RENATA ANDRADE

APPLICATION OF PROXIMAL SENSORS FOR THE PREDICTION OF SOIL CLASSES AND ATTRIBUTES IN BRAZIL

LAVRAS-MG

2022

RENATA ANDRADE

APPLICATION OF PROXIMAL SENSORS FOR THE PREDICTION OF SOIL CLASSES AND ATTRIBUTES IN BRAZIL

APLICAÇÃO DE SENSORES PRÓXIMOS PARA A PREDIÇÃO DE CLASSES E ATRIBUTOS DO SOLO NO BRASIL

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

Prof. Dr. Sérgio Henrique Godinho Silva

Orientador

LAVRAS-MG

2022

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

Andrade, Renata.
Application of proximal sensors for the prediction of soil classes and attributes in Brazil / Renata Andrade. - 2022.
106 p.: il.

Orientador(a): Sérgio Henrique Godinho Silva.

Tese (doutorado acadêmico) - Universidade Federal de Lavras, 2022.
Bibliografia.

1. Prediction models; Machine learning; Soil conservation. 2. PXRF; Sustainable soil management. I. Silva, Sérgio Henrique Godinho. II.

RENATA ANDRADE

APPLICATION OF PROXIMAL SENSORS FOR THE PREDICTION OF SOIL CLASSES AND ATTRIBUTES IN BRAZIL

APLICAÇÃO DE SENSORES PRÓXIMOS PARA A PREDIÇÃO DE CLASSES E ATRIBUTOS DO SOLO NO BRASIL

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

APROVADA em 23 de Fevereiro de 2022.

Dr. Nilton Curi UFLA

Dra. Giovana Clarice Poggere UTFPR

Dr. Salvador Francisco Acunã Guzman UPRM

Dr. Junior Cesar Avanzi UFLA

Prof. Dr. Sérgio Henrique Godinho Silva
Orientador

LAVRAS-MG

2022

Aos meus amigos e irmãos em Cristo, Lidiana Pereira Barrozo Guimarães e Jean Kellison Guimarães dos Santos, por terem sido a expressão das mãos cuidadosas de Jesus em minha vida.

Dedico

AGRADECIMENTOS

A Deus por me conduzir para uma cidade tão abençoada e me permitir continuar os estudos.

A Jesus por todo amor e salvação.

Ao meu amigo, Espírito Santo, por me consolar e ensinar o bom caminho.

À minha família por tudo que fizeram por mim.

À Universidade Federal de Lavras e, principalmente, ao Departamento de Ciência do Solo por contribuírem para a minha formação acadêmica e profissional. O presente trabalho foi realizado com o apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), e Fundação de Amparo à Pesquisa de Minas Gerais (FAPEMIG).

Ao professor orientador Dr. Sérgio Henrique Godinho Silva por todo apoio e orientação. Muito obrigada professor pelo altruísmo, paciência, conselhos, amizade, bondade, benevolência, empatia e dedicação. Sem a sua orientação, eu não teria chegado até o final dessa jornada.

Aos professores Bruno Teixeira Ribeiro, Bruno Montoani Silva, Geraldo César de Oliveira, Eduardo da Costa Severiano e Nilton Curi pelo apoio e ajuda em momentos essenciais da minha carreira acadêmica.

Aos companheiros de laboratório e amigos da Ufla, Dirce de Cássia, Bethânia Mansur, Geila Carvalho, Álvaro José, Anita Fernanda, Fernanda Magno, Luiza Pierangeli, Marcelo Mancini e tantos outros que trilharam essa jornada comigo.

Às companheiras de apartamento, Adriane Braga, Letícia Vaz e Vanessa Ferreira. Muito obrigada por todos os bons momentos e a amizade que ficarão para sempre na minha memória.

À toda a igreja Vale Church por me acolher com tanto amor e cuidado. Aos pastores Luiz Cláudio e Aline Borges por toda dedicação, conselho e apoio. Aos meus irmãos de Grupo Vida, Talita e Gabriel Zugaiar que me acolheram incondicionalmente. Nós seremos para sempre unidos pelo elo perfeito que é o amor fraternal de Jesus. Eu amo muito vocês.

A todos que contribuíram para a minha formação profissional e crescimento na fé,

meus sinceros agradecimentos!

“Prepara-se o cavalo para o dia da batalha, mas o Senhor é que dá a vitória.”

Provérbios 21:31

ABSTRACT

Knowledge of soil properties makes a significant contribution towards sustainable soil management, decision making, and soil conservation. For that, a quick, environmentally friendly, non-invasive, cost-effective, and reliable method for soil properties assessment is desirable. As such, this study used portable X-ray fluorescence (pXRF) spectrometry, visible near-infrared spectroscopy (Vis-NIR) and NixPro^{TM} color sensor data to characterize 1019 Brazilian tropical soils samples, exploring the ability of six machine learning algorithms [ordinary least squares regression (OLS), Support Vector Machine with Linear Kernel (SVMLK), Cubist Regression (CR), XGBoost (XGB), Artificial Neural Network (ANN), and Random Forest (RF)] for prediction of different soil properties. The soil samples were collected in both surface and subsurface horizons of different soil classes, under several land uses, and with varying parent materials. Numerical prediction models were built for surface, and subsurface horizons separately and combined for the following soil properties: total nitrogen (TN), cation exchange capacity (CTC), soil organic matter (SOM), and soil texture (total sand, silt, clay, coarse sand, and fine sand contents). The study also encompasses the categorical prediction of properties, such as soil taxonomic classification at order and suborder levels, and soil textural classes (complete and simplified textural triangles). The NixPro^{TM} color sensor data were scanned under both dry and moist conditions. Four preprocessing methods were applied on the raw Vis-NIR spectra: first derivative, absorbance, smoothed, and binning. Samples were randomly separated into 70% for modeling and 30% for validation. The best approach varied according to the predicted soil property. However, pXRF data were a key information for the accuracy of soil properties prediction, followed by Vis-NIR spectra, and then, NixPro^{TM} color data. The results showed the increase in accuracy via fusion of proximal sensors data for soil properties prediction: TN (R^2 = 0.50), CEC (0.75), SOM (0.56), total sand (0.84), silt (0.83), clay (0.90), coarse sand (0.87), fine sand (0.82), soil order (overall accuracy = 81.19%), soil suborder (74.35%), Family particle size classes (96.55%), and USDA soil texture triangle (82.75%). The results reported in this study for the tropical soils represent alternatives for reducing costs and time needed for assessing such soil properties data, supporting agronomic and environmental strategies.

Keywords: PXRF. Vis-NIR. NixPro^{TM}. Machine learning. Prediction models. Sustainable soil management. Soil conservation.

RESUMO

O conhecimento das propriedades do solo contribui significativamente para o manejo sustentável do solo, tomada de decisões e conservação do solo. Para isso, é desejável um método rápido, ecologicamente correto, não invasivo, econômico e confiável para avaliação das propriedades do solo. Assim, este estudo utilizou a fluorescência de raios X (pXRF) portátil, espectroscopia no infravermelho próximo e visível (Vis-NIR) e dados do sensor de cores NixPro^{TM} para caracterizar 1019 amostras de solos tropicais brasileiros, explorando a capacidade de seis algoritmos de aprendizado de máquina [ordinary least squares regression (OLS), Support Vector Machine with Linear Kernel (SVMLK), Cubist Regression (CR), XGBoost (XGB), Artificial Neural Network (ANN), and Random Forest (RF))] para predição de diferentes propriedades do solo. As amostras de solo foram coletadas em horizontes superficiais e subsuperficiais de diferentes classes de solos, sob diversos usos da terra e com diferentes materiais de origem. Modelos numéricos de predição foram construídos para horizontes de superfície e subsuperfície separadamente e combinados, para as seguintes propriedades do solo: nitrogênio total (NT), capacidade de troca de cátions (CTC), matéria orgânica do solo (MOS) e textura do solo (areia total, silte, argila, areia grossa e areia fina). Este estudo também abrange a predição categórica de propriedades, como classificação taxonômica do solo em níveis de ordem e subordem e classes texturais do solo (triângulos texturais). Os dados do sensor de cores NixPro^{TM} foram adquiridos em condições secas e úmidas. Quatro métodos de pré-processamento foram aplicados nos espectros de Vis-NIR: primeira derivada, absorbância, smoothed e binning. As amostras foram separadas aleatoriamente em 70% para modelagem e 30% para validação. A melhor abordagem variou de acordo com a propriedade do solo a ser predita. No entanto, os dados pXRF foram uma informação chave para a acurácia da predição das propriedades do solo, seguidos pelos espectros do Vis-NIR e, em seguida, os dados de cor do NixPro^{TM}. Os resultados mostraram o aumento da acurácia via fusão de dados de sensores proximais para predição das propriedades do solo: TN (R^2 = 0,50), CEC (0,75), SOM (0,56), areia total (0,84), silte (0,83), argila (0,90), areia grossa (0,87), areia fina (0,82), ordem do solo (precisão geral = 81,19%), subordem do solo (74,35%), *Family particle size classes* (96,55%) e triângulo de textura do solo do *USDA* (82,75%). Os resultados reportados neste estudo para os solos tropicais representam alternativas para redução de custos e tempo necessário para a avaliação dessas propriedades do solo, subsidiando estratégias agronômicas e ambientais.

Palavras-Chave: PXRF. Vis-NIR. NixPro^{TM}. Aprendizado de máquina. Modelos de predição. Manejo sustentável do solo. Conservação do solo.

SUMÁRIO

SECTION 1

General Introduction

Knowledge of soil properties makes a significant contribution towards sustainable soil management, decision making, and soil conservation. For its determination, soil samples have to be collected, submitted to traditional laboratory analyses, which are time-consuming and costly, require chemical reagents, and generate chemical waste, restricting the number of samples that can be processed (BAVER; GARDNER; GARDNER, 1972; BREMNER, 1996; GEE; BAUDER, 1986; NELSON; SOMMERS, 1996). A reliable, novel alternative to traditional approaches is the use of proximal sensors to assess soil properties, such as nutrient contents (FISCHER et al., 2020; PELEGRINO et al., 2021; TOWETT et al., 2020), base saturation (RAWAL et al., 2019), soil organic matter (RAVANSARI et al., 2021), among others. Some works have used proximal sensors to assess other soil aspects, such as its biology (TEIXEIRA et al., 2021), formation processes (STOCKMANN et al., 2016), soil contaminants (FERREIRA et al., 2021; HORTA et al., 2021; JANG, 2010; KIM et al., 2019; TIAN et al., 2018), and determination of heavy metal contents in archeological studies (KENNEDY; KELLOWAY, 2021). Besides optimizing time and reducing costs, proximal sensors also increase the number of samples that can be analyzed, in an environmentally friendly way (ANDRADE et al., 2020a; BENEDET et al., 2020b; FARIA et al., 2020; TEIXEIRA et al., 2020). In this study, three proximal sensors are discussed based on their recent acclaim or novelty: portable X-ray fluorescence (pXRF) spectrometry, visible/near-infrared diffuse reflectance (Vis-NIR) spectroscopy and the NixPro^{TM} color sensor.

The pXRF is the fastest-growing sensor in popularity within the soil science community. In Brazil, its introduction occurred for archeological purposes by (IKEOKA et al., 2012). This method requires minimal sample preparation, and, with adequate calibration, the equipment can be used both in laboratory and in field conditions to analyze elemental contents of soils (from Mg to U) (DIJAIR et al., 2020; RIBEIRO et al., 2017; WEINDORF; BAKR; ZHU, 2014; SILVA et al., 2021). When a sample is irradiated with X-rays, inner shell electrons are displaced. When outer shell electrons cascade down to fill inner orbitals, they give off energy termed fluorescence, which is characteristic of each element, allowing for its identification. The quantification of the element content is related to the intensity of the corresponding fluorescence. Therefore, this device with a single technique can both identify and quantify various elements simultaneously, quickly, without destroying the sample, and without generating any chemical residue (SILVA et al., 2021; WEINDORF; BAKR; ZHU, 2014).

Vis-NIR has the highest number of published studies related to soils among the three mentioned sensors. It requires only a few seconds to measure a soil sample in the visible, near infrared (350-2500 nm) spectral region providing a spectral signature of the analyzed material (CONFORTI et al., 2015; LAZAAR et al., 2021). Quantitative near infrared soil spectral analysis presents more than 2,000 variables, thus requiring techniques to discern the response of soil properties using spectral characteristics. Different approaches have been developed to relate the soil spectrum to soil properties, but usually four methods are considered foremost, *i.e*., Satitzky-Golay first derivative (FD), absorbance (Abs), smoothed (Smo), and binning (Bin) (CONFORTI et al., 2015; STEVENS and RAMIREZ-LOPEZ, 2020), besides raw spectrum.

NixPro^{TM} color sensor has only recently appeared in soil science studies, and hence there is a very limited number of published studies investigating its uses (ANDRADE et al., 2020b; JHA et al., 2021; MANCINI et al., 2020; STIGLITZ et al., 2017; FARIA et al., 2021). It is a portable, rechargeable, extremely fast reading (1-2 seconds) and inexpensive tool that provides color reports in many different numerical color systems, such as RGB, XYZ, CIELAB, LCH, HEX, CMYK, and ACES (STIGLITZ et al., 2016), making soil color determination less subjective (STIGLITZ et al., 2016; MANCINI et al., 2020).

More recently, with the increase in popularity of these proximal sensors, some studies have investigated different approaches to build prediction models aiming to increase their predictive power. One of these approaches is through data fusion, which means gathering different proximal sensor data as explanatory variables. Andrade et al. (2020c), Benedet et al. (2020a, 2020b), Swetha and Chakraborty (2021), Weindorf et al. (2016), Zhang and Hartemink (2020) successfully combined different proximal sensors data resulting in more robust and accurate prediction models. Another approach is through adding auxiliary input data to the explanatory variables. Andrade et al. (2020a) successfully predicted exchangeable Ca^{2+}, Mg^{2+}, and available K^{+} using pXRF plus soil texture as auxiliary input data. Stiglitz et al. (2017) built robust prediction models for soil organic carbon through NixPro^{TM} plus sample depth as auxiliary information.

However, investigations towards data fusion for soil properties prediction in tropical soils are still rare. Moreover, studies across the globe that combine pXRF, Vis-NIR, and NixPro^{TM} data are unknown to date, to the best of the author's knowledge. Such investigations are necessary to reliably guide future predictions on tropical environments. Additionally, considering that the acquisition and maintenance of these sensors are expensive, it is important to evaluate whether combining these sensors provides more robust and accurate

prediction models, supporting their acquisition, compared with using a unique proximal sensor.

Given all the variability that can affect models' accuracy, such as heterogeneity of the analyzed samples, chosen algorithm, and the data preprocessing methods applied, it is necessary to test different approaches for predicting soil properties in tropical regions. Therefore, the aims of this work were:

i) chapter one – to use pXRF data to characterize the Brazilian Coastal Plains (BCP) soils and assess four machine learning algorithms [ordinary least squares regression (OLS), cubist regression (CR), XGBoost (XGB), and random forest (RF)] for prediction of total nitrogen (TN), effective CEC, and SOM using pXRF data;

ii) chapter two - to use NixPro^{TM} colorimetric capacities and pXRF elemental data to characterize seven different soil orders in Brazilian tropical soils and explore the ability of three machine learning algorithms [Support Vector Machine with Linear Kernel (SVMLK), Artificial Neural Network (ANN), and Random Forest (RF)] with and without Principal Component Analysis (PCA) pretreatment for prediction of soil classifications at the order and suborder taxonomic levels for both dry and moist conditions of samples analyzed via NixPro^{TM};

iii) chapter three - to predict soil texture and soil textural classes through the random forest algorithm, evaluating: 1) separately and combined pXRF, Vis-NIR, and NixPro^{TM} data, 2) proximal sensors plus environmental co-variates as explanatory variables, and 3) build prediction models in sub-datasets separated by soil order.

We hypothesize that robust and accurate prediction models will be delivered for the soil properties analyzed by at least one of the abovementioned proximal sensors approaches, even though the dataset presents large variability of soil order, land use, and parent material.

References

ANDRADE, R. et al. Prediction of soil fertility via portable X-ray fluorescence (pXRF) spectrometry and soil texture in the Brazilian Coastal Plains. **Geoderma**, v. 357, p. 113960, jan. 2020a.

ANDRADE, R. et al. Tropical soil order and suborder prediction combining optical and X-ray approaches. **Geoderma Regional**, v. 23, p. e00331, dez. 2020b.

BAVER, L. D.; GARDNER, W. H.; GARDNER, W. R. **Soil physics**. 4.ed. ed. New York: John Wiley & Sons, 1972.

BENEDET, L. et al. Soil subgroup prediction via portable X-ray fluorescence and visible near-infrared spectroscopy. **Geoderma**, v. 365, p. 114212, abr. 2020a.

BENEDET, L. et al. Soil texture prediction using portable X-ray fluorescence spectrometry and visible near-infrared diffuse reflectance spectroscopy. **Geoderma**, v. 376, p. 114553, out. 2020b.

BREMNER, J. M. Nitrogen Total. In: SPARKS, D. L. (Ed.). . **Methods of soil analysis. Part 3. Chemical methods**. [s.l.] Soil Science Society of America, American Society of Agronomy, 1996. p. 1085–1121.

CONFORTI, M. et al. Visible and near infrared spectroscopy for predicting texture in forest soil: an application in southern Italy. **iForest - Biogeosciences and Forestry**, v. 8, n. 3, p. 339–347, 1 jun. 2015.

DIJAIR, T. S. B. et al. Correcting field determination of elemental contents in soils via portable X-ray fluorescence spectrometry. **Ciência e Agrotecnologia**, v. 44, p. e002420, 2020.

FARIA, Á. J. G. DE et al. Soils of the Brazilian Coastal Plains biome: prediction of chemical attributes via portable X-ray fluorescence (pXRF) spectrometry and robust prediction models. **Soil Research**, 2020.

FARIA, A.J.G.; SILVA, S.H.G.; ANDRADE, R.; MANCINI, M.; MELO, L.C.A.; WEINDORF, D.C.; GUILHERME, L.R.G.; CURI, N., 2022. Prediction of soil organic matter content by combining data from Nix Pro^{TM} color sensor and portable X-ray fluorescence spectrometry in tropical soils. Geoderma Reg. 28, e00461.

FERREIRA, G. W. D. et al. Assessment of iron-rich tailings via portable X-ray fluorescence spectrometry: the Mariana dam disaster, southeast Brazil. **Environmental Monitoring and Assessment**, v. 193, n. 4, p. 203, abr. 2021.

FISCHER, S. et al. Soil and farm management effects on yield and nutrient concentrations of food crops in East Africa. **Science of The Total Environment**, v. 716, p. 137078, maio 2020.

GEE, G. W.; BAUDER, J. W. Particle-size analysis. In: **Methods of soil analysis: Part 1 - Physical and mineralogical methods**. 2. ed. [s.l.] Soil Science Society of America, American Society of Agronomy, 1986. v. 5p. 383–411.

HORTA, A. et al. Integrating portable X-ray fluorescence (pXRF) measurement uncertainty for accurate soil contamination mapping. **Geoderma**, v. 382, p. 114712, jan. 2021.

IKEOKA, R. A. et al. PXRF and multivariate statistics analysis of pre-colonial pottery from northeast of Brazil: PXRF and HCA analysis of pre-colonial pottery from Brazil. **X-Ray Spectrometry**, v. 41, n. 1, p. 12–15, jan. 2012.

JANG, M. Application of portable X-ray fluorescence (pXRF) for heavy metal analysis of soils in crop fields near abandoned mine sites. **Environmental Geochemistry and Health**, v. 32, n. 3, p. 207–216, jun. 2010.

JHA, G. et al. Rapid and inexpensive assessment of soil total iron using Nix Pro color sensor. **Agricultural & Environmental Letters**, v. 6, n. 3, jan. 2021.

KENNEDY, S. A.; KELLOWAY, S. J. Heavy Metals in Archaeological Soils: The Application of Portable X-Ray Fluorescence (pXRF) Spectroscopy for Assessing Risk to Human Health at Industrial Sites. **Advances in Archaeological Practice**, v. 9, n. 2, p. 145–159, maio 2021.

KIM, H.-R. et al. Better assessment of the distribution of As and Pb in soils in a former smelting area, using ordinary co-kriging and sequential Gaussian co-simulation of portable X-ray fluorescence (PXRF) and ICP-AES data. **Geoderma**, v. 341, p. 26–38, maio 2019.

LAZAAR, A. et al. The manifestation of VIS-NIRS spectroscopy data to predict and map soil texture in the Triffa plain (Morocco). **Kuwait Journal of Science**, v. 48, n. 1, 2021.

MANCINI, M. et al. Soil parent material prediction for Brazil via proximal soil sensing. **Geoderma Regional**, v. 22, p. e00310, set. 2020.

NELSON, D. W.; SOMMERS, L. E. Total carbon, organic carbon, and organic matter. In: SPARKS, D. L. (Ed.). . **Methods of soil analysis. Part 3. Chemical methods**. [s.l.] Soil Science Society of America, American Society of Agronomy, 1996. p. 961–1010.

PELEGRINO, M. H. P. et al. Prediction of soil nutrient content via pXRF spectrometry and its spatial variation in a highly variable tropical area. **Precision Agriculture**, 11 jun. 2021.

RAVANSARI, R. et al. Rapid PXRF soil organic carbon and organic matter assessment using novel modular radiation detector assembly. **Geoderma**, v. 382, p. 114728, jan. 2021.

RAWAL, A. et al. Determination of base saturation percentage in agricultural soils via portable X-ray fluorescence spectrometer. **Geoderma**, v. 338, p. 375–382, mar. 2019.

RIBEIRO, B. T. et al. Portable X-ray fluorescence (pXRF) applications in tropical Soil Science. **Ciência e Agrotecnologia**, v. 41, n. 3, p. 245–254, jun. 2017.

SILVA, S. H. G. et al. PXRF in tropical soils: methodology, applications, achievements and challenges. In: **Advances in Agronomy**. [s.l.] Elsevier, 2021. v. 167p. 1–62.

STEVENS, A., RAMIREZ-LOPEZ, L., 2020. An introduction to the prospectr package. R package version 0.2.1.

STIGLITZ, R. et al. Evaluation of an inexpensive sensor to measure soil color. **Computers and Electronics in Agriculture**, v. 121, p. 141–148, fev. 2016.

STIGLITZ, R. et al. Using an inexpensive color sensor for rapid assessment of soil organic carbon. **Geoderma**, v. 286, p. 98–103, jan. 2017.

STOCKMANN, U. et al. Utilizing portable X-ray fluorescence spectrometry for in-field investigation of pedogenesis. **Catena**, v. 139, p. 220–231, abr. 2016.

SWETHA, R. K.; CHAKRABORTY, S. Combination of soil texture with Nix color sensor can improve soil organic carbon prediction. **Geoderma**, v. 382, p. 114775, jan. 2021.

TEIXEIRA, A. F. DOS S. et al. Tropical soil pH and sorption complex prediction via portable X-ray fluorescence spectrometry. **Geoderma**, v. 361, p. 114132, mar. 2020.

TEIXEIRA, A. F. DOS S. et al. Soil physicochemical properties and terrain information predict soil enzymes activity in phytophysiognomies of the Quadrilátero Ferrífero region in Brazil. **Catena**, v. 199, p. 105083, abr. 2021.

TIAN, K. et al. In situ investigation of heavy metals at trace concentrations in greenhouse soils via portable X-ray fluorescence spectroscopy. **Environmental Science and Pollution Research**, v. 25, n. 11, p. 11011–11022, abr. 2018.

TOWETT, E. K. et al. Comprehensive nutrient analysis in agricultural organic amendments through non-destructive assays using machine learning. **Plos One**, v. 15, n. 12, p. e0242821, 10 dez. 2020.

WEINDORF, D. C. et al. Simultaneous assessment of key properties of arid soil by combined PXRF and Vis-NIR data: Arid soil assessment by PXRF and Vis-NIR. **European Journal of Soil Science**, v. 67, n. 2, p. 173–183, mar. 2016.

WEINDORF, D. C.; BAKR, N.; ZHU, Y. Advances in portable X-ray fluorescence (PXRF) for environmental, pedological, and agronomic applications. In: **Advances in Agronomy**. [s.l.] Elsevier, 2014. v. 128p. 1–45.

ZHANG, Y.; HARTEMINK, A. E. Data fusion of vis-NIR and PXRF spectra to predict soil physical and chemical properties. **European Journal of Soil Science**, v. 71, n. 3, p. 316–333, maio 2020.

SECTION 2 - ARTICLES

ARTICLE 1 - Assessing models for prediction of some soil chemical properties from portable X-ray fluorescence (pXRF) spectrometry data in Brazilian Coastal Plains.

Article published in Geoderma Journal, v. 357, p. 9, 2020

(https://doi.org/10.1016/j.geoderma.2019.113957)

Renata Andrade[a], Sérgio Henrique Godinho Silva[a], David C. Weindorf[b], Somsubhra Chakraborty[c], Wilson Missina Faria[a], Luiz Felipe Mesquita[d], Luiz Roberto Guimarães Guilherme[a], Nilton Curi[a]

[a]Department of Soil Science, Federal University of Lavras, Lavras, MG, Brazil

[b]Department of Plant and Soil Science, Texas Tech University, Lubbock, TX, USA

[c]Indian Institute of Technology Kharagpur, Kharagpur, India

[d]Suzano Papel e Celulose, Espírito Santo, ES, Brazil

Abstract

Portable X-ray fluorescence (pXRF) spectrometry is becoming increasingly popular for predicting soil properties worldwide. However, there are still very few works on this subject under tropical conditions. Therefore, the objectives of this study were to use pXRF data to characterize the Brazilian Coastal Plains (BCP) soils and assess four machine learning algorithms [ordinary least squares regression (OLS), cubist regression (CR), XGBoost (XGB), and random forest (RF)] for prediction of total nitrogen (TN), cation exchange capacity (CEC), and soil organic matter (SOM) using pXRF data. A total of 285 soil samples were collected from the A and B horizons representing Ultisols, Oxisols, Spodosols, and Entisols. The pXRF reported elements helped in the characterization of the BCP soils. In general, the RF model achieved the best performances for TN (R^2=0.50), CEC (0.75), and SOM (0.56) when A and B horizons were combined, although better results have been reported in the literature for soils from other regions of the world. The results reported here for the BCP soils represent alternatives for reducing costs and time needed for assessing such data, supporting agronomic and environmental strategies.

Keywords: Total nitrogen, Cation exchange capacity, Soil organic matter, Machine learning algorithms, Kaolinitic soils, Cohesive soils

1. Introduction

The Brazilian Coastal Plains (BCP) occupy a large part of the eastern Brazilian coast, representing 20 million ha and encompassing many federation states (Silva et al., 2015). The soils of the BCP are mainly originated from pre-weathered sediments of Barreiras Formation, with cohesive Ultisols and Oxisols as the dominant soil classes (Carvalho Filho et al., 2013; Fonsêca et al., 2007; Ker et al., 2017). They present natural chemical and physical properties, however, limit the management of agricultural/silvicultural crops (Carvalho Filho et al., 2013; Fonsêca et al., 2007). These limitations include low soil organic matter (SOM), cation exchange capacity (CEC), in addition to nutrient contents, and subsuperficial cohesion, which prevent the development of the plant root system mainly in the dry period (Ker et al., 2017; Santos et al., 2014b). The mineralogy of these soils is very uniform and dominated by quartz in the sand and silt fractions, and by kaolinite with very small contents of iron oxides and traces of gibbsite in the clay fraction (Carvalho Filho et al., 2013; Ker et al., 2017).

With the rapid advances in soil proximal sensors, soil scientists are increasingly gaining access to tools that optimize soil properties predictions. One promising tool is the portable X-ray fluorescence (pXRF) spectrometry. Importantly, pXRF is able to quantify a great diversity of elements in soil, in a few seconds, at low cost and without generation of analytical wastes (Horta et al., 2015; Hseu et al., 2016; Ribeiro et al., 2017; Stockmann et al., 2016; Weindorf et al., 2014; Zhu et al., 2011). The quantification of elemental contents occurs through detection of fluorescence energy that is characteristic of each element present in the material analyzed (Ribeiro et al., 2017). Additionally, the pXRF sensor provides adequate analytical accuracy (Stockmann et al., 2016; Wang et al., 2013; West et al., 2013) and have been used successfully in diverse studies (Che et al., 2012; Ribeiro et al., 2018; Silva et al., 2016, 2017; Stockmann et al., 2016; Terra et al., 2014; Wang et al., 2013; Weindorf et al., 2012; Wu et al., 2016).

So far, pXRF has been successfully used to predict several soil properties (Duda et al., 2017; Ribeiro et al., 2018; Sharma et al., 2015, 2014; Silva et al., 2018, 2017; Teixeira et al., 2018; Zhu et al., 2011). The prediction accuracy was improved through the development of several machine learning algorithms. As pXRF generates a large data set, the use of machine learning tools may accelerate the data processing. Besides, the pXRF reported numerical variables can be associated with categorical variables via machine learning algorithms,

improving the soil attribute prediction models. Some of the widely used algorithms in soil science are the Ordinary Least Squares Regression (OLS) and the Random Forest (RF) (Pelegrino et al., 2018; Silva et al., 2017).

Although the use of pXRF for predicting soil properties can be finetuned in combination with several robust algorithms, very few works have focused on the prediction of tropical soil properties so far. Therefore, the objectives of this work were to use pXRF data to characterize the BCP soils and assess four machine learning algorithms [ordinary least squares regression (OLS), cubist regression (CR), XGBoost (XGB), and random forest (RF)] for prediction of total nitrogen (TN), effective CEC, and SOM using pXRF data. We hypothesize that TN, CEC, and SOM in BCP soils correlate well with the soil elemental composition and, therefore, can be successfully predicted from pXRF reported elemental values.

2. Material and Methods

2.1. Study area

This study was carried out using soil samples collected in the northcentral part of Espírito Santo state, southern Bahia state, and northeastern Minas Gerais state, Brazil (Fig. 1). The selected soils for this work represent most soil environments in the BCP across its 20 million ha area. The parent materials of most soils are Tertiary sediments from the Barreiras Formation. The climate is humid tropical with a discrete dry season while the lowest average monthly rainfall is above 60mm (Gomes et al., 2017).

