons of soils in the state of Pará, Brazil.

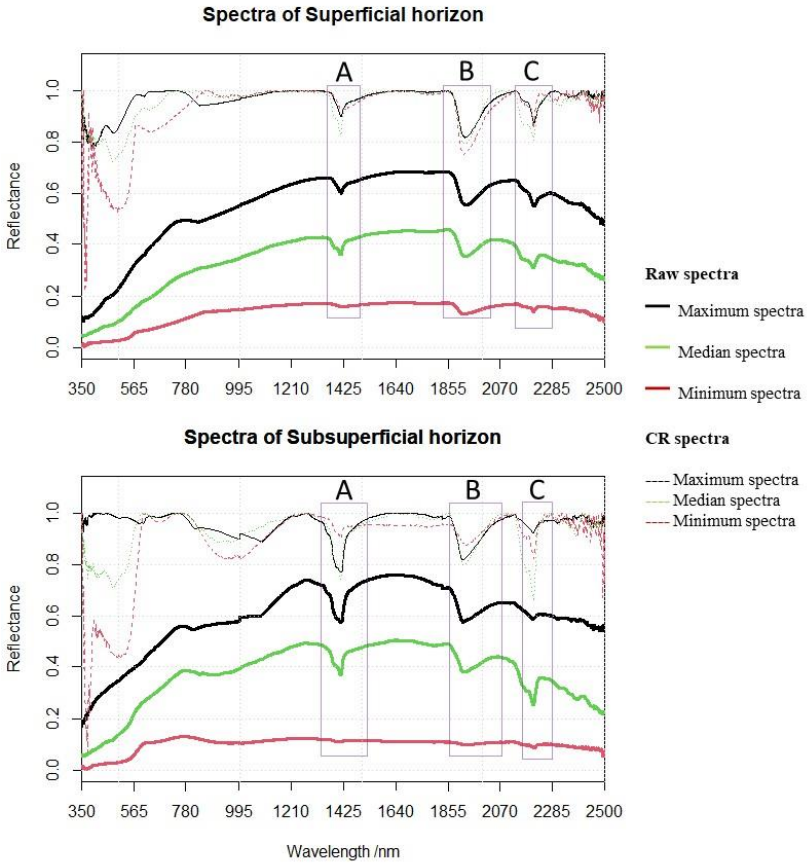

Figure 6 - Maximum, minimum, and median of raw and CR spectra from the Visible-Near Infrared (Vis-NIR) proximal sensor analyses of superficial and subsuperficial horizons of soils in the state of Pará, Brazil.

In order to utilize soil texture as a tool for environmental modeling, it is essential to measure the variation in texture at specific depths, rather than examining only the pedogenetic horizons (Dharumarajan and Hegde, 2022). The superficial horizon appeared to have more sand, which usually leads to higher reflectance values (Benedet et al., 2020a). However, it was noticed that the average reflectance spectrum of the superficial horizon samples was slightly lower than that of subsuperficial horizon (Figure 6). This phenomenon can be explained by the fact that the soil samples were taken from forested areas where trees constantly deposit organic residues on the soil surface. This deposition of organic residues darkens the color of the soil, thereby reducing the reflectance of the matrix. Besides, the subsuperficial horizon has a yellowish hue that increases reflectance when compared to the dark color of the matrix. According to Chaves et al. (2020) and Viscarra Rossel and Hicks, (2015), the amount of organic carbon in the soil is inversely proportional to its reflectance. In simpler terms, as the level of organic carbon in the soil increases, the reflectance of that soil decreases. In this context,

Fernandes et al. (2004) observed a decrease in reflectance due to the presence of organic matter in the second half of the spectrum, from approximately 550 nm onwards.

2.3.2 Modeling and prediction of soil texture through pXRF data

Table 2 presents the results of texture prediction using the RF and SVM algorithms. The RF algorithm showed better results in predicting sand, clay, and silt fractions with pXRF data compared to SVM. It had higher R^2 and RPIQ values, and lower RMSE and MAE. The clay predictive models had higher R^2 values than the sand and silt models in both horizons and the combined horizons. However, the RPIQ was only superior in the subsuperficial horizon, which had a value of 3.71. The prediction of the sand fraction had the same R^2 value of 0.89 in both the superficial and subsuperficial horizons. The RMSE was 12.20 and 12.44 for the superficial and subsuperficial horizons, respectively, while the MAE was 9.83 and 10.38. The combined horizons had RMSE of 11.56 and MAE of 9.28, with lower errors than the individual horizons. However, regarding the clay and silt, the combined horizons showed higher RPIQ in the superficial and combined horizons.

Table 2 - Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Coefficient of Determination (R^2), and Ratio of Performance to Interquartile Distance (RPIQ) for the prediction of texture in Amazonian tropical soils in superficial, subsuperficial horizons, and combined horizons using pXRF and Vis-NIR data (isolated and with combined data).

Proximal sensor	Variable	Horizon	RF				SVM			
			RMSE	MAE	R^2	RPIQ	RMSE	MAE	R^2	RPIQ
pXRF	Sand	Superficial	12.20	9.83	**0.89**	3.09	21.69	16.99	0.31	1.74
		Subsuperficial	12.44	10.38	**0.89**	3.63	21.36	16.84	0.35	2.11
		Combined	11.56	9.28	**0.89**	3.44	21.8	16.89	0.31	1.82
pXRF	Clay	Superficial	8.96	6.66	**0.92**	2.40	17.59	12.21	0.15	1.22
		Subsuperficial	8.80	6.78	**0.91**	3.71	15.06	10.85	0.58	2.17

		Combined	7.78	5.94	**0.92**	3.11	15.58	10.88	0.44	1.55
pXRF	Silt	Superficial	6.72	4.95	**0.89**	2.55	13.05	8.95	0.21	1.31
		Subsuperficial	9.10	7.02	**0.89**	2.20	17.81	11.01	0.15	1.12
		Combined	7.06	5.30	**0.91**	2.48	14.72	9.05	0.18	1.19
Vis-NIR	Sand	Superficial	9.58	7.67	**0.87**	3.94	17.27	13.42	0.55	2.18
		Subsuperficial	10.63	8.04	**0.87**	4.25	13.03	7.55	**0.74**	3.28
		Combined	10.07	7.99	**0.87**	3,95	17.24	13.20	0.56	2.38
Vis-NIR	Clay	Superficial	7.28	5.44	**0.87**	2.95	14.13	9.53	0.50	1.52
		Subsuperficial	7.61	5.53	**0.90**	4.55	14.12	9.30	0.58	2,29
		Combined	7.07	5.25	**0.89**	3.44	14.11	10.44	0.52	1.70
Vis-NIR	Silt	Superficial	9.42	6.58	0.55	1.81	12.51	7.86	0.25	1.37
		Subsuperficial	10.99	7.77	0.67	1.82	17.16	11.49	0.18	1.16
		Combined	10.50	7.54	0.56	1.66	14.12	9.56	0.23	1.22
pXRF + Vis-NIR	Sand	Superficial	7.79	6.11	**0.93**	4.69	16.25	12.14	0.61	2.30
		Subsuperficial	10.10	7.87	**0.90**	4.55	20.48	16.11	0.40	2.20
		Combined	9.06	7.28	**0.91**	4.51	17.57	13.24	0.54	2.25
pXRF + Vis-NIR	Clay	Superficial	5.58	4.09	**0.93**	3.86	10.79	6.71	**0.72**	1.98
		Subsuperficial	6.59	4.77	**0.93**	4.77	12.07	7.91	**0.70**	2.70
		Combined	5.97	4.47	**0.94**	4.13	10.93	6.74	**0.71**	2.03
pXRF + Vis-NIR	Silt	Superficial	5.72	4.12	**0.92**	2,92	12.49	8.32	0.26	1.42
		Subsuperficial	8.41	6.30	**0.91**	2.44	16.72	10.54	0.20	1.13
		Combined	6.73	4.73	**0.92**	2.67	14.80	9.80	0.17	1.18

$R^2 \geq 0.70$ values are given in bold

The results for clay demonstrated R^2 value of 0.92 for the superficial horizon with RMSE of 8.96 and MAE of 6.66. The combined horizons also showed the same R^2 value of 0.92 with lower errors of RMSE 7.78 and MAE 5.94. However, the accuracy of the subsuperficial horizon had better results with a combination of R^2 value of 0.91, RMSE of 8.80, and MAE of 6.78. Additionally, the subsuperficial horizon had a higher RPIQ value of 3.71. The silt prediction for both the superficial and subsuperficial horizons showed an R^2 value of 0.89, with higher RPIQ in the superficial horizon (2.55). The combined horizons prediction revealed an R^2 value of 0.91, with RMSE and MAE values of 7.06 and 5.30, respectively. The superficial horizon had RMSE of 6.72 and MAE of 4.95, while the subsuperficial horizon had an RMSE of 9.10 and MAE of 7.02. The SVM texture prediction had lower R^2 and RPIQ values compared to the RF prediction. The best performance obtained using SVM was shown when predicting clay in the subsuperficial horizon. The model had an R^2 value of 0.58, an RMSE value of 15.06, an MAE value of 10.85, and an RPIQ value of 2.17. However, when predicting soil fractions using pXRF data alone, the RF model was found to be more efficient than SVM. This was because the RF model showed the highest R^2 and RPIQ values with lower RMSE and MAE values. Furthermore, the combination of horizons was found to be more important for predicting sand and clay, which showed the lowest errors when the best model was chosen based on the highest values of R^2 and RPIQ combined with lower values of RMSE.

Silva et al. (2020) indicated that predictive models using pXRF data and robust algorithms can accurately determine the soil texture of tropical soils, regardless of their class, parent material, and degree of weathering. Additionally, among the various algorithms tested, SVM showed a slight advantage in terms of performance. Benedet et al. (2020a) demonstrated that pXRF data alone can provide robust predictions for soil texture, although relevant information and adequate results can also be obtained with Vis-NIR data. Zhang and Hartemink (2020) found that the validation achieved for soil texture prediction with pXRF data was above 0.80, which was higher than the results obtained with Vis-NIR spectral data. However, in the present research, sand, and clay predictions with pXRF data were less accurate (higher RMSE and MAE; and lowest R^2 and RPIQ) compared to predictions with Vis-NIR data. Only the silt prediction was better using only pXRF data.

Many researchers use the R^2 value to determine the accuracy of prediction models. However, this statistical parameter can be misleading as it is influenced not only by the explained variance but also by the variance of the dataset. It is easier to achieve a high R^2 when the dataset is more variable. Although higher R^2 values indicate better predictions, lower RMSE

values are a more accurate measure of the model's performance. RMSE represents the error of the model and should be considered alongside R^2 when evaluating the quality of prediction models (Ahmadi et al., 2021).

2.3.3 Modeling and prediction of texture using Vis-NIR data

RF models for sand, clay, and silt prediction showed higher R^2, and RPIQ values compared to SVM models using Vis-NIR data (Table 2). Similar to the results obtained with pXRF, the clay fraction prediction models with Vis-NIR data generally showed higher R^2, and RPIQ values and lower errors compared to sand and silt, regardless of the approach used to assess the horizon. The RF model for sand prediction had a common R^2 (0.87) in the superficial horizon (RMSE = 9.58; MAE = 7.67, RPIQ = 3.94), subsuperficial (RMSE = 10.63; MAE = 8.04, RPIQ = 4.25), and in the combined horizons (RMSE = 10.07; MAE = 7.99, RIPQ = 3.95). The prediction of clay content using the RF algorithm showed varying R^2 values for the different soil horizons. For the superficial horizon, the R^2 value was 0.87, while for the subsuperficial horizon it was 0.90. The combined horizons demonstrated an R^2 value of 0.89. However, the combination of horizons resulted in the lowest errors, with an RMSE of 7.07 and an MAE of 5.25. Additionally, the RPIQ was high at 3.44. The RF prediction for silt showed the lowest R^2 and RPIQ values of 0.67 in the subsuperficial horizon (RMSE = 10.99, MAE = 7.77) and 0.55 in the superficial horizon (RMSE = 9.42; MAE = 6.58). Silt predictions using Vis-NIR data were inferior to silt predictions by pXRF data. Predictions of the silt fraction are often less precise than those for clay (Viscarra Rossel et al., 2011). Using Vis-NIR data, the combination of soil horizons was relevant only for clay, reducing errors.

SVM models with Vis-NIR data showed better results compared to pXRF data. Sand reached the highest R^2 (0.74) in the subsuperficial horizon (RMSE = 13.03; MAE = 7.55; RPIQ = 3.28), followed by clay with R^2 (0.58), also in the subsuperficial (RMSE = 14.12; MAE = 9.30; RPIQ = 2.29). The model for silt presented lower R^2 values with SVM, ranging from 0.18 to 0.25. Overall, the RF model had higher R^2 values for sand, clay, and silt using pXRF data compared to data obtained with Vis-NIR, but the predictions by Vis-NIR present higher RPIQ for sand and clay. As observed for pXRF data, the RF model with Vis-NIR data was more efficient than SVM in predicting soil texture. This was evidenced by the higher R^2, RPIQ values, and lower RMSE and MAE values. Andrade et al. (2020) achieved well-performing

models for predicting soil particle size fractions, especially for sand and clay, using proximal sensors.

The number of studies using the Vis-NIR spectrometer for predicting soil attributes has increased considerably in the last decade (Silva et al., 2021). Nevertheless, it is essential to conduct more studies to demonstrate the potential of this technique as an alternative to the conventional soil texture analysis method in different environmental conditions (Ahmadi et al., 2021).

2.3.4 Modeling and prediction of texture using the combination of pXRF and Vis-NIR data

RF models for sand, clay, and silt prediction showed higher R^2, and RPIQ values and lower errors compared to SVM models using data from the combined sensors (Table 2). RF predictions for the sand fraction had similar R^2 values in the superficial horizons (R^2 = 0.93), subsuperficial horizons (R^2= 0.90) and in the combined horizons (R^2 =0.91) but the superficial horizon presented the lowest errors and higher RPIQ values. With RF, the clay content prediction model achieved higher R^2, and RPIQ values, and the lowest error values than the other fractions (sand, and silt). In the superficial (RMSE = 5.58; MAE = 4.09) and subsuperficial (RMSE = 6.59; MAE = 4.77) horizons, RF prediction of clay content had a common R^2 (0.93), while in the combined horizons, R^2 was 0.94 (RMSE = 5.97; MAE = 4.47). With RF, silt prediction had the same R^2 (0.92) in the superficial horizon (RMSE = 5.72; MAE = 4.12; RPIQ = 2.92) and in the combined horizons (RMSE = 6.73; MAE = 4.73; RPIQ = 2.67), whereas in the subsuperficial horizon, it had a lower R^2 (0.91) (RMSE = 8.41; MAE = 6.30; RPIQ = 2.44). It is important to highlight that sensor fusion provided better results for silt prediction compared to predictions using pXRF data. The combination of soil horizons was not relevant to reducing errors in any fraction, either with RF or SVM.

SVM models using data resulting from the combination of sensors showed better results in clay prediction compared to separate sensors. R^2 values were different, with the best result in the superficial horizon (R^2 = 0.72; RMSE = 10.79; MAE = 6.71; RPIQ = 1.98), followed by the combination of horizons (R^2 = 0.71; RMSE = 10.93; MAE = 6.74; RPIQ = 2.03) and subsuperficial horizon (R^2 = 0.70; RMSE = 12.07; MAE = 7.91; RPIQ = 2.70). Sand in the superficial horizon showed better R^2 (0.61) and lower errors (RMSE = 16.25; MAE = 12.14; RPIQ = 2.30) with SVM compared to separate sensors. Silt prediction models had the lowest

R^2 values. It is noteworthy that the combination of sensors obtained better models in clay prediction compared to separate sensors. Once again, the RF model was more efficient than SVM in predicting soil fractions, showing higher R^2 and RPIQ values and lower RMSE and MAE values. Due to the Law of Large Numbers, RF does not suffer from overfitting, being an effective prediction tool. Injecting the right kind of randomness makes it an accurate regressor (Breiman, 2001).

With the combination of sensors, better prediction results were obtained compared to the best individual sensor (Vis-NIR). Overall, with better adjustments for clay (Figure 7). In this study, the combination of sensors data was more accurate, especially in terms of lower model errors and the substantial improvement in clay prediction in the SVM algorithm. For O'Rourke et al. (2016), the synergistic use of Vis-NIR and pXRF spectra could increase the predictive capacity of soil properties beyond that using any sensor alone.

According to Benedet et al. (2020a), in general, models with pXRF and pXRF + Vis-NIR data provide slightly better results compared to predictions using only Vis-NIR data. However, the combined use of sensors contributes little or nothing to prediction accuracy when compared to pXRF data alone. The study by Zhang and Hartemink (2020) revealed that both pXRF and Vis-NIR provide good soil texture prediction (sand, silt, and clay) over a large particle size distribution, but the fusion of proximal sensor data has been used to improve prediction accuracy. Andrade et al. (2022), using different proximal sensors and data preprocessing methods, achieved similar results, but showed that pXRF data were more important for soil texture prediction, which suggests that the success of combining proximal sensor data may be driven by the dataset.