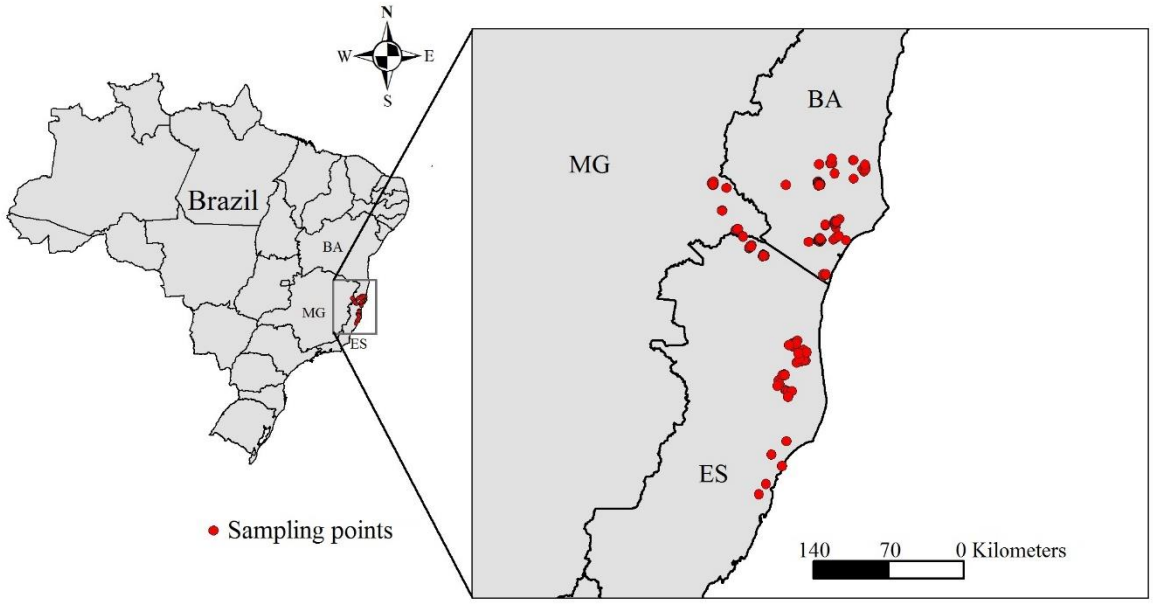

Fig. 1. Study area and sampling sites in Brazil. MG: Minas Gerais state; BA: Bahia state; ES: Espírito Santo state.

2.2. Field sampling and laboratory characterization

A total of 285 soil samples representing Ultisols, Oxisols, Spodosols, and Entisols were collected from the A and B horizons of 121 soil profiles (Fig. 1) (Soil Survey Staff, 2014). All samples were analyzed in the laboratory for TN, effective CEC, and SOM. TN content was measured by the Kjeldhal method using concentrated H_2SO_4, K_2SO_4, and $CuSO_4$ to digest the samples (Bremner, 1996). The CEC was calculated from the sum of total exchangeable bases and the exchangeable acidity (Chapman, 1965). The exchangeable Ca^{2+}, Mg^{2+}, and Al^{3+} were extracted from the soil using a 1 mol L^{-1} KCl solution. Al^{3+} was determined by titration with a 0.0125 mol L^{-1} NaOH solution, whereas Ca^{2+} and Mg^{2+} were analyzed by atomic absorption spectrophotometry (Perkin Elmer® model AAnalyst 800). K^+ was extracted by Mehlich^{-1} solution and determined by flame photometer. SOM content of the samples was determined by dichromate acid oxidation (Nelson and Sommers, 1996).

A pXRF (Bruker® model S1 Titan LE) was used to obtain the contents of the diverse elements. The pXRF contains a 50 KeV and 100 μA X-ray tube, which allows for the detection of elements of the Periodic Table ranging from Mg to U in parts per million (ppm). Scans were performed in triplicate in Trace (dual soil) mode for 60 s using the Geochem software. The following elements/oxides were obtained in all soil samples used in this work: Al_2O_3, As, Ag, Bi, CaO, Cl, Cu, Fe, Hf, K_2O, Mn, Mo, Nb, Ni, Pb, P_2O_5, Rb, Rh, S, SiO_2, Sr, Ti, V, Y, Zn, and Zr.

Furthermore, the pXRF reported elemental contents of two NIST certified standards (2710a and 2711a) and one pXRF manufacturer standard (check sample) were compared with their respective certified values to calculate % recovery (% of recovery = 100 × pXRF reported content/certified content). The recovery values for the elements identified in all samples and used in this work are as follows (2710a/2711a/CS) (0 value indicates no reference value in the certified materials or no elemental detection by pXRF): Al_2O_3 (0.31/0.63/0.91); CaO (0.81/0.48/0); Cl (0/0/0); Cu (0.87/0.71/0.97); Fe (0.71/0.69/0.91); K_2O (0.66/0.49/0.89); Mn (0.72/0.61/0.89); Nb (0/0/0); P_2O_5 (1.36/4.79/0); SiO_2 (0.41/0.51/0.42); Sr (1.00/0.90/0); Ti (0.94/0.70/0); Y (0/0/0); Zn (1.05/0.83/0); Zr (0.99/0/0).

2.3. Statistical analyses and modeling

Prior to the modeling, all pXRF results from scans of soil samples from A and B horizons were centralized (centralized elemental content = (x – mean(x)) / std. dev. (x), where x is the original elemental content) and scaled (scaled elemental content = (x – mean(x)), where x is the original elemental content). Besides scaling and centralizing, there was no other transformation in the data. Then, soil samples were randomly separated into modeling and validation datasets consisting of 70% (n = 197) and 30% (n = 88) of the total data, respectively. The models to predict TN, CEC, and SOM contents from pXRF data were created for three conditions: using only A horizon data (123 samples), only B horizon data (162 samples), and using A and B horizon data combined (A+B) (285 samples).

In order to adjust the prediction models, four machine learning algorithms were tested: ordinary least squares (OLS), cubist regression (CR), XGBoost (XGB), and random forest (RF). The algorithms were developed in R software (version 3.4.4) (R Development Core Team, 2009) through “caret” package (Kuhn, 2008). In the OLS models, the statistically insignificant variables were removed using Akaike's information criterion via step function. It provides a global model of the variable for prediction based on the assumption that the explanatory variables are measured with minimal or no error, allowing the use of single or multiple explanatory variables in the model (Sharma et al., 2013).

Cubist regression (CB) is a rule-based regression technique that was developed by Quinlan (1993). CB retrieves a set of rules associated with sets of multi-variate models. Then, a specific set of predictor variables will choose an actual prediction model based on the rule that best fits the predictors (Appelhans et al., 2015). Tuning hyperparameters of CB were optimized as follows: committees = 47, and neighbors = 6. XGBoost is an implementation of gradient boosted decision trees. It can be used for regression and classification models. It

builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function (Chen and Guestrin, 2016). XGBoost models were performed with the following optimized hyperparameters: n rounds = 53, learning rate (η) = 2.7, lambda = 0.1, and alpha = 0.02.

The RF analysis was performed with the following parameters: number of trees of the model (ntrees) = 500, node size = 5, and number of variables used in each tree (mtry) = 5, equivalent to the number of variables divided by 3, as suggested by Liaw and Wiener (2002). RF also provides the variables importance for the model, i.e., how the prediction accuracy changes as a variable is left out of the model, while other variables are maintained. Therefore, if a variable is removed and the prediction error increases, that variable is more important for the model (Breiman, 2001; Liaw and Wiener, 2002).

Principal component analysis (PCA) is a multivariate statistical technique useful for displaying and analyzing the structure of multivariate data. Basically, this statistical technique represents the original dataset in a new reference system characterized by new orthogonal variables called principal components (PCs). PCA analysis was performed in R software, through the R packages factoextra (Kassambara and Mundt, 2016) and ggplot2 (Wickham, 2016).

2.4. Assessment of prediction accuracy

The accuracy of TN, CEC, and SOM contents predictions by the machine learning algorithms was assessed by comparing the predicted with the observed values through the following statistical indexes: determination coefficient (R^2) and root mean squared error (RMSE) (Eq. 1). The models with greater R^2 and RMSE value closer to 0 were considered the best ones for predicting laboratory analysis using pXRF data.

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

where n: number of observations, yi: estimated value by the model, mi: measured value by the chemical analysis.

3. Results

3.1. Chemical elemental characterization of Brazilian Coastal Plain soils through pXRF

The descriptive statistics of TN, CEC, and SOM values for A and B horizons, separately and combined, for BCP soils show the variability of data for such properties (Table 1), mainly demonstrated by the coefficient of variation (CV%). This high variability of the

data can generate more realistic models for different conditions and land-uses, since the samples represent a wide range of values of the analyzed properties, such as TN ranging from 0.50 to 4.40 g kg^{-1}, the CEC from 0.40 to 12.69 $cmol_c$ kg^{-1}, and the SOM from 0.40 to 69.90 g kg^{-1}.

Table 1. Descriptive statistics of total nitrogen (TN), effective cation exchange capacity (CEC), and soil organic matter (SOM) values for the A and B horizons, and A+B horizons combined, from Brazilian Coastal Plain (BCP) soils.

Soil Property	Soil Horizon	Min	Max	Mean	STD [a]	CV (%) [b]
TN ($g\ Kg^{-1}$)	A	0.70	4.40	1.83	0.75	40.98
	B	0.50	1.70	1.00	0.25	25.00
	A + B	0.50	4.40	1.36	0.67	49.26
CEC ($cmol_c\ Kg^{-1}$)	A	0.66	12.69	4.05	2.41	59.51
	B	0.40	4.72	1.76	0.73	41.48
	A + B	0.40	12.69	2.75	2.02	73.45
SOM ($g\ Kg^{-1}$)	A	6.40	69.90	26.27	13.05	49.68
	B	0.40	62.60	8.04	6.35	78.98
	A + B	0.40	69.90	15.90	13.34	83.90

[a] STD: standard deviation.
[b] CV: coefficient of variation.

Interestingly, PCA revealed that pXRF variables can discriminate the horizons for the soil orders (Fig. 2). The higher contents of CaO, S, P_2O_5, and SiO_2 were found in the A horizon, while the higher contents of K_2O, Fe, Al_2O_3, and Mo were found in the B horizon. Among the elements of greater stability found in tropical soils by Mancini et al. (2019), the higher contents of Ni, V, and Pb were found in the B horizon, while Sr, and P_2O_5 were found in higher contents in the A horizon of BCP soils.

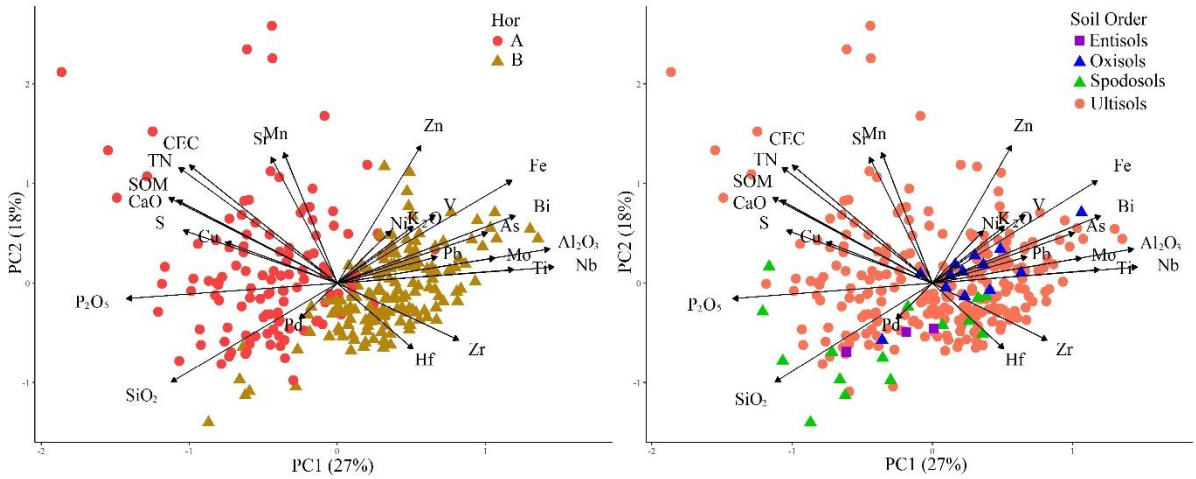

Fig. 2. Principal Component Analysis (PCA) for the chemical soil attributes and pXRF data in A and B horizons, and for soil orders in Brazilian Coastal Plain (BCP) Soils. Chemical attributes: Total nitrogen (g kg^{-1}), cation exchange capacity ($cmol_c$ kg^{-1}), and soil organic matter (g kg^{-1}). PXRF data: Al_2O_3, As, Bi, CaO, Cu, Fe, Hf, K_2O, Mn, Mo, Nb, Ni, P_2O_5, Pb, Pd, S, SiO_2, Sr, Ti, V, Zn, and Zr in parts per million (ppm).

Despite the land use for all soils used in this study is similar, the soil orders were well discriminated by the pXRF information (Fig. 2). The high contents of Al_2O_3 and Fe were found in the more weathered Oxisols. High contents of SiO_2 were found in Spodosols. Due to the ample dominance of Ultisols, their distinction from other soil orders became difficult. Such Ultisols show a marked variation between the sandy surface horizon and the more clayey subsurface horizon. This is reflected in the higher contents of Al_2O_3 along the soil depth, associated with kaolinite, and the higher contents of SiO_2 in the A horizon, associated with quartz. Fig. 3 shows that the variables which less contributed to explaining the variation in PC1 were Hf, Mn, Ni, Pd, and Sr. The most important variables to explain the variation in PC1 were Al_2O_3, CaO, Fe, Nb, P_2O_5, and SiO_2 (Fig. 3). In Fig. 2, the PC1 indicates a clear gradient of increasing soil fertility and decreasing acidity from right to left, better demonstrated by Fig. 3. Most of macro and micronutrients, higher CEC, and the highest contents of SiO_2 are on the left side, the same side represented by the youngest soils (Entisols). The elements with the highest atomic weights, the highest contents of Al_2O_3 and Fe_2O_3 are on the right side, the same side represented by the Oxisols.

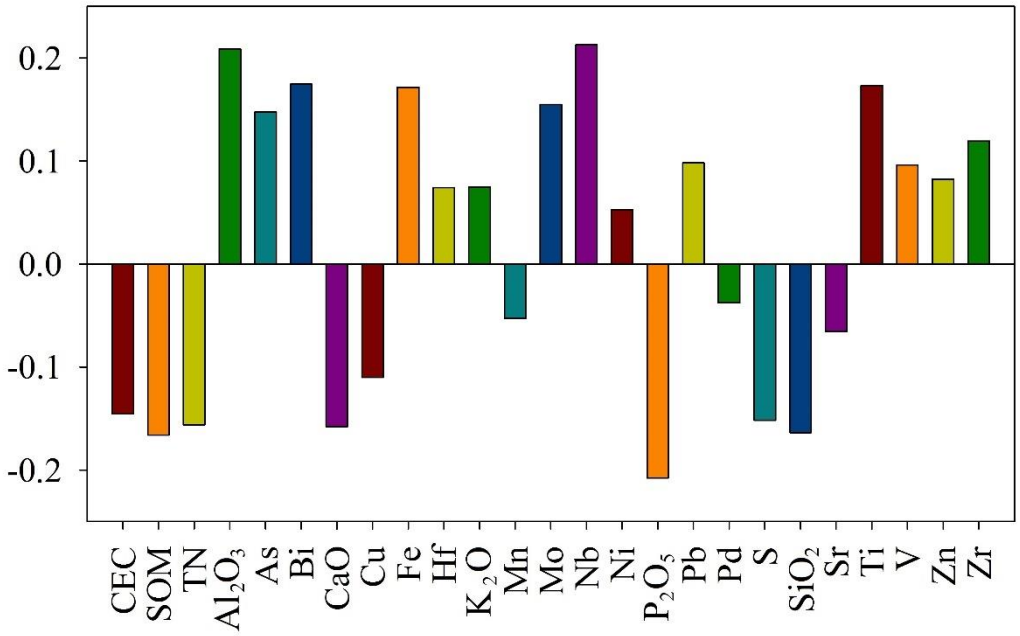

Fig. 3. Loading weight plot of principal component (PC1) of the Principal Component Analysis (PCA) for the chemical soil attributes and pXRF data in Brazilian Coastal Plain (BCP) Soils.

3.2. Models performance for predicting TN, CEC and SOM

Fig. 4 shows the performance of the models using data from A and B horizons, separately and combined (A+B). In general, the soil property prediction models for the combined horizons presented greater R^2 and smaller RMSE than the models for A and B horizon separately.

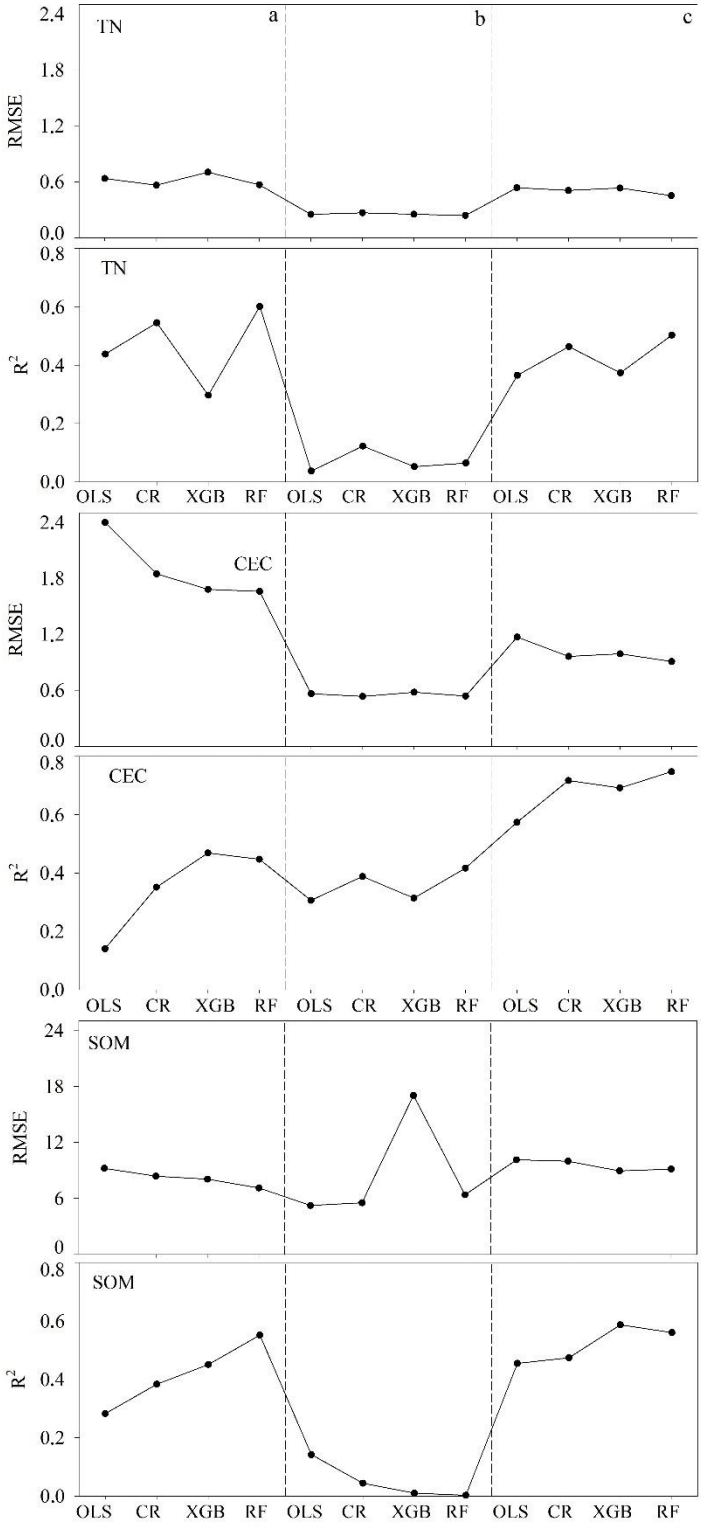

Fig. 4. Determination coefficient (R^2), and root mean square error (RMSE) of the total nitrogen (TN), effective cation exchange capacity (CEC), and soil organic matter (SOM) prediction models in A horizon (a), in B horizon (b), and in A+B horizons combined (c) in the Brazilian Coastal Plain (BCP) soils. OLS - ordinary least square; CR - cubist regression; XGB - XGBoost; RF - random forest.

For CEC in A and B horizons separately, the prediction models achieved maximum R^2 of 0.47 (XGB) and 0.42 (RF), respectively. However, for the combined horizons the RF-based CEC prediction model achieved the highest R^2 of 0.75. For predicting SOM in A horizon and for A+B horizons combined, the highest obtained R^2 remained almost the same (0.55 and 0.59 for RF and XGB, respectively).

However, for predicting SOM in B horizon, the models could not generate reliable predictions (maximum $R^2 = 0.14$ for OLS). The same pattern was observed for predicting TN, where RF models for A horizon and the combined horizons performed better (R^2 of 0.60 and 0.50, respectively) than the models developed for the B horizon (highest $R^2 = 0.12$ for CR).

In general, among the machine learning algorithms, the RF model achieved the best performances. As shown in Fig. 5, the prediction error boxplots showed the low dispersion of errors for RF model compared to the other tested algorithms. The higher R^2 and smaller RMSE yielded for predicting CEC (0.75 and 0.91) and SOM (0.56 and 9.13), when data from A+B horizons combined were used (Fig. 4). The TN prediction model using RF algorithm produced better performance in A horizon (0.60 and 0.57).

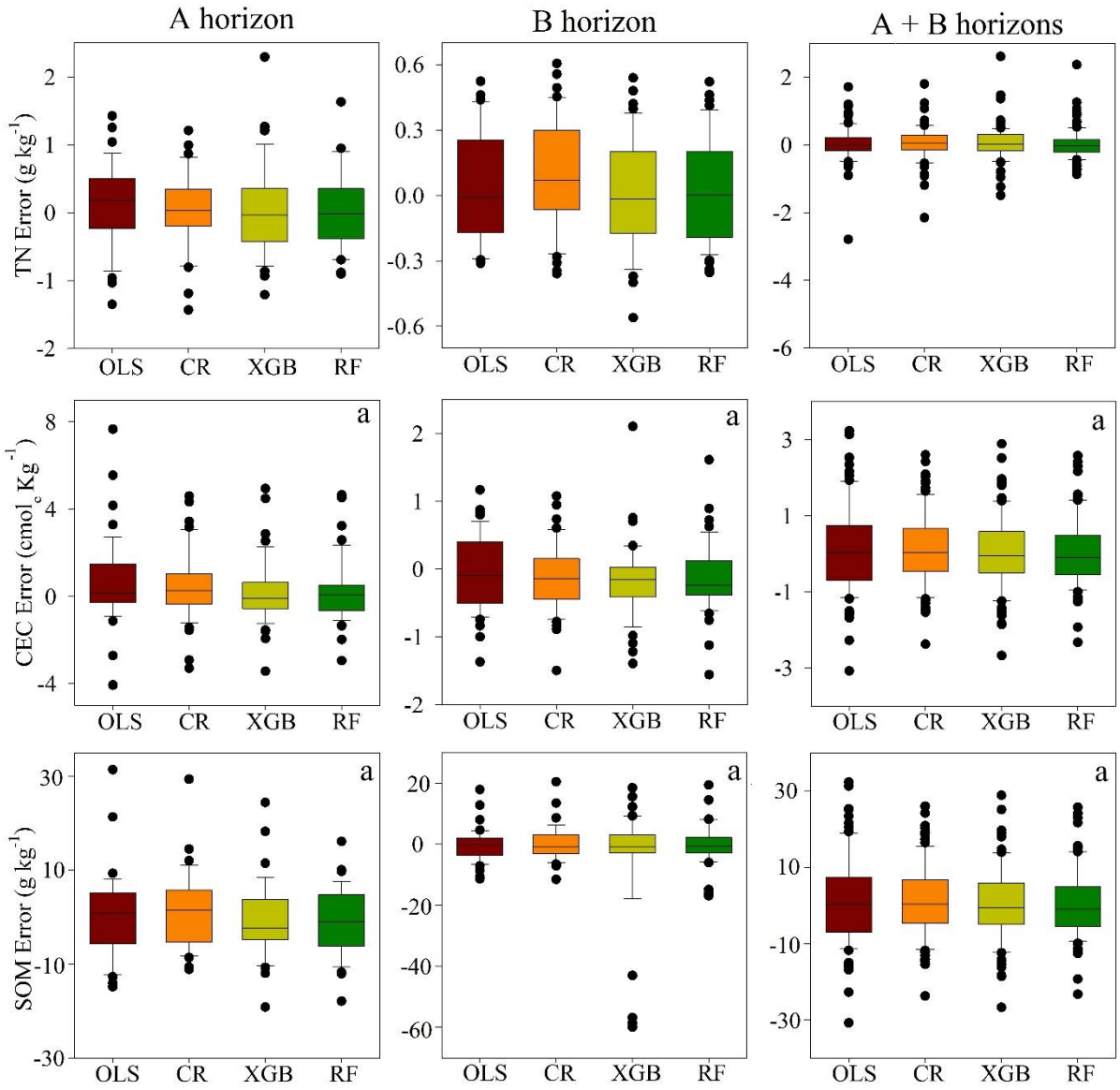

Fig. 5. Prediction error boxplots in the validation dataset for the total nitrogen (TN), effective cation exchange capacity (CEC), and soil organic matter (SOM) contents in A horizon, in B horizon, and in A+B horizons combined in the Brazilian Coastal Plain (BCP) soils.

3.3. Variables importance

The RF variables importance calculated for each soil property prediction model is given in Fig. 6. Percentage of increment of Mean Square Error (%IncMSE), estimated with out-of-bag-cross validation, is a robust and informative measure of the relative importance of one independent variable (Liaw and Wiener, 2002). Importantly, %IncMSE metric is considered adequate by several authors (Grömping, 2009; Ishwaran, 2007; González et al., 2015). The larger the %IncMSE, the higher the importance of the variable for the prediction model (González et al., 2015).

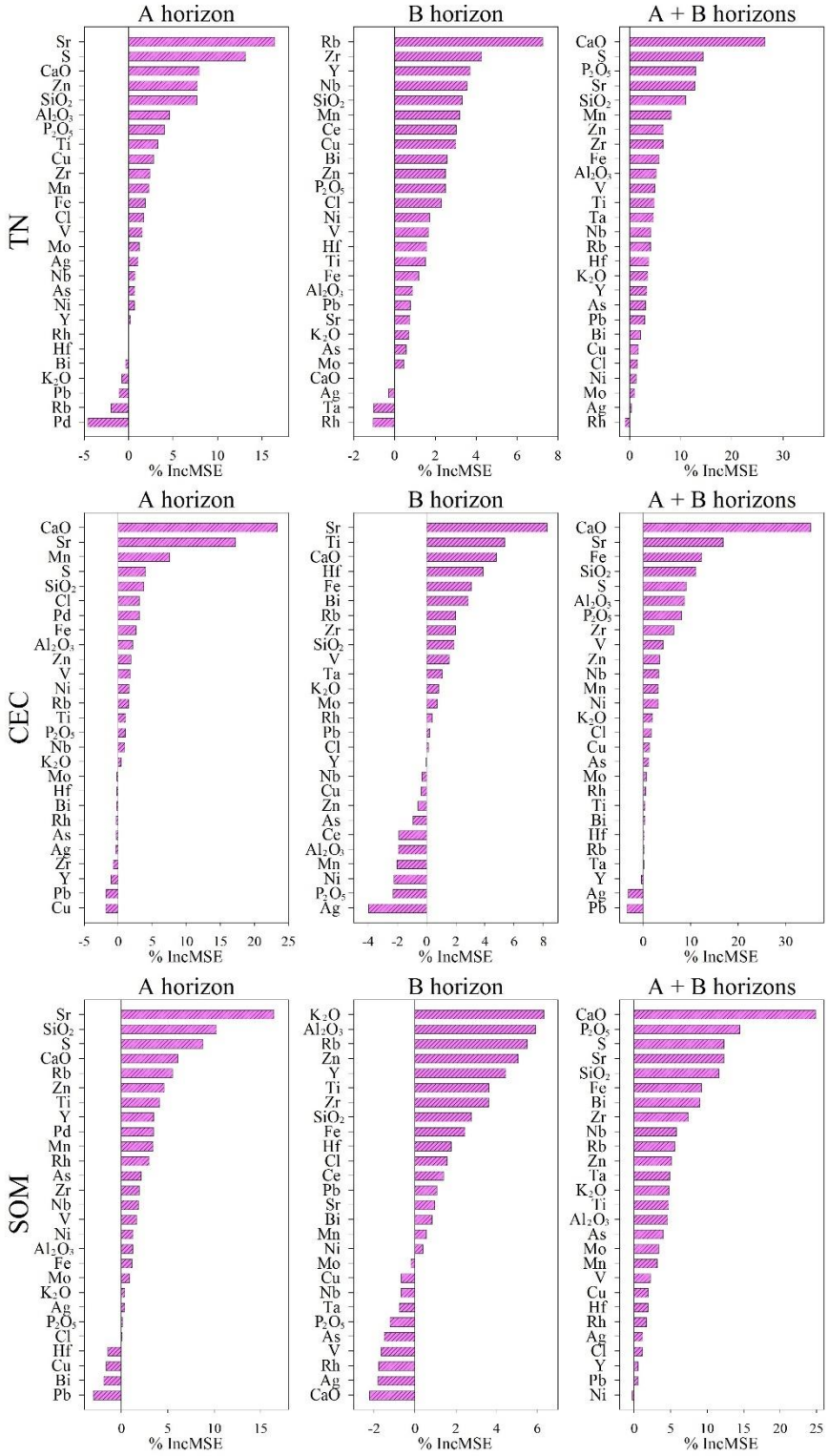

Fig. 6. Plots of relative importance of variables in RF model for total nitrogen (TN), effective cation exchange capacity (CEC), and soil organic matter (SOM) prediction models in Brazilian Coastal Plain (BCP) soils.

In general, CaO, SiO_2, Sr, and S were the five most important variables in all soil horizons studied. These variables, except for Sr, have a very dynamic behavior in the soil environment due to their radical temporal changes caused by the soil management. CaO and S are included in the most widely agricultural inputs used in tropical soils (limestone). Consequently, their behavior in soil is very dynamic and changes with soil management. As Sr appeared in almost all samples and in almost all models, it may be related to the parent material of the studied soils (Gomes et al., 2017).

3.4. SOM influence on CEC and TN predictions

Fig. 7 reveals the relationship between the pXRF data and SOM to influence the TN and CEC prediction behavior. In the A horizon, samples featuring SOM values > 40 g kg^{-1} were underestimated while the same behavior was observed for TN and CEC predictions in samples with SOM > 40 g kg^{-1}. Conversely, samples with SOM < 20 g kg^{-1} were overestimated. The same behavior was observed in TN and CEC predictions with SOM < 20 g kg^{-1}. Although the relation between N and CEC with SOM in the B horizon was not so evident, the same pattern of overestimation and underestimation can be observed. As tropical soils are highly weathered and leached, SOM mineralization plays a major role as the source of CEC and N supply to plants (Patrick et al., 2013).

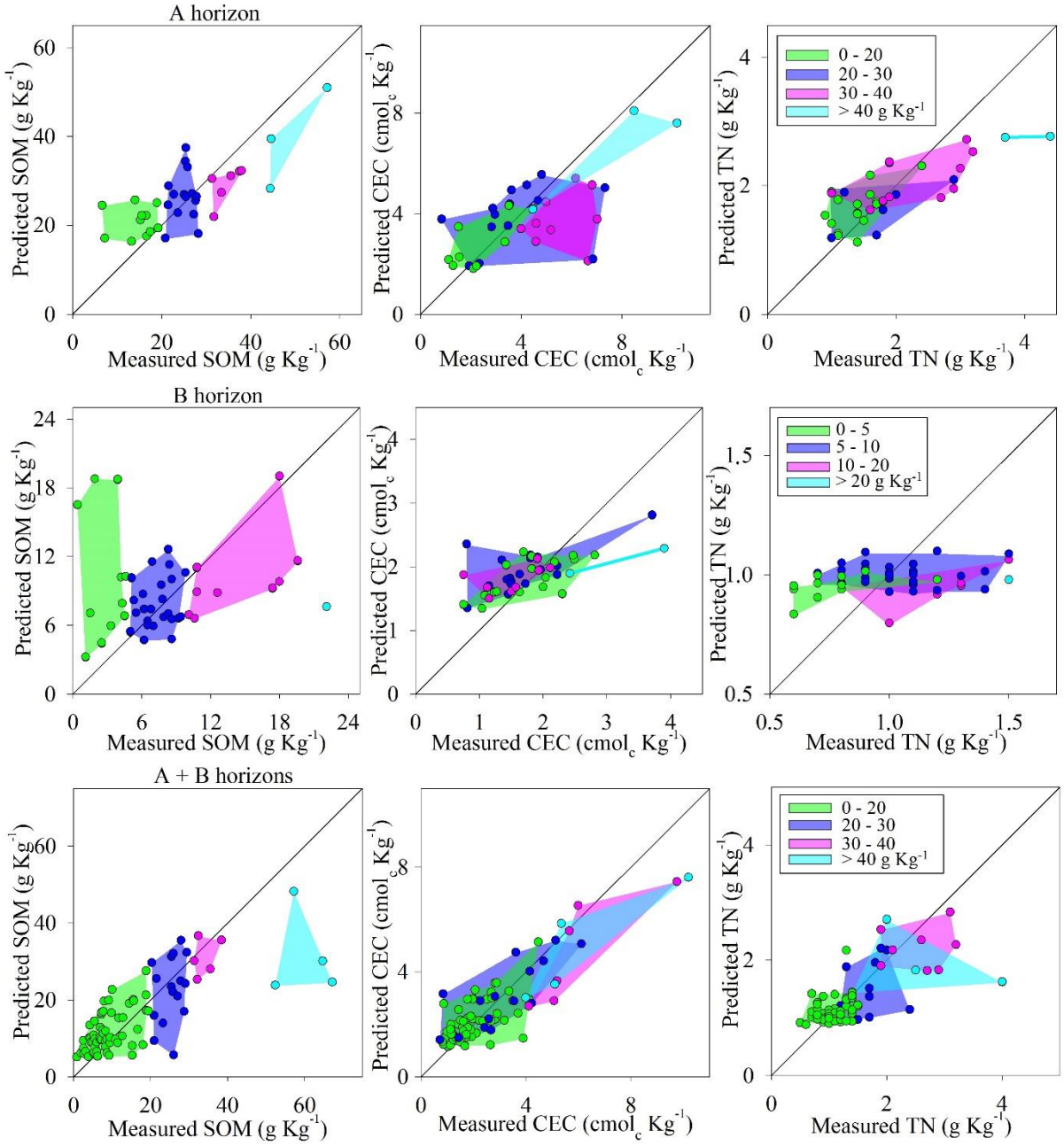

Fig. 7. Prediction scatter plot of total nitrogen (TN), effective cation exchange capacity (CEC), and soil organic matter (SOM) contents by RF models for the validation dataset in A horizon, in B horizon, and in A+B horizons combined in the Brazilian Coastal Plain (BCP) soils based on pXRF data.

4. Discussion

4.1. Chemical characterization of the BCP soils

The range of BCP soil properties analyzed in this study (Table 1) contribute to the generation of more reliable models, since they encompass greater variability of such data, such as suggested by Rawal et al. (2019). Furthermore, these values are similar to those found

by Barros et al. (2013), Santos et al. (2014a), and Silva et al. (2012, 2015) while working in BCP soils. These contents are smaller than those found in temperate regions (Konen et al., 2003) because of the comparatively milder and warmer climate conditions and the lower degree of weathering, which make C and N stabilization mechanisms more effective (Six et al., 2002).

PCA showed the correlations between elemental contents obtained by pXRF and soil classes (Fig. 2), e.g., high contents of SiO_2 were found in Spodosols, while high contents of Al_2O_3 and Fe_2O_3 were observed in Oxisols. This supports Oliveira et al. (2010) who observed that BCP Spodosols are of sandy texture dominated by quartz (SiO_2) and that Fe and Al accumulation occur as oxide minerals in Oxisols. Also, most of these soils are Ultisols, presenting greater amounts of sand, mainly composed of quartz (SiO_2) in the A horizon and greater amounts of clay (mainly kaolinite and Fe/Al-oxides) in the B horizon (Kämpf et al., 2012). Such marked variation of texture in A and B horizons found in BCP soils (Cintra et al., 2004; Ker et al., 2017) corresponds to the SiO_2 versus Al_2O_3 and Fe_2O_3 contents present in the sand and clay fractions, respectively, which are also evident in Fig. 2.

The entire area of study has been cultivated with Eucalyptus, so the soil management that includes the addition of ameliorants explains the higher values of macro- (Ca, S, N, P) and micronutrients (Mn, Cu) in the A horizon in addition to the nutrient recycling. Regarding other elements and compounds, Sr and P_2O_5 were elevated in the A horizon probably due to the addition of ameliorants, since such soils are highly leached. Similar data was reported by Teixeira et al. (2018), who also found higher contents of Sr and P_2O_5 in A horizon of soils cultivated in Brazilian Cerrado, contrasting the contents found in soils under native vegetation. Other studies have also reported the ability of pXRF to characterize different soil environments and properties (Cardelli et al., 2017; Chakraborty et al., 2019; Mancini et al., 2019; Weindorf et al., 2016).

4.2. Models performance for predicting TN, CEC, and SOM

The results of the predictions in this study indicate the feasibility of obtaining TN, CEC, and SOM results from pXRF data in Brazilian soils. It is noteworthy that such tests are still under development in Brazil, requiring further investigation and comparison with other works on this topic (e.g., studies from temperate regions). For instance, only one previous study of Brazilian soils attempted to predict CEC and SOM from pXRF (Silva et al., 2017). That study was developed in Brazilian Cerrado, where soils are completely different from BCP soils (Resende et al., 2014). Silva et al. (2017) found the best predictions were delivered

by RF and reached R^2 and RMSE of 0.96 and 0.89 cmolc dm^{-3} for CEC and 0.78 and 0.80 dag kg^{-1} for SOM predictions, respectively. The disparity in predictive model performance between Brazilian Cerrado soils and those of the BCP reinforce that additional testing is vital toward reaching the optimal methodology to deliver the best results for each soil-environmental condition. The present work is the first attempt to use pXRF for TN prediction in Brazilian soils.

Comparatively, the works developed on temperate regions were able to predict CEC, SOM, and TN with greater accuracy than the present study for BCP soils. Yet eventually, the RMSE values of SOM and TN were lower or closer to those reported in previous studies using pXRF as proximal sensing technique (Cardelli et al., 2017; Wang et al., 2015). Sharma et al. (2015) predicted CEC from pXRF data achieving an R^2 of 0.90 for soils from the United States of America. Duda et al. (2017) also successfully predicted TN [R^2 = 0.73, RMSE (%) = 0.05] and other soil properties in Romania using pXRF data. Xu et al. (2019) achieved an R^2 of 0.53 for predicting SOM in Chinese soils from pXRF data, however, better results (R^2 = 0.81) were obtained after combining the mid-infrared (mid-IR) spectrometer and laser-induced breakdown spectroscopy (LIBS) sensors. These findings indicate the global search for cheaper, more rapid, and environmentally friendly methods to predict soil properties and the need for additional testing under variable soil conditions to verify the efficiency of this methodology. A final large question remaining is whether a single global model can effectively predict the parameter of interest or whether multiple, localized models will produce superior results.

As a summary of the findings of this work, RF tended to deliver the best results. Regarding the comparison of the datasets, the best performances were achieved using the combined dataset (A+B horizon data), which can be attributed to the relatively larger dataset used to produce a more robust RF model. This supports the findings of Souza et al. (2016) and Silva et al. (2017). From a practical aspect, the best results obtained for the combined dataset may facilitate future analysis especially for samples collected without considering the soil horizon (a common practice worldwide). Similarly, Santana et al. (2018) successfully used unknown soil samples (i.e., samples collected in soils without considering the soil class or horizon) to develop models for predicting results of conventional laboratory analysis from pXRF.

4.3. Model applicability

The range of predicted soil properties encompasses the values reported in the literature for BCP soils, and the models for combined A and B horizons generally delivered the best predictions. The models developed herein could be used for predicting TN, CEC, and SOM in soil samples collected at any depth that encompasses the A or B horizons within BCP biome (20 million ha). However, the models developed in this work only involved samples collected under eucalyptus plantations, indicating that other tests should be conducted for soils under other land uses (planted pastures, coffee, etc.). The most important model variables were those with dynamic behavior in the soil environment that are strongly influenced by soil management.

Although the models developed in this study may be applicable to virtually all BCP soils under such conditions, future and continued calibration and validation of models may be required for establishing stronger relationships between elemental concentrations obtained by pXRF and TN, CEC, and SOM under other soil conditions. Finally, while regional differences may require localized standardization of the pXRF with a unique combination of elements germane to the study area, pXRF shows considerable promise as a technique for rapidly assessing such soil properties in addition to reducing the time and amount of chemical waste generated from conventional laboratory analyses (Rawal et al., 2019; Weindorf et al., 2014; Ribeiro et al., 2017). Although works have been published utilizing pXRF to predict soil properties worldwide, the efficiency of this methodology necessitates refinement with deference to soil variability, depth, land use, parent material, and environmental conditions.

5. Conclusions

The elemental contents reported by pXRF provided an adequate characterization of BCP soils. In general, RF algorithm produced better results in predicting TN, CEC, and SOM than OLS, XGBoost, and CR models. Variables such as CaO, SiO_2, and S which exhibit dynamic behavior in the soil environment appeared as influential in RF based prediction models. Because of its intrinsic relationship, the patterns observed in SOM predictions through pXRF data were reflected in the TN and CEC predictions too. Further studies are needed to improve the predictions of chemical properties of BCP soils and other regions of the globe to determine which soil properties are capable of being predicted via pXRF data and if general or specific models are more prone to deliver the best results under contrasting soil and environmental conditions.

6. Acknowledgments

The authors gratefully acknowledge FAPEMIG, CNPq and CAPES Brazilian funding agencies for providing financial resources for this research. The authors gratefully acknowledge the BL Allen Endowment in Pedology at Texas Tech University in conducting this research.

References

Appelhans, T., Mwangomo, E., Hardy, D.R., Hemp, A., Nauss, T., 2015. Evaluating machine learning approaches for the interpolation of monthly air temperature at Mt. Kilimanjaro, Tanzania. Spat. Stat. 14, 91–113. https://doi.org/10.1016/j.spasta.2015.05.008

Barros, J.D. de S., Chaves, L.H.G., Chaves, I. de B., Farias, C.H. de A., Pereira, W.E., 2013. Estoque de Carbono e Nitrogênio em sistemas de manejo do solo, nos Tabuleiros Costeiros Paraibanos. Rev. Caatinga 26, 35–42.

Breiman, L., 2001. Random Forests. Mach. Learn. 45, 5–32.

Bremner, J.M., 1996. Nitrogen Total, in: Methods of Soil Analysis. Part 3. Chemical Methods. Soil Science Society of America, American Society of Agronomy, pp. 1085–1121.

Cardelli, V., Weindorf, D.C., Chakraborty, S., Li, B., De Feudis, M., Cocco, S., Agnelli, A., Choudhury, A., Ray, D.P., Corti, G., 2017. Non-saturated soil organic horizon characterization via advanced proximal sensors. Geoderma 288, 130–142. https://doi.org/10.1016/j.geoderma.2016.10.036

Carvalho Filho, A., Curi, N., Fonseca, S. da, 2013. Avaliação informatizada e validada da aptidão silvicultural das terras dos tabuleiros costeiros Brasileiros para Eucalipto, 1st ed. Editora UFLA, Lavras, MG.

Chakraborty, S., Li, B., Weindorf, D.C., Deb, S., Acree, A., De, P., Panda, P., 2019. Use of portable X-ray fluorescence spectrometry for classifying soils from different land use land cover systems in India. Geoderma 338, 5–13. https://doi.org/10.1016/j.geoderma.2018.11.043

Chapman, H.D., 1965. Cation Exchange Capacity, in: Methods of Soil Analysis. Part 2: Chemical and Microbiological Properties. American Society of Agronomy, pp. 891–901.

Che, V.B., Fontijn, K., Ernst, G.G.J., Kervyn, M., Elburg, M., Van Ranst, E., Suh, C.E., 2012. Evaluating the degree of weathering in landslide-prone soils in the humid tropics: The case of Limbe, SW Cameroon. Geoderma 170, 378–389. https://doi.org/10.1016/j.geoderma.2011.10.013

Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data

Mining - KDD '16. Presented at the the 22nd ACM SIGKDD International Conference, ACM Press, San Francisco, California, USA, pp. 785–794. https://doi.org/10.1145/2939672.2939785

Cintra, F.L.D., Portela, J.C., Nogueira, L.C., 2004. Caracterização físíca e hídrica em solos dos Tabuleiros Costeiros no Distrito de Irrigação Platô de Neópolis. Rev. Bras. Eng. Agríc. E Ambient. 8, 45–50. https://doi.org/10.1590/S1415-43662004000100007

Duda, B.M., Weindorf, D.C., Chakraborty, S., Li, B., Man, T., Paulette, L., Deb, S., 2017. Soil characterization across catenas via advanced proximal sensors. Geoderma 298, 78–91. https://doi.org/10.1016/j.geoderma.2017.03.017

Fonsêca, M.H.P., Guerra, H.O.C., Lacerda, R.D. de, Barreto, A.N., 2007. Uso de propriedades físico-hídricas do solo na identificação de camadas adensadas nos Tabuleiros Costeiros, Sergipe. Rev. Bras. Eng. Agríc. E Ambient. 11, 368–373. https://doi.org/10.1590/S1415-43662007000400004

Gomes, J.B.V., Araújo Filho, J.C., Vidal-Torrado, P., Cooper, M., Silva, E.A. da, Curi, N., 2017. Cemented Horizons and Hardpans in the Coastal Tablelands of Northeastern Brazil. Rev. Bras. Ciênc. Solo 41. https://doi.org/10.1590/18069657rbcs20150453

González, S., Herrera, F., García, S., 2015. Monotonic Random Forest with an Ensemble Pruning Mechanism based on the Degree of Monotonicity. New Gener. Comput. 33, 367–388. https://doi.org/10.1007/s00354-015-0402-4

Grömping, U., 2009. Variable Importance Assessment in Regression: Linear Regression versus Random Forest. Am. Stat. 63, 308–319. https://doi.org/10.1198/tast.2009.08199

Horta, A., Malone, B., Stockmann, U., Minasny, B., Bishop, T.F.A., McBratney, A.B., Pallasser, R., Pozza, L., 2015. Potential of integrated field spectroscopy and spatial analysis for enhanced assessment of soil contamination: A prospective review. Geoderma 241–242, 180–209. https://doi.org/10.1016/j.geoderma.2014.11.024

Hseu, Z.-Y., Chen, Z.-S., Tsai, C.-C., Jien, S.-H., 2016. Portable X-Ray Fluorescence (pXRF) for Determining Cr and Ni Contents of Serpentine Soils in the Field, in: Hartemink, A.E., Minasny, B. (Eds.), Digital Soil Morphometrics. Springer International Publishing, Cham, pp. 37–50. https://doi.org/10.1007/978-3-319-28295-4_3

Ishwaran, H., 2007. Variable importance in binary regression trees and forests. Electron. J. Stat. 1, 519–537. https://doi.org/10.1214/07-EJS039

Kassambara, A., Mundt, F., 2016. Factoextra: Extract and Visualize the Results of Multivariate Data Analyses.

Ker, J.C., Schaefer, C.E.G.R., Romero, R.E., Corrêa, M.M., 2017. Solos dos Tabuleiros Costeiros, in: Pedologia - Solos Dos Biomas Brasileiros. Sociedade Brasileira de Ciência do Solo, Viçosa, MG, pp. 467–492.

Konen, M.E., Burras, C.L., Sandor, J.A., 2003. Organic Carbon, Texture, and Quantitative Color Measurement Relationships for Cultivated Soils in North Central Iowa. Soil Sci. Soc. Am. J. 67, 1823. https://doi.org/10.2136/sssaj2003.1823

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

Liaw, A., Wiener, M., 2002. Classification and Regression by Random Forest. R News 2, 18–22.

Mancini, M., Weindorf, D.C., Chakraborty, S., Silva, S.H.G., dos Santos Teixeira, A.F., Guilherme, L.R.G., Curi, N., 2019. Tracing tropical soil parent material analysis via portable X-ray fluorescence (pXRF) spectrometry in Brazilian Cerrado. Geoderma 337, 718–728. https://doi.org/10.1016/j.geoderma.2018.10.026

Nelson, D.W., Sommers, L.E., 1996. Total carbon, organic carbon, and organic matter, in: Methods of Soil Analysis. Part 3. Chemical Methods. Soil Science Society of America, American Society of Agronomy, pp. 961–1010.

Oliveira, Aline Pacobahyba de, Ker, J.C., Silva, I.R. da, Fontes, M.P.F., Oliveira, Alessandra Pacobahyba de, Neves, A.T.G., 2010. Spodosols pedogenesis under barreiras formation and sandbank environments in the south of Bahia. Rev. Bras. Ciênc. Solo 34, 847–860. https://doi.org/10.1590/S0100-06832010000300026

Patrick, M., Tenywa, J.S., Ebanyat, P., Tenywa, M.M., Mubiru, D.N., Basamba, T.A., Leip, A., 2013. Soil Organic Carbon Thresholds and Nitrogen Management in Tropical Agroecosystems: Concepts and Prospects. J. Sustain. Dev. 6. https://doi.org/10.5539/jsd.v6n12p31

Pelegrino, M.H.P., Weindorf, D.C., Silva, S.H.G., de Menezes, M.D., Poggere, G.C., Guilherme, L.R.G., Curi, N., 2018. Synthesis of proximal sensing, terrain analysis, and parent material information for available micronutrient prediction in tropical soils. Precis. Agric. https://doi.org/10.1007/s11119-018-9608-z

Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Francisco, CA, USA.

R Development Core Team, 2009. R: A Language and Environmental for Statistical Computing. R Found. Stat. Comput.

Ribeiro, B.T., Silva, S.H.G., Silva, E.A., Guilherme, L.R.G., 2017. Portable X-ray fluorescence (pXRF) applications in tropical Soil Science. Ciênc. E Agrotecnologia 41, 245–254. https://doi.org/10.1590/1413-70542017413000117

Ribeiro, B.T., Weindorf, D.C., Silva, B.M., Tassinari, D., Amarante, L.C., Curi, N., Guimarães Guilherme, L.R., 2018. The Influence of Soil Moisture on Oxide Determination in Tropical Soils via Portable X-ray Fluorescence. Soil Sci. Soc. Am. J. 82, 632. https://doi.org/10.2136/sssaj2017.11.0380

Santos, W.J.R., Curi, N., Silva, S.H.G., Fonseca, S. da, Silva, E. da, Marques, J.J., 2014a. Detailed soil survey of an experimental watershed representative of the Brazilian Coastal Plains and its practical application. Ciênc. E Agrotecnologia 38, 50–60. https://doi.org/10.1590/S1413-70542014000100006

Santos, W.J.R., Silva, B.M., Oliveira, G.C., Volpato, M.M.L., Lima, J.M., Curi, N., Marques, J.J., 2014b. Soil moisture in the root zone and its relation to plant vigor assessed by

remote sensing at management scale. Geoderma 221–222, 91–95. https://doi.org/10.1016/j.geoderma.2014.01.006

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, 141–147. https://doi.org/10.1016/j.geoderma.2014.05.005

Sharma, A., Weindorf, D.C., Wang, D., Chakraborty, S., 2015. Characterizing soils via portable X-ray fluorescence spectrometer: 4. Cation exchange capacity (CEC). Geoderma 239–240, 130–134. https://doi.org/10.1016/j.geoderma.2014.10.001

Sharma, V., Rudnick, D.R., Irmak, S., 2013. Development and Evaluation of Ordinary Least Squares Regression Models for Predicting Irrigated and Rainfed Maize and Soybean Yields. Trans. ASABE 56, 1361–1378. https://doi.org/10.13031/trans.56.9973

Silva, A.L.P. da, Silva, A.P. da, Souza, A.P. de, Santos, D., Silva, S. de M., Silva, V.B. da, 2012. Resposta do abacaxizeiro "Vitória" a doses de nitrogênio em solos de tabuleiros costeiros da Paraíba. Rev. Bras. Ciênc. Solo 36, 447–456. https://doi.org/10.1590/S0100-06832012000200014

Silva, E. da, Curi, N., Ferreira, M.M., Volpato, M.M.L., Santos, W.J.R. dos, Silva, S.H.G., 2015. Pedotransfer functions for water retention in the main soils from the Brazilian Coastal Plains. Ciênc. E Agrotecnologia 39, 331–338. https://doi.org/10.1590/S1413-70542015000400003

Silva, S., Poggere, G., Menezes, M., Carvalho, G., Guilherme, L., Curi, N., 2016. Proximal Sensing and Digital Terrain Models Applied to Digital Soil Mapping and Modeling of Brazilian Latosols (Oxisols). Remote Sens. 8, 614. https://doi.org/10.3390/rs8080614

Silva, S.H.G., Silva, E.A., Poggere, G.C., Guilherme, L.R.G., Curi, N., 2018. Tropical soils characterization at low cost and time using portable X-ray fluorescence spectrometer (pXRF): Effects of different sample preparation methods. Ciênc. E Agrotecnologia 42, 80–92. https://doi.org/10.1590/1413-70542018421009117

Silva, S.H.G., Teixeira, A.F. dos S., Menezes, M.D. de, Guilherme, L.R.G., Moreira, F.M. de S., Curi, N., 2017. Multiple linear regression and random forest to predict and map soil properties using data from portable X-ray fluorescence spectrometer (pXRF). Ciênc. E Agrotecnologia 41, 648–664. https://doi.org/10.1590/1413-70542017416010317

Six, J., Feller, C., Denef, K., Ogle, S.M., de Moraes, J.C., Albrecht, A., 2002. Soil organic matter, biota and aggregation in temperate and tropical soils - Effects of no-tillage. Agronomie 22, 755–775. https://doi.org/10.1051/agro:2002043

Soil Survey Staff, 2014. Keys to Soil Taxonomy, 12th ed. USDA, Washington, DC.

Souza, E. de, Fernandes Filho, E.I., Schaefer, C.E.G.R., Batjes, N.H., Santos, G.R. dos, Pontes, L.M., 2016. Pedotransfer functions to estimate bulk density from soil properties and environmental covariates: Rio Doce basin. Sci. Agric. 73, 525–534. https://doi.org/10.1590/0103-9016-2015-0485

Stockmann, U., Cattle, S.R., Minasny, B., McBratney, A.B., 2016. Utilizing portable X-ray fluorescence spectrometry for in-field investigation of pedogenesis. CATENA 139, 220–231. https://doi.org/10.1016/j.catena.2016.01.007

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 uses. Ciênc. E Agrotecnologia 42, 501–512. https://doi.org/10.1590/1413-70542018425017518

Terra, J., Sanches, R.O., Bueno, M.I.M.S., Melquiades, F.L., 2014. Análise Multielementar de solos: uma proposta envolvendo equipamento portátil de fluorescência de raios X. Semina Ciênc. Exatas E Tecnológicas 35, 207. https://doi.org/10.5433/1679-0375.2014v35n2p207

Wang, D., Chakraborty, S., Weindorf, D.C., Li, B., Sharma, A., Paul, S., Ali, Md.N., 2015. Synthesized use of VisNIR DRS and PXRF for soil characterization: Total carbon and total nitrogen. Geoderma 243–244, 157–167. https://doi.org/10.1016/j.geoderma.2014.12.011

Wang, S., Li, W., Li, J., Liu, X., 2013. Prediction of Soil Texture Using FT-NIR Spectroscopy and PXRF Spectrometry With Data Fusion: Soil Sci. 178, 626–638. https://doi.org/10.1097/SS.0000000000000026

Weindorf, D.C., Bakr, N., Zhu, Y., 2014. Advances in Portable X-ray Fluorescence (PXRF) for Environmental, Pedological, and Agronomic Applications, in: Advances in Agronomy. Elsevier, pp. 1–45. https://doi.org/10.1016/B978-0-12-802139-2.00001-9

Weindorf, D.C., Chakraborty, S., Herrero, J., Li, B., Castañeda, C., Choudhury, A., 2016. Simultaneous assessment of key properties of arid soil by combined PXRF and Vis-NIR data: Arid soil assessment by PXRF and Vis-NIR. Eur. J. Soil Sci. 67, 173–183. https://doi.org/10.1111/ejss.12320

Weindorf, D.C., Zhu, Y., McDaniel, P., Valerio, M., Lynn, L., Michaelson, G., Clark, M., Ping, C.L., 2012. Characterizing soils via portable x-ray fluorescence spectrometer: 2. Spodic and Albic horizons. Geoderma 189–190, 268–277. https://doi.org/10.1016/j.geoderma.2012.06.034

West, M., Ellis, A.T., Potts, P.J., Streli, C., Vanhoof, C., Wegrzynek, D., Wobrauschek, P., 2013. 2013 Atomic spectrometry update—A review of advances in X-ray fluorescence spectrometry. J. Anal. At. Spectrom. 28, 1544. https://doi.org/10.1039/c3ja90046k

Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

Wu, Z., Lin, C., Su, Z., Zhou, S., Zhou, H., 2016. Multiple landscape “source–sink” structures for the monitoring and management of non-point source organic carbon loss in a peri-urban watershed. CATENA 145, 15–29. https://doi.org/10.1016/j.catena.2016.05.020

Xu, D., Zhao, R., Li, S., Chen, S., Jiang, Q., Zhou, L., Shi, Z., 2019. Multi-sensor fusion for the determination of several soil properties in the Yangtze River Delta, China: Predictions based on multi-sensor fusion. Eur. J. Soil Sci. 70, 162–173. https://doi.org/10.1111/ejss.12729

Zhu, Y., Weindorf, D.C., Zhang, W., 2011. Characterizing soils using a portable X-ray fluorescence spectrometer: 1. Soil texture. Geoderma 167–168, 167–177. https://doi.org/10.1016/j.geoderma.2011.08.010

ARTICLE 2 - Tropical soil order and suborder prediction combining optical and X-ray approaches

Article published in Geoderma Regional Journal, v. 23, p. 13, 2020

(https://doi.org/10.1016/j.geodrs.2020.e00331)

Renata Andrade[a], Sérgio Henrique Godinho Silva[a], David C. Weindorf[b],

Somsubhra Chakraborty[c], Wilson Missina Faria[a],

Luiz Roberto Guimarães Guilherme[a], Nilton Curi[a]

[a]Department of Soil Science, Federal University of Lavras, Caixa Postal 3037, Lavras, MG Postal Code 37200-000, Brazil

[b]Department of Earth and Atmospheric Sciences, Central Michigan University, Mount Pleasant, MI 48859, USA

[c]Agricultural and Food Engineering Department, Indian Institute of Technology, Kharagpur, West Bengal Postal Code 721302, India

Abstract

Proper soil taxonomic classification makes a significant contribution toward sustainable soil management, decision making, and soil conservation. For that, a quick, environmentally-friendly, non-invasive, cost-effective, and reliable method for soil class assessment is desirable. As such, this study used NixPro color and portable X-ray fluorescence (pXRF) data to characterize seven different soil orders in Brazilian tropical soils, exploring the ability of three machine learning algorithms [Support Vector Machine with Linear Kernel (SVMLK), Artificial Neural Network (ANN), and Random Forest (RF)] with and without Principal Component Analysis (PCA) pretreatment for prediction of different soils at the order and suborder taxonomic levels under both dry and moist conditions. In total, 734 soil samples were collected from surface and subsurface horizons encompassing twelve suborders. The soil profiles were morphologically described, and taxonomy classified per the Brazilian Soil Classification System and the approximate correspondence was made with the US Soil Taxonomy. Soil samples were separated into modeling (70%) and validation (30%) sub-datasets, overall accuracy and Cohen's Kappa coefficient evaluated model quality. Models

generated from B horizon sample with pXRF and NixPro (moist samples) data combined delivered the best accuracy for order (81.19% overall accuracy and 0.71 Kappa index) and suborder predictions (74.35% overall accuracy and 0.65 Kappa index) through RF algorithm without PCA pretreatment. Summarily, the use of these two portable sensor systems was shown effective at accurately predicting different soil orders and suborders in tropical soils. Future works should extend the results of this study to temperate regions to corroborate the conclusions presented herein.

Keywords: pXRF, NixPro color sensor, Soil classification, Machine learning, Kappa coefficient.

1. Introduction

It is well known that a proper soil taxonomic classification makes a significant contribution to sustainable soil management (Resende et al., 2014). Once separated into different classes, soils are expected to be similar in terms of their intrinsic chemical and physical characteristics, and therefore, more useful in soil management, including cultivation planning and soil conservation. Worldwide, two international systems are most prominently used for soil classification: US Soil Taxonomy (USDA, 1999) and the World Reference Base for Soil Resources (FAO, 2014). But there are also many national systems that are used locally for soil classification; for example, The Australian Soil Classification from Australia (Isbell, 2016), Classification and Diagnostic System of Russian Soils from Russia (Shishov et al., 2004), Chinese Soil Taxonomy from China (CRGCST, 2001), Romanian System of Soil Taxonomy from Romania (Florea and Munteanu, 2000), Dutch Soil Classification System from the Netherlands (de Bakker and Schelling, 1966), the Canadian System of Soil Classification from Canada (Soil Classification Working Group, 1998), and the French Soil Classification System from France (Baize and Girard, 2009). Similarly, to most of those systems, the Brazilian Soil Classification System (dos Santos et al., 2018), currently in its 5th edition, reflects local pedological descriptions and consists of six levels: order, suborder, great group, sub-group, family, and series, very similar to the US Soil Taxonomy though family and series levels are still under development.

The top level (order) features thirteen soil classes. For the main soil orders occurring in Brazil (Latossolos and Argissolos – equivalent to Oxisols and mostly Ultisols, respectively) in addition to other less expressive orders, soil color is utilized to define the suborder level, similar to the Yearbook of Agriculture 1938 (USDA, 1938), which featured soil classification

based on color. Soil color has strong influence in this classification system with inferences about soil genesis, natural fertility, texture, and management. In Brazil, color correlates well with most mineralogical, physical, and chemical characteristics of the soil.

Soil color is a consequence of soil mineral and organic constituents. Regarding mineralogy, it is also an indicator of the presence of iron oxides (Fe), revealing the nature of these minerals and providing information on pedogenetic conditions. For example, hematite (Fe_2O_3) and goethite (FeOOH) occur in the clay fraction, contain Fe^{3+}, and provide red and yellow colors, respectively (Resende et al., 2014). Gray colors in soil are caused by primary and secondary soil minerals in the absence of Fe(III) compounds. Conversely, darker color is promoted by organic matter (Schaetzl and Anderson, 2005).