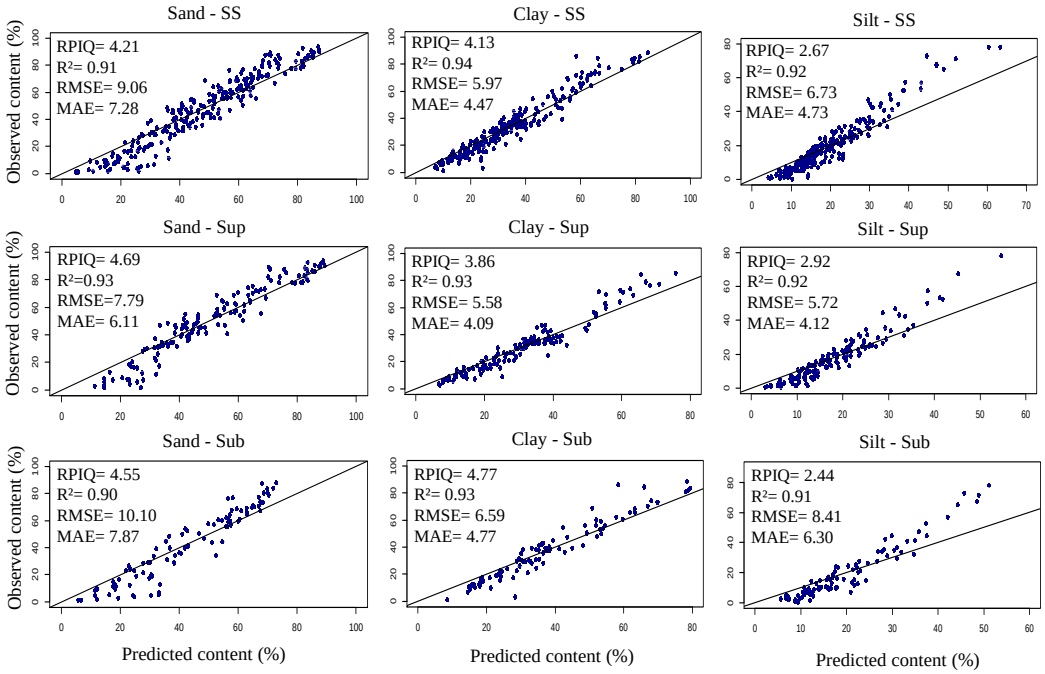

Figure 7 - Performance of models for clay, silt, and sand contents obtained from conventional analyses and contents predicted by the RF algorithm, using data from the sensor fusion of portable X-ray fluorescence (pXRF) and visible and near-infrared spectrosmetry (Vis-NIR) analysis. SS - Superficial and Subsuperficial Horizons, Sup - Superficial Horizon, and Sub - Subsuperficial Horizon.

2.3.5 Importance of pXRF and Vis-NIR variables

To determine the importance of variables in predicting sand, silt, and clay in the soil, only the results obtained from the RF algorithm were considered, as it produced the most accurate predictions. The analysis of the pXRF data (Figure 8) revealed that, for sand, the most important element in the combined horizons was Fe, followed by Mn and Ti. In the superficial horizon, the most important elements were Fe, V, and Ti. In the subsuperficial horizon, the most important elements were Zr, Fe, and Ti, followed by Mn.

Variable	Sand (SS)	Clay (SS)	Silt (SS)	Sand (Sup)	Clay (Sup)	Silt (Sup)	Sand (Sub)	Clay (Sub)	Silt (Sub)
Fe	100	100	100	100	100	70	83	100	x
Ti	44	61	x	44	57	x	72	34	x
Sr	x	14	x	x	x	x	x	28	x
V	36	x	x	59	51	x	70	19	x
Mn	58	7	x	41	37	x	72	15	x
Si	x	15	68	38	49	100	x	13	10
Al	x	3	90	x	43	53	x	x	1
Mg	x	4	x	x	30	x	x	x	x
Zr	8	x	28	x	x	x	100	x	100
Cu	x	x	x	x	x	86	x	x	x

Figure 8 - Importance of variables in the model generated with pXRF data. Values in percentage (%). SS - Combined horizons, Sup - Superficial Horizon, and Sub - Subsuperficial Horizon.

The Fe was found to be the most important variable in predicting clay and was followed by Ti and Si (combined horizon); Ti and V (superficial horizon), and Ti and Sr (subsuperficial horizon). For silt in the combined horizons, the important elements were Fe, Al, and Si. In the superficial horizon, the important elements were Si, Cu, and Fe. In the subsuperficial horizon, only Zr and Si were important.

The importance of Vis-NIR variables is presented in Figure 9. In this study, generally, the important spectra for sand prediction were the range of 635 and 711 nm and in the range 2390 to 2475 nm. For clay prediction, it was around 635 and 708 nm and between 2475 and 2478 nm. Silt had spectra in the range of 356 and 634 nm. These wavelengths provide understanding of soil mineralogy (Demattê et al., 2021). For example, wavelengths in the range of 635 and 711 nm, observed for all fractions, can be associated with iron oxides, indicating the presence of hematite and goethite, minerals that give red and yellow coloration to soils (Pereira et al., 2020). In tropical soil samples, absorption features were identified at 425 nm, 480 nm, 513 nm, 650 nm, 903 nm, and 1000 nm due to the microscopic interaction between iron oxides and electromagnetic radiation (Terra et al., 2018).

Variable	Sand (SS)	Clay (SS)	Silt (SS)	Sand (Sup)	Clay (Sup)	Silt (Sup)	Sand (Sub)	Clay (Sub)	Silt (Sub)
Wave2474	100	x	x	46	x	x	65	x	x
Wave2390	90	x	x	76	x	x	64	x	x
Wave711	86	x	x	1	x	x	44	x	x
Wave2417	82	x	x	64	x	x	70	x	x
Wave649	81	x	x	36	x	x	23	x	x
Wave635	81	7	x	67	20	x	67	13	x
Wave2475	79	100	x	69	100	x	33	100	x
Wave660	74	x	x	0	x	x	100	0	x
Wave2418	73	x	x	100	x	x	73	0	x
Wave650	72	x	x	35	x	x	20	0	x
Wave699	63	x	x	31	x	x	24	0	x
Wave2477	54	x	x	56	x	x	x	0	x
Wave2439	34	x	x	7	x	x	100	0	x
Wave678	x	x	x	15	x	x	45	0	x
Wave2498	x	31	x	x	51	x	x	89	x
Wave708	x	4	x	x	3	x	x	7	x
Wave707	x	x	x	x	x	x	x	x	x
Wave634	x	x	100	x	x	100	x	x	x
Wave356	x	x	x	x	x	x	x	x	100

Figure 9 - Importance of variables in the model generated with Vis-NIR data. Values in percentage (%). SS - Superficial and Subsuperficial Horizons, Sup - Superficial Horizon, and Sub - Subsuperficial Horizon.

Brazilian soils' clay fraction is mainly composed of minerals like kaolinite, hematite, goethite, and gibbsite (Silva et al., 2020). Despite hematite and goethite being the most widespread iron oxides in tropical soils, there is also the occurrence of magnetite and maghemite in *pedons* derived from Fe-rich rocks (Fabris et al., 1998, 1997). Fe^{3+} ions found in Amazonian soil may originate from goethite and/or hematite minerals, and can be retained on the surface of kaolinite. This retention may lead to the formation of gibbsite through adsorption and/or isomorphic substitution in the octahedral layer of the kaolinite structure (Couceiro and Santana, 1999).

The importance of Zr can be justified by the presence of zircon, and Ti by titanite, ilmenite, rutile, and anatase. The Amazon has heavy minerals like rutile, anatase, and zircon despite the prevalence of quartz and kaolinite (Fritsch et al., 2002). Zr showed a positive correlation with sand in tropical soils (Silva et al., 2021). The zircon ($ZrSiO_4$) and titanite are minerals highly resistant to weathering (through chemical and physical processes) and found in Brazilian soils. These tetravalent cations crystallize in silicate, phosphate, and oxide minerals (Marques et al., 2004; Wilford, 2012). The importance of Ti is associated with the mineral ilmenite ($FeTiO_3$) as it is usually an accessory mineral to gabbro, and in Fe oxides Ti can replace

Fe, especially in soils originating from mafic rocks (Fabris et al., 1998, 1997; Poggere et al., 2018).

The results obtained through the use of pXRF data, coupled with the Vis-NIR ranges, provide valuable insights into the types of minerals present in the analyzed samples. In the state of Pará, the most common rock types are sedimentary rocks such as shales, siltstones, and sandstones, while igneous rocks like basalt and diabase are found to a lesser extent (De Souza et al., 2018). Other studies support the findings of the present work. For instance, the prediction of soil texture in three Brazilian states using pXRF data showed that Fe, Si, and Zn were the most important variables for predicting sand and clay content (Benedet et al., 2020a).

The presence of V can be attributed to weathering in the region. It shows a strong correlation with the clay content of soils due to its ability to be incorporated into octahedral sites in the structure of clay minerals such as kaolinite, gibbsite, hematite, and goethite. (Marques et al., 2004). The importance of Mn can be attributed to its presence in oxide minerals. The composition of oxide soils in the state of Pará indicates that titanium (Ti) and Mn are constituents of these soils (De Souza et al., 2018). Furthermore, the sampling area encompassed municipalities in the Transamazon region, which exhibited rocks predominantly composed of quartz, feldspar, muscovite, and chlorite, with additional minerals such as calcite, tourmaline, and opaque minerals. These minerals are decomposed, stained by iron and manganese oxides, and exhibit a greenish hue (Falesi, 1974).

Just like Fe, Al, Ti, V, Zr, Mn, and Sr were important elements in this study, revealing the chemical composition of forest-covered soil samples, de Souza et al. (2018), evaluating Amazonian soils, found high average values of Al, Co, Cr, Fe, Hg, Pb, Ti, V, Zr, Ca, K, Mg, Mn, Na, Ni, P, Sr, and Zn. The heterogeneity of variables important for predicting the fractions reveals the variability of the studied soils, which certainly reflects different parent materials, topography, vegetation, and natural pedogenetic processes.

In the study by Andrade et al. (2020) on Brazilian tropical soils, Fe and Si were more relevant for sand and clay prediction models, while for silt, Cu (in both horizons) and Mn (for horizon B) were more important. Cu was only significant for silt prediction, as observed in this study. The importance of Cu in silt modeling can be explained by isomorphic substitution of Cu in iron oxides with magnetite, a mineral commonly found in soils derived from gabbro and other mafic rocks (Lu et al., 2009; Santos et al., 2016). The importance of Si and Al can be associated with the presence of muscovite ($KAl_2(AlSi_3O_{10})(F,OH)_2$) (Andrade et al., 2020), especially in poorly drained conditions. The weathering of muscovite leads to the slow and

continuous release of Si and the redistribution of Al across various mineral and organic phases (Fritsch et al., 2007).

The importance of Sr for predicting the clay fraction in the subsuperficial is likely due to climatic and geological factors. Strontium is the 15th most abundant element in the earth's crust (Dayani et al., 2021) and the two principal Sr minerals occurring naturally are strontianite ($SrCO_3$) and celestite ($SrSO_4$) (Martińez and Uribe, 1995). These elements can be released into soil by natural weathering and by mining activities in the region (if any) (Kaleem et al., 2021). Sr is released into the soil solution and can be retained in clay minerals or organic matter. O'Rourke et al. (2016) found a negative relationship between Sr and the sand fraction, indirectly reinforcing the possible existence of a positive interaction between clay content and Sr. The presence of Sr in the subsuperficial horizon also occurs through the flow of Sr deposited with rainfall and leached from the litter to the first 50 cm of soil. Thus, both rainfall and weathering can create a depth gradient (Poszwa et al., 2002). As leaching is common in tropical soils, the presence of this element is justified.

The bands located around 2200–2400 nm are associated with the presence of 1:1 and 2:1 minerals (Ben-Dor, 2002; Demattê et al., 2015). Terra et al. (2018) reported absorption in the range of 2355 and 2448 nm due to Al-OH in the illite mica and absorption resulting from organic compounds at 2316 nm and 2382 nm. Felix et al. (2016), studying samples from different soil layers developed from basalt, found absorption peaks in the range between 2160-2300 nm and in the region around 2400 nm. These wavelengths were associated with the presence of carbon-oxygen bonds (O-H linked to kaolinite) (Netto and Baptista, 2000).

The spectral bands located around 2439–2475 nm are associated with the presence of organic matter. This importance is understandable given that the sampled study areas are primary and secondary forests. According to Inda et al. (2014), organic carbon levels are higher in soil under forests. Wavelengths near 2488 nm were linked to overtones of C-H, N-H, and R-OH molecules (Xiaobo et al., 2010), and overtones and combination vibrations of functional organic groups of cellulose were observed near wavelengths of 2304 to 2488 nm (Viscarra Rossel and Hicks, 2015).

2.4 Limitations and strengths of methods for soil texture determination and future studies

Both wet-chemistry and proximal sensor-based methods present limitations and strengths, which require some considerations in order to decide which one to adopt. First, traditional texture analysis, based on wet-chemistry, is considered the reference method for soil texture determination and it is well established worldwide, allowing for comparison of results among different laboratories and research groups. Variations of the methods exist (e.g., pipette, hydrometer, etc.,) and it requires several equipment during the different phases of the analysis (Gee and Or, 2018). Also, it may take several days, given the need for a long time for dispersion of clay particles, especially for some tropical soils and, depending on the method, the separation of silt and clay fractions. There are some alternatives involving different equipment, such as laser light scattering and S-ray absorption, which may provide a faster estimate of clay content, although these instruments are expensive, limiting their application in many laboratories (Hobley and Prater, 2019). In other cases, such as the sedimentation and laser techniques, several sample pretreatments are mandatory (peroxide, sodium hypochlorite, sodium hexametaphosphate, sonication intensity) (Parent et al., 2021). In addition to the time needed, effluents are produced in such analysis, which requires treatments prior to disposal. It is important to mention that, in some regions of the world, especially in developing countries, there is a scarcity of laboratories capable of conducting such analysis. As an example, in most of the Brazilian Amazon region, there are serious limitations related to laboratory infrastructure for soil analysis, and very few laboratories are capable of conducting soil texture analysis (Embrapa, 2023, 2020). Finally, the costs of the analyses per sample in Brazilian laboratories, without including shipping, is around USD4.00/sample. In this study, the total amount of expenses on soil texture by conventional laboratory analysis was USD932.00, which may correspond to a large amount of money considering the country currency.

Conversely, proximal sensors have been increasingly adopted for soil texture predictions given the successful results achieved worldwide recently (Zhu et al., 2011, Silva et al., 2020; Gozukara et al., 2022). The advantages they provide include minimal sample preparation, no chemical reagent needed, delivery of a large number of results in a few seconds, which allows for the analysis of hundreds of samples per day, and application of the results for multiple purposes, e.g., direct characterization of samples and calibration of models to predict other related properties, such as soil texture in case of soil samples (Nocita et al., 2015;

O'Rourke and Holden, 2011). However, drawbacks involve the high acquisition cost of such sensors, although it can be gradually reduced over time, the need for calibration models either to predict other properties or to correlate their results with other reference methods. Moreover, since such proximal sensors approach is relatively recent, especially for pXRF, there is still a need for standardization of analytical procedures to make results obtained from different equipment comparable, which is still being investigated (Silva et al., 2021). Thus, efforts are still required to make such an approach applicable to multiple soil conditions. In this aspect, this study contributes to demonstrate the capability of proximal sensors data to predict soil texture for the Amazonian region, which has not been explored yet, providing the possibility to assess if such an approach can be reliably used in such conditions.

Specifically regarding pXRF and Vis-NIR, the combination of data from pXRF and Vis-NIR, due to the large number of wavelengths in which reflectance is determined in Vis-NIR analysis, a large number of variables is generated (> 2,000 variables). This requires longer time needed for data processing than for pXRF, which delivers fewer variables (up to 45 elemental contents). Another difference is that Vis-NIR data can be submitted to several preprocessing techniques, compared with pXRF results more directly interpretable (elemental contents present in the sample) and relatable to soil attributes (Benedet et al., 2020b; Dasgupta et al., 2022). In addition, pXRF is less expensive than Vis-NIR. The choice of the sensor to be used needs evaluation on the type of data of interest (elemental contents or reflectance in the Vis-NIR range).

Considering the application of Vis–NIR and pXRF for soil properties prediction, both sensors have become attractive due to the positive results. However, the decision to use the combination of both sensors' data for the prediction of soil texture (and other properties), must be based not only on speed, ease, and cost, but also on the accuracy of the analysis results. In this sense, combination of both sensors' data, although potentially more time-consuming and costly, can offer significantly greater accuracy (Wan et al., 2020), which was also observed for Amazonian soils, justifying the evaluation of the devices both separately and together. The findings of this current study contrasted with other reports on pXRF and Vis-NIR capacity to predict texture of Brazilian soils from other regions, when pXRF outperformed Vis-NIR and the combination of sensors data was not needed to achieve the highest accuracies (Benedet et al., 2020a). It demonstrates the need for regional tests to determine the best approach to be applied at different locations. Obviously, if only one sensor meets the needs for a particular region or purpose, costs for these analyses are minimized, especially in countries with limited

funds available for sensors acquisition. It is important to notice that predictions based on Vis-NIR data might be influenced by the different pretreatments of the spectral data, which can change the accuracy of the predictions (Benedet et al., 2020a, Mancini et al., 2022; Teixeira et al., 2022). Thus, we encourage further studies evaluating this aspect.