Traditionally, a soil classification combines a soil surveyor's specialized knowledge, field descriptions, and laboratory analysis (Vasques et al., 2014). However, with the increasing demand for precision agriculture, alternative methods to facilitate soil classification are sought. For instance, laboratory analyses are generally time-consuming, expensive, invasive, and produce chemical waste decreasing the number of soil samples feasibly collected and analyzed in the lab, consequently constraining soil survey. Therefore, quick, environmentally-friendly, non-invasive, cost-effective and reliable soil methodologies for classification are desirable (Benedet et al., 2020).

A reliable, novel alternative to traditional approaches is the use of proximal sensors to assess soil properties (Lemière, 2018; Silva et al., 2019). Inexpensive, non-destructive, and less time consuming, the use of such tools to assess soil information have been applied to predict different soil attributes, such as pH, base saturation, soil texture, cation exchange capacity (Rawal et al., 2019; Sharma et al., 2014; Silva et al., 2017), soil organic carbon, and total nitrogen (Mikhailova et al., 2017; Stiglitz et al., 2018). Furthermore, elemental contents, nutrients, and heavy metals have been assessed (Chakraborty et al., 2019; Hu et al., 2017;Mukhopadhyay et al., 2020; Teixeira et al., 2018). However, comparatively few efforts have been made applying proximal sensors for soil classification. In the United States, enhanced pedon horizonation has been documented in nondescript soils (Weindorf et al., 2012a) and some subsoil diagnostic horizons have been identified via proximal sensors (Weindorf et al., 2012b). More recently, proximal sensing was used to distinguish chernozems from phaeozems in the Transylvanian Plain of Romania (Acree et al., 2020). In doing so, Acree et al. (2020) advocated for changes to the forthcoming third edition of Soil Taxonomy, whereby proximal sensor data could be directly used for taxonomic differentiation. Portable X-ray fluorescence (pXRF) spectrometry is one of the most capable proximal sensors,

providing multi-elemental quantification from Mg to U in the periodic table simultaneously (Hseu et al., 2016; Ribeiro et al., 2017; Weindorf et al., 2014). It requires minimal sample preparation and, with adequate calibration, the equipment can be used both in the laboratory or in field conditions (Ribeiro et al., 2017; Weindorf et al., 2014). Spatial and temporal monitoring of soil and plant properties are also easily accomplished (Kincey et al., 2018; Pelegrino et al., 2018).

A complementary proximal sensor, the NixPro™ color sensor, has been recently applied in soil science for assessing soil color enabling correlation with soil organic carbon and total nitrogen (Mikhailova et al., 2017; Stiglitz et al., 2018; Stiglitz et al., 2017). It is a quick, portable, inexpensive, and rechargeable color sensor providing color reports in many different numerical color systems, being less subjective in soil color determination than the Munsell color chart (Stiglitz et al., 2016; Mancini et al., 2020).

Some studies have shown that combining pXRF elemental data and NixPro™ color data may optimize the accuracy of the predictive models (Kagiliery et al., 2019; Mukhopadhyay et al., 2020). Given the success of such combined sensor approaches, their application and testing on soil order and suborder assessment appears timely, especially in developing countries where financial resources are increasingly scarce for soil characterization. Thus, the objectives of this research were to use NixPro™ colorimetric capacities and pXRF elemental composition to characterize seven different soil orders in Brazilian tropical soils and explore the ability of three machine learning algorithms [Support Vector Machine with Linear Kernel (SVMLK), Artificial Neural Network (ANN), and Random Forest (RF)] with and without Principal Component Analysis (PCA) pretreatment for prediction of different soils at the order and suborder taxonomic levels for both dry and moist conditions. Our hypothesis is that the combination of pXRF and NixPro™ will prove useful for rapid and accurate assessment of soil order and suborder in tropical soils.

2. Material and Methods

2.1. General occurrence and features

This study was conducted using soil samples from four Brazilian states (Fig. 1). Mean annual temperature ranges from 20 to 22 °C, with 1300 to 1600 mm of annual rainfall. The Köppen-Gieger climate classification (Alvares et al., 2013; Kottek et al., 2006) of the sampled areas are tropical with dry winter (Aw) in Rio de Janeiro state, humid subtropical with dry summer (Csb) in São Paulo state, humid subtropical with temperate summer (Cfb) in Santa Catarina state, tropical with dry winter (Aw) and humid subtropical with dry winter and rainy

summer (Cwa) in Minas Gerais state. The samples collected included twelve soil suborders classified according to the Brazilian Soil Classification System (SiBCS) (dos Santos et al., 2018); their correspondence to US Soil Taxonomy (USDA, 1999) is given in Table 1. Soils were collected under different land uses, encompassing native vegetation and crop areas featuring pasture, eucalyptus, coffee, tobacco, and corn. During the soil survey in the field, variable parent materials were identified therein inclusive of gneiss, basalt, gabbro, sandstone, alluvial and colluvial sediments, slate, phyllite, tuffite, quartzite, and Tertiary and Quaternary sediments. The cultivated areas have received application of limestone and fertilizers; no management practices have been conducted in areas under native vegetation.

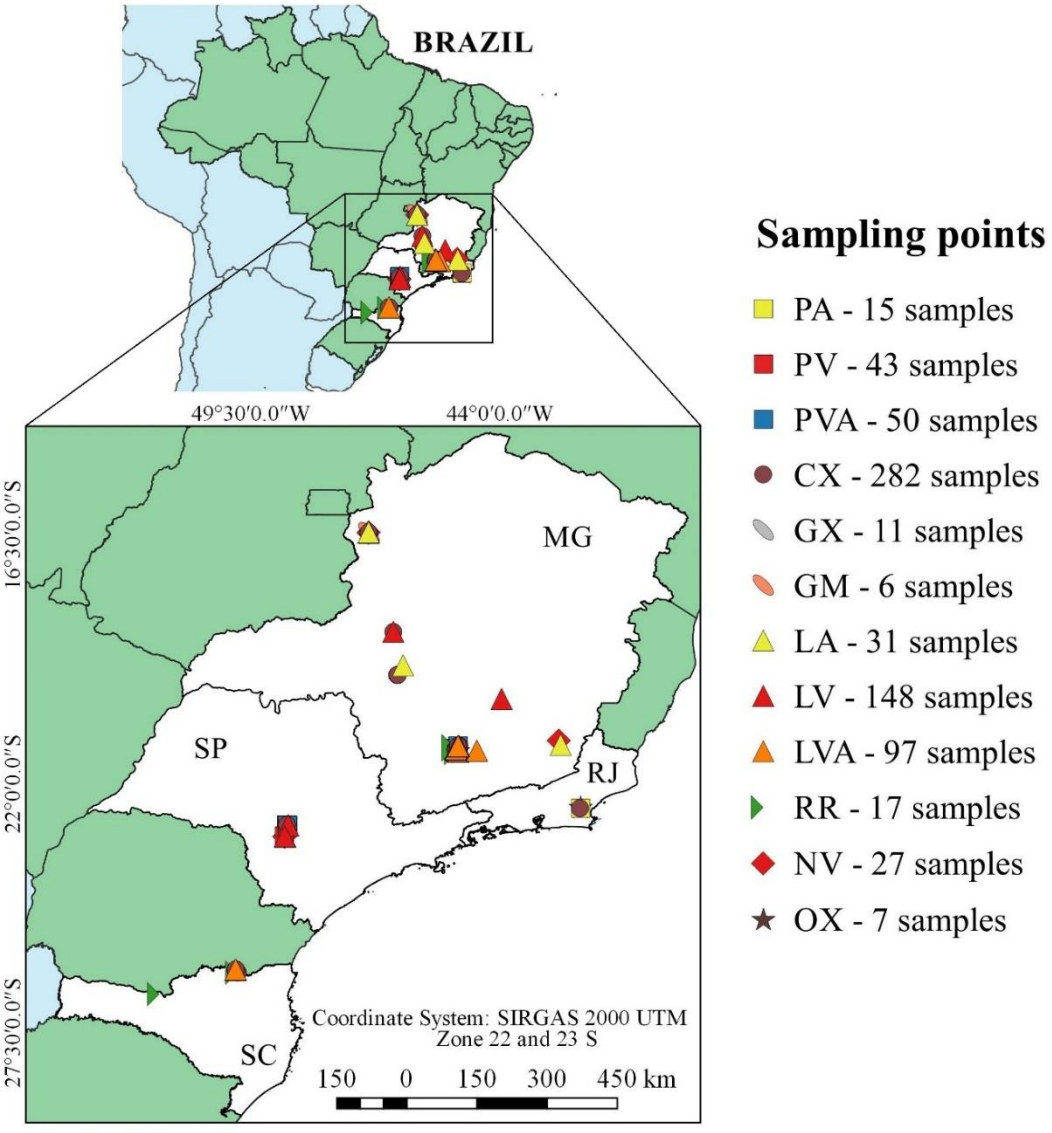

Fig. 1. Distribution of the collected samples in the different Brazilian states: Minas Gerais (MG), São Paulo (SP), Rio de Janeiro (RJ), and Santa Catarina (SC). Abbreviations see Table 1.

Table 1. Soil suborders according to the *Brazilian Soil Classification System* (SiBCS) (Santos et al., 2018) and corresponding *US Soil Taxonomy* (USDA, 1999) classes.

Identif.	SiBCS classification	Soil Taxonomy classification
PA	Argissolo Amarelo	Typic Hapludult
PV	Argissolo Vermelho	Rhodic Kandiudult
PVA	Argissolo Vermelho-Amarelo	Typic Hapludult
CX	Cambissolo Háplico	Typic Dystrustept
GX	Gleissolo Háplico	Typic Endoaquent
GM	Gleissolo Melânico	Mollic Endoaquent
LA	Latossolo Amarelo	Xanthic Hapludox
LV	Latossolo Vermelho	Anionic Acrudox
LVA	Latossolo Vermelho-Amarelo	Typic Hapludox
RR	Neossolo Regolítico	Typic Udorthent
NV	Nitossolo Vermelho	Typic Rhodudult
OX	Organossolo Háplico	Typic Udifolist

2.2. Soil sampling and laboratory analyses

Samples were collected from the following surficial or subsurficial horizons: A (311 samples), O (5 samples), E (1 sample), B (403 samples), and C (14 samples), per dos Santos et al. (2015), constituting a total of 734 samples. Samples were air-dried, passed through a 2-mm sieve, and analyzed in the laboratory for soil texture via the pipette method (Gee and Bauder, 1986). Soil pH (1:2.5 - v:v, soil:water) was measured with an electrometric pH-meter (Donagema et al., 2011). Exchangeable Ca^{2+}, Mg^{2+}, and Al^{3+} were extracted with 1 mol L^{-1} KCl (McLean et al., 1958), and available K^{+} and P were extracted using Mehlich^{-1} solution (Mehlich, 1953). Al^{3+} was determined by titration with a 0.0125 mol L^{-1} NaOH solution, whereas Ca^{2+} and Mg^{2+} were analyzed by atomic absorption spectrophotometry (Perkin Elmer® model AAnalyst 800). K^{+} was determined by flame photometer and quantification of available P was made through molecular absorption spectrometry. Soil organic carbon was determined via titration per Walkley and Black (1934). Finally, soils were classified according to the *Brazilian Soil Classification System* and the approximate correspondence was made with the *US Soil Taxonomy*.

2.3. PXRF analysis

A pXRF spectrometer (Bruker® model S1 Titan LE) was used to scan all samples and obtain the elemental composition per Weindorf and Chakraborty (2016). The equipment

features a Rh X-ray tube and was operated at 50 kV and 100 μA online power (220 VAC) with an integrated silicon drift detector (145 eV), which allows for the detection of elements ranging from Mg to U. Scans were performed in triplicate in Trace (dual soil) mode for 60 s using the Geochem software. The following elements were obtained in all soil samples used in this work: Al, As, Ca, Cr, Cu, Fe, K, Mn, Ni, P, Pb, Rb, S, Si, Sr, Ti, V, Y, and Zn.

To ensure the quality and accuracy of the data generated by pXRF, prior to use the equipment was calibrated using the manufacturer calibration alloy coin. PXRF-reported elemental contents from National Institute of Standards and Technology (NIST) certified standards (2710a and 2711a) and one pXRF manufacturer standard (check sample) were compared with their respective certified values to calculate recovery (% of recovery = 100 x pXRF reported content/certified content) (Koch et al., 2017). The recovery values for the elements identified in all samples and used in this work are as follows (2710a/2711a/CS) (0 value indicates no reference value in the certified materials or no elemental detection by pXRF): Al (85/72/90), As (92/63/0), Ca (32/44/0), Cr (0/110/0), Cu (77/69/89), Fe (71/65/90), K (52/48/85), Mn (69/58/86), Ni (0/65/91), P (372/512/0), Pb (112/105/97), Rb (99/102/0), S (0/0/0), Si (57/54/95), Sr (240/210/0), Ti (80/67/0), V (48/24/0), Y (0/0/0), and Zn (91/83/0).

2.4. NixPro™ analysis

An inexpensive NixPro™ color sensor (Hamilton, Ontario, Canada) was used to collect numerical color data from each soil sample. The sensor is controlled wirelessly by a smartphone or tablet through Bluetooth and has its own light-emitting diode (LED) light source located within the concave base of the sensor about 1 cm above the field of view. Spectral acquisition range is 380 – 730 nm. The sensor produces scan results in various color system codes, such as RGB, XYZ, CIELAB, LCH, HEX, CMYK, and ACES; all are interrelated and can be easily converted to uniquely identify individual colors of matrix being scanned. Kagiliery et al. (2019) provides an overview of the various color systems acquired by the NixPro™ sensor. The sensor is also rechargeable, easily accessible because of its small size, and can be recalibrated easily.

All samples were scanned by placing the sensor on each soil sample that was leveled to give the sensor a flat area to rest directly on, completely covering the base of the sensor, allowing no outside light to enter the scan area. Each sample was scanned under both dry and moist soil conditions following the methods described previously by Stiglitz et al. (2017; 2016). Samples were moistened using a water dropper to the point of no more color change in the soil.

All CIE (1978) color coordinates (lightness (L*), redness (a*), and yellowness (b*) colorimetrics) were collated and then used to calculate hue angle (H°) and chroma (C*) (dos Santos et al., 2015). The H° and C* were calculated based on the following equations: H° = arctg (b*/a*) and $C^*=(a^{*2}+b^{*2})^{1/2}$. The total color change (ΔE) for each soil suborder was then calculated as per AMSA (2012) by the following equation $\Delta E = [(\Delta L^*)^2 + (\Delta a^*)^2 + (\Delta b^*)^2]^{0.5}$. All numerical color data collected from the color system codes, H° and C* were utilized as explanatory variables for the prediction models.

2.5. Data analysis and modeling

Prior to analyses, principal component analysis (PCA) was applied for data dimensionality reduction. PCA is a mathematical orthogonal transformation that creates new uncorrelated variables that successively maximizes variance. It was executed using R software version 3.6.1 (R Development Core Team, 2018). The principal components (PC) derived from PCA that explain most of the data variance were applied to all prediction models in order to test the performance of models built with and without dimensionality reduction.

In order to predict both soil order and suborder, soil samples were separated into modeling and validation datasets consisting of 70% and 30% of the total data, respectively. Three types of predictions were evaluated: 1) models using only pXRF data, 2) models using only NixPro™ color sensor data (Color systems, H° and C*) (dry and moist samples), and 3) models with both data combined (general) (Table 2). Moreover, the models were created separately for three conditions: using surface horizon data (316 samples), using subsurface horizon data (418 samples), and with both horizons combined (734 samples); these were split by soil profile in order to avoid autocorrelation. The ability of the generated models to correctly predict soil order and suborder was tested by an independent validation dataset (30% of the full dataset).

Table 2. Summary of the order and suborder prediction models showing the explanatory variables in each model.

Models	Explanatory variables	Number of predictor variables
1	pXRF	19
2	Nix (dry samples)	27
3	Nix (moist samples)	27
4	pXRF + Nix (dry samples)	46
5	pXRF + Nix (moist samples)	46

Support Vector Machine with Linear Kernel (SVMLK), Artificial Neural Network (ANN), and Random Forest (RF) algorithms were utilized in each data set to generate the soil order and suborder prediction models, using R software (version 3.5.3) (R Development Core Team, 2018), through the "caret" package (Kuhn, 2008). The SVMLK parameter cost (Cost) was set to default. ANN models were created with size (hidden units), and decay (weight decay) parameters set to default. The RF parameter ntree (number of trees in the model) was set to 5000, mtry (number of variables used in each tree) corresponded to one third the number of predictors (Liaw and Wiener, 2002), and nodesize (minimum node size) was set to default. RF does not provide a final equation, but the variable importance for the model can be assessed. The more important the variable, the more the prediction error increases as this variable is left out of the model, while other variables are maintained (Breiman, 2001; Liaw and Wiener, 2002). One of the metrics released by RF algorithm is percentage of increment of mean square error (%IncMSE). %IncMSE is a robust and informative metric about the relative importance of each independent variable (Grömping, 2009; Ishwaran, 2007) and prevents bias (González et al., 2015). It represents the percentage of increase in mean square error (MSE) of predictions (estimated with out-of-bag cross validation) as a certain variable is permuted while others are maintained in trees during RF modeling (Liaw and Wiener, 2002). The larger the %IncMSE value, the higher the importance of the variable for the prediction model (González et al., 2015).

2.6. Evaluating model performance

The predicted and measured soil orders and suborders were arranged into an error matrix. Then, overall accuracy, ranging from 0 to 100%, was calculated by the sum of the major diagonal (the correctly classified samples) divided by the total number of samples in the entire error matrix (the closer to 1, the greater the accuracy). The performance of the models was also assessed using Cohen's Kappa coefficient (Cohen, 1960; Sim and Wright, 2005) (Eq. 1). Kappa values range from −1 to 1, indicating increasing accuracy as the values move closer to 1.

$$Kappa = \frac{Po-Pe}{1-Pe} \quad (1)$$

where Po is the proportion of correctly classified sites and Pe is the probability of random agreement (Landis and Koch, 1977).

3. Results and discussion

3.1. Tropical soil characterization via proximal sensors

Boxplots showing the elemental distribution for the twelve soil suborders in Brazilian tropical soils are shown in Fig. 2. Elemental composition provided by pXRF was able to detect differences between the soil suborders. Certain elements, such as Al, Fe, and Zn, display a clear disparity in their range of content, being able to discriminate at least two soil suborders among themselves. For instance, Al clearly discriminates GX and OX. Zn, in turn, clearly discriminates OX from GX, and NV. Also, some elements presented a marked range for certain soil suborders, for instance, greater contents of As, Ca, Cu, K, Mn, Sr, and Zn were found in CX. Greater contents of Fe, Cr, Ni, Pb, and Y were found in LV, caused by mafic parent rocks. These relationships may be related to soil management, mineralogy, and parent material. Mancini et al. (2019), also working in tropical conditions, and Stockamann et al. (2016), working in a temperate region, found remarkable differences in elemental composition provided by pXRF for a myriad of soils derived from several parent materials, such as phyllite, basalt, dolerite, mudstone, and sandstone, indicating that the differences found in this study may also be a result of parent material influence.

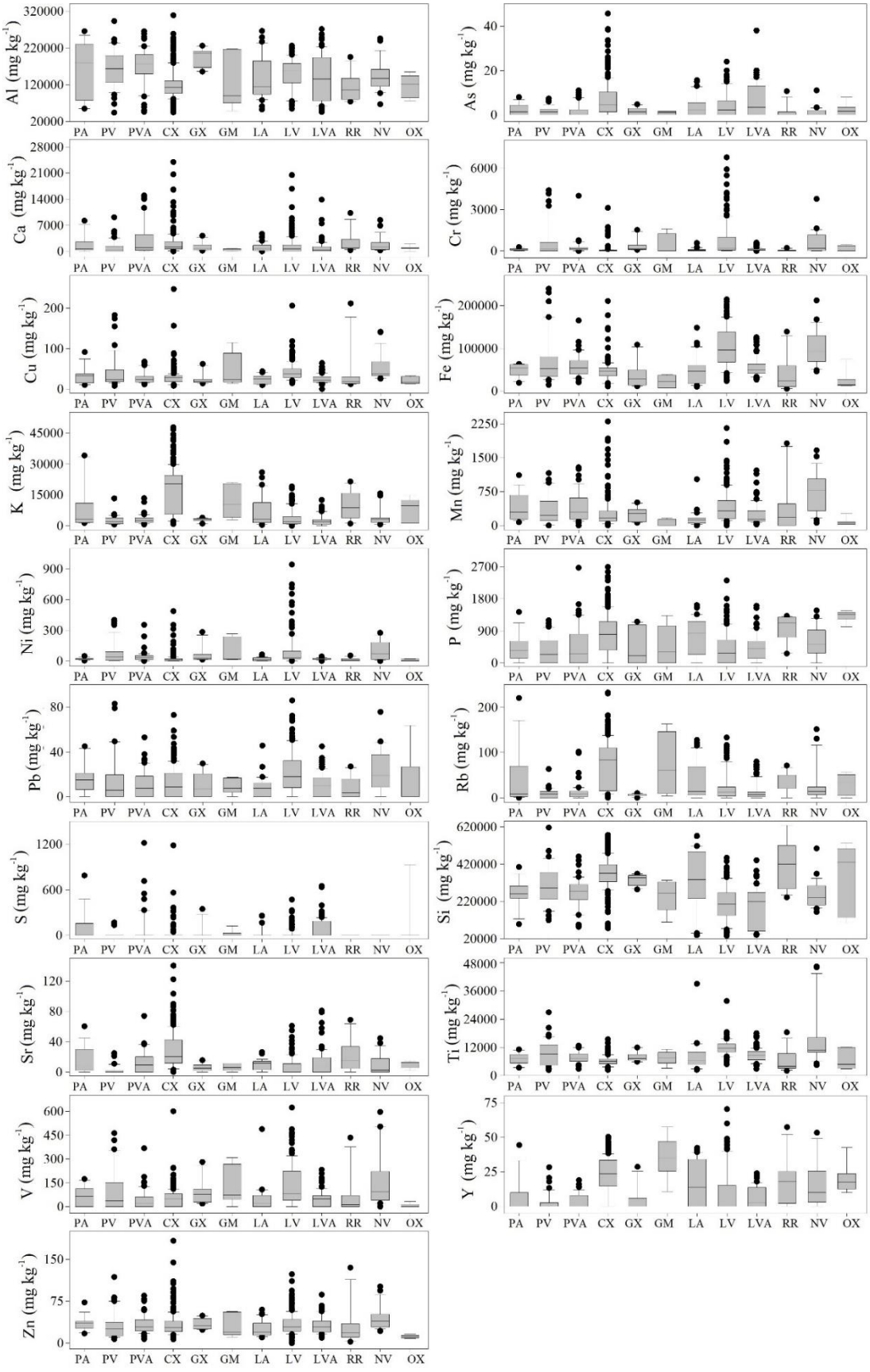

Fig. 2. Boxplots comparing twelve soil suborders through elemental contents obtained via portable X-ray fluorescence (pXRF) spectrometry in Brazilian tropical soils.

A clear distinction among soil orders and suborders based on pXRF data was not attained. The diverse land uses of the evaluated soils, including distinct soil management practices with applications and doses of fertilizers and conditioners at different depths alter the original chemical composition of soils, especially for the macro- and micronutrients. Some patterns related to soil mineralogy were identified from the elemental contents shown in Fig. 2. For that, recall the mineralogy of most Brazilian soils is basically composed of quartz and muscovite in the coarse fractions, and kaolinite, hematite, goethite, and gibbsite in the clay fraction, in different proportions (Kämpf et al., 2012; Brinatti et al., 2010). Thus, red soils (LV, NV and PV) presented the greatest Fe contents since they were derived from gabbro, basalt, melanocratic gneiss, etc., which contributes to the greater amount of hematite formation under oxidizing conditions and hence, their red color. This pattern was somewhat followed by Ti, Pb, and Ni. Conversely, GM, GX, and OX had low Fe contents, due to their occurrence under saturated conditions that favor the reduction of Fe(III) compounds, facilitating Fe(II) removal from these soil systems. So, they are basically composed of quartz, kaolinite, and gibbsite (Resende et al., 2011; Resende et al., 2014). Si and Al variability rely on soil parent materials and the degree of weathering leaching. Both elements, mainly Si, tend to be removed from soils as the degree of weathering advances (Kämpf et al., 2012), as demonstrated by the lowest Si contents found in Oxisols (LV and LVA). An elemental analysis by pXRF including only soils under natural conditions may deliver an optimized pattern between certain elements and soil orders and suborders.

PCA revealed that pXRF and NixPro™ color sensor combined can modestly discriminate soils that feature color in the second level (suborder) of soil classification from those which do not (Fig. 3). The PC1 and PC2 combined explained 60.45% of the total variance. Soils that feature color at suborder level presented the highest content of elements of greater stability in the soil, such as Ni, V, and Pb (Mancini et al., 2019). The highest content of Fe was also found in those soils, probably due to the presence of iron oxides which can confer color to the soil matrix. Hematite (Fe_2O_3) and goethite (FeOOH) occur in the clay fraction, and provide red and yellow colors, respectively (Resende et al., 2014).

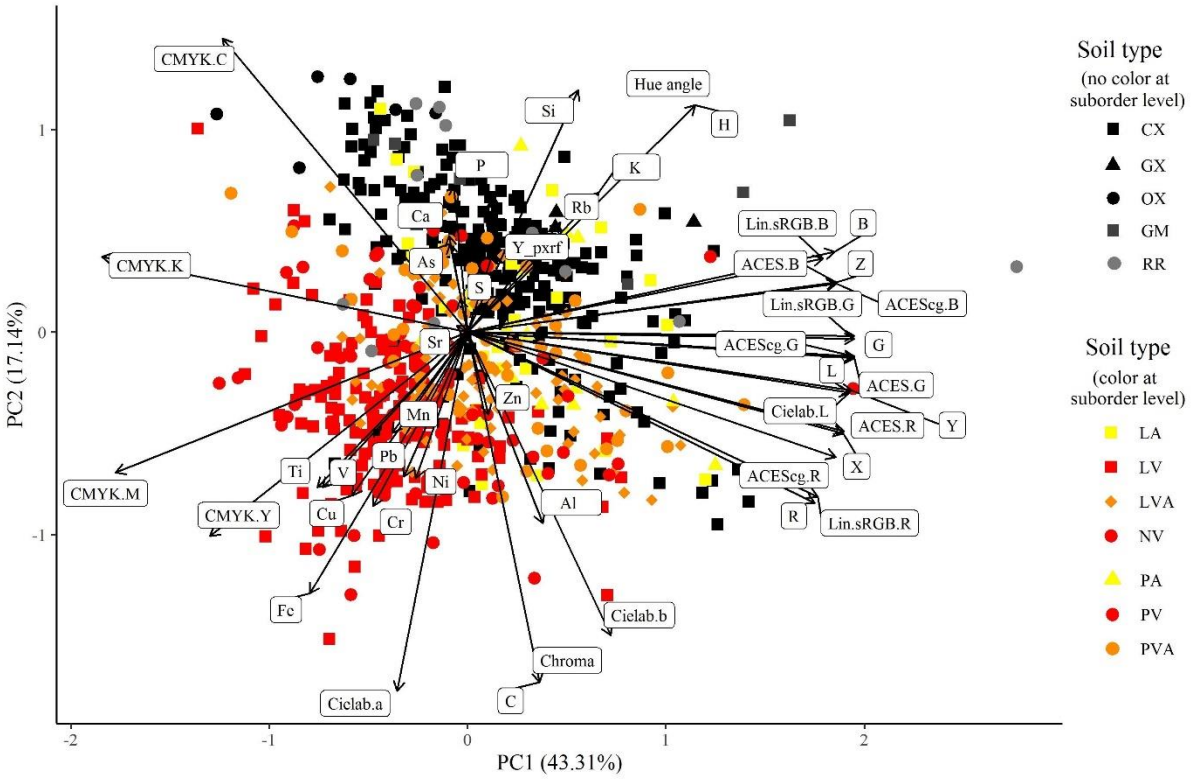

Fig. 3. Principal component analysis (PCA) for twelve soil suborders through elemental composition provided by portable X-ray fluorescence (pXRF) and numerical color information provided by NixPro™ color sensor in Brazilian tropical soils. Abbreviations see Table 1.

Soils that feature yellow color at the suborder level were the least distinguished from the colorless soils at the suborder level, followed by red-yellow, and finally red. Red soils presented the greatest disparity between soils with color at suborder level, with the highest values of a* (from CIELab color system, −green:red+), M (from CMYK color system, representing magenta contribution), and Y (from CMYK color system, representing yellow contribution).

3.2. Classification performance at soil order level

The performance of the models using data from A and B horizons, separately and combined (A + B), for the prediction of first level (order) of soil classification without and with PCA pretreatment are shown in Figs. 4 and 5, respectively. The best model was generated using B horizon data via pXRF and NixPro™ (moist samples) data through RF algorithm, resulting in an overall accuracy of 81.19% and Kappa index of 0.71.

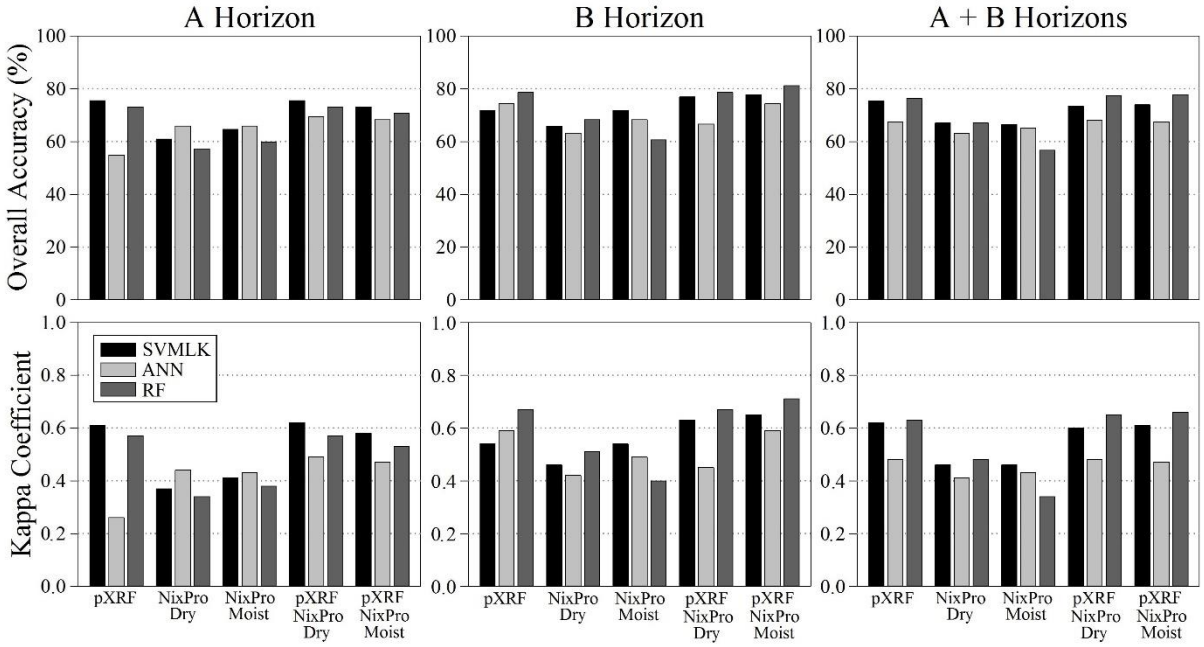

Fig. 4. Overall accuracy and Kappa coefficient values for the prediction models in the first level (order) of soil classification for Brazilian tropical soils. SVMLK – Support Vector Machine with Linear Kernel; ANN – Artificial Neural Network; RF – Random Forest.

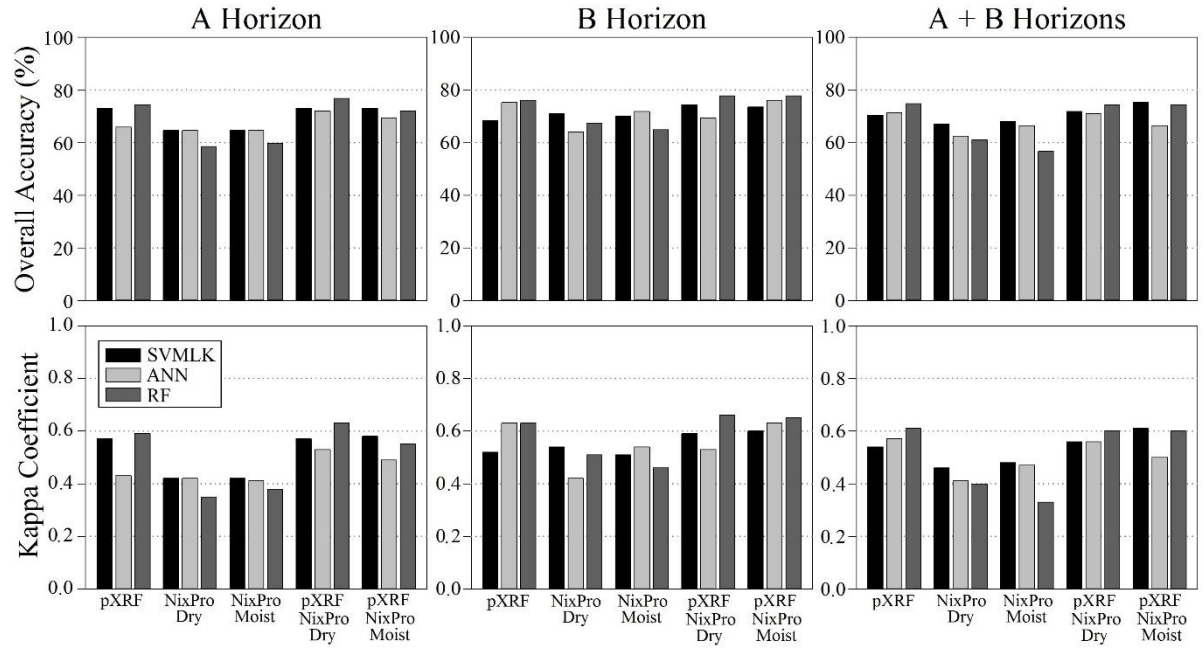

Fig. 5. Overall accuracy and Kappa coefficient values for the prediction models with principal component analysis (PCA) pretreatment in the first level (order) of soil classification for Brazilian tropical soils. SVMLK – Support Vector Machine with Linear Kernel; ANN – Artificial Neural Network; RF – Random Forest.

The prediction models presented a slight increase in accuracy when fitting data from both proximal sensors in B horizon samples. Studying the prediction of S content in lignite, Kagiliery et al. (2019) found a modest increase in predictive accuracy (pXRF = 0.80 R^2, pXRF + NixPro™ = 0.85 R^2) afforded by utilizing combined pXRF and NixPro™ color data. The authors emphasize that NixPro™ color system data is a worthwhile addition since it is inexpensive and quick (<2.5 s).

In general, models generated using data from A horizons and A + B horizons combined provided at least one prediction model, in all different sub-datasets, with overall accuracy surpassing 60%. Surface horizons were generally more affected by the soil management practices such as fertilizers, liming and gypsum application. In addition, soil management can lead to changes in soil organic matter content which may affect soil color. Moreover, nutrient recycling and weathering are also responsible for changing the elemental composition of the soil. The high elemental and color variability in the surface horizons may be the reason for comparatively lower accuracy of the A, and A + B prediction models.

Models developed from NixPro™ color data in isolation delivered lower accuracy; overall accuracy ranged between 56.65% and 68.37%, and Kappa coefficients ranged between 0.34 and 0.51. Dimensional reduction through PCA (Fig. 5) did not perform well, as none of the models achieved 80% overall accuracy.

Analyzing the confusion matrix delivered by the best prediction model (B horizon sub-dataset via pXRF and NixPro™ data with RF algorithm) (Table 3), it is noteworthy that with exception of R, all other soil orders achieved at least 50% classification accuracy. O, C, and L soil orders achieved the best classification accuracy (100, 95.74, and 82.22%, respectively).

Table 3. Confusion matrix and classification accuracy obtained from the best prediction model (pXRF + NixPro™ (moist samples) data) for the first level (order) prediction delivered by Random Forest model in B horizons without principal component analysis (PCA) pretreatment in Brazilian tropical soils.

From\To	P	C	G	L	N	R	O	Total	% correct
P[a]	9	1	0	0	7	0	0	17	52.94
C	1	45	0	1	0	0	0	47	95.74
G	0	1	1	0	0	0	0	2	50.00
L	3	3	0	37	2	0	0	45	82.22
N	0	0	0	2	2	0	0	4	50.00
R	0	0	0	1	0	0	0	1	0.00
O	0	0	0	0	0	0	1	1	100.00

[a] P – Argissolos (Ultisols), C – Cambissolos (Inceptisols), G – Gleissolos (Entisols), L – Latossolos (Oxisols), N – Nitossolos (Ultisols), R – Neossolos (Entisols), O – Organossolos (Histosols).

3.3. Classification performance at the suborder level

In general, the performance of the prediction models for the second level (suborder) of soil classification (Figs. 6 and 7) were modestly worse. The best prediction model was generated using B horizon data via pXRF and NixPro™ (moist samples) data through RF algorithm, resulting in overall accuracy of 74.35% and a Kappa index of 0.65. From the twelve soil suborders, four soils presented 0.0 of classification accuracy, PVA, GM, LA, and RR (Table 4). Conversely, OX and GX achieved 100.0% classification accuracy. Within the order of the Ultisols (Argissolos, P), PV achieved the best classification accuracy (75.0%). Within the order of Oxisols (Latossolos, L), LV and LVA were better classified (79.17 and 81.25%, respectively).

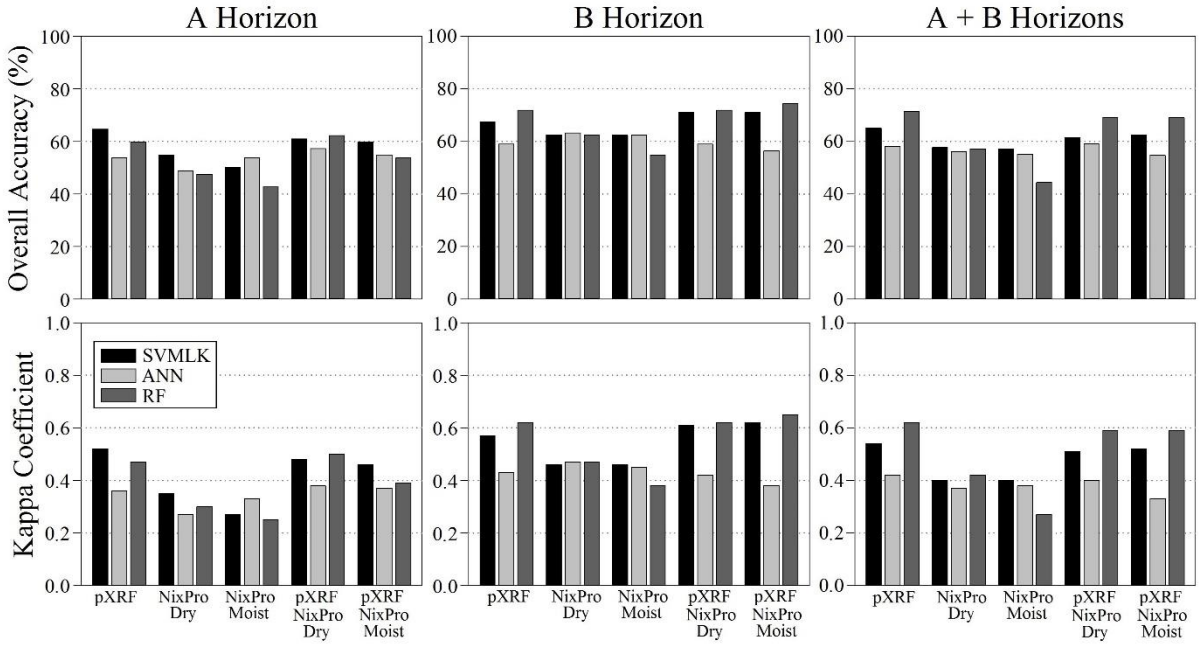

Fig. 6. Overall accuracy and Kappa coefficient values for the prediction models in the second level (suborder) of soil classification for Brazilian tropical soils. SVMLK – Support Vector Machine with Linear Kernel; ANN – Artificial Neural Network; RF – Random Forest.

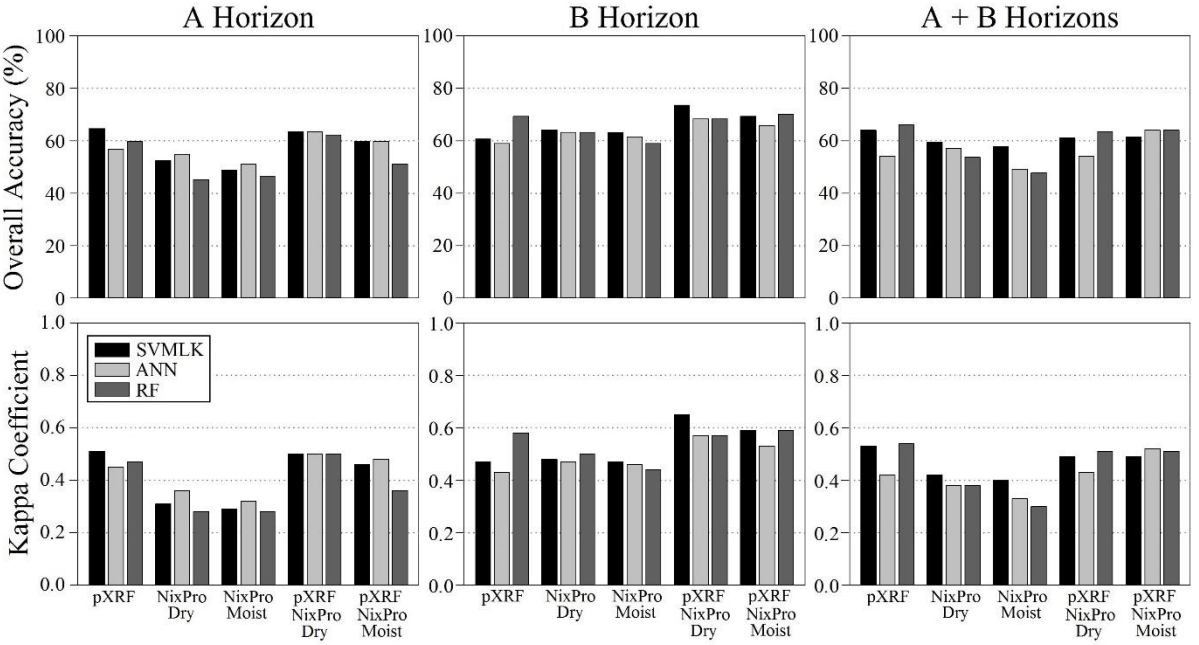

Fig. 7. Overall accuracy and Kappa coefficient values for the prediction models with principal component analysis (PCA) pretreatment in the second level (suborder) of soil classification for Brazilian tropical soils. SVMLK – Support Vector Machine with Linear Kernel; ANN – Artificial Neural Network; RF – Random Forest.

Table 4. Confusion matrix and classification accuracy obtained from the best prediction model (pXRF + NixPro™ (moist samples) data) for the second level (suborder) prediction delivered by Random Forest model in B horizons without principal component analysis (PCA) pretreatment in Brazilian tropical soils.

From\To	PA	PV	PVA	CX	GX	GM	LA	LV	LVA	NV	RR	OX	Total	% correct
PA[a]	1	0	0	1	0	0	0	0	0	0	0	0	2	50.00
PV	0	6	0	0	0	0	0	2	0	0	0	0	8	75.00
PVA	0	1	0	1	0	0	0	0	5	0	0	0	7	0.00
CX	0	0	1	45	0	0	0	0	1	0	0	0	47	95.74
GX	0	0	0	0	1	0	0	0	0	0	0	0	1	100.00
GM	0	0	0	1	0	0	0	0	0	0	0	0	1	0.00
LA	0	0	0	2	0	0	0	0	3	0	0	0	5	0.00
LV	0	2	0	0	0	0	0	19	1	2	0	0	24	79.17
LVA	0	0	0	3	0	0	0	0	13	0	0	0	16	81.25
NV	0	0	1	0	0	0	0	2	0	1	0	0	4	25.00
RR	0	0	0	1	0	0	0	0	0	0	0	0	1	0.00
OX	0	0	0	0	0	0	0	0	0	0	0	1	1	100.00

[a] Abbreviations see Table 1.

The SVMLK models performed slightly better than ANN models with overall accuracy ranging between 50.0 and 70.94% for SVMLK, and between 48.78% and 63.24% for ANN. The RF algorithm delivered the most accurate predictions for suborder models, showing the highest overall accuracy and Kappa coefficient values in almost all sub-datasets (Fig. 6). Other research also found that RF was the best algorithm in predicting TN, CEC, and SOM (Andrade et al., 2020b), exchangeable Ca^{2+} and Mg^{2+}, and available K^{+} (Andrade et al., 2020a), and also, mapping exchangeable Ca^{2+}, Mg^{2+}, Al^{3+}, pH, SOM, base saturation, potential and effective CEC, and P-rem (Silva et al., 2017). Mancini et al. (2019) found that RF was the most consistent algorithm in predicting parent material in tropical soils. Considering the difficulties of soil survey in tropical areas (e.g., scarce financial resources) RF could be an adequate alternative for building reasonable soil order and suborder predictive models.

Although the prediction accuracy for both soil order and suborder were acceptable, the best models present some limitations. In this study, prediction models were built with different subsuperficial soil horizons (E, B, and C). However, the models only apply for the soils listed in Table 1. Further studies are encouraged to apply RF algorithm for the prediction of the thirteen soil orders in Brazilian Soil Classification System (SiBCS) (dos Santos et al., 2018).

3.4. Variable importance

The most important variables for the best soil order and suborder prediction models (pXRF + NixPro™ (moist samples) data) are shown in Fig. 8. K, Si, Rb, and Ti were the most important pXRF variables in both order and suborder prediction models.

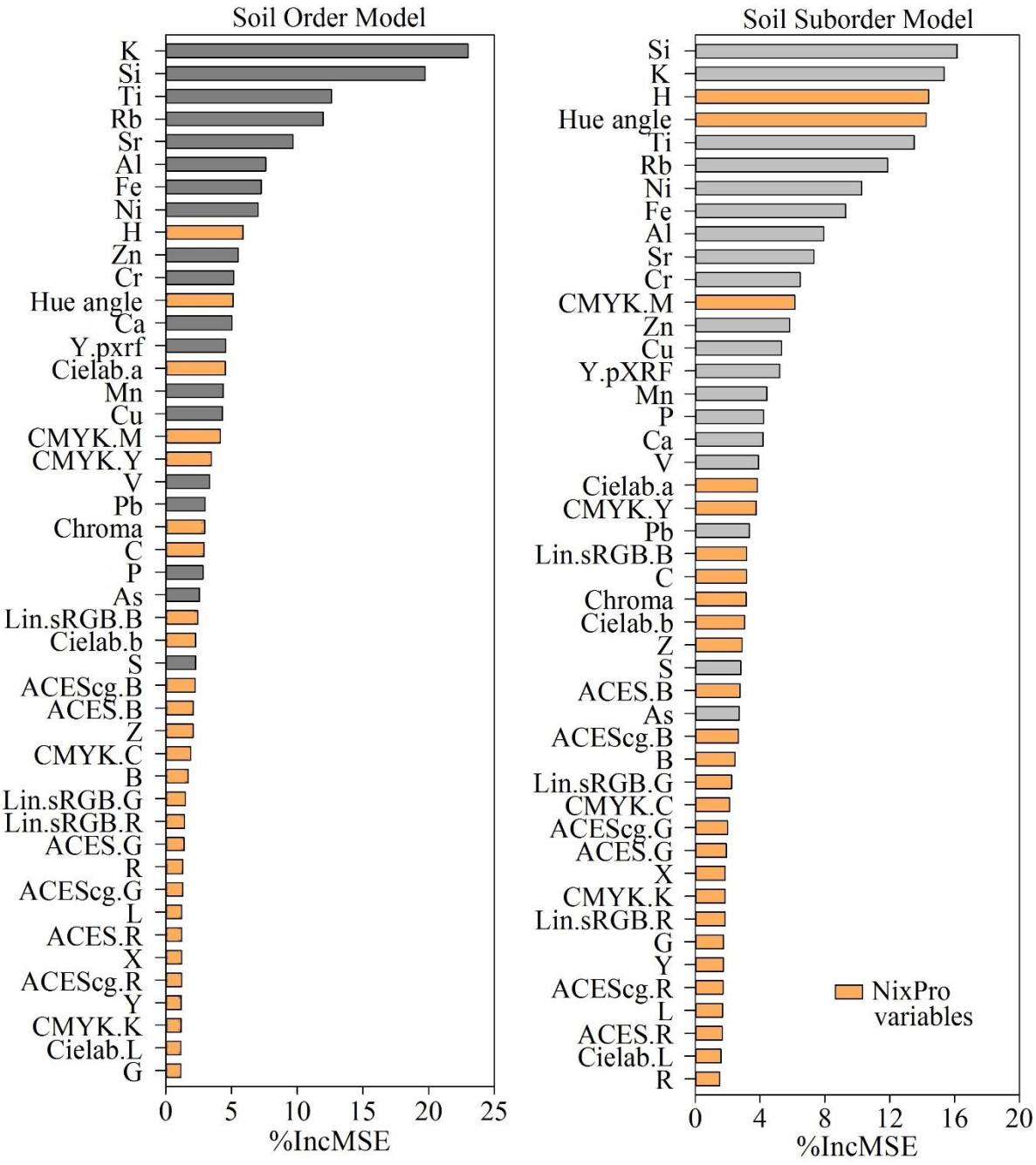

Fig. 8. Variables importance for the best prediction model (portable X-ray fluorescence pXRF + NixPro™ (moist samples) data) for order and suborder predictions delivered by Random Forest model in B horizons without principal component analysis (PCA) pretreatment in Brazilian tropical soils.

Si and K have a very dynamic behavior in the soil environment due to their radical temporal changes caused by soil management. However, even though K is a macronutrient and soil management may be the major influencer in elemental K content (Andrade et al., 2020a), there is some soils that can potentially present kaolinite intergraded with other soil minerals such as mica. Having K as a component of the crystalline structure of the mineral makes this element a marker for the soils that feature such mineralogy, thus assisting algorithms in accuracy predictions. Conversely, Rb and Ti have been considered fingerprint elements for distinguishing soil parent material (Carvalho Filho et al., 2015), since these elements tend to be more stable in soils (Gomes et al., 2017; Schaetzl and Anderson, 2005).

Relating these results to the elemental distribution for different soil suborders (Fig. 2), it is noteworthy that RF models were able to identify as important variables those elements presenting very distinct elemental content distributions. For example, K and Si did not show a clear disparity in their range content; the same was true for Ti and Rb.

The most important numerical color variables for soil order prediction were H (from LCH color system), hue angle (H*, calculated from the CIELab color system), and a* (from CIELab color system). LCH uses cylindrical coordinates instead of rectangular coordinates whereby H reflects hue expressed in degrees. By comparison, a* ranges from greenness (−) to redness (+). For suborder prediction models the most important variables were H, Hue angle (H*), and M (from CMYK color system). CMYK uses a percentage scale (0–100%) to define the contributions of the colors, whereby M represents magenta contribution. The H (from LCH color system) was also the most important NixPro™ color variable for Kagiliery et al. (2019) in predicting lignite S content.

As found by Kagiliery et al. (2019), pXRF provided the most important variables for the prediction models, followed by NixPro™ numerical color information which supports it as an auxiliary proximal sensor. NixPro™ color variables were more important for suborder prediction models (Fig. 8). This was likely because soil color is utilized to define the second level (suborder) of soil classification.

3.5. NixPro™ improvement of soil order and suborder predictions

The inclusion of NixPro™ data in the prediction models helped to improve the accuracy of P, L, and O soil orders, and PA, LV, LVA, and OX soil suborders (Table 5). The major improvements occurred in OX and PA soil suborders, which presented 0.00% accuracy

using only pXRF data, and achieved 100.0 and 50.0% of accuracy respectively, when NixPro™ data (moist samples) were added.

Table 5. Classification accuracy for the best (pXRF + NixPro™ (moist samples) data) and second-best prediction models (pXRF data) for order and suborder predictions delivered by Random Forest model in B horizons without principal component analysis (PCA) pretreatment in Brazilian tropical soils.

Soil order	% correct[b]	% correct[c]
P[a]	47.06	**52.94**
C	95.74	95.74
G	50.00	50.00
L	80.00	**82.22**
N	50.00	50.00
R	0.00	0.00
O	0.00	**100.00**

Soil suborder	% correct[b]	% correct[c]
PA[a]	0.00	**50.00**
PV	75.00	75.00
PVA	14.29	0.00
CX	97.87	95.74
GX	100.00	100.00
GM	0.00	0.00
LA	0.00	0.00
LV	66.67	**79.17**
LVA	75.00	**81.25**
NV	50.00	25.00
RR	0.00	0.00
OX	0.00	**100.00**

Bold numbers represent improved accuracy when adding NixPro™ color data to the prediction models.

[a] Abbreviations see Table 1.

[b] Second-best model.

[c] Best model.

Table 5. Classification accuracy for the best (pXRF + NixPro™ (moist samples) data) and second-best prediction models (pXRF data) for order and suborder predictions delivered by Random Forest model in B horizons without principal component analysis (PCA) pretreatment in Brazilian tropical soils.

P and L feature color in the second level of classification (suborder) in Brazilian Soil Classification System (dos Santos et al., 2018). O, by its proper conditions, feature a darker color due the greater content of soil organic matter (Resende et al., 2014). Thus, although the increase in predictive accuracy afforded by utilizing NixPro™ color data were modest (2.56% for Overall accuracy and 0.03 for Kappa coefficient in soil suborder prediction models), PA and OX only could be more accurately predicted by adding soil color information in the prediction models. This establishes NixPro™ as an important auxiliary proximal sensor in order to improve the accuracy of different soil suborder prediction when the elemental

composition provided by pXRF are no longer efficient in proving soil information for distinguishing soil classes.

3.6. NixPro™ numerical color data: Dry or moist samples?

A CIELab plot for dry and moist soil samples is shown in Fig. 9. Moisture made soil samples appeared darker, decreasing the lightness variable (L*) range, which was from 8.44 to 79.97 for dry samples, and from 10.02 to 58.25 for moist samples. This occurred because the presence of moisture in soil decreases its reflectance. Lobell and Asner (2002) reported that although the soil reflectance changed according to soil order, the oven and air dry soil samples featured a greater reflectance factor, and as moisture increased, the reflectance factor for all soils decreased.

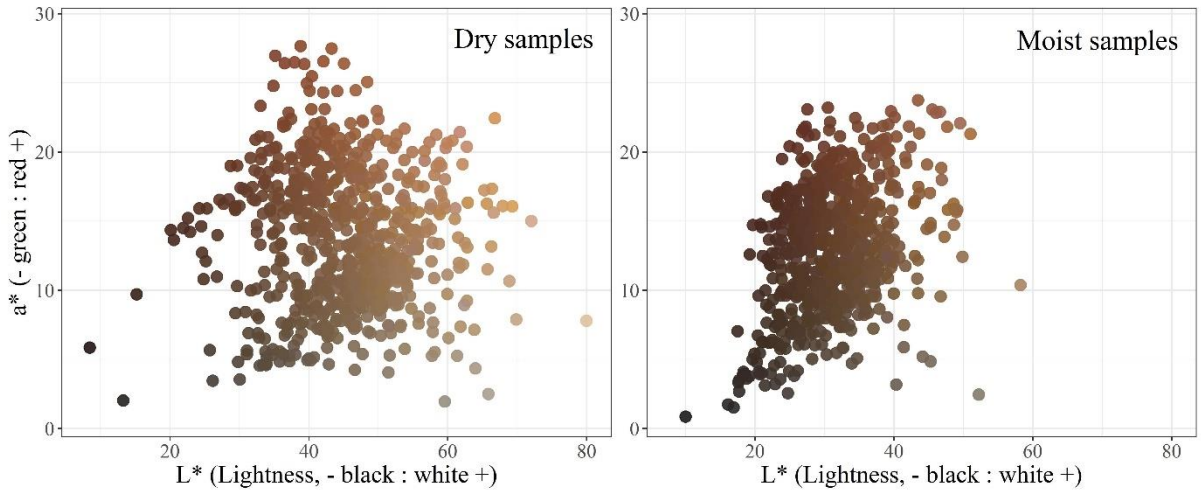

Fig. 9. CIELab color system plot for dry and moist samples of Brazilian tropical soils.

As soil moisture increases, water covers soil particle surfaces, and then proceeds to fill micro and macropores (Hillel, 1998). However, Lobell and Asner (2002) stated that once soil water is sufficient to cover most of the particle surfaces, additional water that fills large pore spaces will no longer cause a large effect on reflectance factor. Other factors like soil texture (sand, silt, and clay contents), surface roughness, the presence of iron oxides and organic matter may also affect soil spectral reflectance (Benedet et al., 2020).

The total color change index (ΔE) was also affected by moisture (Table 6). ΔE reports the total color range variation; the larger the ΔE, the greater the variation in color range (Cakmak et al., 2018; Holman et al., 2019). It is noteworthy that the ΔE index presented greater values under dry conditions, i.e., dry samples delivered a greater range of numerical

soil information for the same soil class which makes prediction difficult for the algorithms. Conversely, soil scans delivered more homogeneous numerical colors when samples were moist, as evidenced by a lower ΔE.

Table 6. Total color change (Δ*E*) for dry and moist condition scans of samples in Brazilian tropical soils.

Soil suborder	Δ*E*[b]	Δ*E*[c]
PA[a]	32.11	14.89
PV	54.32	30.95
PVA	58.14	35.81
CX	57.04	50.39
GX	32.33	22.89
GM	33.38	32.24
LA	43.70	45.15
LV	63.97	36.33
LVA	41.80	35.09
NV	44.83	22.85
RR	50.95	46.91
OX	43.85	28.69

[a] Abbreviations see Table 1.

[b] Dry samples.

[c] Moist samples.

PA, NV, LV, PV, PVA, and OX presented the greatest decreases in ΔE comparing dry and moist samples (53.6, 49.0, 43.2, 43.0, 38.4, and 34.6%, respectively). PA, LV, and OX showed more prediction accuracy when NixPro™ numerical color information from moist samples were added to the prediction models. LVA, even though presenting less ΔE reduction (16.1%), was also more accurately predicted through moist color information.

Stiglitz et al. (2016) reported that the NixPro™ color sensor determined the true color of a soil sample regardless of moisture content, i.e., the results were nearly identical for both dry and moist conditions. Raeesi et al. (2019) found that dry and moist samples presented a different correlation coefficient between the soil color variables and soil organic matter (SOM). Stiglitz et al. (2017) compared the performance of SOC prediction models built with dry and moist sample color information and reported that the results delivered by dry scans presented higher accuracy. In the present study, moist samples featured lower ΔE within the same soil suborder compared with dry samples allowing the algorithms to improve the accuracy when using numerical color data from moist samples (Figs. 4 and 6). Therefore, for

the prediction of soil classes in tropical regions, moist samples deliver more efficient numerical color data. However, what will decide whether the most appropriate dataset is dry or moist is the soil characteristic or attribute to be predicted.

4. Conclusions

Although the models developed in this study likely have limitations, the achieved accuracy for soil order and suborder predictions were moderately reliable. The soil samples used in this study were from a variety of different soils, with different weathering degrees, developed from a myriad of parent materials, with different SOM contents, and a high range of iron content. Even so, the prediction models delivered acceptable results. It is likely that independent models developed for soils of different land-uses may achieve greater accuracy. Also, soils that are not used agriculturally may impact SOM content and the accuracy of the models based on soil color may vary.

Soil order and suborder prediction models offer a rapid and inexpensive method for soil physicochemical assessment which could assist in determining best management practices, most appropriate land-use, and appropriate taxonomic classification. The best prediction models highlight the considerable promise of using the pXRF and NixPro™ techniques for rapidly assessing soil order (81.19% of overall accuracy and 0.71 of Kappa coefficient), and suborder (74.35% and 0.65) through RF algorithm. Although the increase in predictive accuracy afforded by utilizing NixPro™ color data were modest (2.56% for Overall accuracy and 0.04 for Kappa coefficient), PA and OX could only be accurately predicted with soil color information included in the prediction models. Further studies to predict soil order and suborder in field conditions are advisable. Since pXRF and NixPro™ can be used as rapid field assessment tools, this can enable a much higher spatial density of samples which can make prediction models more robust and may improve the understanding of the distribution of soils across landscapes.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors gratefully acknowledge FAPEMIG, CNPq and CAPES Brazilian funding agencies for providing financial resources for this research.

References

Acree, A., Weindorf, D.C., Paulette, L., Van Gestle, N., Chakraborty, S., Man, T., Jordan, C., Prieto, J.L., 2020. Soil classification in Romanian catenas via advanced proximal sensors. Geoderma 377, 114587. https://doi.org/10.1016/j.geoderma.2020.114587

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

Andrade, R., Faria, W.M., Silva, S.H.G., Chakraborty, S., Weindorf, D.C., Mesquita, L.F., Guilherme, L.R.G., Curi, N., 2020a. Prediction of soil fertility via portable X-ray fluorescence (pXRF) spectrometry and soil texture in the Brazilian Coastal Plains. Geoderma 357, 113960. https://doi.org/10.1016/j.geoderma.2019.113960

Andrade, R., Silva, S.H.G., Weindorf, D.C., Chakraborty, S., Faria, W.M., Mesquita, L.F., Guilherme, L.R.G., Curi, N., 2020b. Assessing models for prediction of some soil chemical properties from portable X-ray fluorescence (pXRF) spectrometry data in Brazilian Coastal Plains. Geoderma 357, 113957. https://doi.org/10.1016/j.geoderma.2019.113957

Baize, D., Girard, M.C., 2009. Référentiel Pédologique. Ed. Quae, Versailles, France.