Besides the efficiency of these two sensors, the algorithm used for predictions and the input data (explanatory variables) also largely influences the results, as widely investigated (Parent et al., 2021; Benedet et al., 2020a; Mancini et al., 2022). Other studies suggest that the generated prediction models can be improved by incorporating other sensors data (Aitkenhead et al., 2013), such as a color sensor (Andrade et al., 2022), and other information, as the soil classification and soil parent material. For example, according to Andrade et al. (2022), the separation of data sets by soil class or adding parent material information had a positive impact on soil texture predictions. Also, development of regional/local models (encompassing a small region and, thus, lower variability of soils) versus more general models (containing more heterogeneous soils) have also affected models' performance (Faria et al. 2020), encouraging such regional tests, as performed herein for Amazon region and that indicated a different behavior than that found for other tropical soils (Benedet et al., 2020a). Finally, the soil samples used in this study were from a wide variety of soils from Amazon region. The good results indicate the possibility of using this proximal sensor approach for texture prediction of Amazonian soils (Nocita et al., 2015).

2.5 Conclusion

The soil texture of the Brazilian Amazon region studied herein is mainly loamy, followed by sandy. The data from proximal sensors (pXRF and Vis-NIR), along with the RF algorithm, proved effective in predicting the texture of tropical soils in the Amazon region. Among the soil fractions, clay had the most accurate predictions (i.e. highest values of R^2 and RPIQ and lowest values of RMSE). Predictions only based on Vis-NIR data outperformed those ones using only pXRF data for sand and clay fractions, while pXRF delivered the best predictions for the silt fraction (lowest RMSE, and higher R^2 and RPIQ).

Regarding the effect of soil horizon on the predictions, texture predictions for the superficial horizons were more accurate than those for the subsuperficial horizon. However, the combination of data from both horizons did not deliver the best predictions. Thus, the best predictions considering all the possible variations of datasets (i.e., proximal sensors and

horizons combined or separate, and algorithms) were achieved using sensor fusion (pXRF + Vis-NIR) for clay in the superficial horizon (R^2 = 0.93; RPIQ = 3.86; and RMSE = 5.58); for silt in the superficial horizon (R^2 = 0.92; RPIQ = 2.92; and RMSE = 5.72); and for sand in the superficial horizon (R^2 = 0.93; RPIQ = 4.69; and RMSE = 7.79). Considering the separate sensors data, the prediction results using only Vis-NIR data were reasonable for the three particle size fractions, except for silt predictions, in the soil samples from combined and separate horizons.

The variables generated by pXRF and Vis-NIR that reflect the mineralogy of the soils were ranked as the most important ones for the prediction models. For instance, Fe was the most important variable for predicting sand and clay, reflecting the mineralogy of Pará State soils, which are rich in iron oxides. Similarly, Ti was the second most important variable for sand and clay predictions. These elements are commonly present in minerals such as magnetite, maghemite, titanite, ilmenite, rutile, and anatase, commonly found in the region. Other important elements for all fractions were Si, Al, V, Cu and Zr, corresponding to a mineralogical result of a region with sedimentary rocks and, to a lesser extent, igneous rocks such as basalt and diabase.

Finally, studies on the cost/benefit ratio and improvement in prediction accuracy of working with both Vis-NIR and pXRF sensors are encouraged, since their efficiency tend to be site-specific. In addition, further research on this approach using other sensors and additional variables, such as soil classification and parent material, is strongly encouraged to improve the prediction of soil texture in other tropical regions.

Acknowledgment: The authors would like to thank the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior for the scholarship (CAPES - 88887.598616/2021-00), Conselho Nacional de Desenvolvimento Científico e Tecnológico for provide research grants (CNPq # 310283/2019-1 P.W.M.S-F).

Data availability statement: Data are available upon request to the corresponding author.

Conflict of Interest: The authors declare that they have no conflict of interest.

REFERENCES

Ahmadi, A., Emami, M., Daccache, A., He, L., 2021. Soil properties prediction for

precision agriculture using visible and near-infrared spectroscopy: A systematic review and meta-analysis. Agronomy. https://doi.org/10.3390/agronomy11030433.

Aitkenhead, M.J., Coull, M., Towers, W., Hudson, G., Black, H.I.J., 2013. Prediction of soil characteristics and colour using data from the National Soils Inventory of Scotland. Geoderma 200–201. https://doi.org/10.1016/j.geoderma.2013.02.013.

Alvares, C.A., Stape, J.L., Sentelhas, P.C., De Moraes Gonçalves, J.L., Sparovek, G., 2013. Köppen's climate classification map for Brazil. Meteorol. Zeitschrift 22. https://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.

Andrade, R., Silva, S.H.G., Faria, W.M., Poggere, G.C., Barbosa, J.Z., Guilherme, L.R.G., Curi, N., 2020. Proximal sensing applied to soil texture prediction and mapping in Brazil. Geoderma Reg. 23. https://doi.org/10.1016/j.geodrs.2020.e00321.

Azcarate, S.M., Ríos-Reina, R., Amigo, J.M., Goicoechea, H.C., 2021. Data handling in data fusion: Methodologies and applications. TrAC - Trends Anal. Chem. https://doi.org/10.1016/j.trac.2021.116355.

Bellon-Maurel, V., Fernandez-Ahumada, E., Palagos, B., Roger, J.M., McBratney, A., 2010. Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy. TrAC - Trends Anal. Chem. https://doi.org/10.1016/j.trac.2010.05.006.

Ben-Dor, E., 2002. Quantitative remote sensing of soil properties. Adv. Agron. https://doi.org/10.1016/s0065-2113(02)75005-0.

Ben Dor, E., Ong, C., Lau, I.C., 2015. Reflectance measurements of soils in the

laboratory: Standards and protocols. Geoderma 245–246. https://doi.org/10.1016/j.geoderma.2015.01.002.

Benedet, L., Faria, W.M., Silva, S.H.G., Mancini, M., Demattê, J.A.M., Guilherme, L.R.G., Curi, N., 2020a. Soil texture prediction using portable X-ray fluorescence spectrometry and visible near-infrared diffuse reflectance spectroscopy. Geoderma 376. https://doi.org/10.1016/j.geoderma.2020.114553.

Benedet, L., Faria, W.M., Silva, S.H.G., Mancini, M., Guilherme, L.R.G., Demattê, J.A.M., Curi, N., 2020b. Soil subgroup prediction via portable X-ray fluorescence and visible near-infrared spectroscopy. Geoderma 365. https://doi.org/10.1016/j.geoderma.2020.114212.

Bezdicek, D.F., Papendick, R.I., Lal, R., 2015. Introduction: Importance of soil quality to health and sustainable land management. Methods Assess. Soil Qual. https://doi.org/10.2136/sssaspecpub49.introduction.

Birani, S.M., Fernandes, A.R., de Souza Braz, A.M., Pedroso, A.J.S., Alleoni, L.R.F., 2015. Available contents of potentially toxic elements in soils from the Eastern Amazon. Chemie der Erde 75. https://doi.org/10.1016/j.chemer.2015.01.001.

Brasil, E.C., Cravo, M. da S., Viegas, I. de J.M., 2020. Recomendações de calagem e adubação para o estado do Pará., 2nd. ed. rev. ed. Embrapa, Brasília, DF.

Breiman, L., 2001. Random forests. Mach. Learn. 45. https://doi.org/10.1023/A:1010933404324.

Carpenter, D.N., Bockheim, J.G., Reich, P.F., 2014. Soils of temperate rainforests of the North American Pacific Coast. Geoderma 230–231. https://doi.org/10.1016/j.geoderma.2014.04.023.

Chaves, S.F. da S., Gama, M.A.P., Alves, R.M., de Oliveira, R.P., Pedroza Neto, J.L., Lima, V.M.N., 2020. Evaluation of physicochemical attributes of a yellow latosol under agroforestry system as compared to secondary forest in the Eastern Amazon. Agrofor. Syst. 94.

https://doi.org/10.1007/s10457-020-00513-6.

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

Coblinski, J.A., Giasson, É., Demattê, J.A.M., Dotto, A.C., Costa, J.J.F., Vašát, R., 2020. Prediction of soil texture classes through different wavelength regions of reflectance spectroscopy at various soil depths. Catena 189. https://doi.org/10.1016/j.catena.2020.104485.

Cortes, C., Vapnik, V., 1995. Support-Vector Networks. Mach. Learn. 20. https://doi.org/10.1023/A:1022627411411.

Couceiro, P.R. da C., Santana, G.P., 1999. Caulinita em solo da Amazônia: caracterização e permutabilidade. Acta Amaz. 29. https://doi.org/10.1590/1809-43921999292275.

Dayani, N., AmirArjmand, A., Nouri-Khezrabad, M., Hasani, S., 2021. High-efficiency mechanochemical synthesis of Strontium carbonate nanopowder from Celestite. Int. J. Appl. Ceram. Technol. 18. https://doi.org/10.1111/ijac.13612.

Dasgupta, S., Chakraborty, S., Weindorf, D.C., Li, B., Silva, S.H.G., Bhattacharyya, K., 2022. Influence of auxiliary soil variables to improve PXRF-based soil fertility evaluation in India. Geoderma Reg. 30. https://doi.org/10.1016/j.geodrs.2022.e00557.

De Souza, E.S., Fernandes, A.R., De Souza Braz, A.M., De Oliveira, F.J., Alleoni, L.R.F., Campos, M.C.C., 2018. Physical, chemical, and mineralogical attributes of a representative group of soils from the eastern Amazon region in Brazil. SOIL 4. https://doi.org/10.5194/soil-4-195-2018.

de Souza, E.S., Fernandes, A.R., de Souza Braz, A.M., Sabino, L.L.L., Alleoni, L.R.F., 2015. Potentially toxic elements (PTEs) in soils from the surroundings of the Trans-Amazonian Highway, Brazil. Environ. Monit. Assess. 187. https://doi.org/10.1007/s10661-014-4074-1.

de Souza, J.J.L.L., Fontes, M.P.F., Gilkes, R., da Costa, L.M., de Oliveira, T.S., 2018. Geochemical signature of Amazon tropical rainforest soils. Rev. Bras. Cienc. do Solo 42. https://doi.org/10.1590/18069657rbcs20170192.

Delarmelinda, E.A., Souza Júnior, V.S. de, Wadt, P.G.S., Deng, Y., Campos, M.C.C., Câmara, E.R.G., 2017. Soil-landscape relationship in a chronosequence of the middle Madeira River in southwestern Amazon, Brazil. Catena 149. https://doi.org/10.1016/j.catena.2016.09.021.

Demattê, J.A.M., Araújo, S.R., Fiorio, P.R., Fongaro, C.T., Nanni, M.R., 2015. Espectroscopia VIS-NIR-SWIR na avaliação de solos ao longo de uma topossequência em Piracicaba (SP). Rev. Cienc. Agron. 46. https://doi.org/10.5935/1806-6690.20150054.

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

Demattê, J.A.M., Terra, F. da S., Bellinaso, H., Poppiel, R.R., Marques Júnior, J., Braz, K.F., Milori, D.M.B.P., Villas-Boas, P.R., Guimarães, C.C.B., Silva, S.H.G., Campos, J.R. da R., Tavares, T.R., 2021. Fundamentos do sensoriamento próximo de solos, in: Souza-Filho, L.F., Silva, R.C. da, César, F.R.C.F., Souza, C.M.M. (Eds.), Tópicos Em Ciência Do Solo.

SBCS, Viçosa, pp. 413–486.

Dharumarajan, S., Hegde, R., 2022. Digital mapping of soil texture classes using Random Forest classification algorithm. Soil Use Manag. 38. https://doi.org/10.1111/sum.12668

dos Santos, D.S., Ribeiro, P.G., Andrade, R., Silva, S.H.G., Gastauer, M., Caldeira, C.F., Guedes, R.S., Dias, Y.N., Souza Filho, P.W.M., Ramos, S.J., 2024. Clean and accurate soil quality monitoring in mining areas under environmental rehabilitation in the Eastern Brazilian Amazon. Environ. Monit. Assess. 2024 1964 196, 1–18. https://doi.org/10.1007/S10661-024-12495-4.

Embrapa, 2023. Laboratório de Solos e Plantas da Embrapa mantém Certificado de Excelência 2022/2023 [WWW Document]. URL https://www.embrapa.br/en/busca-de-noticias/-/noticia/82628249/laboratorio-de-solos-e-plantas-da-embrapa--mantem-certificado-de-excelencia-20222023.

Embrapa, 2020. Laboratório de Solos e Plantas [WWW Document]. URL https://www.embrapa.br/documents/1354300/1527966/Folder+de+serviços+e+valores+2020+-+Laboratório+de+Solos/4e1bdea7-c01f-e837-b08e-a17012fd69f1.

Fabris, J.D., Coey, J.M.D., Mussel, W.D.N., 1998. Magnetic soils from mafic lithodomains in Brazil. Hyperfine Interact. 113.

Fabris, J.D., De Jesus Filho, M.F., Coey, J.M.D., Mussel, W.D.N., Goulart, A.T., 1997. Iron-rich spinels from Brazilian soils. Hyperfine Interact. 110.

Falesi, I.C., 1974. Soils of the Brazilian Amazon, in: Man in the Amazon.

Faria, Á.J.G. de, Silva, S.H.G., Melo, L.C.A., Andrade, R., Mancini, M., Mesquita, L.F., Teixeira, A.F. dos 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 Res. 58, 683–695. https://doi.org/10.1071/SR20136.

Felix, J.C., Vendrame, P.R.S., Marchão, R.L., de Oliveira, J.F., Guimarães, M. de F., Brossard, M., Becquer, T., Brito, O.R., 2016. Predição de fósforo, carbono e nitrogênio em solos de basalto, por meio de espectroscopia NIR. Pesqui. Agropecu. Bras. 51. https://doi.org/10.1590/S0100-204X2016000900039.

Fernandes, R.B.A., Barrón, V., Torrent, J., Fontes, M.P.F., 2004. Quantificação de óxidos de ferro de Latossolos brasileiros por espectroscopia de refletância difusa. Rev. Bras. Ciência do Solo 28. https://doi.org/10.1590/s0100-06832004000200003.

Fritsch, E., Herbillon, A.J., Do Nascimento, N.R., Grimaldi, M., Melfi, A.J., 2007. From Plinthic Acrisols to Plinthosols and Gleysols: Iron and groundwater dynamics in the tertiary sediments of the upper Amazon basin. Eur. J. Soil Sci. 58. https://doi.org/10.1111/j.1365-2389.2006.00877.x.

Fritsch, E., Montes-Lauar, C.R., Boulet, R., Melfi, A.J., Balan, E., Magat, P., 2002. Lateritic and redoximorphic features in a faulted landscape near Manaus, Brazil. Eur. J. Soil Sci. 53. https://doi.org/10.1046/j.1351-0754.2002.00448.x.

Gee, Bauder, 1986. Particle size analysis. Methods of soil analysis. Am Soc Agron. 7.

Gee, G.W., Or, D., 2018. 2.4 Particle-Size Analysis, in: Methods of Soil Analysis: Part 4 Physical Methods. SSSA, Madison, pp. 255–293. https://doi.org/10.2136/sssabookser5.4.c12.

Gonçalves, D.A.M., da Silveira Pereira, W.V., Johannesson, K.H., Pérez, D.V., Guilherme, L.R.G., Fernandes, A.R., 2022. Geochemical Background for Potentially Toxic Elements in Forested Soils of the State of Pará, Brazilian Amazon. Minerals 12. https://doi.org/10.3390/min12060674.

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.

Hobley, E.U., Prater, I., 2019. Estimating soil texture from vis–NIR spectra. Eur. J. Soil

Sci. 70. https://doi.org/10.1111/ejss.12733.

Hossain, M.S., Rahman, G.K.M.M., Alam, M.S., Rahman, M.M., Solaiman, A.R.M., Mia, M.A.B., 2018. Modelling of soil texture and its verification with related soil properties. Soil Res. 56. https://doi.org/10.1071/sr17252.

IBGE, 2022. Brasil | Cidades e Estados | IBGE - Instituto Brasileiro de Geografia e Estatística [WWW Document]. URL https://www.ibge.gov.br/cidades-e-estados/pa.html (accessed 3.31.24).

Inda, A.V., Tomasi, C.A., de Oliveira, J.S., Fink, J.R., 2014. Óxidos de ferro e área superficial de Latossolo subtropical sob campo e floresta nativa. Cienc. Rural. https://doi.org/10.1590/S0103-84782014000200015.

Kaleem, M., Naseem, S., Bashir, E., Shahab, B., Rafique, T., 2021. Discrete geochemical behavior of Sr and Ba in the groundwater of Southern Mor Range, Balochistan, a tracer for igneous and sedimentary rocks weathering and related environmental issues. Appl. Geochemistry 130. https://doi.org/10.1016/j.apgeochem.2021.104996.

Karatzoglou, A., Smola, A., Hornik, K., 2023. _kernlab: Kernel-Based Machine Learning Lab_. R package version 0.9-32 [WWW Document].

Kassambara, A., Mundt, F., 2020. _factoextra: Extract and Visualize the Results of Multivariate Data Analyses_. R package version 1.0.7, [WWW Document]. URL https://cran.r-project.org/package=factoextra.

Kuhn, M., 2008. Building predictive models in R using the caret package. J. Stat. Softw. 28. https://doi.org/10.18637/jss.v028.i05.

Kursa, M.B., Rudnicki, W.R., 2010. Feature selection with the boruta package. J. Stat. Softw. 36. https://doi.org/10.18637/jss.v036.i11.

Lê, S., Josse, J., Husson, F., 2008. FactoMineR: An R package for multivariate analysis.

J. Stat. Softw. 25. https://doi.org/10.18637/jss.v025.i01.

Liaw, A., Wiener, M., 2002. Classification and Regression by randomForest. R News 2.

Lu, S.G., Chen, Y.Y., Shan, H.D., Bai, S.Q., 2009. Mineralogy and heavy metal leachability of magnetic fractions separated from some Chinese coal fly ashes. J. Hazard. Mater. 169. https://doi.org/10.1016/j.jhazmat.2009.03.078.

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., 2024. Multinational prediction of soil organic carbon and texture via proximal sensors. Soil Sci. Soc. Am. J. 88, 8–26. https://doi.org/10.1002/SAJ2.20593.

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 Reg. 30. https://doi.org/10.1016/j.geodrs.2022.e00573.

Marques, J.J., Schulze, D.G., Curi, N., Mertzman, S.A., 2004. Trace element geochemistry in Brazilian Cerrado soils. Geoderma 121. https://doi.org/10.1016/j.geoderma.2003.10.003.

Martińez, A.L., Uribe, A.S., 1995. Interfacial properties of celestite and strontianite in aqueous solutions. Miner. Eng. 8. https://doi.org/10.1016/0892-6875(95)00064-W.

Minasny, B., Hartemink, A.E., 2011. Predicting soil properties in the tropics. Earth-Science Rev. https://doi.org/10.1016/j.earscirev.2011.01.005.

Moeys, J., 2018. _soiltexture: Functions for Soil Texture Plot, Classification and Transformation_. R 727 package version 1.5.1, [WWW Document]. URL https://cran.rproject.org/package=soiltexture.

Naimi, S., Ayoubi, S., Di Raimo, L.A.D.L., Dematte, J.A.M., 2022. Quantification of some intrinsic soil properties using proximal sensing in arid lands: Application of Vis-NIR, MIR, and pXRF spectroscopy. Geoderma Reg. 28. https://doi.org/10.1016/j.geodrs.2022.e00484.

Nawar, S., Buddenbaum, H., Hill, J., Kozak, J., Mouazen, A.M., 2016. Estimating the soil clay content and organic matter by means of different calibration methods of vis-NIR diffuse reflectance spectroscopy. Soil Tillage Res. 155. https://doi.org/10.1016/j.still.2015.07.021.

Netto, J.D.S.M., Baptista, G.M. de M., 2000. Spectral reflectance of soils.

Nocita, M., Stevens, A., van Wesemael, B., Aitkenhead, M., Bachmann, M., Barthès, B., Dor, E. Ben, Brown, D.J., Clairotte, M., Csorba, A., Dardenne, P., Demattê, J.A.M., Genot, V., Guerrero, C., Knadel, M., Montanarella, L., Noon, C., Ramirez-Lopez, L., Robertson, J., Sakai, H., Soriano-Disla, J.M., Shepherd, K.D., Stenberg, B., Towett, E.K., Vargas, R., Wetterlind, J., 2015. Soil Spectroscopy: An Alternative to Wet Chemistry for Soil Monitoring. Adv. Agron. 132. https://doi.org/10.1016/bs.agron.2015.02.002.

O'Rourke, S.M., Holden, N.M., 2011. Optical sensing and chemometric analysis of soil organic carbon - a cost effective alternative to conventional laboratory methods? Soil Use Manag. 27. https://doi.org/10.1111/j.1475-2743.2011.00337.x.

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. https://doi.org/10.1016/j.geoderma.2016.05.005.

Parent, E.J., Parent, S.É., Parent, L.E., 2021. Determining soil particle-size distribution from infrared spectra using machine learning predictions: Methodology and modeling. PLoS One 16. https://doi.org/10.1371/journal.pone.0233242.

Pereira, T.T.C., Fábio, ;, Oliveira, S., Freitas, D.F., Damasceno, B.D., Dias, A.C., 2020.

A Mineralogia dos Solos Tropicais: Estado da Arte e Relação com o Uso e Manejo. Geonomos 28.

Peres, C.A., Gardner, T.A., Barlow, J., Zuanon, J., Michalski, F., Lees, A.C., Vieira, I.C.G., Moreira, F.M.S., Feeley, K.J., 2010. Biodiversity conservation in human-modified Amazonian forest landscapes. Biol. Conserv. 143. https://doi.org/10.1016/j.biocon.2010.01.021.

Poggere, G.C., Inda, A.V., Barrón, V., Kämpf, N., de Brito, A.D.B., Barbosa, J.Z., Curi, N., 2018. Maghemite quantification and magnetic signature of Brazilian soils with contrasting parent materials. Appl. Clay Sci. 161. https://doi.org/10.1016/j.clay.2018.05.014.

Poszwa, A., Dambrine, E., Ferry, B., Pollier, B., Loubet, M., 2002. Do deep tree roots provide nutrients to the tropical rainforest? Biogeochemistry 60. https://doi.org/10.1023/A:1016548113624.

Quesada, C.A., Lloyd, J., Schwarz, M., Patiño, S., Baker, T.R., Czimczik, C., Fyllas, N.M., Martinelli, L., Nardoto, G.B., Schmerler, J., Santos, A.J.B., Hodnett, M.G., Herrera, R., Luizão, F.J., Arneth, A., Lloyd, G., Dezzeo, N., Hilke, I., Kuhlmann, I., Raessler, M., Brand, W.A., Geilmann, H., Filho, J.O.M., Carvalho, F.P., Filho, R.N.A., Chaves, J.E., Cruz, O.F., Pimentel, T.P., Paiva, R., 2010. Variations in chemical and physical properties of Amazon forest soils in relation to their genesis. Biogeosciences 7. https://doi.org/10.5194/bg-7-1515-2010.

R Core Team 2023. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna.https://www.R-project.org/

Resende, M., Curi, N., Rezende, S.B., Corrêa, G.F., Ker, J.C., 2014. Pedologia: Base para distinção de ambientes, 6th ed. Editora UFLA, Lavras.

Richter, N., Jarmer, T., Chabrillat, S., Oyonarte, C., Hostert, P., Kaufmann, H., 2009. Free Iron Oxide Determination in Mediterranean Soils using Diffuse Reflectance Spectroscopy. Soil Sci. Soc. Am. J. 73. https://doi.org/10.2136/sssaj2008.0025.

Santos, A. do C., Pereira, M.G., dos Anjos, L.H.C., Bernini, T.A., Cooper, M., 2016. Genesis of soils formed from mafic igneous rock in the Atlantic forest environment. Rev. Bras. Cienc. do Solo 40. https://doi.org/10.1590/18069657rbcs20150056.

Santos, H.G. dos, Jacomine, P.K.T., Anjos, L.H.C. dos, Oliveira, V.A. de, Lumbreras, J.F., Coelho, M.R., Almeida, J.A. de, Araujo Filho, J.C. de, Oliveira, J.B. de, Cunha, T.J.F., 2018. Sistema brasileiro de classificação de solos, Embrapa Solos.

Schaetzl, R.J., Anderson, S., 2005. Soil: Genesis and Geomorphology, 1st ed. Cambridge University Press, New York.

Schoeneberger, P.J., Wysocki, D.A., Benham, E.C., Soil Staff, S., 2012. Field book for describing and sampling soils, Version 3.0. Natural Resources Conservation Service, National Soil Survey Center, Lincoln.

Shahriari, M., Delbari, M., Afrasiab, P., Pahlavan-Rad, M.R., 2019. Predicting regional spatial distribution of soil texture in floodplains using remote sensing data: A case of southeastern Iran. Catena 182. https://doi.org/10.1016/j.catena.2019.104149.

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. https://doi.org/10.1016/bs.agron.2020.12.001.

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. https://doi.org/10.1016/j.geoderma.2019.114136.

Siqueira, R.G., Moquedace, C.M., Francelino, M.R., Schaefer, C.E.G.R., Fernandes-Filho, E.I., 2023. Machine learning applied for Antarctic soil mapping: Spatial prediction of soil texture for Maritime Antarctica and Northern Antarctic Peninsula. Geoderma 432. https://doi.org/10.1016/j.geoderma.2023.116405.

Smith, P., House, J.I., Bustamante, M., Sobocká, J., Harper, R., Pan, G., West, P.C., Clark, J.M., Adhya, T., Rumpel, C., Paustian, K., Kuikman, P., Cotrufo, M.F., Elliott, J.A., Mcdowell, R., Griffiths, R.I., Asakawa, S., Bondeau, A., Jain, A.K., Meersmans, J., Pugh, T.A.M., 2016. Global change pressures on soils from land use and management. Glob. Chang. Biol. 22. https://doi.org/10.1111/gcb.13068.

Soil Survey Staff, 2022. Keys to Soil Taxonomy, 13th edition. United States Dep. Agric. Nat. Resour. Conserv. Serv.

Song, J., Shi, X., Wang, H., Lv, X., Zhang, W., Wang, J., Li, T., Li, W., 2024. Combination of feature selection and geographical stratification increases the soil total nitrogen estimation accuracy based on vis-NIR and pXRF spectral fusion. Comput. Electron. Agric. 218, 108636. https://doi.org/10.1016/J.COMPAG.2024.108636.

Stevens, A., Ramirez Lopez, L., 2014. An introduction to the prospectr package. R Packag. Vignette, Rep. No. R Packag. Version 0.1.

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. J. South Am. Earth Sci. 116. https://doi.org/10.1016/j.jsames.2022.103873.

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. https://doi.org/10.1016/j.geoderma.2017.10.053.

Viscarra Rossel, R.A., Adamchuk, V.I., Sudduth, K.A., McKenzie, N.J., Lobsey, C., 2011. Proximal Soil Sensing. An Effective Approach for Soil Measurements in Space and Time. Adv. Agron. 113. https://doi.org/10.1016/B978-0-12-386473-4.00010-5.

Viscarra Rossel, R.A., Hicks, W.S., 2015. Soil organic carbon and its fractions estimated by visible-near infrared transfer functions. Eur. J. Soil Sci. 66.

https://doi.org/10.1111/ejss.12237.

Wan, M., Hu, W., Qu, M., Li, W., Zhang, C., Kang, J., Hong, Y., Chen, Y., Huang, B., 2020. Rapid estimation of soil cation exchange capacity through sensor data fusion of portable XRF spectrometry and Vis-NIR spectroscopy. Geoderma 363. https://doi.org/10.1016/j.geoderma.2019.114163.

Wickham, H., 2016. ggplot2 Elegant Graphics for Data Analysis (Use R!). Springer.

Wickham, H., François, R., Henry, L., Müller, K. Vaughan D., 2023. _dplyr: A Grammar of Data 802 Manipulation_. R package version 1.1.2, [WWW Document]. URL https://cran.rproject.org/package=dplyr.

Wilford, J., 2012. A weathering intensity index for the Australian continent using airborne gamma-ray spectrometry and digital terrain analysis. Geoderma 183–184. https://doi.org/10.1016/j.geoderma.2010.12.022.

Xiaobo, Z., Jiewen, Z., Povey, M.J.W., Holmes, M., Hanpin, M., 2010. Variables selection methods in near-infrared spectroscopy. Anal. Chim. Acta. https://doi.org/10.1016/j.aca.2010.03.048.

Zhang, Y., Hartemink, A.E., 2020. Data fusion of vis–NIR and PXRF spectra to predict soil physical and chemical properties. Eur. J. Soil Sci. 71. https://doi.org/10.1111/ejss.12875.

Zhang, Y., Hartemink, A.E., 2019. Soil horizon delineation using vis-NIR and pXRF data. Catena 180. https://doi.org/10.1016/j.catena.2019.05.001.

Zhu, Y., Weindorf, D.C., Zhang, W., 2011. Characterizing soils using a portable X-ray fluorescence spectrometer: 1. Soil texture. Geoderma 167–168. https://doi.org/10.1016/j.geoderma.2011.08.010.

O tópico 3 corresponde ao segundo artigo resultado da tese, submetido na revista Computers and Electronics in Agriculture.

3 CAN pXRF AND VIS-NIR SPECTROMETRY BE APPLIED TO THE PREDICTION OF NATURAL SOIL FERTILITY IN THE AMAZON?

RESUMO

A avaliação da fertilidade do solo com elevada acurácia é crucial para a gestão ambiental dos solos florestais, bem como para a sua conservação e preservação, mas as análises químicas necessárias são demoradas, dispendiosas e geram resíduos químicos tóxicos. No entanto, os dados de sensores proximais integrados com algoritmos de aprendizagem de máquina têm sido bem sucedidos na avaliação da fertilidade do solo. Para validar essa abordagem, é necessário testar esses sensores em diversos tipos de solo, especialmente em regiões tropicais, onde há grande variabilidade de solo. Neste contexto, esta pesquisa empregou sensores portáteis próximos de espectrometria de fluorescência de raios X (pXRF) e espectrometria no visível e infravermelho próximo (Vis-NIR) para predizer o pH, a matéria orgânica do solo (MOS) e a capacidade de troca catiônica (CTC) de solos tropicais amazônicos, em duas profundidades de amostragem, no estado do Pará, Brasil. Foram estabelecidos os seguintes objetivos: i) comparação da eficácia de dados Vis-NIR e pXRF isolados e combinados para a predição de atributos químicos, utilizando o algoritmo Random Forest (RF); ii) comparação entre dois métodos (Boruta e PCA) para reduzir a dimensionalidade dos dados dos sensores. As amostras de solo foram coletadas a 0-20 cm e 80-100 cm, correspondentes a superfície e subsuperfície, em áreas florestais. Os métodos de seleção de variáveis Boruta e PCA produziram bons resultados na predição de pH do solo, MOS e CTC. O sensor pXRF apresentou a maior eficácia na predição do pH do solo para ambas profundidades pelo Boruta, enquanto que a maior precisão na predição da MOS foi apenas nas amostras de superfície pela PCA, já o sensor Vis-NIR demonstrou a maior precisão na predição para amostras da subsuperfície pelo Boruta. A CTC teve melhores resultados pela PCA para ambas profundidades, a superfície pelos dados do pXRF e a subsuperfície pelos dados do Vis-NIR. A integração dos dados dos sensores produziu predições mais exatas apenas para o pH em ambas profundidades e para a MOS na

subsuperfície do que a utilização de dados de sensores individuais. Os menores valores de RMSE e maiores valores de RPIQ foram observados para a superfície, com Boruta, para pH do solo (RMSE: 0,26; RPIQ: 3,48), MOS (RMSE: 3,27; RPIQ: 3,19) e para CTC a maior precisão foi observada nas amostras de subsuperfície com a PCA (RMSE: 1,80; RPIQ: 2,71), porém numericamente menor que o sensor Vis-NIR isolado. Este estudo contribuiu para o avanço das pesquisas sobre a aplicação de sensores proximais na predição da fertilidade do solo e, consequentemente, na caracterização do solo do bioma amazônico. Além disso, essa abordagem de sensores próximos aplicada a áreas de vegetação nativa, especialmente na Amazônia, é rara e se mostrou bem sucedida para tais condições.

Palavras-chave: aprendizagem de máquinas; sensores próximos; floresta aleatória

ABSTRACT

Assessing soil fertility with high precision is crucial for the environmental management of forest soils, as well as for their conservation and preservation, but chemical analyses are time-consuming, expensive and generate toxic chemical residues. However, proximal sensors data integrated with machine learning algorithms has been successful to assess soil fertility. To validate this approach, it is necessary to test these sensors in diverse soil types, especially in tropical regions, where there is great soil variability. In this context, this research employed portable proximal X-ray fluorescence spectrometry (pXRF) and visible and near-infrared spectrometry (Vis-NIR) sensors to predict the soil pH, soil organic matter (SOM), and cation exchange capacity (CEC) of tropical Amazonian soils, at two sampling depths, in the state of Pará, Brazil. The following objectives were set: i) a comparison of the efficacy of isolated and combined Vis-NIR and pXRF data for the prediction of chemical attributes, using the Random Forest (RF) algorithm; ii) a comparison between two methods (Boruta and PCA) for reducing the dimensionality of sensor data. Soil samples were collected at 0-20 cm and 80-100 cm, corresponding to surface and subsurface, respectively, in forest areas. The Boruta and PCA variable selection methods yielded good results in predicting soil pH, SOM and CEC. The pXRF sensor demonstrated the greatest effectiveness in predicting soil pH for both depths, as determined by Boruta. PCA, on the other hand, exhibited the highest accuracy in predicting MOS, but only for surface samples. The Vis-NIR sensor, however, exhibited the greatest accuracy in predicting subsurface samples, as determined by Boruta. The CEC was better

predicted by PCA for both depths, surface samples were predicted by pXRF data, and subsurface samples were predicted by Vis-NIR data. The integration of sensor data yielded more accurate predictions for pH at both depths and for MOS in the subsurface than using data from individual sensors. The lowest RMSE values and highest RPIQ values were observed for the surface, with Boruta, for soil pH (RMSE: 0.26; RPIQ: 3.48), SOM (RMSE: 3.27; RPIQ: 3. 19) and for CEC, the highest accuracy was observed in subsurface samples with PCA (RMSE: 1.80; RPIQ: 2.71), but numerically lower than the Vis-NIR sensor alone. This study contributed to the advancement of research into applying proximal sensors for predicting soil fertility, and consequently in soil characterization of the Amazonian biome. Moreover, this proximal sensor approach applied to areas under native vegetation, especially in the Amazon, is rare and was proved successful for such conditions.

Keywords: machine learning; proximal sensors; Random forest

3.1 Introduction

A large part of the Brazilian territory is occupied by the Amazon biome, an ecosystem with high forest density and a rich diversity of fauna, flora, rivers, and soils (Moreira et al., 2018). However, its soils are generally infertile, acidic, and nutrient-poor (Souza et al., 2018; Lopes and Guimarães Guilherme, 2016). The "fingerprint" of this tropical forest soil is defined by climatic factors, parent material (Prietzel and Christophel, 2014), and by some processes that can alter soil chemical parameters, such as weathering and leaching (Sujatha and Jaidhar, 2024).