Benedet, L., Faria, W.M., Silva, S.H.G., Mancini, M., Guilherme, L.R.G., Demattê, J.A.M., Curi, N., 2020. Soil subgroup prediction via portable X-ray fluorescence and visible near-infrared spectroscopy. Geoderma 365, 114212. https://doi.org/10.1016/j.geoderma.2020.114212

Breiman, L., 2001. Random Forests. Mach. Learn. 45, 5–32.

Brinatti, A.M., Mascarenhas, Y.P., Pereira, V.P., Partiti, C.S.D., Macedo, A., 2010. Mineralogical characterization of a highly-weathered soil by the Rietveld Method. Sci. Agric. 67, 454–464. https://doi.org/10.1590/s0103-90162010000400013

Cakmak, H., Kumcuoglu, S., Tavman, S., 2018. Production of edible coatings with twin-nozzle electrospraying equipment and the effects on shelf-life stability of fresh-cut apple slices. J. Food Process Eng. 41, e12627. https://doi.org/10.1111/jfpe.12627

Carvalho Filho, A. de, Inda, A.V., Fink, J.R., Curi, N., 2015. Iron oxides in soils of different lithological origins in Ferriferous Quadrilateral (Minas Gerais, Brazil). Appl. Clay Sci. 118, 1–7. https://doi.org/10.1016/j.clay.2015.08.037

Chakraborty, S., Li, B., Weindorf, D.C., Deb, S., Acree, A., De, P., Panda, P., 2019. Use of portable X-ray fluorescence spectrometry for classifying soils from different land use land cover systems in India. Geoderma 338, 5–13. https://doi.org/10.1016/j.geoderma.2018.11.043

CIE, 1978. Recommendations on uniform color spaces - color equations, psychometric color terms. Commission Internationale de l'eclairage, Paris, France.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20, 37–46. https://doi.org/10.1177/001316446002000104

CRGCST, 2001. Chinese soil taxonomy, (3rd Edition). ed. Science Press, Beijing-New York.

de Bakker, H., Schelling, J., 1966. A System of Soil Classification for The Netherlands, the Higher Levels. Pudoc, Wageningen

Donagema, G.K., Campos, D.V.B., Calderano, S.B., Teixeira, W.G., Viana, J.H.M., 2011. Manual de métodos de análise de solo, 2.ed. ed. Embrapa Solos, Rio de Janeiro.

FAO, 2014. World reference base for soil resources 2014. World Soil Resources Reports, Rome.

Florea, N., Munteanu, I., 2000. Romanian system of soil taxonomy. Editura Universitatii Al. I. Cuza, Iassy,.

Gee, G.W., Bauder, J.W., 1986. Particle-size analysis. In: Methods of Soil Analysis: Part 1 - Physical and Mineralogical Methods. Soil Science Society of America, American Society of Agronomy, pp. 383–411. https://doi.org/10.2136/sssabookser5.1.2ed.c15

Gomes, J.B.V., Araújo Filho, J.C., Vidal-Torrado, P., Cooper, M., Silva, E.A. da, Curi, N., 2017. Cemented horizons and hardpans in the Coastal Tablelands of Northeastern Brazil. Rev. Bras. Ciênc. Solo 41. https://doi.org/10.1590/18069657rbcs20150453

González, S., Herrera, F., García, S., 2015. Monotonic random forest with an ensemble pruning mechanism based on the degree of monotonicity. New Gener. Comput. 33, 367–388. https://doi.org/10.1007/s00354-015-0402-4

Grömping, U., 2009. Variable importance assessment in regression: Linear regression versus random forest. Am. Stat. 63, 308–319. https://doi.org/10.1198/tast.2009.08199

Hillel, D., 1998. Environmental soil physics. Academic Press, San Diego, CA.

Holman, B.W.B., Kerr, M.J., Morris, S., Hopkins, D.L., 2019. The identification of dark cutting beef carcasses in Australia, using Nix Pro Color SensorTM colour measures, and their relationship to bolar blade, striploin and topside quality traits. Meat Sci. 148, 50–54. https://doi.org/10.1016/j.meatsci.2018.10.002

Hseu, Z.-Y., Chen, Z.-S., Tsai, C.-C., Jien, S.-H., 2016. Portable X-Ray fluorescence (pXRF) for determining Cr and Ni contents of serpentine soils in the field. In: Hartemink, A.E., Minasny, B. (Eds.), Digital soil morphometrics. Springer International Publishing, Cham, pp. 37–50. https://doi.org/10.1007/978-3-319-28295-4_3

Hu, B., Chen, S., Hu, J., Xia, F., Xu, J., Li, Y., Shi, Z., 2017. Application of portable XRF and VNIR sensors for rapid assessment of soil heavy metal pollution. Plos One 12, e0172438. https://doi.org/10.1371/journal.pone.0172438

Isbell, R., 2016. The Australian Soil Classification. CSIRO Publishing, Collingwood, Victoria. https://doi.org/10.1071/9781486304646

Ishwaran, H., 2007. Variable importance in binary regression trees and forests. Electron. J. Stat. 1, 519–537. https://doi.org/10.1214/07-EJS039

Kämpf, N., Marques, J.J., Curi, N., 2012. Mineralogia de solos brasileiros, in: Ker, J.C., Curi, N., Schaefer, C.E.G.R., Vidal-Torrado, P. (Eds.), Pedologia - Fundamentos. SBCS, Viçosa, pp. 81–146.

Kagiliery, J., Chakraborty, S., Acree, A., Weindorf, D.C., Brevik, E.C., Jelinski, N.A., Li, B., Jordan, C., 2019. Rapid quantification of lignite sulfur content: combining optical and X-ray approaches. Int. J. Coal Geol. 216, 103336. https://doi.org/10.1016/j.coal.2019.103336

Kincey, M., Warburton, J., Brewer, P., 2018. Contaminated sediment flux from eroding abandoned historical metal mines: Spatial and temporal variability in geomorphological drivers. Geomorphology 319, 199–215. https://doi.org/10.1016/j.geomorph.2018.07.026

Koch, J., Chakraborty, S., Li, B., Kucera, J.M., Van Deventer, P., Daniell, A., Faul, C., Man, T., Pearson, D., Duda, B., Weindorf, C.A., Weindorf, D.C., 2017. Proximal sensor analysis of mine tailings in South Africa: An exploratory study. J. Geochem. Explor. 181, 45–57. https://doi.org/10.1016/j.gexplo.2017.06.020

Kottek, M., Grieser, J., Beck, C., Rudolf, B., Rubel, F., 2006. World map of the Köppen-Geiger climate classification updated. Meteorol. Z. 15, 259–263. https://doi.org/10.1127/0941-2948/2006/0130

Kuhn, M., 2008. Building predictive models in R using the caret package. J. Stat. Softw. 28, 1–26. https://doi.org/10.18637/jss.v028.i05

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33, 159–174.

Lemière, B., 2018. A review of pXRF (field portable X-ray fluorescence) applications for applied geochemistry. J. Geochem. Explor. 188, 350–363. https://doi.org/10.1016/j.gexplo.2018.02.006

Liaw, A., Wiener, M., 2002. Classification and regression by random forest. R News 2, 18–22.

Lobell, D.B., Asner, G.P., 2002. Moisture effects on soil reflectance. Soil Sci. Soc. Am. J. 66, 722–727. https://doi.org/10.2136/sssaj2002.7220

Mancini, M., Weindorf, D.C., Chakraborty, S., Silva, S.H.G., dos Santos Teixeira, A.F., Guilherme, L.R.G., Curi, N., 2019. Tracing tropical soil parent material analysis via portable X-ray fluorescence (pXRF) spectrometry in Brazilian Cerrado. Geoderma 337, 718–728. https://doi.org/10.1016/j.geoderma.2018.10.026

Mancini, M., Weindorf, D.C., Monteiro, M.E.C., de Faria, Á.J.G., dos Santos Teixeira, A.F., de Lima, W., de Lima, F.R.D., Dijair, T.S.B., Marques, F.D.A., Ribeiro, D., Silva, S.H.G., Chakraborty, S., Curi, N., 2020. From sensor data to Munsell color system: Machine learning algorithm applied to tropical soil color classification via NixTM Pro sensor. Geoderma 375, 114471. https://doi.org/10.1016/j.geoderma.2020.114471

McLean, E.O., Heddleson, M.R., Bartlett, R.J., Holowaychuk, N., 1958. Aluminum in soils: I. Extraction methods and magnitudes in clays and Ohio soils. Soil Sci. Soc. Am. Proc. 22, 382-387. https://doi.org/10.2136/sssaj1958.03615995002200050005x

Mehlich, A., 1953. Determination of P, Ca, Mg, K, Na and NH4 by North Carolina Soil Testing Laboratories. Raleigh, NC.

Mikhailova, E.A., Stiglitz, R.Y., Post, C.J., Schlautman, M.A., Sharp, J.L., Gerard, P.D., 2017. Predicting soil organic carbon and total nitrogen in the Russian Chernozem from depth and wireless color sensor measurements. Eurasian Soil Sci. 50, 1414–1419. https://doi.org/10.1134/S106422931713004X

Mukhopadhyay, S., Chakraborty, S., Bhadoria, P.B.S., Li, B., Weindorf, D.C., 2020. Assessment of heavy metal and soil organic carbon by portable X-ray fluorescence spectrometry and NixProTM sensor in landfill soils of India. Geoderma Reg. 20, e00249. https://doi.org/10.1016/j.geodrs.2019.e00249

Pelegrino, M.H.P., Weindorf, D.C., Silva, S.H.G., de Menezes, M.D., Poggere, G.C., Guilherme, L.R.G., Curi, N., 2018. Synthesis of proximal sensing, terrain analysis, and parent material information for available micronutrient prediction in tropical soils. Precis. Agric. https://doi.org/10.1007/s11119-018-9608-z

R Development Core Team, 2018. R: A language and environmental for statistical computing. R Found. Stat. Comput.

Raeesi, M., Zolfaghari, A.A., Yazdani, M.R., Gorji, M., Sabetizade, M., 2019. Prediction of soil organic matter using an inexpensive colour sensor in arid and semiarid areas of Iran. Soil Res. 57, 276. https://doi.org/10.1071/SR18323

Rawal, A., Chakraborty, S., Li, B., Lewis, K., Godoy, M., Paulette, L., Weindorf, D.C., 2019. Determination of base saturation percentage in agricultural soils via portable X-ray fluorescence spectrometer. Geoderma 338, 375–382. https://doi.org/10.1016/j.geoderma.2018.12.032

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

Resende, M., Curi, N., Ker, J.C., Rezende, S.B. 2011. Mineralogia de Solos Brasileiros - Interpretações e aplicações, 2. ed. Editora UFLA, Lavras, MG.

Ribeiro, B.T., Silva, S.H.G., Silva, E.A., Guilherme, L.R.G., 2017. Portable X-ray fluorescence (pXRF) applications in tropical Soil Science. Ciênc. E Agrotecnologia 41, 245–254. https://doi.org/10.1590/1413-70542017413000117

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

Santos, R.D. dos, Santos, H.G. dos, Ker, J.C., Cunha dos Anjos, L.H., Shimizu, S.H., 2015. Manual de descrição e coleta de solo no campo, 7° Ed. ed. Sociedade Brasileira de Ciência do Solo, Viçosa, MG.

Schaetzl, R.J., Anderson, S., 2005. Soils – genesis and geomorphology. Cambridge University Press: Cambridge.

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, 141–147. https://doi.org/10.1016/j.geoderma.2014.05.005

Shishov, L.L., Tonkonogov, V.D., Lebedeva, I.I., Gerasimova, V.I., 2004. Classification and diagnostic system of Russian soils. (Eds), Oikumena, Smolensk.

Silva, E.A., Weindorf, D.C., Silva, S.H.G., Ribeiro, B.T., Poggere, G.C., Carvalho, T.S., Gonçalves, M.G.M., Guilherme, L.R.G., Curi, N., 2019. Advances in tropical soil characterization via portable X-ray fluorescence spectrometry. Pedosphere 29, 468–482. https://doi.org/10.1016/S1002-0160(19)60815-5

Silva, S.H.G., Teixeira, A.F. dos S., Menezes, M.D. de, Guilherme, L.R.G., Moreira, F.M. de S., Curi, N., 2017. Multiple linear regression and random forest to predict and map soil properties using data from portable X-ray fluorescence spectrometer (pXRF). Ciênc. E Agrotecnologia 41, 648–664. https://doi.org/10.1590/1413-70542017416010317

Sim, J., Wright, C.C., 2005. The Kappa statistic in reliability studies: Use, interpretation, and sample size requirements. Phys. Ther. 257–268. https://doi.org/10.1093/ptj/85.3.257

Soil classification working group, 1998. The Canadian System of Soil Classification, 3rd edition, NRC Research Press, Ottawa, Canada.

Stiglitz, R., Mikhailova, E., Post, C., Schlautman, M., Sharp, J., 2017. Using an inexpensive color sensor for rapid assessment of soil organic carbon. Geoderma 286, 98–103. https://doi.org/10.1016/j.geoderma.2016.10.027

Stiglitz, R., Mikhailova, E., Post, C., Schlautman, M., Sharp, J., 2016. Evaluation of an inexpensive sensor to measure soil color. Comput. Electron. Agric. 121, 141–148. https://doi.org/10.1016/j.compag.2015.11.014

Stiglitz, R., Mikhailova, E., Sharp, J., Post, C., Schlautman, M., Gerard, P., Cope, M., 2018. Predicting soil organic carbon and total nitrogen at the farm scale using quantitative color sensor measurements. Agronomy 8, 212. https://doi.org/10.3390/agronomy8100212

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 uses. Ciênc. E Agrotecnologia 42, 501–512. https://doi.org/10.1590/1413-70542018425017518

USDA, 1999. Soil taxonomy: A basic system of soil classification for making and interpreting soil surveys, Second Edition. ed. USDA, U.S. Gov. Print. Office, Washington, DC.

USDA, 1938. Soils and men: The 1938 yearbook in agriculture. US Department of Agriculture, USA.

Vasques, G.M., Demattê, J.A.M., Viscarra Rossel, R.A., Ramírez-López, L., Terra, F.S., 2014. Soil classification using visible/near-infrared diffuse reflectance spectra from multiple depths. Geoderma 223–225, 73–78. https://doi.org/10.1016/j.geoderma.2014.01.019

Walkley, A., Black, I.A., 1934. An examination of the Degtjareff method for determinig soil organic matter, and a proposed modification of the chromic acid titration method. Soil Sci. 37, 29–38. https://doi.org/10.1097/00010694-193401000-00003

Weindorf, D.C., Bakr, N., Zhu, Y., 2014. Advances in portable X-ray fluorescence (PXRF) for environmental, pedological, and agronomic applications. In: Advances in Agronomy. Elsevier, pp. 1–45. https://doi.org/10.1016/B978-0-12-802139-2.00001-9

Weindorf, D.C., Chakraborty, S., 2016. Portable X-ray fluorescence spectrometry analysis of soils. In: Methods of Soil Analysis. Madison: Soil Science Society of America, pp. 1–8. https://doi.org/10.2136/methods-soil.2015.0033

Weindorf, D.C., Zhu, Y., Haggard, B., Lofton, J., Chakraborty, S., Bakr, N., Zhang, W., Weindorf, W.C., Legoria, M., 2012a. Enhanced pedon horizonation using portable X-ray fluorescence spectrometry. Soil Sci. Soc. Am. J. 76, 522–531. https://doi.org/10.2136/sssaj2011.0174

Weindorf, D.C., Zhu, Y., McDaniel, P., Valerio, M., Lynn, L., Michaelson, G., Clark, M., Ping, C.L., 2012b. Characterizing soils via portable X-ray fluorescence spectrometer: 2. Spodic and Albic horizons. Geoderma 189–190, 268–277. https://doi.org/10.1016/j.geoderma.2012.06.034

ARTICLE 3 - Proximal sensor data fusion for tropical soil property prediction 1. Soil Texture

This article was prepared following the format style of Geoderma Journal

Renata Andrade[a], Marcelo Mancini[a], Anita Fernanda dos Santos Teixeira[a], Sérgio Henrique Godinho Silva[a], David C. Weindorf[b], Somsubhra Chakraborty[c], Luiz Roberto Guimarães Guilherme[a], Nilton Curi[a]

[a]Department of Soil Science, Federal University of Lavras, Caixa Postal 3037, Lavras, MG Postal Code 37200-000, Brazil

[b]Department of Earth and Atmospheric Sciences, Central Michigan University, Mount Pleasant, MI 48859, USA

[c]Agricultural and Food Engineering Department, Indian Institute of Technology, Kharagpur, West Bengal Postal Code 721302, India

Abstract

Soil texture is a primary variable influencing many soil chemical-physical-biological processes, providing important information for decision-making regarding sustainable soil management. The standard traditional methods for determining soil texture, however, are performed manually and are time-consuming, costly, and generate chemical wastes. As an alternative, portable X-ray fluorescence (pXRF) spectrometry and visible near-infrared spectroscopy (Vis-NIR) have been increasingly used worldwide to predict soil attributes. Other sensors (e.g., NixPro^{TM} color sensor) are also promising, but less evaluated to date. Thus, investigations towards proximal sensor data fusion for prediction of soil textural separates (clay, silt, and total, coarse, and fine sand contents) and soil textural classes (loam, loamy sand, etc) in tropical soils are rare. Therefore, this study aimed to evaluate proximal sensor data for predicting soil particle size fractions and soil textural classes (both Family particle size classes and USDA soil texture triangle) via random forest algorithm in tropical regions. A total of 464 soil samples were collected from A (n = 208) and B (n = 256) horizons

in Brazil. Soil samples were submitted to laboratory analyses for soil texture and proximal sensor (pXRF, Vis-NIR, and NixPro^{TM}) scanning. Samples were randomly split into 70% for modeling and 30% for validation. The best approach varied according to the predicted attribute; however, pXRF data were key information for soil texture prediction accuracy. The best results delivered highly accurate predictions via the aforementioned proximal sensors for rapid assessment of soil texture (total sand R^2 = 0.84, RMSE = 7.60%; silt 0.83, 6.11%; clay 0.90, 5.64%; coarse sand 0.87, 6.30%; fine sand 0.82, 5.27%). Categorical prediction accuracy for soil textural classes (Family particle size classes, overall accuracy = 0.97, Kappa index = 0.95; USDA soil texture triangle, 0.83, 0.73) was enhanced when the predictions were made by soil order sub-datasets. Smoothed Vis-NIR preprocessing and dry NixPro^{TM} color data positively influenced the results. The results reported here represent alternatives for reducing costs and time needed for evaluating soil texture, supporting agronomic and environmental strategies in Brazilian conditions. Further works should extend the results of this study to temperate regions to corroborate the conclusions presented herein regarding the fusion of these three proximal sensors.

Keywords: pXRF, Vis-NIR, NixPro^{TM}, machine learning, prediction models, tropical soils.

1. Introduction

Soil texture represents the mineral particle size distribution and is usually described as the contents of sand (2-0.05 mm), silt (0.05-0.002 mm), and clay (<0.002 mm) (Resende et al., 2014). Sand particles can be divided into coarse sand (0.5–2 mm) and fine sand (0.05–0.5 mm), being the balance of these two fractions highly influential in water retention and infiltration capacity. Coarse-sand-textured soils present greater aeration and water conduction under saturated conditions, while fine-sand-textured soils present smaller pores with more water retention and lower conduction under unsaturated conditions (Parahyba et al., 2019; Resende et al., 2014). Soil texture can also be placed in textural classes, which can reflect a soil's potential uses and limitations, besides being highly relevant for soil classification (Almagro et al., 2021; Groenendyk et al., 2015). Moreover, soil texture provides insights into pedogenetic processes and factors of soil formation and is also related to water infiltration and storage, plant nutrient availability and uptake, soil temperature, erosion susceptibility, ease of compaction (Tümsavaş et al., 2019), cation exchange capacity, soil organic matter content, soil aeration, soil aggregation, and land suitability for different crops (Phogat et al., 2015).

To determine soil particle size distribution, soil samples are subjected to traditional laboratory analysis. The most widespread methodologies are the hydrometer and pipette methods (Baver et al., 1972; Gee and Bauder, 1986), which are time-consuming, costly, and require chemical reagents, restricting the number of samples that can be processed. To overcome these issues, proximal sensors have been used to predict soil properties, including soil texture, optimizing time, reducing costs, and increasing the number of samples analyzed in an environmentally friendly way (Andrade et al., 2020b; Benedet et al., 2020a). Besides the increasing adoption of proximal sensors on this matter, questions remain regarding the need for fusing data of multiple sensors, selection of the data preprocessing method that provides optimal results, extra explanatory variables that may enhance the predictions, etc.

Portable X-ray fluorescence (pXRF) spectrometry, a technique that identifies and quantifies the chemical elements on the analyzed material, and visible near-infrared spectroscopy (Vis-NIR), a technique that provides a spectral signature of the analyzed material, have been successfully utilized for the prediction of soil texture. Soil texture predictions achieved R^2 values ranging from 0.73 to 0.88 via pXRF (Silva et al., 2020), and from 0.70 to 0.81 via Vis-NIR (Conforti et al., 2015). Although these works have been successful in predicting sand, silt, and clay contents, the literature still does not report the use of these sensors, neither separately nor combined, for the prediction of coarse sand and fine sand. Also, other increasingly-used sensors have not been evaluated for such predictions, such as the NixPro^{TM} color sensor.

The NixPro^{TM} color sensor is an inexpensive tool that provides color reports in many different numerical color systems such as RGB, XYZ, CIELAB, LCH, HEX, CMYK, and ACES (Stiglitz et al., 2016). Since it is less subjective and extremely fast reading (1-2 seconds), it has been more used for soil prediction-related studies lately. Depending on the variable to be predicted, the color information can be used directly for the prediction or be used as complementary information. Stiglitz et al. (2018), when predicting ln(Total Nitrogen), reached R^2 0.67 and RMSE 0.53 ln(%) using only CIELAB color system. Still, Andrade et al. (2020c) used pXRF data to predict soil order and suborder *plus* NixPro^{TM} color as auxiliary information, which slightly increased model's accuracy. Although color itself is not correlated with color, as clay-, silt-, and sand-sized particles can assume varied colors depending on the minerals present therein, NixPro^{TM} color data has not yet been tested for soil texture prediction, since in specific conditions color may be able to indirectly help predictions of soil texture in combination with other proximal sensors.

More recently, as proximal sensor popularity has grown, some studies have investigated different approaches to build prediction models aiming to increase their predictive power. One of these approaches is through data fusion, which means gathering different proximal sensor data as explanatory variables aiming to encompass more information about the soil property being predicted. Andrade et al. (2020c), Benedet et al. (2020a, 2020b), Swetha and Chakraborty (2021), Weindorf et al. (2016), Zhang and Hartemink (2020) successfully combined different proximal sensors data resulting in more robust and accurate prediction models. However, investigations towards data fusion for soil texture prediction in tropical soils are still rare. Moreover, studies across the globe that combine pXRF, Vis-NIR, and NixPro^{TM} data for soil texture prediction are unknown to date, to the best of the authors' knowledge. Such investigations are necessary to reliably guide future predictions on tropical environments. Additionally, since the acquisition of pXRF and Vis-NIR spectrometers are expensive, it is important to evaluate in tropical conditions whether combining these sensors provides more robust and accurate prediction models than a single sensor in isolation, since in temperate soils it has already been established (Weindorf and Chakraborty, 2018).

Another approach to attempt to build powerful prediction models is through auxiliary input data, which could be a cheaper and faster solution instead of combining different proximal sensor data. Andrade et al. (2020a) successfully predicted exchangeable Ca^{2+}, Mg^{2+}, and available K^{+} using pXRF *plus* soil texture as auxiliary input data. Stiglitz et al. (2017) built robust prediction models for soil organic carbon through NixPro^{TM} *plus* sample depth as auxiliary information. However, whether this approach works and which auxiliary input data works best for soil texture prediction is not yet known. For instance, soil order, soil horizon, and parent material may enhance such soil texture predictions due to the relations between these factors and particle size distribution (Schaetzl and Anderson, 2015).

Given all the variability that can affect model accuracy (e.g., heterogeneity of the analyzed samples, chosen algorithm, and the applied data preprocessing methods) it is necessary to test different approaches for predicting soil textural separates and soil textural classes in tropical regions. Therefore, this work aimed to predict both on tropical soils through the random forest algorithm, evaluating the feasibility of the following approaches: 1) separately and combined pXRF, Vis-NIR, and NixPro^{TM} data, 2) proximal sensors *plus* environmental co-variates as explanatory variables, and 3) prediction models in sub-datasets separated by soil order. We hypothesize that robust and accurate prediction models will be delivered for soil textural separates (total sand, silt, clay, coarse sand, and fine sand contents)

and soil textural classes by at least one of the aforementioned approaches, despite large variability in soil order, land use, and parent material in the dataset. Furthermore, we hypothesize that pXRF alone will provide results comparable with the data fusion of the three sensors.

2. MATERIAL AND METHODS

2.1 General occurrence and features

This study was conducted using soil samples from four Brazilian states, Minas Gerais (MG), São Paulo (SP), Rio de Janeiro (RJ), and Santa Catarina (SC). Mean annual temperature ranges from 20 to 22 °C, with 1300 to 1600 mm of annual rainfall. The Köppen climate classification (Alvares et al., 2013; Köppen, 1936; Kottek et al., 2006) of the sampled areas are tropical with dry winter (Aw) in Rio de Janeiro state, humid subtropical with dry summer (Csb) in São Paulo state, humid subtropical with temperate summer (Cfb) in Santa Catarina state, tropical with dry winter (Aw) and humid subtropical with dry winter and rainy summer (Cwa) in Minas Gerais state. The soils collected included Ultisols (85 samples), Inceptisols (135), Mollisols (3), Entisols (23), Oxisols (211), Alfisols (2), and Histosols (5) (Soil Survey Staff, 2014). The soil profiles were morphologically described and classified per the *Brazilian Soil Classification System* (SiBCS) (Santos et al., 2018); then, the approximate correspondence was made with the *US Soil Taxonomy* (Soil Survey Staff, 2014).

Soil samples were collected under different parent materials, encompassing amphibole (2 samples), slate (60), Botucatu sandstone (27), Mafra-Formation sandstone (19), basalt (8), limestone (1), charnokite (1), Rio-do-Sul-Formation shale (37), gabbro (42), granite-gneiss (215), itabirite (1), mica-schist (3), pelitic rocks (12), quartzite (23), alluvial sediments (10), and tuffite (3) with varying land uses better described in Andrade et al. (2021). During the soil survey in the field, the sampling points were chosen in areas covering a complete description of the spatial variability of soil classes and properties, besides soil parent material and land use.

2.2 Soil sampling and laboratory analyses

Samples were collected from surface (208 samples) and subsurface (256) horizons, comprising a total of 464 samples. The collected samples were air-dried, disaggregated to pass a 2-mm sieve, and analyzed by pipette method to determine the particle size distribution per the pipette method (Gee and Bauder, 1986) (Fig. 1). The sand fraction was separated into coarse sand (0.5–2 mm) and fine sand (0.05–0.5 mm) through sieving. Soil samples were

classified according to the Family particle size classes and USDA soil texture triangle (Soil Survey Staff, 2014), which were adapted to the Brazilian conditions (Fig. 1).

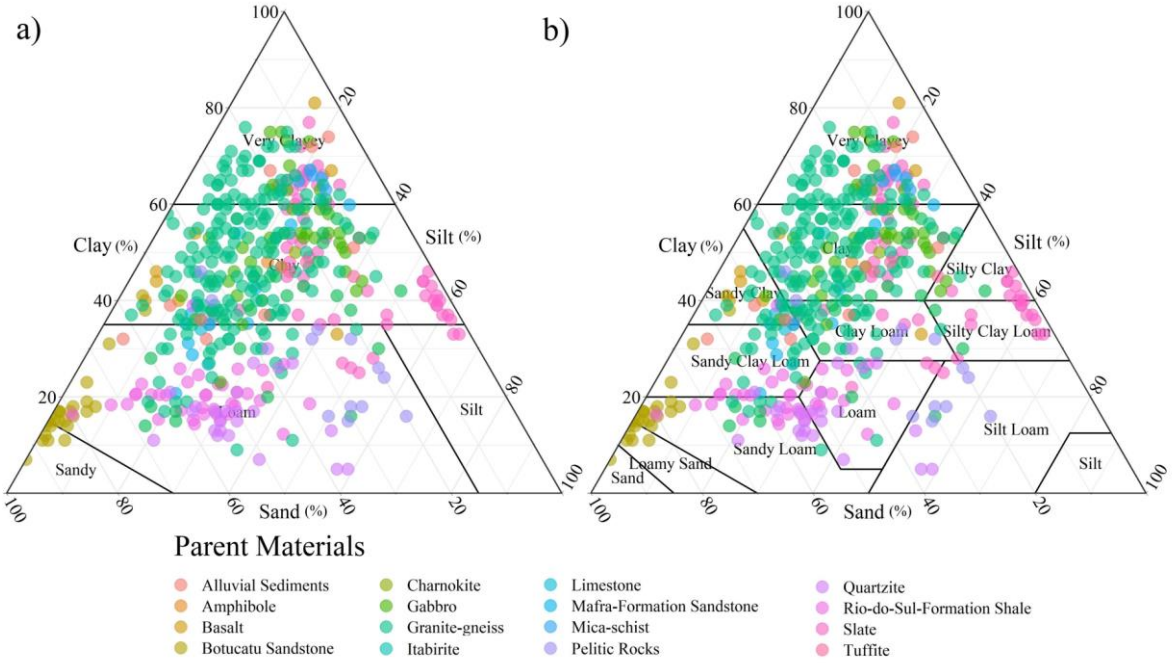

Fig. 1. Soil textural classes of the collected samples and their distribution per parent material of Brazilian soils. a) Family particle size classes, b) USDA soil texture triangle. Both adapted from *US Soil Taxonomy* (Soil Survey Staff, 2014).

2.3 Proximal sensors scanning

2.3.1 pXRF analysis

A pXRF spectrometer (model S1 Titan, Bruker Analytical Instrumentation, Billerica, MA, USA) was applied to scan all soil samples and acquire the elemental composition per Weindorf and Chakraborty (2016). The pXRF features a 50 KeV and 100 μA X-ray tube, which perform the scans of the elements of the Periodic Table ranging from Mg to U in mg kg^{-1}. Scans were performed in air dry samples, in triplicate in Trace (dual soil) mode for 60 s using the Geochem software. In total, the pXRF spectrometer detected the concentration of twenty elements in all soil samples and they were all used in this work: Al, As, Ca, Cl, Cr, Cu, Fe, K, Mn, Ni, P, Pb, Rb, Si, Sr, Ti, V, Y, Zn, and Zr.

To guarantee the quality of the data generated by pXRF, National Institute of Standards and Technology (NIST) certified references (2710a and 2711a) and the pXRF manufacturer check sample had their contents acquired by pXRF contrasted to certified

reference values. From these certified materials, the recovery values obtained by the equipment per element were calculated (recovery value = elemental content obtained by pXRF/certified elemental content). The elements obtained through the pXRF analyses and their recovery values (2710a/2711a/check sample) were: Al (0.88/0.74/0.91), As (0.86/0.64/--), Ca (0.35/0.45/--), Cl (--/--/--), Cr (--/1.19/--), Cu (0.75/0.65/0.91), Fe (0.74/0.66/0.93), K (0.54/0.51/0.83), Mn (0.70/0.61/0.88), Ni (--/0.62/0.90), P (3.62/5.05/--), Pb (1.12/1.05/1.07), Rb (1.01/1.04/--), Si (0.59/0.57/0.94), Sr (2.35/2.07/--), Ti (0.78/0.70/--), V (0.48/0.20/--), Y (--/--/--), Zn (0.91/0.85/--), and Zr (1.08/--/--). Dashes (--) indicates either that the element has no certified content in the reference material or was not detected by pXRF.