Soil fertility includes several factors such as nutrient ratios, soil pH, soil organic matter (SOM), cation exchange capacity (CEC), soil texture, and soil structure. Soils with low pH have fewer binding sites for nutrients and are therefore generally less fertile (Kouadio et al., 2018). On the other hand, SOM increases soil water-holding capacity, aeration, CEC, nutrient availability, and reduces soil erosion (Kouadio et al., 2018; Ramos et al., 2018). Accurate and timely information on soil fertility is fundamental for decision-making (Li et al., 2024), especially in the Amazon region, which is seeking innovations for the sustainable use and conservation of its soils (BRASIL, 2012), to the detriment of the misuse of this resource due to deforestation and anthropogenic activities (Akerman et al., 2021). In this context, researchers have incorporated X-ray fluorescence spectrometry (XRF) and visible and near-infrared diffuse reflectance spectrometry (Vis-NIR) into soil analysis.

The XRF technique employs the fluorescence of specific energies of atomic species that are excited when irradiated with X-rays. This allows the identification and quantification of the soil chemical composition (Viscarra Rossel et al., 2011). Conversely, Vis-NIR is a commonly used method to assess the reflectance of the sample at different wavelengths of the visible and near-infrared spectral range, which is strongly correlated to soil organic and mineral constituents (Wan et al., 2020). The correlations between the specific characteristics of the spectral curves and the chemical, physical, and mineralogical attributes of the soil allow such sensors to discriminate soils (Demattê et al., 2015). As a result, the use of proximal sensors is becoming increasingly widespread, allowing the information derived from these sensors to be used to predict soil attributes (Huang et al., 2023; Silvero et al., 2023).

The use of sensor data for fertility prediction has been investigated using a variety of machine learning methods, including partial least squares (PLS) regression, random forest (RF), stepwise generalized linear models (SGLM), multiple linear regression (MLR), support vector machine (SVM), and cubist regression (Al Masmoudi et al., 2022; Andrade et al., 2020; Helfer et al., 2020; Hounkpatin et al., 2022). Among them, RF has outperformed (Benedet et al., 2021; Teixeira et al., 2022) due to its ability to handle a large number of predictor variables in a dataset (Speiser et al., 2019), however, this can also adversely affect the accuracy of the predictions (Zhao et al., 2023). Therefore, it is fundamental to determine the most important predictors to be included in a reduced and parsimonious model (Saidy et al., 2012; Speiser et al., 2019) which can be done by different methods, such as Boruta (BO), principal component analysis (PCA), simulated annealing (SA), and genetic algorithm (GA) (Arjasakusuma et al., 2020). In addition, prediction performance is also influenced by many variables, such as soil depth, land use, algorithm, sample preparation methods, data pre-processing methods, among others (Teixeira et al., 2022).

There is little research in the Amazon region with proximal sensors, especially on soils under native vegetation, beyond the uncertainties about the applicability of this approach for predicting different soil attributes (Teixeira et al., 2022). Therefore, additional studies involving a variety of soil types (including mineralogy, source material, and soil class) are essential to improve the reliability of proximal sensors for soil property prediction (Benedet et al., 2021). In this context, the aims of this study were to i) compare the use of isolated and combined Vis-NIR and pXRF data for predicting SOM, soil pH, and CEC of tropical Amazonian soils at two depths (236 samples), using the RF algorithm and ii) to compare two methods for reducing the dimensionality of sensor data. It is hypothesized that the combined data will provide more

accurate models for predicting these attributes than when used separately. Furthermore, a data reduction method will be specified. The results of this study will provide valuable insight into the feasibility of using these sensors for the fertility prediction of Brazilian tropical soils under native vegetation.

3.2 Material and methods

3.2.1 Sampling area and laboratory analyses

The soil samples were obtained from the soil bank of the Laboratory of Trace Elements in the Environment (LETAM) of the Federal Rural University of Amazon. These samples were collected from natural areas with forest cover, including primary and secondary forests, in 61 municipalities of the Brazilian state of Pará (Figure 1). More details can be found in Gonçalves et al. (2022) and Souza et al. (2018). The state and municipal governments of Pará have been particularly active in their efforts to reduce deforestation, and have been recognized as a reference model for other parts of the Amazon (Nunes et al., 2016). The state is classified as having a tropical climate, according to the Köppen climate classification system, with 66.6% of the territory classified as Am (monsoon), 28.4% as Af (without a dry season), 4.9% as Aw (with dry winter), and < 0.01% of occurrence as As (with dry summer). The mean annual temperature varies between 24°C and 26°C, while the mean annual rainfall varies between 1,700 and 2,500 mm (Alvares et al., 2013).

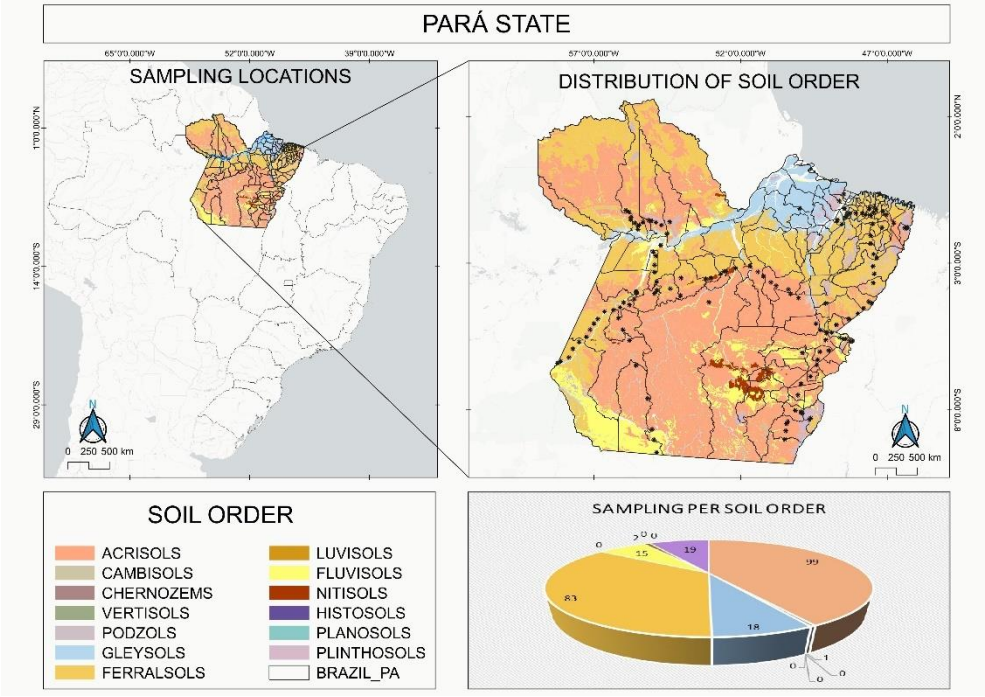

Figure 1 - Map of the study area, indicating the locations of soil samples and the distribution of soil samples by soil classification following World Reference Base system (IUSS Working Group, 2015) in the state of Pará, Brazil.

The sampling was carried out in two soil depths: 0-20 cm (141 samples) and 80-100 cm, (95 samples), somewhat equivalent to soil in the surface and subsurface, respectively. The samples include eight soil orders, following the World Reference Base (IUSS Working Group, 2015) and the Brazilian Soil Classification System (Santos et al., 2018): Acrisols, Ferralsols, Plinthosols, Gleysols, Leptsols, Arenosols, Nitisols, and Cambisols. Acrisols and Ferralsols are the most prevalent (Table 1). The soil classification map of the state of Pará was downloaded from the IBGE website (IBGE, 2024) and used in QGIS software version 3.16.11 to determine the number of samples per soil class. The classification according to the Brazilian Soil Classification System (SiBCS) was aligned with the World Reference Base for Soil Resources (WRB/FAO) classification system (IUSS Working Group WRB, 2015).

Table 1- Classification according to the Brazilian Soil Classification System (SiBCS) and the World Reference Base for Soil Resources (WRB/FAO) classification system

SiBCS	WRB/FAO	Number of samples
Argissolos	Acrisols	97
Cambissolos	Cambisols	1
Gleissolos	Gleysols	18
Latossolos	Ferralsols	83
Nitossolos	Nitisols	2
Plintossolos	Plinthosols	19
Neossolos Litólicos	Leptsols	3
Neossolos quartzarênicos	Arenosols	13

Soil samples were prepared by standard procedures, including air-drying and sieving through a 2 mm mesh, for both chemical and proximal sensors analysis. Soil pH, SOM, and CEC at pH 7.0 were determined in triplicate according to Teixeira et al. (2017). In this method, pH was quantified using an electrode immersed in a suspension with a soil:water ratio of 1:2.5. The SOM content was determined by oxidation with a potassium dichromate mixture, followed by titration with an ammoniacal ferrous sulphate solution. The CEC was calculated as the sum of the concentrations of exchangeable cations (Ca^{2+}, Mg^{2+}, K^{+}, and Al^{3+}), whereby exchangeable calcium, magnesium, and aluminum were determined using a 1 mol L^{-1} potassium chloride (KCl) extractant solution and exchangeable potassium was extracted with Mehlich-1 solution. This represents the cationic ions bound to the exchange complex under the conditions found in the study area, together with potential acidity (H^{+} + Al^{3+}), extracted with buffered calcium acetate at pH 7.0 and determined volumetrically with NaOH solution in the presence of phenolphthalein as an indicator.

3.2.2 Vis-NIR spectral acquisition

The spectra of the samples were acquired from the Brazilian Soil Spectral Library database (Piracicaba, SP) (Demattê et al., 2019). The Vis-NIR spectroradiometer (Analytical Spectral Devices, Boulder, CO, USA) was employed to obtain the spectra of samples previously dried at 45 °C and sieved through a 2 mm mesh. This instrument has a spectral range extending

from the visible (Vis) to near-infrared (NIR) wavelengths (350-2500 nm) and a spectral resolution of 1 nm from 350 to 700 nm, 3 nm from 700 to 1400 nm, and 10 nm from 1400 to 2500 nm (Demattê et al., 2019). The data output sampling interval was 1 nm, with 2151 channels reported. The sensor was positioned 8 cm from the sample surface distributed among Petri dishes, scanning an area of approximately 2 cm^2. The analysis involved two replicates for each sample, and after one replicate the sample was rotated 180°. A standard Spectralon white plate was scanned at 20-minute intervals. The mean values from two replicates were employed for each sample. The mean of 100 readings was taken from each spectrum over a 10-second interval. The soil spectra were obtained according to the protocol proposed by Ben Dor et al. (2015).

3.2.3 pXRF total element contents acquisition

Samples were analyzed in triplicate on the S1 TITAN 800 portable X-ray fluorescence spectrometer (pXRF) (Bruker, Billerica, MA, USA), with each reading lasting 60 s. The pXRF was calibrated prior analysis using the instrument manufacturer's certified material (check sample) and two certified reference materials (2710a and 2711a) from the National Institute of Standards & Technology (NIST). The triplicate averages were used for each sample. The recovery values were calculated using the following formula: recovery in % = 100 (pXRF content/content of reference samples). For further analysis, all the elements that had values above the detection limit for all the samples were used. These elements were magnesium (Mg), aluminum (Al), silicon (Si), phosphorus (P), calcium (Ca), titanium (Ti), vanadium (V), manganese (Mn), iron (Fe), copper (Cu), zinc (Zn), strontium (Sr), and zirconium (Zr). The recovery values were satisfactory (Table S1). A missing value indicates that the equipment was unable to detect the element or that the element has no certified value.

3.2.4 Exploratory analysis

To gain a deeper understanding of the soils, the data were first subjected to descriptive analysis and boxplots. The results of the SOM, soil pH, CEC, and total element contents obtained by pXRF from both depths were subjected to Spearman's correlation analysis

("corrpolot" package) and to principal component analysis (PCA), ("FactoMineR" and "factoextra") packages. The PCA's objective of these analyses was to ascertain whether the variables exhibited any pattern in relation to the surface and subsurface samples. All figures were created using "ggplot2". The development of both exploratory analysis and modeling was facilitated by the use of R software (R Core Team, 2023).

3.2.5 Dimensionality reduction

Prior to modeling, two data reduction techniques were performed separately to compare which method contributes to the best predictions: principal component analysis (PCA) and the Boruta algorithm (Kursa and Rudnicki, 2010). The pXRF, Vis-NIR, and pXRF + Vis-NIR data corresponding to samples from surface and subsurface were processed independently. PCA technique identifies and extracts the most significant information from the data and represents it as a set of new orthogonal variables (principal components, PCs) (Abdi and Williams, 2010). In contrast, the Boruta algorithm selects relevant variables, operating as a wrapper around a RF algorithm, iteratively removing features that have been proven by a statistical test to be less relevant than random probes (Kursa and Rudnicki, 2010). The number of principal components chosen was based on the cumulative proportion of ≥99% and a minimum of 0,05% of the proportion of variance. On the other hand, Boruta reduces the dimensionality of the dataset and defines the variables as confirmed, tentative, or rejected (Kursa and Rudnicki, 2010). Only variables that were confirmed were considered.

3.2.6 Modeling

The total element contents obtained by pXRF and the spectral reflectance from Vis-NIR (independent variables) were used to predict the dependent variables soil pH, SOM, and CEC based on the RF algorithm using the "RandomForest" and "caret" packages (Kuhn, 2008; Liaw and Wiener, 2002). A total of 13 elements (Mg, Al, Si, P, Ca, Ti, V, Mn, Fe, Cu, Zn, Sr, and Zr) and the spectral reflectance for the 2,150 wavelengths (ranging from 350 to 2,500 nm) provided by pXRF and Vis-NIR, respectively, were analyzed.

For modeling, we separated the data by depth and employed them separately. The datasets included: I) raw pXRF data reduced by PCA, II) raw pXRF data reduced by Boruta, III) raw Vis-NIR data reduced by PCA, IV) raw Vis-NIR data reduced by Boruta, V) the concatenation of pXRF + Vis-NIR data reduced by PCA, and VI) the concatenation of pXRF + Vis-NIR data reduced by Boruta. Cross-validation (*leave-one-out*, LOOCV) was performed, and the model accuracy was assessed by various metrics: root mean square error - RMSE (Eq.1), mean absolute error – MAE (Eq.2), coefficient of determination - R^2 (Eq.3), residual prediction deviation - RPD (Eq.4), and ratio of performance to interquartile distance – RPIQ (Eq.5).

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

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}\right| \quad \text{(Eq. 2)}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad \text{(Eq. 3)}$$

$$RPD = \frac{SD}{RMSE} \quad \text{(Eq. 4)}$$

$$RPIQ = \frac{\mathrm{IQ}}{RMSE} \quad \text{(Eq. 5)}$$

where, *n*: number of observations; y_i: value measured by laboratory analysis; $\hat{y}_i$: value estimated by the model; $\bar{y}$: average of all the values measured by the chemical analysis; SD: standard deviation; IQ: difference obtained between the value referring to the 3rd and 1st quartile of the data distribution.

The RPD metric is particularly valuable for comparing models and assessing their quality, while the RPIQ index is also effective for evaluating model performance, especially when soil samples exhibit skewed distributions (Bellon-Maurel et al., 2010; Ge et al., 2020). The accuracy of the predictions was evaluated according to the lowest RMSE values along with the highest RPIQ values.

3.3 Results and discussion

3.3.1 Descriptive Statistics

Considering the two depths, the soil pH values varied from 2.46 to 6.94 (Figure 2, Table S2). The mean soil pH values were 4.51 for the surface and 4.81 for the subsurface. The SOM content varied from 0.90 to 66.96 g kg^{-1}. Considering both depths, respectively, the mean values were 12.54 and 6.53 g kg^{-1} with a CV varying in 65.80% to 123.32%. The high CV is a consequence of the variability of the soil classes and parent materials. The CEC in the depths also showed a wide range, from 1.40 to 183.23 $cmol_c$ dm^{-3}, with mean values of 6.33 for the subsurface and 25.25 $cmol_c$ dm^{-3} for the surface.

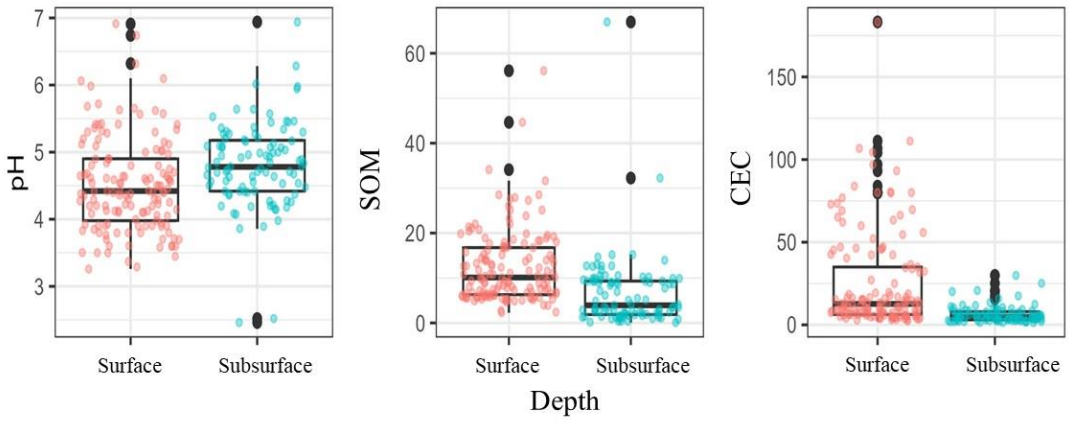

Figure 2 - Boxplots of soil pH, soil organic matter - SOM (g kg^{-1}) and cation exchange capacity -CEC ($cmol_c$ dm^{-3}) in soil samples from the surface and subsurface

The values obtained from conventional analyses demonstrate the diversity of soils. The high variability in SOM content can be attributed to the soil sampling site, soil depth, and forest composition, which varies with soil type and granulometry, as well as climatic variation within the region. Furthermore, forest composition in secondary forest areas is diverse, primarily due to its formation resulting from various causes, including burning and deforestation. Forest fires have been shown to reduce aboveground biomass, potentially leading to long-term legacy effects on soil carbon (C) stocks (Ling et al., 2021). These fires create an unstable ecosystem with limited resources, often resulting in significant soil carbon loss (Ling et al., 2021). Furthermore, it is well documented that certain forest species contribute to the accumulation of

SOM (Weber et al., 2020) and that the content of SOM in forest areas is typically high (Teixeira et al., 2018).

Summarizing, the surface exhibited the highest mean SOM and CEC values and the lowest soil pH when compared to the subsurface. The elevated surface average SOM value can be attributed to the greater accumulation of organic material on the surface, which is common in forest soils (Osman, 2013), reflecting in higher CEC values (Schaetzl and Anderson, 2005). In this sense, the presence of SOM with negative charges favors increased CEC, especially in tropical soils (Jiang et al., 2017; Ramos et al., 2018). Additionally, the formation of litter under tree canopies may contribute to an increase in the spatial variability of soil attributes, including CEC (Weber et al., 2020). Notwithstanding the elevated mean clay content in the subsurface horizon of numerous soil types within the data set, which affects CEC (Pavão et al., 2024) and the impact of the charges provided by SOM (Ramos et al., 2018), the elevated CEC values at the surface are attributed to the high mean value of the MOS content. The lower mean soil pH values observed in the surface horizons can be attributed to the decomposition of SOM, which acidifies the soil due to greater quantity of humic substances, which are rich in carboxylic (-COOH) and phenolic (-OH) groups (Ramos et al., 2018).

The observed correlations between soil fertility attributes and total pXRF elements were relatively low (Figure S1), as pXRF determined the total contents of elements, contrary to soil fertility assessment (Weindorf et al., 2014). Principal component analysis revealed that 49.7% (the two first components) of the total variance were explained by the different depths (Figure 3), as the samples had more similarities, for it is predominantly composed of Acrisols, and Ferralsols. The PCA showed minimal segregation between soil samples from the two depths. However, each soil depth had specific characteristics that render it a subject of investigation, such as higher SOM content in surface soils and lower CEC and nutrient contents in subsurface soils. These differences can be attributed to a number of factors, including land use history, climatic conditions, and the various soil formation factors. Due to the considerable vertical variability, assessment at separate depths is pertinent for the inventory of soil resources or agriculture.

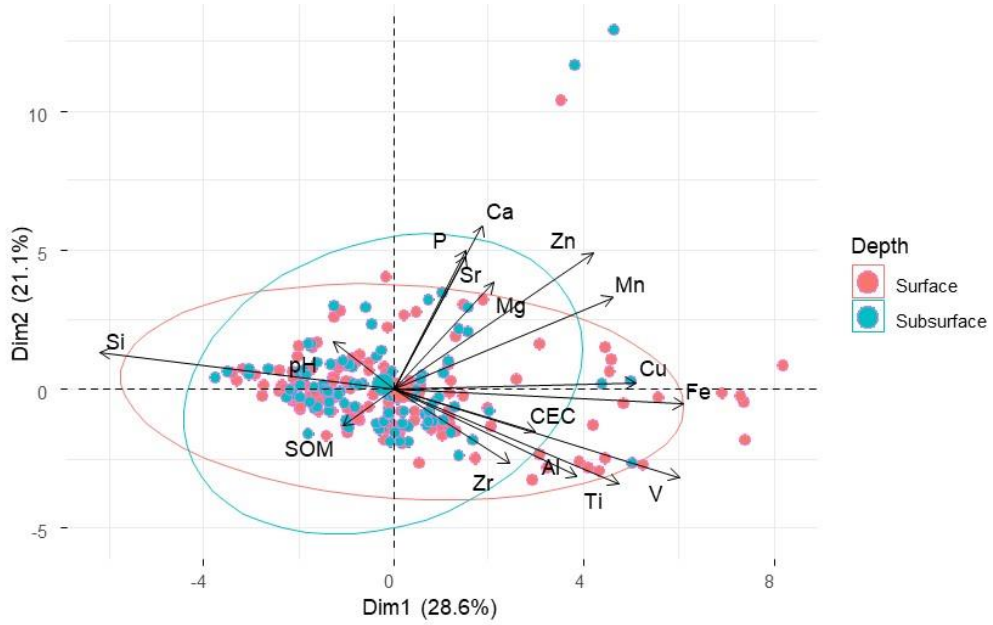

Figure 3 - Principal Component Analysis (PCA) of soil fertility attributes and portable X-ray fluorescence spectrometry (pXRF) data from samples of the surface and subsurface.

3.3.2. Prediction of soil pH

The results of the prediction of soil pH, SOM, and CEC and their respective R^2, RMSE, MAE, RPD, and RPIQ values, using the RF algorithm with two-dimensionality reduction techniques (Boruta and PCA), are presented in Table 4. The soil pH exhibited favorable predictions ($R^2 \geq 0.89$; RMSE ≤ 0.30 and RPIQ ≥ 2.48) when employing pXRF, Vis-NIR, and pXRF + Vis-NIR data, with Boruta and PCA (metric values remained proximate), across both depths. The lowest error values and highest RPIQ values were found using the pXRF + Vis-NIR data when Boruta was used, both in the superficial (RMSE=0.26; RPIQ=3.48) and subsurface (RMSE= 0.27, RPIQ=2.78). In comparison to the utilization of sensor data alone, the pXRF exhibited the most optimal combination of errors and RPIQ for both depths, with shallow (RMSE= 0.27; RPIQ=3.41) and subsurface (RMSE=0.29; RPIQ=2.63) depths. This observation was consistent when Boruta was employed. The application of pXRF data (surpassing Vis-NIR) and pXRF + Vis-NIR data using Boruta exhibited superior results in terms of RMSE and RPIQ values when compared to PCA for both horizons (Figure 5). Selection using Boruta demonstrated numerically superior prediction values compared to PCA. This study underscores the impact of variable selection techniques on model performance.

In contrast to other soil attributes, soil pH does not exhibit a direct spectral response (Ge et al., 2020), as it is a measure of proton activity (molar concentration of hydrogen ions in solution), however, it is predicted by association with active soil attributes, such as SOM and clay (Stenberg et al., 2010). Gozukara et al. (2022a) demonstrated that Vis-NIR data exhibited superior predictive performance in estimating soil pH relative to pXRF data (it was applied to the spectra). On the other hand, in the present study, data obtained by pXRF demonstrated satisfactory results (superior to Vis-NIR) for soil pH prediction, as well as in tropical soils (R^2 = 0.84) (Teixeira et al., 2018) and temperate soils (R^2 = 0.77 and RMSE = 0.68) (Sharma et al., 2014).

Table 4 - Coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), residual prediction deviation (RPD), and ratio of performance to interquartile range (RPIQ) of the prediction of soil pH, soil organic matter (SOM) and cation exchange capacity (CEC) of tropical soils in the surface and subsurface, using pXRF and Vis-NIR data (isolated and combined).

Variable	Sensor	Depth	BORUTA[a]					PCA[b]				
			R^2	RMSE	MAE	RPD	RPIQ	R^2	RMSE	MAE	RPD	RPIQ
pH	pXRF	Surface	0.92	0.27	0.20	2.62	3.41	0.94	0.30	0.23	2.33	3.03
		Subsurface	0.90	0.29	0.20	2.23	2.63	0.93	0.30	0.20	2.11	2.48
	Vis-NIR	Surface	0.90	0.29	0.22	2.40	3.12	0.91	0.29	0.22	2.42	3.15
		Subsurface	0.89	0.29	0.22	2.23	2.63	0.92	0.30	0.22	2.14	2.53
	pXRF + Vis-NIR	Surface	0.93	**0.26**	0.20	2.67	**3.48**	0.92	0.28	0.21	2.50	3.26
		Subsurface	0.90	**0.27**	0.19	2.36	**2.78**	0.93	0.29	0.21	2.21	2.60
SOM[c]	pXRF	Surface	0.92	3.40	2.38	2.43	3.06	0.90	3.24	2.26	2.55	3.22
		Subsurface	0.76	4.12	2.03	1.96	1.80	0.89	4.18	1.96	1.93	1.78
	Vis-NIR	Surface	0.87	3.58	2.48	2.31	2.91	0.92	3.55	2.67	2.33	2.94
		Subsurface	0.85	3.95	2.14	2.04	1.88	0.87	4.18	2.21	1.93	1.78
	pXRF + Vis-NIR	Surface	0.90	**3.27**	2.29	2.53	**3.19**	0.93	3.47	2.61	2.38	3.00
		Subsurface	0.86	**3.53**	1.66	2.28	**2.10**	0.86	4.36	2.23	1.85	1.70
CEC[d]	pXRF	Surface	0.88	12.85	8.22	2.31	2.24	0.91	**11.22**	7.34	2.64	**2.56**
		Subsurface	0.89	2.17	1.46	2.32	2.25	0.94	2.37	1.66	2.12	2.06
	Vis-NIR	Surface	0.83	14.92	10.54	1.99	1.92	0.89	13.34	9.33	2.22	2.15
		Subsurface	0.89	1.96	1.32	2.57	2.49	0.94	**1.79**	1.23	2.80	**2.72**
	pXRF + Vis-NIR	Surface	0.88	12.92	8.29	2.30	2.22	0.89	12.85	8.68	2.31	2.24
		Subsurface	0.93	1.97	1.29	2.55	2.47	0.94	1.80	1.23	2.79	2.71

[a] Algorithm feature selection (Kursa and Rudnicki, 2010); [b] Principal component analysis; [c] soil organic matter; [d] cation exchange capacity at pH = 7

Recent studies estimating soil fertility have employed data from Vis-NIR and pXRF sensors, either individually or in combination. The results of both approaches have been considered satisfactory (Teixeira et al., 2022). However, it is unclear whether a single technique is as effective as their combined use for estimating each fertility parameter (Liu et al., 2021). This is due to the intrinsic soil characteristics and the attributes (*e.g.*, soil pH, CEC, SOM) that each technique is capable of measuring, which make it challenging for a single spectrometer to characterize all soil attributes (Li et al., 2024). The investigation considering the soil depth is particularly important for monitoring carbon, nutrient status, soil pH, and salinity (Stenberg et al., 2010). Gozukara et al. (2022b) propose that splitting predictive models by horizons may provide superior prediction performance for certain soil attributes (i.e. soil sand, clay and silt content) when compared to combining horizons. Likewise, splitting predictive models by depths may achieve high prediction accuracy for clay, sand, and silt when using XRF data and vis-NIR spectra combined with the Random Forest (Pavão et al., 2024). This is also related to soil fertility issues, as some crops require liming practices at depth, while others require only at the surface. It is of great importance to consider the general characteristics of the soil for agricultural or conservation purposes (e.g., depth) in order to ensure that fertility recommendations are carried out properly (Teixeira et al., 2018).

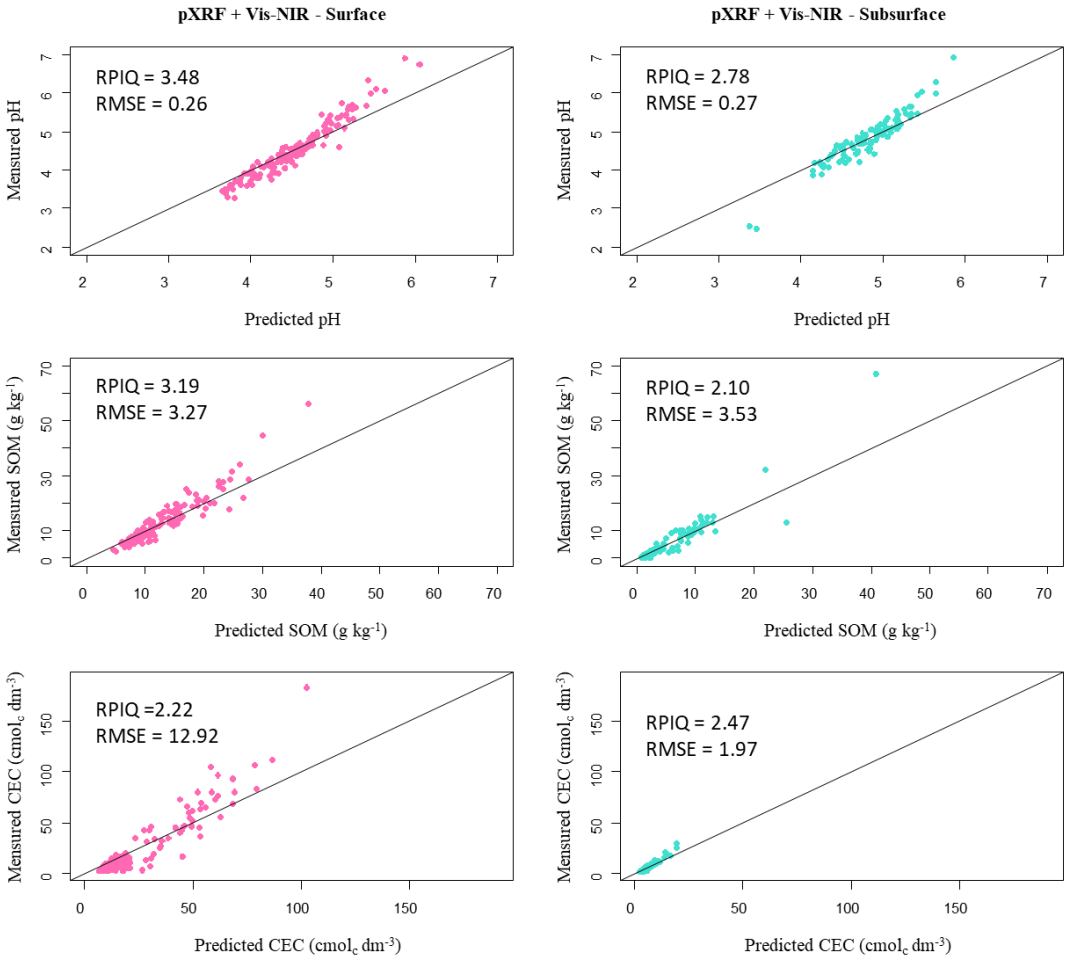

Figure 5 - Performance of models for soil pH, soil organic matter (SOM) and cation exchange capacity (CEC) obtained from conventional analyses and predicted by the RF algorithm, using data from the combination of portable X-ray fluorescence (pXRF) and visible and near-infrared spectrometry (Vis-NIR) analysis.

3.3.3 Prediction of SOM

The application of pXRF, Vis-NIR, and pXRF + Vis-NIR data indicated the potential for accurate SOM predictions at both depths for both data reduction methods (R^2>= 0.76, RMSE≤4.36, and RPIQ≥1.70). When the data sensors were considered separately, the SOM exhibited the lowest RMSE (3.24) and the highest RPIQ (3.22) through PCA in the surface by pXRF. In the subsurface samples, the lowest error (RMSE=3.95) and the highest RPIQ (1.88) was achieved through Boruta by Vis-NIR data.

The combination of pXRF + Vis-NIR data with the Boruta algorithm yielded lower RMSE (3.53) and higher RPIQs (2.10) only for the subsurface, as in the surface samples the values were lower than for the pXRf sensor alone through PCA. Liu et al. (2021) reported that Vis-NIR data alone could predict soil organic carbon and that the combination of Vis-NIR and pXRF did not significantly enhance prediction accuracy. The soil organic fraction is composed of functional groups, including CH, CO, and CN, which have well-documented spectral signals in the Vis-NIR region (O'Rourke et al., 2016). It can be postulated that the Vis-NIR technology demonstrated superior results than pXRF due to the abundance of chemical compound information present in its spectra (e.g., C—H, N—H, C=O, and O—H) (Jiang et al., 2017). Furthermore, SOM exhibits spectral activity throughout the entire Vis-NIR region, with a particularly strong response in the visible region (Liu et al., 2018).

3.3.4 Prediction of CEC

The CEC was accurately predicted using pXRF, Vis-NIR, and pXRF + Vis-NIR data, both via Boruta and PCA, in both depths. The highest prediction performance was identified in the surface layer using pXRF data when PCA was employed with the lowest value RMSE (11.22) and a higher RPIQ (2.56). In contrast, the utilization of Vis-NIR data yielded higher accuracy in the subsurface, with an RMSE of 1.79 and an RPIQ of 2.72, once again employing PCA. The PCA-based prediction model, incorporating data from different sensors, demonstrated higher RPIQ values and lower RMSE compared to the Boruta approach for the surface. The PCA method yielded superior results for CEC, with the exception of the subsurface by pXRF data, which exhibited better performance through Boruta (RMSE: 2.17). A single data sensor proved inadequate for both depths, and the integration of multiple data sensors did not yield superior results for either depth. The results obtained with the combination of data sensors were numerically inferior compared to those obtained with a single sensor (pXRF), exhibiting lower RMSE values and high RPIQ for surface data and Vis-NIR data for subsurface.