2.3.2 NixPro^{TM} analyses

A NixPro^{TM} color sensor (Hamilton, Ontario, Canada) was utilized to collect numerical color data from all soil sample. The sensor is controlled wirelessly by a smartphone or tablet through Bluetooth and has its own light-emitting diode (LED) light source located within the concave base of the sensor about 1 cm above the field of view. The color sensor is inexpensive, and the spectral acquisition range is 380–730 nm. The color system codes reported are interrelated and can be used uniquely identify individual colors of the scanned matrix. The sensor is also rechargeable, easily accessible due its small size, and can be readily recalibrated (Kagiliery et al., 2019).

All samples were leveled to give the sensor a flat area to rest directly on, completely covering the base area, allowing no outside light to enter the scanned sample (Stiglitz et al., 2017, 2016). All samples were scanned under both air dry and moist soil conditions following the methods described previously by Andrade et al. (2020), given the different prediction capacity of samples analyzed in both conditions (Stiglitz et al., 2017). For moist scans, samples were moistened with distilled water using a water dropper to the point of no further color change in the soil. All numerical color data collected from the color system codes were utilized as explanatory variables for the prediction models.

2.3.3 Vis-NIR analysis

Reflectance spectra were acquired in air dry samples, utilizing a model PSR-3500 spectroradiometer (Spectral Evolution, Haverhill, MA, USA) at 1 nm intervals with the spectral bands covering 350 to 2,500 nm. For Vis-NIR scanning, soil samples were homogeneously distributed into Petri dishes. A handheld contact probe featuring a 5W halogen lamp was laid onto the uniformly leveled sample, eliminating the entry of outside

light while acquiring the spectrum. All samples were scanned three times by physically moving the probe between scans; thereafter, the three scans were averaged to ensure homogeneity of scanning. The mean values of the triplicates were used to build the prediction models. Before the scan, for each soil sample, the spectroradiometer was calibrated using a white 12.7-by-12.7-cm NIST traceable radiance calibration panel made of polytetrafluoroethylene (PTFE).

To test the effect of different spectra preprocessing methods on prediction results, as reported by different studies (Benedet et al., 2020b; Conforti et al., 2015; Lazaar et al., 2021; Zhang and Hartemink, 2020), four methods were applied on the raw spectra: i) Savitzky-Golay first-order derivative transformation (FD) - the spectrum is expressed as reflectance (R) as a function of wavelength (λ), [First-order, $dR/d\lambda = f'(\lambda)$] (7 smoothing points); ii) absorbance transformation (Abs) by taking the reciprocal logarithm of the spectrum [Abs = log (1/R)]; iii) smoothed spectrum (Smo) using 2nd-order Savitzky-Golay smoothing method with 11 smoothing points (Savitzky and Golay, 1964); and iv) the reduction from 2151 to 250 spectral variables through binning (Bin) - a method that substitutes spectrum values contained in smaller intervals (original resolution) by mean central values encompassing larger intervals. Thus, including raw spectra, five types of spectral data were tested (Fig. 2). All preprocessing methods were applied via the “prospectr” package (Stevens and Ramirez-Lopez, 2020) in R software (R Development Core Team, 2018). An overview of the raw spectra and all preprocessed spectral curves is given in Fig. 2.

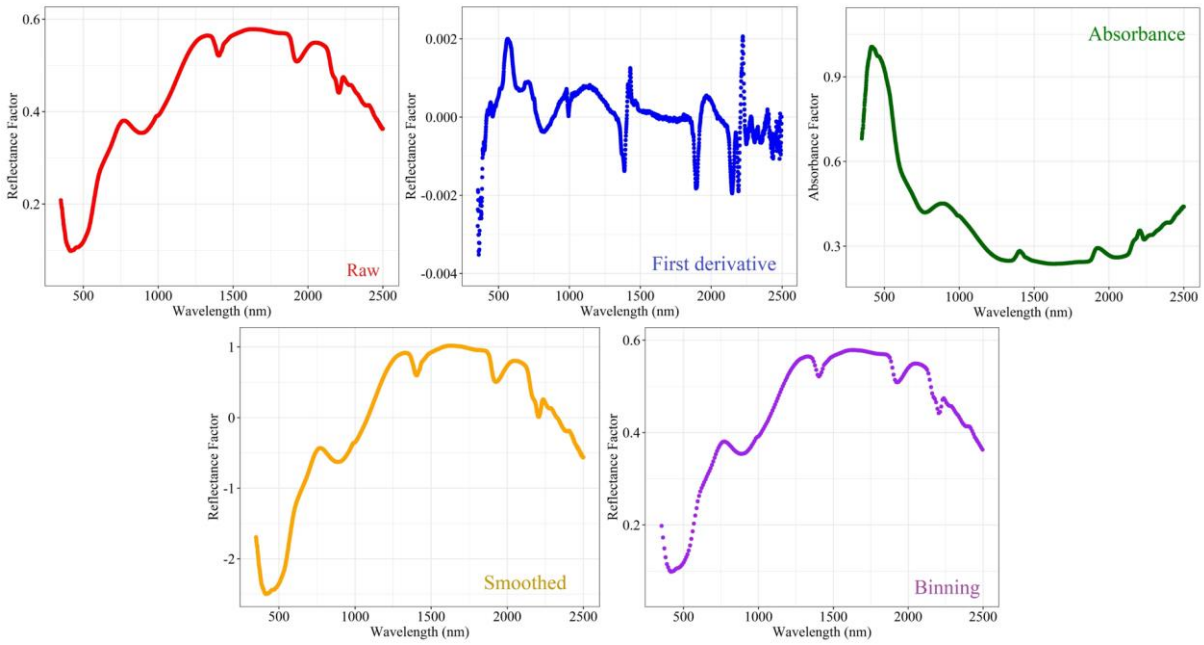

Fig. 2. Overview of raw and preprocessed Vis-NIR mean spectral curves for all soil samples of Brazilian soils.

2.4 Data analyses and modeling

Soil data were randomly separated into modeling and validation sub-datasets consisting of 70% and 30% of the total samples, respectively. The splitting scheme was performed considering all samples from the same soil profile together in the modeling or validation sub-datasets to maintain independence. Initially, 35 different data fusion approaches were used to build the prediction models using proximal sensor data alone and combined (pXRF, Vis-NIR, and NixPro^{TM}) and their respective data pre-treatment (Vis-NIR - Raw, FD, Abs, Smo, Bin) and condition (NixPro^{TM} - dry and moist) (Table 1). Prediction models were built using only A horizon data (208 samples), only B horizon data (256 samples), and using A and B horizon data combined (A+B) (464 samples) totaling 105 prediction models for each soil property.

Table 1. Generated models, preprocessing treatments applied, and the number of explanatory variables utilized to build proximal sensors data fusion prediction models through the random forest algorithm in Brazilian soils.

Models	Predictor variables	Preprocessing[1]	Condition[2]	Number of predictor variables
1	Vis-NIR + NixPro^{TM} + pXRF	Raw		2196
2		First derivative		2190
3		Absorbance	dry	2196
4		Smoothed		2196
5		Binning		295
6	Vis-NIR + NixPro^{TM} + pXRF	Raw		2196
7		First derivative		2190
8		Absorbance	moist	2196
9		Smoothed		2196
10		Binning		295
11	NixPro^{TM} + pXRF	-	dry	45
12	NixPro^{TM} + pXRF	-	moist	45
13	NixPro^{TM} + Vis-NIR	Raw		2176
14		First derivative		2170
15		Absorbance	dry	2176
16		Smoothed		2176
17		Binning		275
18	NixPro^{TM} + Vis-NIR	Raw		2176
19		First derivative		2170
20		Absorbance	moist	2176
21		Smoothed		2176
22		Binning		275
23	NixPro^{TM}	-	dry	25
24	NixPro^{TM}	-	moist	25
25	pXRF	-	-	20
26	pXRF + Vis-NIR	Raw	-	2171
27		First derivative	-	2165
28		Absorbance	-	2171
29		Smoothed	-	2171
30		Binning	-	270
31	Vis-NIR	Raw	-	2151
32		First derivative	-	2145
33		Absorbance	-	2151
34		Smoothed	-	2151
35		Binning	-	250

[1]Preprocessing refers to the Vis-NIR data; [2]Condition in which NixPro^{TM} data was obtained.

Aiming to verify accuracy improvement, the models were built separated for the three main soil orders in the study: Ultisols, Inceptisols, and Oxisols adding more than 315 different prediction models for each soil property. In another approach, additional environmental covariates were used as explanatory variables in the prediction models: soil horizon (SH), soil

order (SO), parent material (PM), and their combinations: SH+SO, SH+PM, SO+PM, and SH+SO+PM, adding 245 more different prediction models.

The same data modeling was performed for categorical predictions of textural classes provided by Family particle size classes and USDA soil texture triangle, both adapted from *US Soil Taxonomy* (Soil Survey Staff, 2014). A flowchart was designed to simplify the understanding of the whole methodological process implemented in this work (Fig. 3).

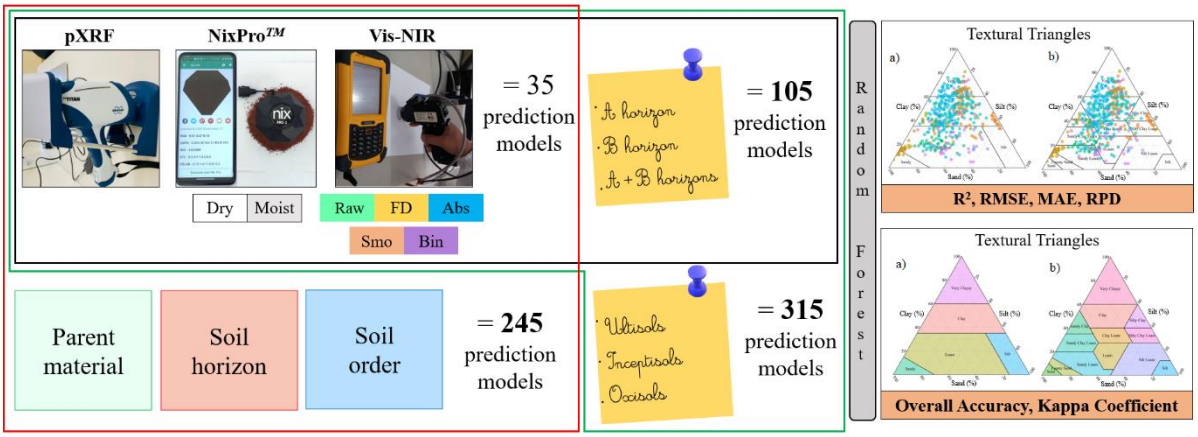

Fig. 3. Flowchart illustrating the different approaches for predicting soil textural separates (total sand, silt, clay, coarse sand, and fine sand content), and soil textural classes (loam, loamy sand, etc) via proximal sensors (pXRF, Vis-NIR, and NixPro^{TM}), separately and combined, through random forest algorithm in Brazilian soils. a) Family particle size classes, b) USDA soil texture triangle. Both adapted from *US Soil Taxonomy* (Soil Survey Staff, 2014). FD - first derivative; Abs - absorbance; Smo - smoothed; Bin - binning. pXRF - portable X-ray fluorescence spectrometry; Vis-NIR - visible near-infrared spectroscopy.

In total, numerical predictions accomplish 3,325 different prediction models (665 for each soil property), and the categorical prediction, 1,330 (665 for each textural triangle). All models were adjusted with the random forest (RF) algorithm in R software (Version 3.4.4) (R Development Core Team, 2018) through the 'caret' package (Kuhn, 2008).

2.5 Evaluating models performance

The accuracy of the predicted total sand, silt, clay, coarse sand, and fine sand numerical contents by random forest algorithm was evaluated by comparing the predicted with the observed values through the coefficient of determination (R^2), root mean square error (RMSE) (Eq. 1), mean absolute error (MAE) (Eq. 2), and residual prediction deviation (RPD)

(Eq. 3). Both MAE and RPD are presented in the Supplementary Material. The equations are given as:

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

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

$$RPD = SD/RMSE \quad (3)$$

where, n: number of observations, y_i: estimated value by the model, m_i: measured value by the chemical analysis, SD: standard deviation of the observed values. RPD has been characterized into three classes: RPD > 2, prediction models delivering accurate results, 1.4 ≤ RPD ≤ 2, prediction models providing moderately accurate results, and RPD <1.4, prediction models being non-reliable (Chang et al., 2001). The models with greater R^2 and smaller RMSE were considered optimal for predicting laboratory analysis.

The validation of predicted textural classes (loam, loamy sand, etc) was performed by overall accuracy and the Cohen's Kappa coefficient calculated by Eqs. (4) and (5), respectively, in a confusion matrix as:

$$Overall\ Accuracy = \frac{Pc}{N} \quad (4)$$

$$Kappa = \frac{Po - Pe}{1 - Pe} \quad (5)$$

where, Pc is the sum of the confusion matrix's main diagonal (predicted soil textural classes), N is the total number of validation samples, Po is the observed agreement, and Pe is the probability of random agreement (Landis and Koch, 1977).

The overall accuracy, ranging from 0 to 1, was calculated by the sum of the correctly classified samples (the major diagonal) divided by the total number of samples in the entire error matrix (the closer to 1, the greater the accuracy) (Congalton, 1991). Kappa coefficient, ranging from −1 to 1, takes into account the number of correctly classified samples, the total number of samples and the misclassifications to deliver the results of the predictions (the closer to 1, the greater the prediction reliability) (Cohen, 1960; Landis and Koch, 1977).

2.6 Relative improvement

To visualize how model accuracy performed in comparison with pXRF-based models, the RMSE increase percentage was calculated (Eq. 6). To calculate this index, the RMSE values achieved by the most accurate prediction model (lowest RMSE) were used to assess the relative improvement (RI) from models built using only pXRF data. Thus, it was possible to assess if other sensor data combinations could outperform pXRF-based prediction models.

In cases that pXRF data alone delivered the best prediction model, the second most accurate prediction model was used for this comparison. The RI equation follows:

$$\mathrm{RI} = \frac{RMSE_{\ compared} - RMSE_{\ pXRF}}{RMSE_{\ pXRF}} * 100 \quad (6)$$

where $RMSE_{\ pXRF}$ is the RMSE of pXRF models, and $RMSE_{\ compared}$ is the RMSE value from models built by other sensor data being compared to the performance of pXRF-based models.

3. RESULTS AND DISCUSSION

3.1 Descriptive statistics for soil texture in tropical soils

The descriptive statistics of total sand, silt, clay, coarse sand, and fine sand contents for A and B horizons separately and combined for the study area indicates the variability of such data (Table 2), mainly demonstrated by the coefficient of variation (CV%). The coarse sand fraction presented the greatest variability with a CV > 82.8% for A and B horizons separately and combined. In this study, the total sand, silt, and clay contents ranged, respectively, from 1.0 to 93.0%, 0.0 to 65.0%, and 5.0 to 81.0%, which is similar to those found by Pinheiro et al. (2018) and Silva et al. (2016) also working in soils from tropical areas.

Table 2. Descriptive statistics of total sand, silt, clay, coarse sand, and fine sand contents (%) for the A and B horizons separately and combined in Brazilian soils.

Particle size class	Horizon	Min.	Max.	Mean	SD[1]	CV (%)[2]
Total sand	A	2.0	93.0	36.4	19.2	52.8
	B	1.0	86.0	33.9	19.6	57.9
	A+B	1.0	93.0	35.0	19.4	55.5
Silt	A	0.0	65.0	24.9	12.8	51.4
	B	0.0	64.0	21.5	13.9	64.6
	A+B	0.0	65.0	23.0	13.5	58.7
Clay	A	7.0	73.0	38.7	15.0	38.7
	B	5.0	81.0	44.7	18.6	41.7
	A+B	5.0	81.0	42.0	17.3	41.3
Coarse sand	A	0.0	67.0	19.2	15.9	82.8
	B	0.0	64.7	18.3	16.7	91.1
	A+B	0.0	67.0	18.7	16.3	87.1
Fine sand	A	1.0	60.0	16.9	12.8	76.0
	B	1.0	65.0	17.5	13.7	78.5
	A+B	1.0	65.0	17.2	13.3	77.3

[1] SD: standard deviation.
[2] CV: coefficient of variation.

The soil samples of the study area covered most of the textural classes (except for sand and silt classes) (Fig. 1). Soil samples came from seven soil orders at different weathering degrees and with varying parent materials, resulting in different proportions of total sand, silt, and clay contents. Due to the greater amount of mafic minerals, and/or a lesser amount of quartz and/or a higher weathering degree, soils derived from gabbro, granite-gneiss, and slate tended to be more clayey and with a high proportion of fine sand in relation to coarse sand (Fig. 4). The opposite occurs for Botucatu sandstone and Rio-do-Sul-Formation shale, which features a large amount of quartz resulting in more sandy soils with greater proportions of coarse sand.

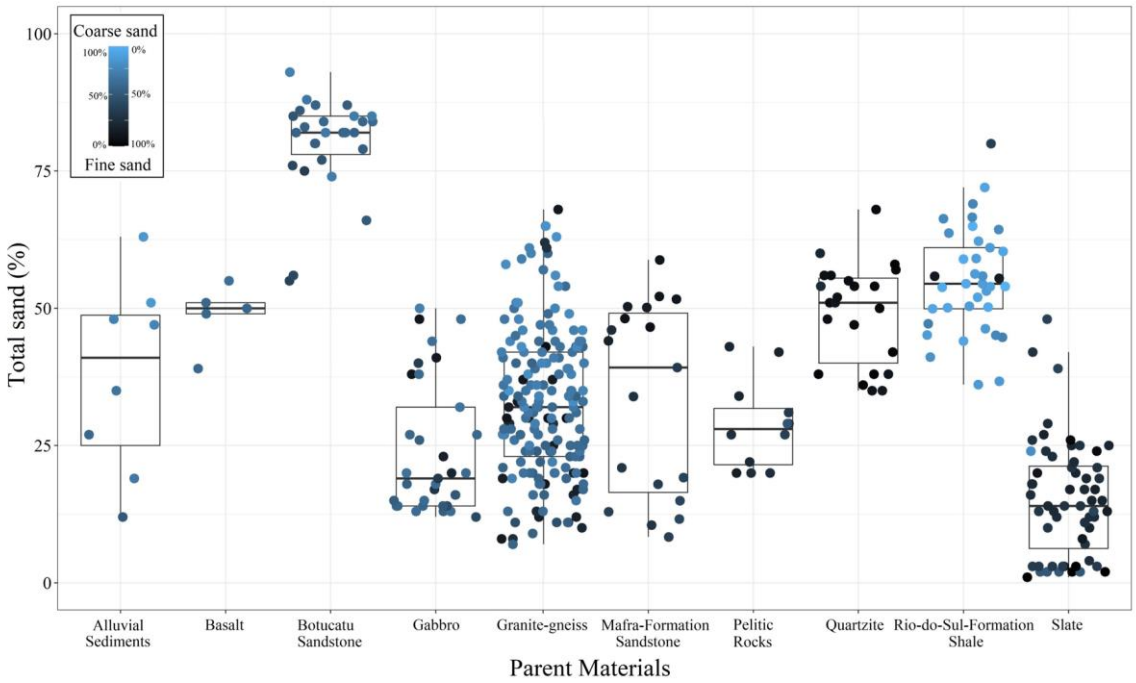

Fig. 4. Coarse and fine sand boxplot separated by soil parent materials of Brazilian soils.

3.2 Prediction models accuracy

3.2.1 Assessing proximal sensors data fusion prediction models

Fig. 5 shows the yielded results of the models using data from A and B horizons, separately and combined, through the random forest algorithm. Overall, the prediction models for the combined horizons (A+B) presented greater R^2 and smaller RMSE (total sand R^2 = 0.84, RMSE = 7.67%; silt 0.83, 6.11%; clay 0.89, 5.75%; coarse sand 0.84, 7.23%; fine sand 0.75, 6.14%) than the models for A and B horizons separately. Models built with combined samples from both surface and subsurface horizons allow for greater applicability of the

prediction models since these models can predict soil texture from samples at any unknown soil horizon, i.e., these models can be used in cases when horizon information is not available. Silva et al. (2020) also found more accurate results for predicting silt and sand contents from the surface and subsurface data combined through the random forest algorithm.

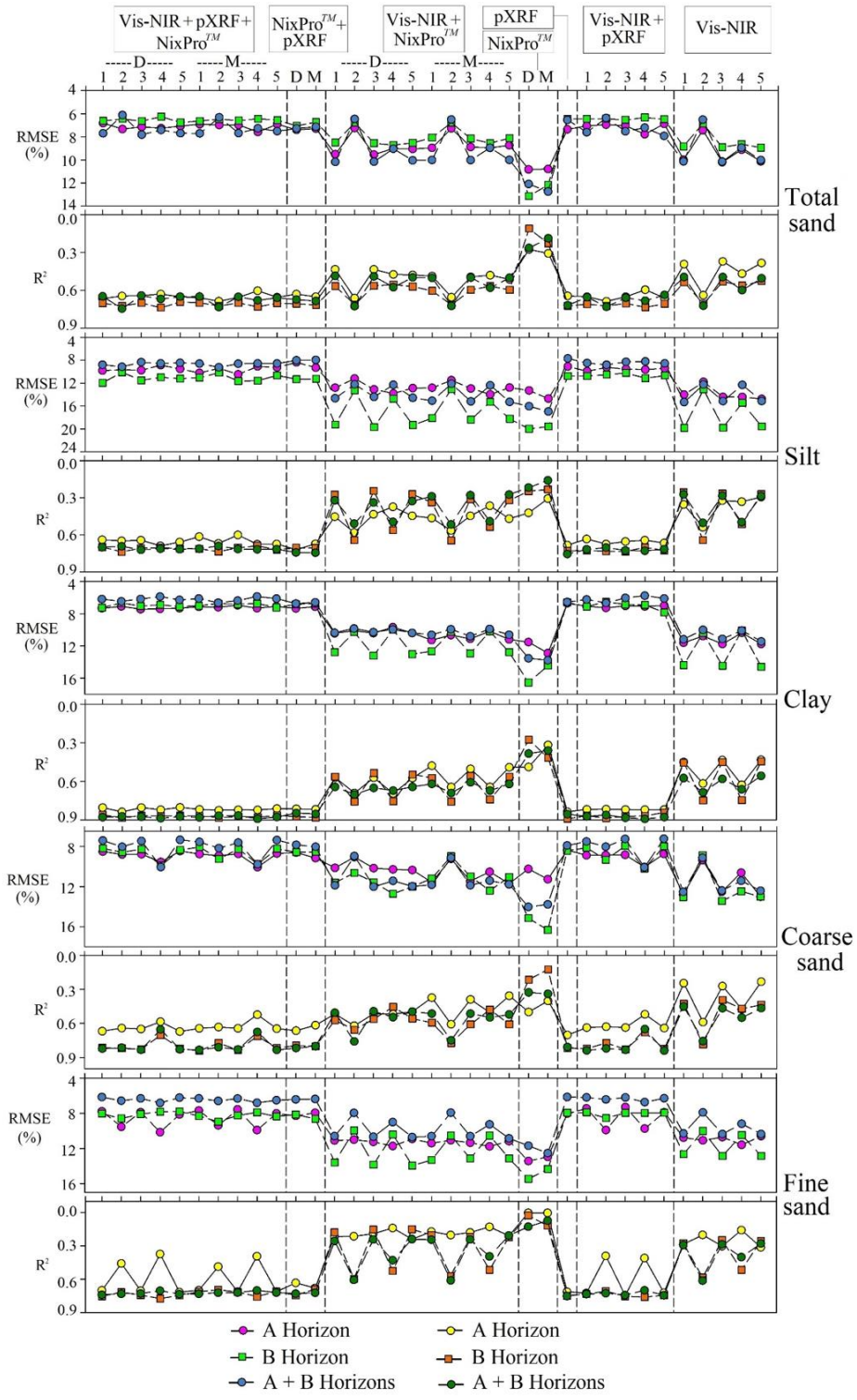

Fig. 5. Coefficient of determination (R^2) and root mean square error (RMSE) for soil texture (total sand, silt, clay, coarse sand, and fine sand contents) prediction models for the A and B horizons separately and combined obtained from the random forest algorithm in Brazilian soils. The numbers refer to Vis-NIR data pretreatment: 1: Raw, 2: First derivative, 3: Absorbance, 4: Smoothed, 5: Binning. The letters refer to NixPro^{TM} scanning condition: D: dry, M: moist. pXRF - portable X-ray fluorescence spectrometry; Vis-NIR - visible near-infrared spectroscopy.

The best proximal sensor data for soil texture prediction, separately or combined, varied according to which soil horizon dataset (A, B, and A+B) the models were trained. For A+B prediction models, the following proximal sensors delivered the most accurate results: total sand pXRF, silt Vis-NIR_{FD}+$NixPro^{TM}_{D}$+pXRF, clay Vis-NIR_{Smo}+pXRF, coarse sand Vis-NIR_{Bin}+pXRF, and fine sand pXRF. Thus, pXRF was the unique sensor that appeared in the best dataset for all the soil properties (5 times), followed by Vis-NIR (3 times) and Nix Pro^{TM} (once). Since each particle size fraction features different mineral composition, different techniques (elemental composition, spectral signature, and soil color) enable the random forest algorithm to make accurate predictions. The relatively simple composition of the sand fraction, dominated by quartz (SiO_2) and muscovite ($KAl_2(AlSi_3O_{10})(F,OH)_2$) in Brazilian soils (Kämpf et al., 2012), may be one of the reasons why elemental composition delivered by pXRF data produced good validation scores. The clay fraction tends to be more diverse in its mineralogy and, hence, chemical composition, thus requiring elemental composition and spectral signature techniques be combined to deliver optimal results. For instance, soils developed from gneiss tend to have kaolinite as the prevalent clay mineral and gibbsite, goethite and hematite as main oxide minerals; while soils derived from gabbro, besides the aforementioned minerals, tend to have a greater proportion of maghemite in the clay fraction (Costa et al., 1999). The silt fraction, intermediate in size between the sand and clay fractions, featured elements/minerals present in the two other fractions and required all three sensors to deliver accurate results.

Although the aforementioned results (Fig. 5) already indicated elemental composition provided by pXRF were the main information used for soil texture prediction, followed by Vis-NIR, the RI of RMSE (RMSE%) highlights a clearer visualization in this regard (Table 3). When comparing the models based only on pXRF data *versus* proximal sensor data combinations, 11 out of 15 RMSE values decreased when other proximal sensor information was added to the modeling. Five decreases in RMSE occurred due to the addition of Vis-NIR data, resulting in reductions from -0.9% to -12.1%. The other six reductions in the RMSE values were due to the incorporation of Vis-NIR and $NixPro^{TM}$ data into the modeling, resulting in reductions from -1.7 to -7.8% in the RMSE values, compared to modeling only with pXRF data. This aspect is also highlighted by Fig. 6, where only models featuring pXRF data delivered the most accurate results, followed by Vis-NIR and then $NixPro^{TM}$.

Table 3. Relative root mean square error improvement (RMSE%) of the best proximal sensor (PS) data fusion prediction models compared with pXRF data-based prediction models through the random forest algorithm in Brazilian soils. Positive and negative RMSE% values indicate decrease and increase, respectively, in RMSE relative to pXRF results.

Dataset	pXRF-based models		PS-based models
		Total sand	
A horizon	pXRF	7.8	Vis-NIR + NixPro^{TM} + pXRF
B horizon	pXRF	6.2	Vis-NIR + NixPro^{TM} + pXRF
A+B horizons	pXRF	-3.4	Vis-NIR + NixPro^{TM} + pXRF
		Silt	
A horizon	pXRF	7.2	Vis-NIR + NixPro^{TM} + pXRF
B horizon	pXRF	3.2	Vis-NIR + NixPro^{TM} + pXRF
A+B horizons	pXRF	6.5	Vis-NIR + NixPro^{TM} + pXRF
		Clay	
A horizon	pXRF	-3.8	Vis-NIR + NixPro^{TM} + pXRF
B horizon	pXRF	0.9	Vis-NIR + pXRF
A+B horizons	pXRF	12.1	Vis-NIR + pXRF
		Coarse sand	
A horizon	pXRF	-1.1	Vis-NIR + NixPro^{TM} + pXRF
B horizon	pXRF	5.7	Vis-NIR + pXRF
A+B horizons	pXRF	8.4	Vis-NIR + pXRF
		Fine sand	
A horizon	pXRF	9.2	Vis-NIR + pXRF
B horizon	pXRF	1.7	Vis-NIR + NixPro^{TM} + pXRF
A+B horizons	pXRF	-0.3	Vis-NIR + NixPro^{TM} + pXRF

pXRF - portable X-ray fluorescence spectrometry; Vis-NIR - visible near-infrared spectroscopy.

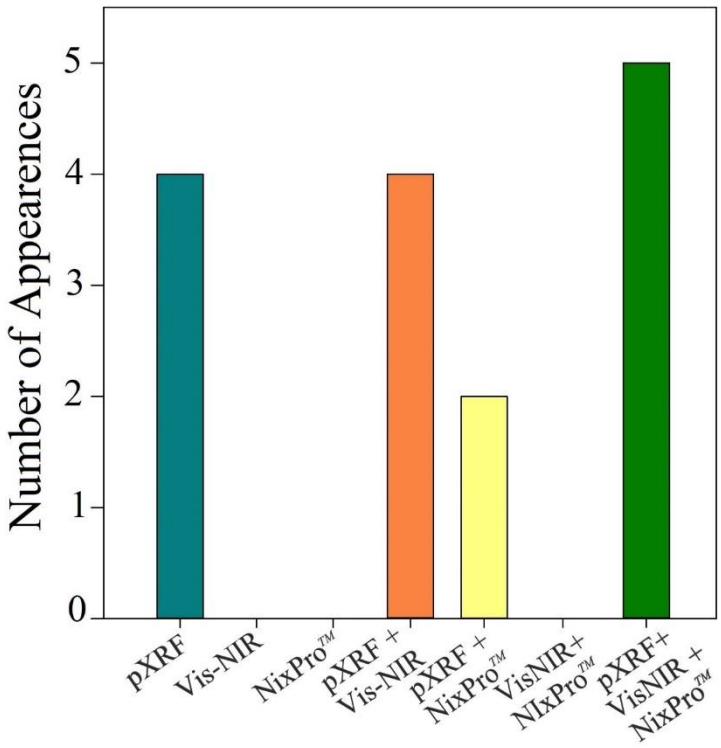

Fig. 6. Number of times each sensor or sensors combined delivered the most accurate prediction model from the random forest algorithm in Brazilian soils. pXRF - portable X-ray fluorescence spectrometry; Vis-NIR - visible near-infrared spectroscopy.