A reliable prediction can be made based on Vis-NIR sensor data, which is indicative of the sensor's ability to discern soil minerals and soil organic matter (SOM), both of which contribute to the cation exchange capacity (CEC). The CEC is subject to the influence of clay mineral activity and, consequently, the negative charges associated with these minerals (Stenberg et al., 2010). Studies of soil mineralogy using Vis-NIR spectroscopy have

demonstrated a direct relationship between minerals and their spectra. This relationship has enabled the use of Vis-NIR to predict CEC with greater accuracy than other methods, as the diagnostic absorption bands of soil minerals fall within the Vis-NIR region (Fang et al., 2018).

3.3.5 Data dimensionality reduction and predictors variable's importance

The total number of Vis-NIR predictor variables was 2151 wavelengths, while the pXRF variables were 13 elements. This quantity of variables has the effect of slowing down the modeling process and, in some cases, may result in a compromise to the model's predictive capacity. This highlights the importance of reaching a parsimonious model that exhibits high accuracy through a minimal number of variables employed. However, it is challenging to assess the trade-off between dispersion and precision (Sanchez-Pinto et al., 2018).

Following the application of PCA, 2151 wavelengths of the Vis-NIR data were reduced to between five and six principal components (PCs), while 13 elements of the pXRF data were reduced to between ten and eleven PCs. In contrast, the Vis-NIR + pXRF data exhibited 12 PCs (Table S3). Viscarra Rossel et al. (2022) also employed the spectra PCA technique to address the challenge of a large number of variables in modeling a target variable, applying principal components. Gozukara et al. (2022b) reported that 10 PCs explained approximately 100% of the cumulative variance in visible-near infrared (Vis-NIR) spectra in the prediction of soil attributes (*e.g.*, soil pH) by the Cubist model. The Boruta algorithm identified between seven and eight elements from the pXRF data set, between two and nine wavelengths from the Vis-NIR data set, and between nine and seventeen variables from the combination of the pXRF + Vis-NIR sensor data set (Table S4). The reduction in the number of variables has an impact on the operating time (computing efficiency) for modeling (Xu et al., 2013), additionally, it results in a more streamlined prediction model that maintains precision while approaching a parsimonious model (Sanchez-Pinto et al., 2018).

The reduction of data dimensionality is a crucial element of the pre-processing phase. The application of both PCA and Boruta yielded satisfactory prediction outcomes. Overall, the integration of sensor data exhibited enhanced predictive capabilities for soil pH, SOM, and CEC in both the A and B horizons when utilizing Boruta, with the exception of CEC in the B horizon, which was subjected to PCA.

The importance of the variables for the prediction of soil pH, SOM, and CEC was determined by analyzing concatenated sensor data with Boruta, as it produced the most accurate predictions for soil pH and SOM on the subsurface (Figure 6). In surface and subsurface, the presence of the elements V, Ti, Fe, Al, Sr, Si, Zr, and Mn, as well as wavelengths in the visible and infrared (359, 360, 369, 371, 396, 2107, 2112, 2117, 2197, 2436, 2453, 2495) range, was observed.

Soil pH and elemental content are often related, and there is a known relationship between acidic soils and elements such as Al, Fe, and Mn (Teixeira et al., 2018). The observed wavelengths are associated with other soil constituents, such as SOM and clay, which affect the pH of the soil. Although soil pH has no active spectral properties, the intrinsic relationship of the primary soil constituents (clay minerals, Fe oxides, SOM, carbonates, and water) with other soil attributes, enables the measurement of soil pH in the Vis-NIR spectrum (Ge et al., 2020). Clay minerals exhibit absorption bands in the near-infrared region at approximately 2200 nm, while SOM is absorbed at different wavelengths in the Vis-NIR spectrum (Viscarra Rossel et al., 2016). The importance of the contents of Al, Ti, V, and Fe was observed in soil pH, which may be associated with the solubility of these elements in the soil solution. The oxide minerals aluminum, titanium, and iron have very low solubility at ordinary soil pH (Grison et al., 2021).

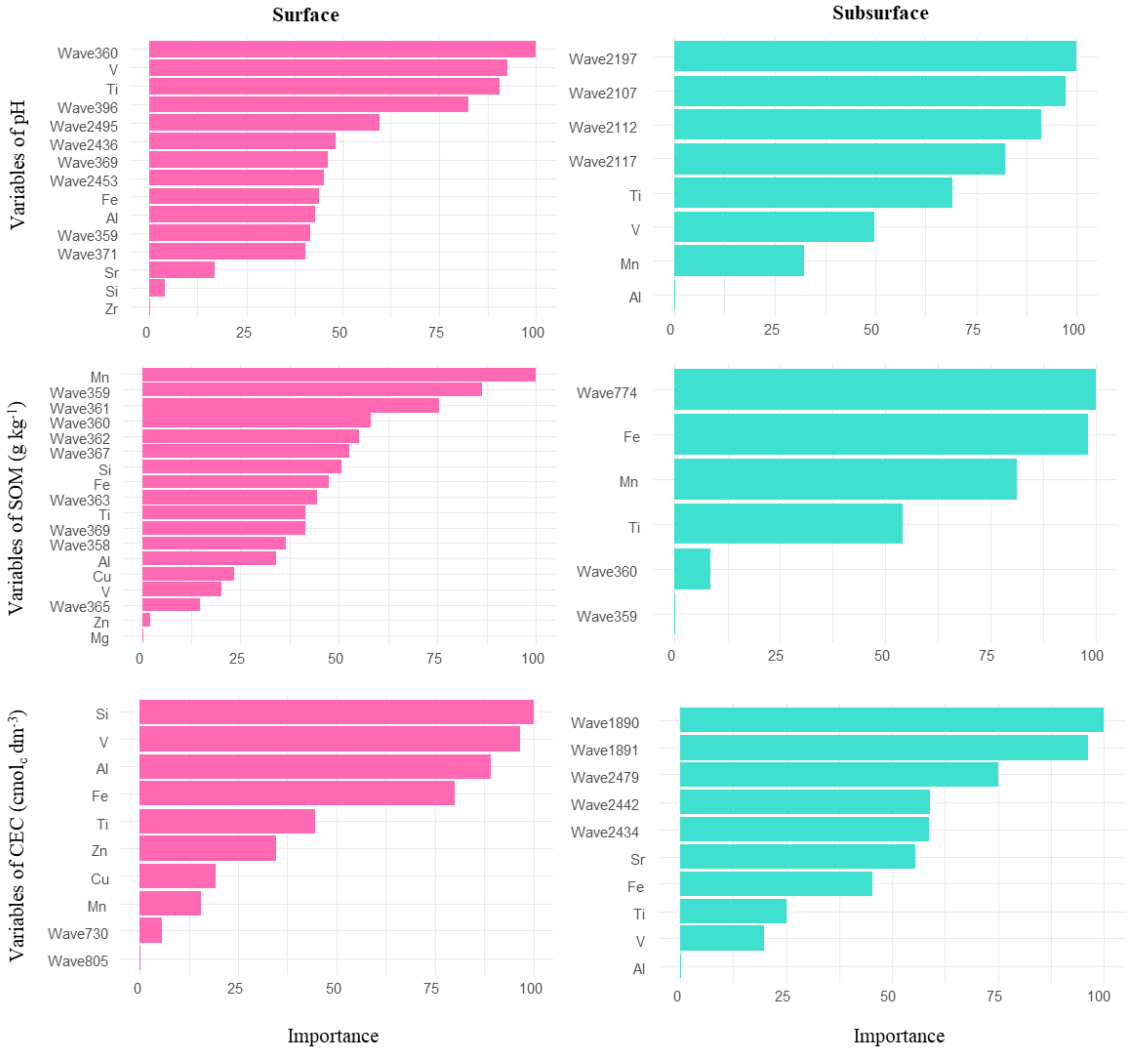

Figure 6 - Importance of the variables in the model generated from the fusion of pXRF and Vis-NIR for the attributes soil pH, soil organic matter (SOM), and cation exchange capacity (CEC).

In order to predict SOM, it was observed the common presence of the elements Mn, Ti, Fe, V, Si, Cu, Al, Zn, and Mg and wavelengths in the 358, 359, 360, 361, 362, 363, 365, 367, 369 and774 nm range. These elements can be explained by the association between SOM and minerals. The importance of Mg, Al, Ca, Mn, Fe, and Zn contents to SOM can be associated by the potential of these elements to form complexes within the SOM system. Specifically, SOM can be stabilized by interaction with the mineral phase (regulating the mineralization of SOM) (Feller and Beare, 1997). In tropical regions, this is attributed to Fe and Al oxides. The element iron is more prevalent in tropical soil (Akerman et al., 2021), furthermore, the

importance of Al and Fe may be attributed to the adsorption of SOM to amorphous substances, such as Al and Fe hydroxides (Saidy et al., 2012). In this regard, Kaiser and Guggenberger (2000) demonstrated that metal oxides are the most effective sorbents of dissolved SOM. Soil with a high humus content is conducive to the process of Fe and Al accumulation, which is a further consequence of this process (Grison et al., 2021). Ramos et al. (2018) observed a positive correlation between total organic carbon stock and the concentration of Ca^{2+} and Mg^{2+}. To predict the CEC, the presence of the elements (Si, V, Al, Fe, Ti, Zn, Cu, Mn and Sr) and wavelengths in the visible and infrared (730, 805, 1890, 1891, 2434, 2442, and 2479 nm) range were observed. This may be attributed to the CEC representing the ability to bind exchangeable cations, including Cu, Zn, and Fe (Sharma et al., 2014). Furthermore, Sharma et al., (2015) demonstrated that the most significant contributors to the CEC predictive model were V and Cu for soils from the USA. The observed wavelengths for CEC can be associated with soil minerals, such as Fe-oxide minerals and illite, which confer CEC. This is because the mineralogical nature addressed through spectral measurement focuses on several bands (e.g., 350-400, ∼1900, ∼2200 and 2450-2500 nm) related to soil minerals (Fang et al., 2018, Souza et al., 2018).

3.4 Conclusions

The RF model was employed to predict the soil pH, soil organic matter (SOM), and cation exchange capacity (CEC) of the surface and subsurface samples of soils. This was accomplished by utilizing data from individual and concatenated proximal sensors (Vis-NIR and pXRF). The pXRF data achieved the most accurate predictions of soil pH for both depths, while the predictions of SOM and CEC were achieved by pXRF only on the surface and by Vis-NIR data in the subsurface. Although only one sensor delivered optimal predictions, the integration of data from multiple sensors yielded superior outcomes compared to the use of individual sensors for soil pH and SOM (only for subsurface). However, for CEC, the combination of data sensors yielded inferior values compared to a unique sensor. The Boruta and PCA variable selection methods yielded favorable results; however, Boruta exhibited superior results in terms of RPD and RPIQ, as well as lower RMSE, suggesting that Boruta is a more effective variable selection algorithm for pH soil and SOM. The findings of this research demonstrate the advancement of using proximal sensor data to predict the soil fertility of Brazilian tropical soils under native vegetation, a subject that has been rarely studied, especially

regarding the Amazon biome. Accurate prediction of soil fertility is imperative, as it can lead to a reduction in the generation of toxic waste in laboratories, decreased costs of fertilizers and soil amendments, and ultimately, enhanced soil conservation in tropical regions.

Acknowledgment: The authors would like to thank the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior for the scholarship (CAPES - 88887.598616/2021-00), Conselho Nacional de Desenvolvimento Científico e Tecnológico for provide research grants (CNPq # 310283/2019-1 P.W.M.S-F, CNPq #304560/2023-5 S.J.R, and CNPq #405089/2021-0 A.R.F).

Data availability statement: Data is available upon request to the corresponding author.

Conflict of Interest: The authors declare that they have no conflict of interest.

References

Abdi, H., Williams, L.J., 2010. Principal component analysis. Wiley Interdiscip. Rev. Comput. Stat. https://doi.org/10.1002/wics.101

Akerman, A., Oliva, P., Poitrasson, F., Boaventura, G. R., da Silva Souza, V., & Seyler, P. (2021). Impact of deforestation on soil iron chemistry and isotope signatures in Amazonia. *Chemical Geology*, *577*. https://doi.org/10.1016/j.chemgeo.2020.120048

Al Masmoudi, Y., Bouslihim, Y., Doumali, K., Hssaini, L., Ibno Namr, K., 2022. Use of machine learning in Moroccan soil fertility prediction as an alternative to laborious analyses. Model. Earth Syst. Environ. 8. https://doi.org/10.1007/s40808-021-01329-8

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

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. https://doi.org/10.1016/j.geoderma.2019.113960

Arjasakusuma, S., Kusuma, S.S., Phinn, S., 2020. Evaluating variable selection and machine learning algorithms for estimating forest heights by combining lidar and hyperspectral data. ISPRS Int. J. Geo-Information 9. https://doi.org/10.3390/ijgi9090507

Bellon-Maurel, V., Fernandez-Ahumada, E., Palagos, B., Roger, J.M., McBratney, A., 2010. Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy. TrAC - Trends Anal. Chem. https://doi.org/10.1016/j.trac.2010.05.006

Ben Dor, E., Ong, C., Lau, I.C., 2015. Reflectance measurements of soils in the laboratory: Standards and protocols. Geoderma 245–246. https://doi.org/10.1016/j.geoderma.2015.01.002

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, 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. https://doi.org/10.1016/j.catena.2020.105003

BRASIL, M. do M.A., 2012. Código Florestal - LEI Nº 12.651 de 25 de maio de 2012. Diário

Of. da União.

Breiman, L., 2001. Random forests. Mach. Learn. 45. https://doi.org/10.1023/A:1010933404324

Demattê, J.A.M., Araújo, S.R., Fiorio, P.R., Fongaro, C.T., Nanni, M.R., 2015. Espectroscopia VIS-NIR-SWIR na avaliação de solos ao longo de uma topossequência em Piracicaba (SP). Rev. Cienc. Agron. 46. https://doi.org/10.5935/1806-6690.20150054

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

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. J. Spectrosc. https://doi.org/10.1155/2018/3168974

Feller, C., Beare, M.H., 1997. Physical control of soil organic matter dynamics in the tropics, in: Geoderma. https://doi.org/10.1016/S0016-7061(97)00039-6

Ge, Y., Morgan, C.L.S., Wijewardane, N.K., 2020. Visible and near-infrared reflectance spectroscopy analysis of soils. Soil Sci. Soc. Am. J. 84. https://doi.org/10.1002/saj2.20158

Gonçalves, D. A. M., da Silveira Pereira, W. V., Johannesson, K. H., Pérez, D. V., Guilherme, L. R. G., & Fernandes, A. R. (2022). Geochemical Background for Potentially Toxic Elements in Forested Soils of the State of Pará, Brazilian Amazon. *Minerals*, *12*(6). https://doi.org/10.3390/min12060674

Gozukara, G., Altunbas, S., Dengiz, O., Adak, A., 2022a. Assessing the effect of soil to water ratios and sampling strategies on the prediction of EC and pH using pXRF and Vis-NIR spectra. Comput. Electron. Agric. 203. https://doi.org/10.1016/j.compag.2022.107459

Gozukara, G., Zhang, Y., Hartemink, A.E., 2022b. Using pXRF and vis-NIR spectra for predicting properties of soils developed in loess. Pedosphere 32. https://doi.org/10.1016/S1002-0160(21)60092-9

Grison, H., Petrovsky, E., & Hanzlikova, H. (2021). Assessing anthropogenic contribution in highly magnetic forest soils developed on basalts using magnetic susceptibility and concentration of elements. *Catena*, *206*. https://doi.org/10.1016/j.catena.2021.105480

Helfer, G.A., Victória Barbosa, J.L., Santos, R. dos, da Costa, A. Ben, 2020. A computational model for soil fertility prediction in ubiquitous agriculture. Comput. Electron. Agric. 175. https://doi.org/10.1016/j.compag.2020.105602

Hounkpatin, K.O.L., Bossa, A.Y., Yira, Y., Igue, M.A., Sinsin, B.A., 2022. Assessment of the soil fertility status in Benin (West Africa) – Digital soil mapping using machine learning. Geoderma Reg. 28. https://doi.org/10.1016/j.geodrs.2021.e00444

Huang, Y. C., Huang, C. Y., Minasny, B., Chen, Z. S., & Hseu, Z. Y. (2023). Using pXRF and Vis-NIR for characterizing diagnostic horizons of fine-textured podzolic soils in subtropical forests. *Geoderma, 437*. https://doi.org/10.1016/j.geoderma.2023.116582

IBGE, 2024. Instituto Brasileiro de Geografia e Estatística [WWW Document].

IUSS Working Group, 2015. World reference base for soil resources 2014 International soil classification system, World Soil Resources Reports No. 106.

Janovský, M. P., Karlík, P., Horák, J., Šmejda, L., Asare Opare, M., Beneš, J., & Hejcman, M. (2020). Historical land-use in an abandoned mountain village in the Czech Republic is reflected by the Mg, P, K, Ca, V, Cr, Mn, Fe, Ni, Cu, Zn, Rb, Zr, and Sr content in contemporary soils. *Catena, 187*. https://doi.org/10.1016/j.catena.2019.104347

Jiang, Q., Li, Q., Wang, X., Wu, Y., Yang, X., Liu, F., 2017. Estimation of soil organic carbon and total nitrogen in different soil layers using VNIR spectroscopy: Effects of spiking on model applicability. Geoderma 293. https://doi.org/10.1016/j.geoderma.2017.01.030

Kaiser, K., Guggenberger, G., 2000. The role of DOM sorption to mineral surfaces in the preservation of organic matter in soils, in: Organic Geochemistry. https://doi.org/10.1016/S0146-6380(00)00046-2

Kouadio, L., Deo, R.C., Byrareddy, V., Adamowski, J.F., Mushtaq, S., Phuong Nguyen, V., 2018. Artificial intelligence approach for the prediction of Robusta coffee yield using soil fertility properties. Comput. Electron. Agric. 155. https://doi.org/10.1016/j.compag.2018.10.014

Kuhn, M., 2008. Building predictive models in R using the caret package. J. Stat. Softw. 28. https://doi.org/10.18637/jss.v028.i05

Kursa, M.B., Rudnicki, W.R., 2010. Feature selection with the boruta package. J. Stat. Softw. 36. https://doi.org/10.18637/jss.v036.i11

Li, X., Pan, W., Li, D., Gao, W., Zeng, R., Zheng, G., Cai, K., Zeng, Y., Jiang, C., 2024. Can fusion of vis-NIR and MIR spectra at three levels improve the prediction accuracy of soil nutrients? Geoderma 441, 116754. https://doi.org/10.1016/j.geoderma.2023.116754

Liaw, A., Wiener, M., 2002. Classification and Regression by randomForest. R News 2.

Ling, L., Fu, Y., Jeewani, P. H., Tang, C., Pan, S., Reid, B. J., Gunina, A., Li, Y., Li, Y., Cai, Y., Kuzyakov, Y., Li, Y., Su, W. qin, Singh, B. P., Luo, Y., & Xu, J. 2021. Organic matter chemistry and bacterial community structure regulate decomposition processes in post-fire forest soils. Soil Biology and Biochemistry, *160*. https://doi.org/10.1016/j.soilbio.2021.108311

Liu, Y., Wang, C., Xiao, C., Shang, K., Zhang, Y., Pan, X., 2021. Prediction of multiple soil fertility parameters using VisNIR spectroscopy and PXRF spectrometry. Soil Sci. Soc. Am. J. 85. https://doi.org/10.1002/saj2.20223

Liu, Y., Xie, X., Wang, M., Zhao, Q., Pan, X., 2018. Removing the Effects of Iron Oxides from Vis-NIR Spectra for Soil Organic Matter Prediction. Soil Sci. Soc. Am. J. 82. https://doi.org/10.2136/sssaj2017.07.0235

Lopes, A.S., Guimarães Guilherme, L.R., 2016. A career perspective on soil management in the Cerrado region of Brazil, in: Advances in Agronomy. https://doi.org/10.1016/bs.agron.2015.12.004

Mancini, M., Weindorf, D.C., Silva, S.H.G., Chakraborty, S., Teixeira, A.F. dos 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. https://doi.org/10.1016/j.geoderma.2019.113885

Moreira, L. J. D., da Silva, E. B., Fontes, M. P. F., Liu, X., & Ma, L. Q. (2018). Speciation, bioaccessibility and potential risk of chromium in Amazon forest soils. *Environmental Pollution*, *239*. https://doi.org/10.1016/j.envpol.2018.04.025

Nunes, S., Gardner, T., Barlow, J., Martins, H., Salomão, R., Monteiro, D., Souza, C., 2016. Compensating for past deforestation: Assessing the legal forest surplus and deficit of the state of Pará, eastern Amazonia. Land use policy 57. https://doi.org/10.1016/j.landusepol.2016.04.022

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. https://doi.org/10.1016/j.geoderma.2016.05.005

Osman, K. T. (2013). Organic Matter of Forest Soils. In *Forest Soils*. https://doi.org/10.1007/978-3-319-02541-4_4

Pavão, Q. S., Ribeiro, P. G., Maciel, G. P., Silva, S. H. G., Araújo, S. R., Fernandes, A. R., Demattê, J. A. M., Souza Filho, P. W. M. e, & Ramos, S. J. (2024). Texture prediction of natural soils in the Brazilian Amazon through proximal sensors. *Geoderma Regional*, *37*, e00813. https://doi.org/10.1016/j.geodrs.2024.e00813

Prietzel, J., & Christophel, D. (2014). Organic carbon stocks in forest soils of the German alps. *Geoderma*, *221–222*. https://doi.org/10.1016/j.geoderma.2014.01.021

R Core Team, 2023. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.

Ramos, F.T., Dores, E.F. de C., Weber, O.L. do. S., Beber, D.C., Campelo, J.H., Maia, J.C. d.

S., 2018. Soil organic matter doubles the cation exchange capacity of tropical soil under no-till farming in Brazil. J. Sci. Food Agric. 98. https://doi.org/10.1002/jsfa.8881

Saidy, A.R., Smernik, R.J., Baldock, J.A., Kaiser, K., Sanderman, J., Macdonald, L.M., 2012. Effects of clay mineralogy and hydrous iron oxides on labile organic carbon stabilisation. Geoderma 173–174. https://doi.org/10.1016/j.geoderma.2011.12.030

Sanchez-Pinto, L.N., Venable, L.R., Fahrenbach, J., Churpek, M.M., 2018. Comparison of variable selection methods for clinical predictive modeling. Int. J. Med. Inform. 116. https://doi.org/10.1016/j.ijmedinf.2018.05.006

Santos, H.G., Jacomine, P.K.T., Anjos, L.H.C., Oliveira, V.A., Lumbreras, J.F., Coelho, M.R., Almeida, J.A., Araujo Filho, J.C., Oliveira, J.B., Cunha, T.J.F., 2018. Sistema Brasileiro de Classificação de Solos. 5th. Embrapa Solos, Brasília revista e ampliada ed.

Schaetzl, R.J., Anderson, S., 2005. Soil: Genesis and Geomorphology, 1st ed. Cambridge University Press, New York.

Sharma, A., Weindorf, D.C., Man, T., Aldabaa, A.A.A., Chakraborty, S., 2014. Characterizing soils via portable X-ray fluorescence spectrometer: 3. Soil reaction (pH). Geoderma 232–234. https://doi.org/10.1016/j.geoderma.2014.05.005

Sharma, A., Weindorf, D.C., Wang, D.D., Chakraborty, S., 2015. Characterizing soils via portable X-ray fluorescence spectrometer: 4. Cation exchange capacity (CEC). Geoderma 239. https://doi.org/10.1016/j.geoderma.2014.10.001

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. (2023). Sensing technologies for characterizing and monitoring soil functions: A review. In *Advances in Agronomy* (Vol.

177). https://doi.org/10.1016/bs.agron.2022.08.002

Souza, E.S. de, Fernandes, A.R., De Souza Braz, A.M., Oliveira, F.J. de, Alleoni, L.R.F., Campos, M.C.C., 2018. Physical, chemical, and mineralogical attributes of a representative group of soils from the eastern Amazon region in Brazil. SOIL 4, 195–212. https://doi.org/10.5194/soil-4-195-2018

Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. In *Expert Systems with Applications* (Vol. 134). https://doi.org/10.1016/j.eswa.2019.05.028

Stenberg, B., Viscarra Rossel, R.A., Mouazen, A.M., Wetterlind, J., 2010. Visible and Near Infrared Spectroscopy in Soil Science, Advances in Agronomy. https://doi.org/10.1016/S0065-2113(10)07005-7

Sujatha, M., Jaidhar, C.., 2024. Machine learning-based approaches to enhance the soil fertility—A review. Expert Syst. Appl. 240.

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. J. South Am. Earth Sci. 116. https://doi.org/10.1016/j.jsames.2022.103873

Teixeira, A.F. dos S., Weindorf, D.C., Silva, S.H.G., Guilherme, L.R.G., Curi, N., 2018. Portable x-ray fluorescence (pXRF) spectrometry applied to the prediction of chemical attributes in inceptisols under different land use. Cienc. e Agrotecnologia 42. https://doi.org/10.1590/1413-70542018425017518

Teixeira, P.C., Donagemma, G.K., Fontana, A., Teixeira, W.G., 2017. Manual de métodos de análise de solo, Embrapa.

Viscarra Rossel, R.A., Adamchuk, V.I., Sudduth, K.A., McKenzie, N.J., Lobsey, C., 2011. Proximal Soil Sensing. An Effective Approach for Soil Measurements in Space and Time. Adv. Agron. 113. https://doi.org/10.1016/B978-0-12-386473-4.00010-5

Viscarra Rossel, R.A., Behrens, T., Ben-Dor, E., Brown, D.J., Demattê, J.A.M., Shepherd, K.D., Shi, Z., Stenberg, B., Stevens, A., Adamchuk, V., Aïchi, H., Barthès, B.G., Bartholomeus, H.M., Bayer, A.D., Bernoux, M., Böttcher, K., Brodský, L., Du, C.W., Chappell, A., Fouad, Y., Genot, V., Gomez, C., Grunwald, S., Gubler, A., Guerrero, C., Hedley, C.B., Knadel, M., Morrás, H.J.M., Nocita, M., Ramirez-Lopez, L., Roudier, P., Campos, E.M.R., Sanborn, P., Sellitto, V.M., Sudduth, K.A., Rawlins, B.G., Walter, C., Winowiecki, L.A., Hong, S.Y., Ji, W., 2016. A global spectral library to characterize the world's soil. Earth-Science Rev. https://doi.org/10.1016/j.earscirev.2016.01.012

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

Wan, M., Hu, W., Qu, M., Li, W., Zhang, C., Kang, J., Hong, Y., Chen, Y., Huang, B., 2020. Rapid estimation of soil cation exchange capacity through sensor data fusion of portable XRF spectrometry and Vis-NIR spectroscopy. Geoderma 363. https://doi.org/10.1016/j.geoderma.2019.114163

Weber, O.B., da Silva, M.C.B., da Silva, C.F., de Sousa, J.A., Taniguch, C.A.K., dos Santos Garruti, D., Romero, R.E., 2020. Biological and chemical attributes of soils under forest species in Northeast Brazil. J. For. Res. 31. https://doi.org/10.1007/s11676-019-00982-1

Weindorf, D.C., Bakr, N., Zhu, Y., 2014. Advances in portable X-ray fluorescence (PXRF)

for environmental, pedological, and agronomic applications. Adv. Agron. 128, 1–45. https://doi.org/10.1016/B978-0-12-802139-2.00001-9

Xu, Z., Kusner, M. J., Weinberger, K. Q., & Chen, M. (2013). Cost-sensitive tree of classifiers. *30th International Conference on Machine Learning, ICML 2013, PART 1*.

Zhao, L., Tan, K., Wang, X., Ding, J., Liu, Z., Ma, H., Han, B., 2023. Hyperspectral Feature Selection for SOM Prediction Using Deep Reinforcement Learning and Multiple Subset Evaluation Strategies. Remote Sens. 15. https://doi.org/10.3390/rs15010127

4. CONCLUSÕES GERAIS

A avaliação física e química dos solos é um processo básico para o manejo, conservação e preservação dos solos, tanto em áreas agrícolas como as áreas naturais. Os métodos de análises tradicionais são consolidados e confiáveis, porém ainda existem algumas desvantagens crucias como o custo, tempo e geração de resíduos tóxicos. Esses desafios precisam ser superados, pois a sociedade almeja cada vez mais um planeta limpo para as futuras gerações, sendo assim, em todas as áreas, inclusive para a ciência dos solos devem ser adotados métodos em conformidade com os objetivos do desenvolvimento sustentável. Os achados dessa tese mostraram as análises pela espectrometria portátil XRF e Vis-NIR nos solos naturais da Amazônia são úteis para a predição da sua textura e fertilidade. Os resultados corroboram para que, cada vez mais, essas técnicas venham serem aprimoradas e auxiliem nas análises rápidas e menos dispendiosas, preenchendo uma lacuna na carência de laboratórios tradicionais equipados para fornecer essas análises em tempo hábil e com fácil acesso. Assim, esse trabalho constitui-se um avanço nas pesquisas com uso de métodos alternativos eficientes para análises de solos.

Supplementary material

Table S1- Recovery of elements based on the certified values of the soil check samples, 2710a and 2711a, in percentages (%).

Element	Mg	Al	Si	P	Ca	Ti	V	Mn	Fe	Cu	Zn	Sr
Check sample	--	--	--	96.7	101.6	--	--	99.4	99.6	101.8	107.9	100.7
2710a	113.5	93.5	93.7	33.8	87.5	93.4	--	93.9	98.3	96.6	101.4	101.1
2711a	122.8	93.1	95.7	55.9	98.8	97.8	114.2	104.2	99.9	97.9	104.5	97.4

Figure S1- Spearman's correlation of soil fertility attributes and portable X-ray fluorescence spectrometry (pXRF) data from samples of the surface and subsurface.

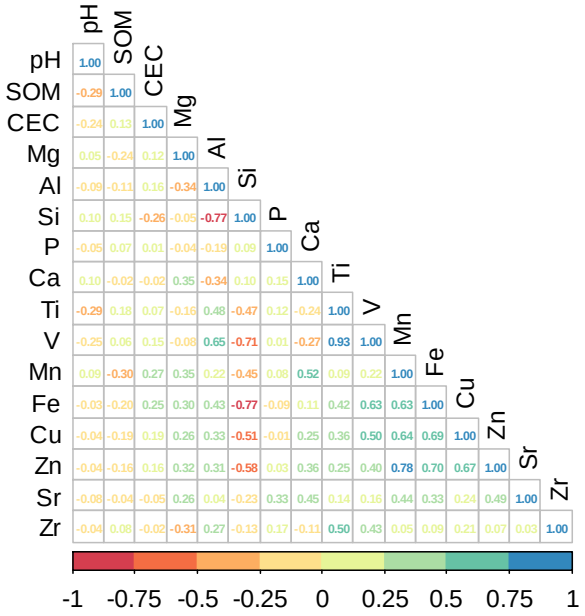

Table S2 - Descriptive statistics for soil pH, SOM (g kg^{-1}) and CEC ($cmol_c$ dm^{-3}) in the surface and subsurface (n= 236)

Depth	Variable	Min[1]	Max[2]	Mean	Median	SD[3]	CV[4]
Surface	pH	3.26	6.91	4.51	4.42	0.70	15.66
	SOM[5]	2.30	56.11	12.54	10.11	8.25	65.80
	CEC[6]	2.35	183.23	25.25	12.68	29.64	117.43
Subsurface	pH	2.46	6.94	4.81	4.78	0.64	13.30
	SOM	0.90	66.96	6.53	3.93	8.06	123.32
	CEC	1.40	29.92	6.33	4.84	5.02	79.23

[1]Minimum values, [2]Maximum values, [3]Standard deviation, [4]Coefficient of variation (%), [5]Soil organic matter, [6]Cation exchange capacity

Table S3 – Principal components (PCs) extracted by PCA and employed to predict pH, soil organic matter (SOM), and cation exchange capacity (CEC).

Sensor	Soil attribute	Depth	Numbers of PCs
pXRF	pH	Surface	11
		Subsurface	10
	SOM	Surface	11
		Subsurface	10
	CEC	Surface	11
		Subsurface	10
Vis-NIR	pH	Surface	5
		Subsurface	6
	SOM	Surface	5
		Subsurface	6
	CEC	Surface	5
		Subsurface	6
pXRF + Vis-NIR	pH	Surface	6
		Subsurface	7
	SOM	Surface	6
		Subsurface	7
	CEC	Surface	6
		Subsurface	7

Table S4 – Variables selected by Boruta employed to predict pH, soil organic matter (SOM), and cation exchange capacity (CEC).

Sensor	Soil attribute	Depth	Numbers of variables	Selected Variables
pXRF	pH	Surface	7	Zr+Sr+Fe+V+Ti+Si+Al
		Subsurface	4	Mn+V+Ti+Al
	SOM	Surface	9	Zn+Cu+Fe+Mn+V+Ti+Si+Al+Mg
		Subsurface	3	Fe+Mn+Ti
	CEC	Surface	8	Zn+Cu+Fe+Mn+V+Ti+Si+Al
		Subsurface	5	Sr+Fe+V+Ti+Al
Vis-NIR	pH	Surface	8	Wave359+Wave360+Wave369+Wave371+Wave396+ Wave2436+Wave2453+Wave2495
		Subsurface	4	Wave2107+Wave2112+Wave2117+Wave2197
	SOM	Surface	9	Wave358+Wave359+Wave360+Wave361+Wave362+Wave363+Wave365+Wave367+Wave369
		Subsurface	3	Wave359+Wave360+Wave774
	CEC	Surface	2	Wave730+Wave805
		Subsurface	5	Wave1890+Wave1891+Wave2434+Wave2442+Wave2479
pXRF + Vis-NIR	pH	Surface	15	Zr+Sr+Fe+V+Ti+Si+Al+Wave359+Wave360+Wave369+Wave371+Wave396+Wave2436+Wave2453+Wave2495
		Subsurface	8	Mn+V+Ti+Al+Wave2107+Wave2112+Wave2117+Wave2197
	SOM	Surface	18	Zn+Cu+Fe+Mn+V+Ti+Si+Al+Mg+Wave358+Wave359+Wave360+Wave361+Wave362+Wave363+Wave365+Wave367+Wave369
		Subsurface	6	Fe+Mn+Ti+Wave359+Wave360+Wave774
	CEC	Surface	10	Zn+Cu+Fe+Mn+V+Ti+Si+Al+Wave730+Wave805
		Subsurface	10	Sr+Fe+V+Ti+Al+Wave1890+Wave1891+Wave2434+Wave2442+Wave2479