Comparable results were found by Benedet et al. (2020a) and Zhang and Hartemink (2020) testing pXRF and Vis-NIR data separately and combined to predict soil texture. Both studies concluded that pXRF data could be uniquely used to predict soil texture, the former study worked with 315 soil samples on Brazilian conditions, and the latter study used 197 soil samples from Wisconsin, USA. Contrariwise, Wang et al. (2013) working with 276 soil samples in China, found that the most accurate model for the prediction of clay and sand contents was calibrated by fused elemental composition and spectral signature information. Also, Conforti et al. (2015), Curcio et al. (2013), and Lazaar et al. (2021) working on Italy, Sicily, and Morocco, respectively, accurately predicted soil texture fractions using only spectral signature data. Thus, due to variations in the mineral composition of soil fractions according to weathering degree and parent material, the techniques needed to build strong prediction models for soil texture may change. In Brazil, the elemental composition provided by pXRF, followed by spectral signature from Vis-NIR, and NixPro^{TM} numerical color information when available, seems to be a solid base for soil texture predictions according to the obtained results on this research.

Reasonable results were found from Vis-NIR-data based models (total sand R^2 = 0.56, silt 0.80, clay 0.69, coarse sand 0.76, and fine sand 0.61) (Fig. 5). However, for models based on NixPro^{TM} data, none of the predictions delivered accurate results (total sand 0.24, silt 0.29, clay 0.38, coarse sand 0.33, fine sand 0.13). Andrade et al. (2020c) concluded that although the increase in predictive power afforded by applying NixPro^{TM} color data was fairly small (2.56% for Overall accuracy and 0.04 for Kappa coefficient), some soil suborders could only be accurately predicted when soil color information was included in the prediction models, using color data as complementary information to refine model accuracy.

3.2.2 Assessing additional explanatory variables

The validation indexes for soil texture prediction models based on proximal sensors *plus* auxiliary input data are shown in Fig. 7. All soil textural fractions (except for silt) had their RMSE values decreased and R^2 increased when at least one of the natural environmental co-variates was added as explanatory variables to the prediction models [(total sand R^2 = 0.84, RMSE = 7.67% pXRF → 0.84, 7.60% pXRF+PM); (clay 0.89, 5.75% pXRF+Vis-NIR → 0.90, 5.64% pXRF+Vis-NIR+NixPro^{TM}+PM); (coarse sand 0.84, 7.23% pXRF+Vis-NIR → 0.87, 6.30% pXRF+NixPro^{TM}+SH+PM); (fine sand 0.75, 6.14% pXRF → 0.82, 5.27% pXRF+PM)]. For total sand, clay, and fine sand, the best auxiliary input data was PM, and for coarse sand, the best auxiliary input data was SH and PM, which support findings that parent material considerably influences all soil texture fractions (Greve et al., 2012). Stiglitz et al. (2018) used soil depth as auxiliary input information to increase soil organic carbon prediction accuracy.

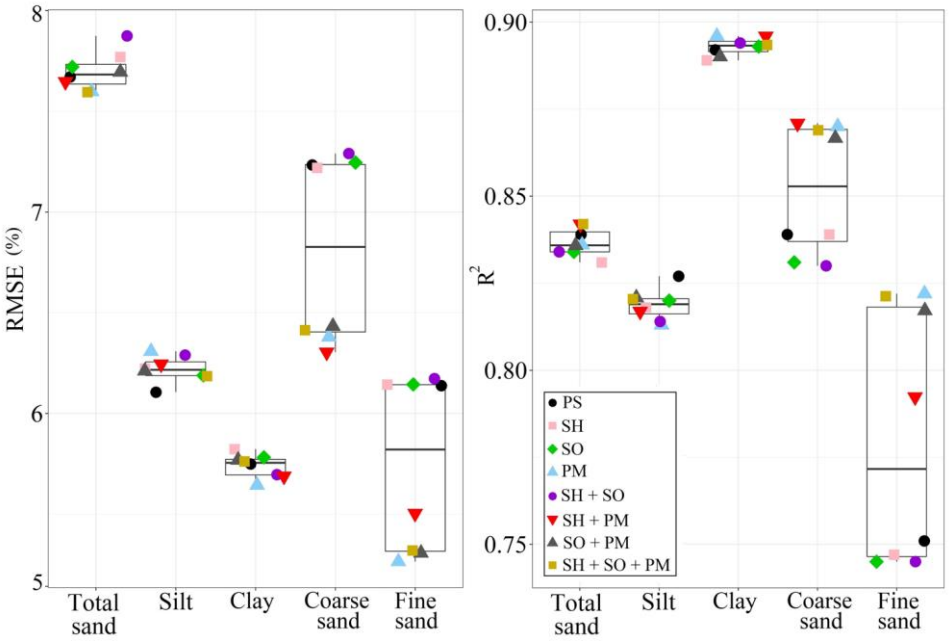

Fig. 7. Determination coefficient (R^2) and root mean square error (RMSE) for the best proximal sensor result *versus* proximal sensors *plus* additional explanatory variables for soil texture (total sand, silt, clay, coarse sand, and fine sand contents) prediction models obtained from the random forest algorithm in Brazilian soils. PS: proximal sensors, SH: soil horizon, SO: soil order, PM: parent material.

3.2.3 Influence of soil order on prediction accuracy

Separating data through a more homogenous condition delivered more accurate results in some sub-datasets. The best improvement was found for the total sand fraction in which higher R^2 and lower RMSE values were found for all soil orders (Fig. 8) compared with the general models. Lower RMSE values were also found for the silt fraction when using the Oxisol sub-datasets, and for fine sand using Ultisol and Oxisol sub-datasets. Thus, dividing soil samples into specific soil orders before model training increases prediction accuracy for total sand, silt, coarse sand, and fine sand textural fractions. Inceptisols delivered the least accurate results.

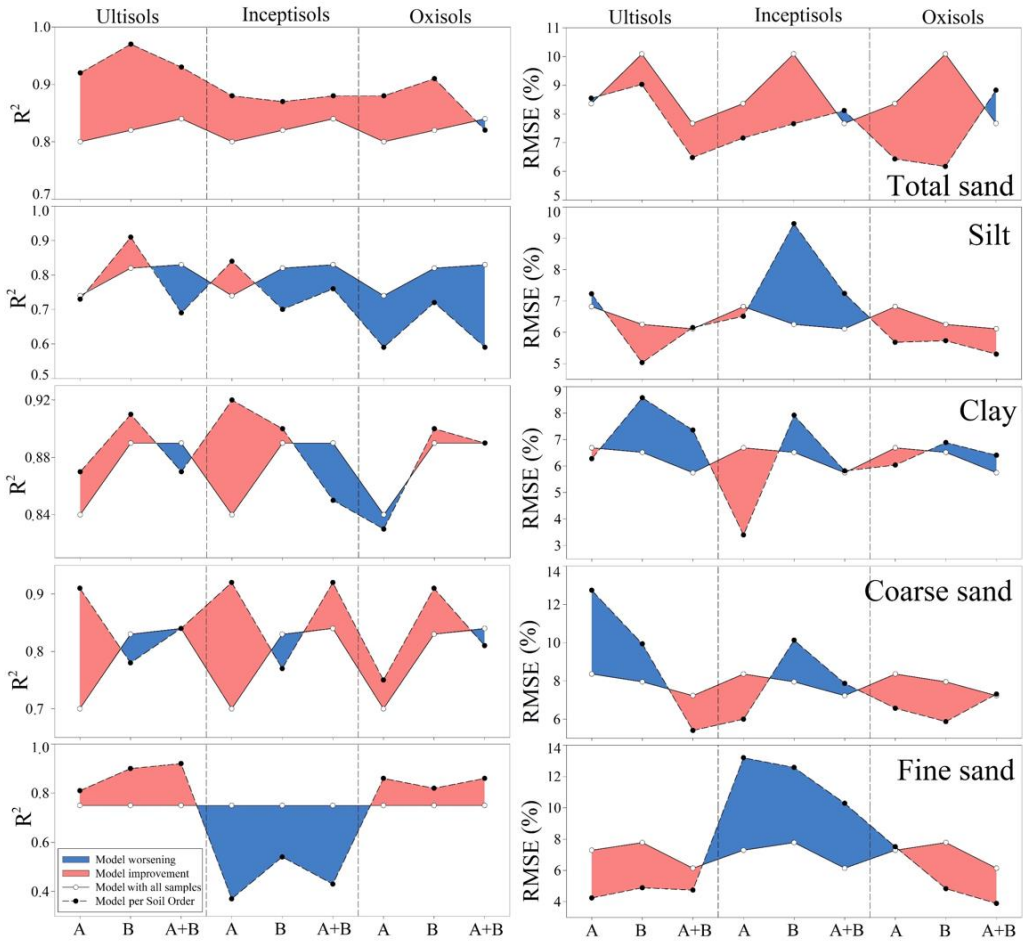

Fig. 8. Determination coefficient (R^2) and root mean square error (RMSE) for soil texture (total sand, silt, clay, coarse sand, and fine sand contents) prediction models for sub-datasets separated by soil order through the random forest algorithm in Brazilian soils.

Stiglitz et al. (2018) also made soil organic carbon (SOC) predictions for the whole dataset, and for soil samples from Alfisols and Entisols separately. The authors found that the highest R^2 was delivered by specific models trained for Alfisol samples (0.81). Thus, dividing the dataset into more homogeneous sub-datasets can lead to more accurate results.

3.2.4 Categorical prediction: A better option?

The predictions for soil textural classes via Family particle size classes and USDA soil texture triangle were made through proximal sensor data fusion, proximal sensor *plus* auxiliary input data, and in sub-datasets separated by soil order. The different approaches resulted in different results for soil textural class predictions. The Family particle size classes, as they present only 5 categories, was easier to predict correctly than the USDA soil texture triangle. This could be observed through the high overall accuracy (correct prediction of soil

textural classes) for the Family particle size classes (0.85) when compared with reasonable accuracy for the USDA soil texture triangle (0.69) (Fig. 9). The overall accuracy of the categories was higher in the Family particle size classes than in the USDA soil texture triangle. The overall accuracy indicated the practical performance of the models while considering the soil textural class, reflecting soil water retention, soil aeration, cation exchange capacity, and many other soil properties (Phogat et al., 2015; Resende et al., 2014; Tümsavaş et al., 2019).

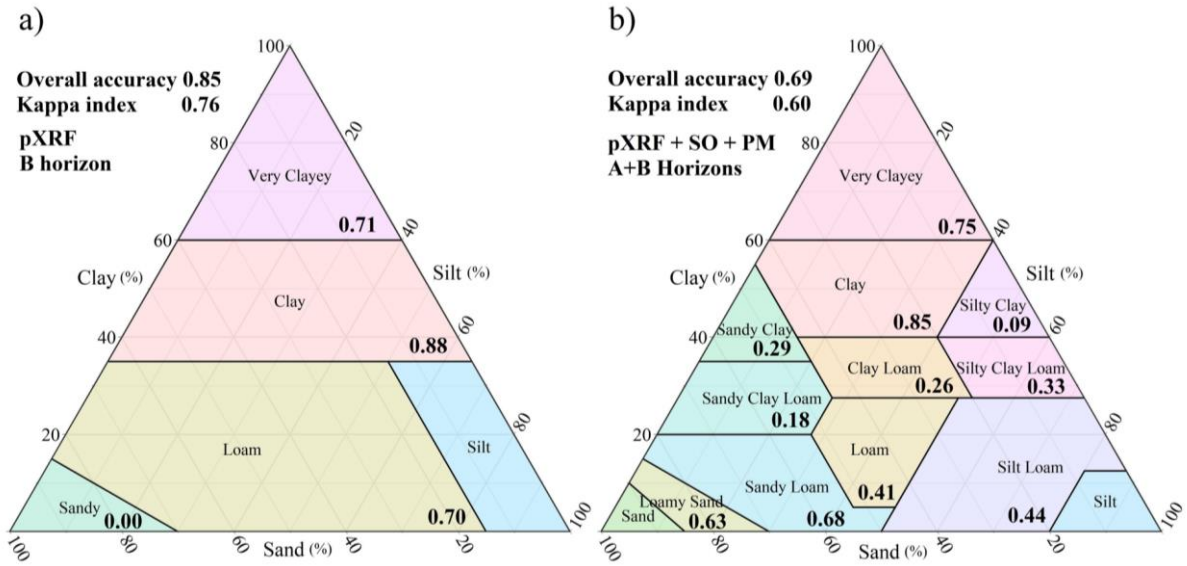

Fig. 9. Overall accuracy and Kappa coefficient values for the best soil texture triangle prediction models through the random forest algorithm in Brazilian soils. Inside bold numbers refer to class accuracy percentage. Classes without accuracy numbers had no samples. a) Family particle size classes, b) USDA soil texture triangle. Both adapted from *US Soil Taxonomy* (Soil Survey Staff, 2014). pXRF - portable X-ray fluorescence spectrometry, SO: soil order, PM: parent material.

The best prediction model for the Family particle size classes was built through pXRF data alone (overall accuracy = 0.85, Kappa coefficient = 0.76) indicating that is possible to predict the soil textural class based only on the soil elemental composition. The prediction models separated by soil order sub-datasets reached even higher accuracy values for Ultisols (0.89, 0.81) and Oxisols (0.97, 0.95) (Table 4), suggesting that in more homogeneous conditions (i.e., samples separated per soil order), the categorical prediction models for soil texture are even more powerful.

Table 4. Overall accuracy and Kappa coefficient values for the best-adjusted models for Family particle size classes and USDA soil texture triangle prediction models through the random forest algorithm in Brazilian soils.

Dataset	Family particle size classes			USDA soil texture triangle		
	Explanatory Variables	Overall Accuracy	Kappa index	Explanatory Variables	Overall Accuracy	Kappa index
A Horizon	pXRF	0.82	0.68	Vis-NIR + NixPro^{TM} + pXRF	0.55	0.41
B Horizon	pXRF	**0.85**	**0.76**	pXRF	0.66	0.57
A+B Horizons	pXRF	0.84	0.74	Vis-NIR + pXRF	0.66	0.55
	Vis-NIR + pXRF + SH	0.84	0.73	pXRF + SH	0.66	0.55
	Vis-NIR + pXRF + SO	0.84	0.73	pXRF + SO	0.67	0.57
	pXRF + PM	0.84	0.74	pXRF + PM	0.68	0.58
	pXRF + SH + SO	0.84	0.73	pXRF + SH + SO	0.68	0.57
	pXRF + SH + PM	0.84	0.73	Vis-NIR + pXRF + SH + PM	0.66	0.55
	NixPro^{TM} + pXRF + SO + PM	0.84	0.73	pXRF + SO + PM	**0.69**	**0.60**
	pXRF + SH + SO + PM	0.83	0.72	pXRF + SH + SO + PM	0.66	0.56
	Ultisols			Ultisols		
A Horizon	Vis-NIR + NixPro^{TM} + pXRF	**0.89**	**0.81**	Vis-NIR + NixPro^{TM} + pXRF	0.56	0.42
B Horizon	pXRF	0.64	0.48	pXRF	0.45	0.33
A+B Horizons	Vis-NIR + NixPro^{TM} + pXRF	0.68	0.52	Vis-NIR + NixPro^{TM} + pXRF	**0.54**	**0.44**
	Inceptisols			Inceptisols		
A Horizon	NixPro^{TM} + pXRF	**0.88**	**0.71**	Vis-NIR + NixPro^{TM} + pXRF	0.59	0.50
B Horizon	Vis-NIR + NixPro^{TM}	0.82	0.68	Vis-NIR + pXRF	**0.59**	**0.53**
A+B Horizons	Vis-NIR + pXRF	0.83	0.66	pXRF	0.51	0.44
	Oxisols			Oxisols		
A Horizon	pXRF	0.89	0.70	Vis-NIR + pXRF	0.65	0.34
B Horizon	pXRF	**0.97**	**0.95**	pXRF	**0.83**	**0.73**
A+B Horizons	Vis-NIR + NixPro^{TM} + pXRF	0.86	0.73	Vis-NIR + NixPro^{TM} + pXRF	0.73	0.58

Optimal values obtained for each soil attribute are given in bold. pXRF - portable X-ray fluorescence spectrometry; Vis-NIR - visible near-infrared spectroscopy. SH: soil horizon, SO: soil order, PM: parent material.

For the USDA soil texture triangle, besides pXRF data, auxiliary input data of SO and PM were needed to achieve the optimal result (0.69, 0.60) (Fig. 9, Table 4). Only sub-dataset prediction models for Oxisols delivered higher accuracy (0.83, 0.73) than the model built with the entire dataset (considering all soil orders). Prediction models built for Inceptisols did not deliver high accuracy values (0.59, 0.53).

Recently, some studies have tried to predict continuous soil properties categorically. Andrade et al. (2021) effectively predicted available B (Overall accuracy = 0.86), Cu^{2+} (0.62), and Mn^{2+} (0.75) availability classes using only pXRF data. Wan et al. (2019), when estimating the classification risk of soil heavy metal contamination, achieved a 0.91 Kappa coefficient through pXRF and Vis-NIR data combined. The advantage of using categorical prediction models is highlighted when only a mixed dataset (numerical and categorical observations) or a large range in values are available to build the prediction models (Weiss, 2015). Categorical models may also be an alternative to surpass dataset outliers (Congalton, 1991). Moreover, since almost all soil properties can be categorized or separated into availability classes, which is largely used by decision-makers from a practical point of view, categorical prediction models could be an elegant solution to rapidly and inexpensively predict soil properties.

3.3 Evaluating Vis-NIR data preprocessing and NixPro^{TM} scanning conditions

Fig. 10 shows the number of times each Vis-NIR preprocessing and NixPro^{TM} condition delivered the most accurate prediction model for the proximal sensors data fusion approach. The best spectrum preprocessing method for soil texture predictions under this study was found to be smoothed, which helped to best cover data variability, delivering a fivefold increase in prediction model accuracy (when considering only Vis-NIR-data based prediction models).

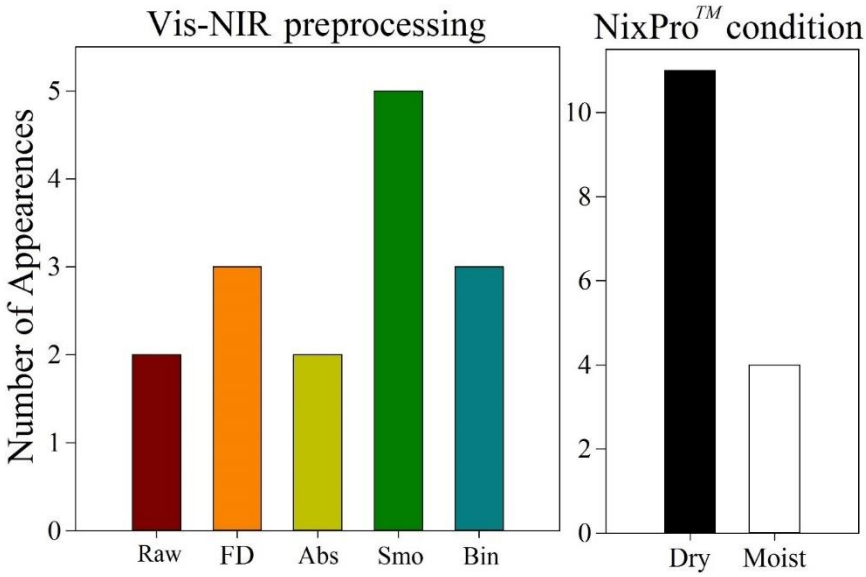

Fig. 10. Number of times each Vis-NIR preprocessing and NixPro^{TM} condition delivered the most accurate prediction model through the random forest algorithm in Brazilian soils. FD - first derivative, Abs - Absorbance, Smo - Smoothed, Bin - binning. Vis-NIR - visible near-infrared spectroscopy.

Vis-NIR data preprocessing is done to improve data quality before modeling, reducing noise, and enhancing the reflectance/absorption frequences. The first derivative removes equally additive and multiplicative influences in the spectra and compensates for instrumental drift (Stevens and Ramirez-Lopez, 2020). Absorbance is used to decrease noise, offset effects, and improve the linearity between the measured absorbance and soil properties (Conforti et al., 2015). Smoothed preprocessing optimizes the signal to noise ratio and reduces the impact of light scattering during spectral acquisition. Binning decreases the impact of noises and errors from the scanning process (Fig. 2) (Stevens and Ramirez-Lopez, 2020).

Some studies using Vis-NIR data to predict soil texture have used other data preprocessing methods to build stronger prediction models, such as multiplicative scatter correction, baseline offset correction, 1st- and 2nd-order detrending (Wang et al., 2013), continuum removal (Benedet et al., 2020a), partial least-squares regression (Curcio et al., 2013), and principal component analysis (Zhang and Hartemink, 2020). This highlights that each dataset, even though predicting the same soil property, requires a specific exploratory data analysis and preprocessing evaluation in order to enhance prediction accuracy.

The best NixPro^{TM} condition for data acquisition was found to be dry, delivering eleven times the most accurate prediction models (when considering only prediction models based on NixPro^{TM} data) compared with just four times for moist conditions. As the samples

with higher clay contents are the ones that present the highest red values when dry (Fig. 11), the models trained with data coming from this condition were able to better distinguish the more clayey ones from the more sandy ones. A CIELab plot for dry and moist soil samples of this study shows how moisture drastically changes soil color, decreasing L (lightness, -black: white+) and a amplitudes (-green: red+) (Fig. 11). Predicting soil order and suborder, Andrade et al. (2020c) found that the best models were built with pXRF *plus* moist NixPro^{TM} data. Conversely when predicting soil organic carbon, Stiglitz et al. (2017) reported that the results delivered by dry scans presented higher accuracy. Thus, whether to use dry or moist conditions for NixPro^{TM} data acquisition seems to vary according to the predicted soil property, dataset variability, and environmental conditions.

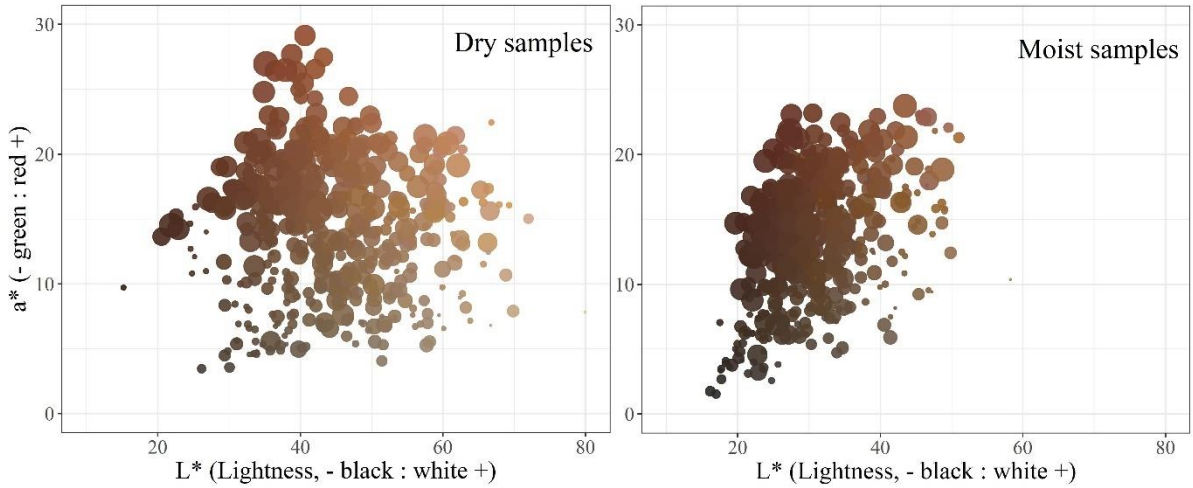

Fig. 11. CIELab color system plot for dry and moist samples of Brazilian soils. The higher the clay content, the larger the point size.

3.4 Operational aspects related to data processing

Initially, a suitable outcome for soil texture prediction was found through the simplest approach, i.e. linear equation (R^2 = 0.75, RMSE = 6.7% for clay content (Broge et al., 2004)). In seventeen years of research, it was possible to build powerful prediction models through machine learning algorithms (R^2 = 0.92, RMSE = 3.4% for clay content (present study)), requiring not only software that can provide resources for such complex modeling, such as the R language (R Development Core Team, 2018), but also computational power and human resources that can master such techniques.

For Vis-NIR data preprocessing methods, a complete package that delivers many preprocessing techniques is very recent (Stevens and Ramirez-Lopez, 2020). This shows that

research towards the operational aspects related to data processing are needed in order to make this novel way to assess soil properties more democratic and accessible to all, and not only for the academic community.

3.5 Models applicability and perspectives

The term *hybrid laboratory* was used by Demattê et al. (2019) to combine soil analytical quality control by both traditional and spectroscopy techniques. Such an approach combines alacrity of processing with cost savings and low environmental impact. Soil analyses provide requisite information in agriculture and environmental monitoring. Thus, a *hybrid laboratory* could perform 20% of the analyses via traditional laboratory analysis and 80% by proximal sensing techniques allowing data acquisition from substantial amounts of samples.

In this sense, the efforts of the academic community to test different approaches to build more robust and accurate prediction models are making *hybrid laboratories* a growing reality around the world. From the first publication on soil texture prediction using proximal sensor data through simple linear approaches (Broge et al., 2004) to the most recent one with more advanced data processing techniques and different machine learning algorithms (Dhawale et al., 2021), many questions were resolved, opening a new frontier with a range of possibilities to be explored. However, other questions were raised, which explains the efforts made herein on the advancement of the best methods to achieve the best predictions (fusion of sensors, modeling per more specific dataset, usage of other auxiliary variables, etc).

Thus, in cases when pXRF, Vis-NIR, and NixPro^{TM} proximal sensors are available, accurate prediction models can be trained for silt (pXRF+Vis-NIR$_{FD}$+NixPro$^{TM}_{d}$), clay (pXRF+Vis-NIR$_{Smo}$+NixPro$^{TM}_{m}$), and coarse sand (pXRF+Vis-NIR$_{Bin}$+NixPro$^{TM}_{d}$). Conversely, pXRF data alone is capable of yielding accurate predictions as well as requiring less data preprocessing, with the support of Andrade et al. (2020b), Benedet et al. (2020b), Gholizadeh et al. (2018), Heggemann et al. (2017), Mahmood et al. (2012), Piikki et al. (2016), Silva et al. (2020), Wang et al. (2013), and Zhu et al. (2011).

4. Conclusions

From the different prediction models trained in this work, the most accurate prediction models highlighted that whether or not to combine proximal sensor data depends on the variable to be predicted (total sand: pXRF; silt: Vis-NIR$_{FD}$+NixPro$^{TM}_{D}$+pXRF; clay: Vis-

NIR_{Smo}+pXRF; coarse sand: Vis-NIR_{Bin}+pXRF; fine sand: pXRF). However, the elemental composition provided by pXRF was central to creating accurate prediction models. The best results achieved for total sand (R^2 = 0.84), silt (0.83), clay (0.90), coarse sand (0.87), and fine sand (0.82) confirm the stated hypothesis that robust and accurate prediction models would be delivered for soil texture by at least one of the tested approaches, even though the dataset included large variability of soil order, land use, and parent material.

Soil texture prediction models built with proximal sensors *plus* auxiliary input data turned out to be effective. All soil textural fractions (except for silt) had their RMSE values decreased and R^2 increased when at least one of the natural environmental co-variates (soil horizon, soil order, and parent material) was added as explanatory variables, with PM being the most important. Creating specific models per soil order also improved the results, mainly for Oxisols (total sand R^2 = 0.91; silt 0.72; clay = 0.90; coarse sand = 0.91; and fine sand = 0.86) and Ultisols (total sand R^2 = 0.97; silt = 0.91; clay = 0.91; coarse sand = 0.91; and fine sand = 0.92).

Categorical predictions are an alternative for soil textural classes prediction (Family particle size classes: overall accuracy = 0.85, Kappa = 0.76; USDA soil texture triangle: 0.69, 0.60), mainly when only a mixed dataset (numerical and categorical observations) or with a large range of values (with outliers) is available to build the prediction models. The best Vis-NIR spectra preprocessing method was Smoothed, followed by First Derivative and Binning. The best NixPro^{TM} scanning condition was on dry soil samples.

The models created in this study can be used to accurately predict soil texture (total sand, silt, clay, coarse sand, and fine sand contents) and soil textural classes (loam, loamy sand, etc). They can be applied to tropical regions including weathered soils developed from sixteen different parent materials. Additionally, predictive models suited very well for soil samples on A+B horizons combined. Further studies are encouraged to extend the findings of this study to other tropical regions with similar soils, in order to rapidly and costly predict soil texture in a large number of samples in an environmentally friendly way.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank the National Council for Scientific and Technological Development (CNPq), Coordination for the Improvement of Higher Education Personnel (CAPES), and Foundation for Research of the State of Minas Gerais (FAPEMIG) for the financial support to develop this research. Also, the authors would like to thank the Texas Tech University High Performance Computing Center for all the support provided.

References

Almagro, A., Oliveira, P.T.S., Meira Neto, A.A., Roy, T., Troch, P., 2021. CABra: a novel large-sample dataset for Brazilian catchments. Hydrol. Earth Syst. Sci. 25, 3105–3135. https://doi.org/10.5194/hess-25-3105-2021

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

Andrade, R., Faria, W.M., Silva, S.H.G., Chakraborty, S., Weindorf, D.C., Mesquita, L.F., Guilherme, L.R.G., Curi, N., 2020a. Prediction of soil fertility via portable X-ray fluorescence (pXRF) spectrometry and soil texture in the Brazilian Coastal Plains. Geoderma 357, 113960. https://doi.org/10.1016/j.geoderma.2019.113960

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

Andrade, R., Silva, S.H.G., Weindorf, D.C., Chakraborty, S., Faria, W.M., Guilherme, L.R.G., Curi, N., 2021. Micronutrients prediction via pXRF spectrometry in Brazil: Influence of weathering degree. Geoderma Reg. 27, e00431. https://doi.org/10.1016/j.geodrs.2021.e00431

Andrade, R., Silva, S.H.G., Weindorf, D.C., Chakraborty, S., Faria, W.M., Guilherme, L.R.G., Curi, N., 2020c. Tropical soil order and suborder prediction combining optical and X-ray approaches. Geoderma Reg. 23, e00331. https://doi.org/10.1016/j.geodrs.2020.e00331

Baver, L.D., Gardner, W.H., Gardner, W.R., 1972. Soil physics, 4.ed. ed. John Wiley & Sons, New York.

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, 114553. 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, 114212. https://doi.org/10.1016/j.geoderma.2020.114212

Broge, N.H., Thomsen, A.G., Greve, M.H., 2004. Prediction of topsoil organic matter and clay content from measurements of spectral reflectance and electrical conductivity. Acta Agric. Scand. Sect. B - Soil Plant Sci. 54, 232–240. https://doi.org/10.1080/09064710410035668

Chang, C.-W., Laird, D.A., Mausbach, M.J., Hurburgh, C.R., 2001. Near-infrared reflectance spectroscopy–principal components regression analyses of soil properties. Soil Sci. Soc. Am. J. 65, 480. https://doi.org/10.2136/sssaj2001.652480x

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20, 37–46. https://doi.org/10.1177/001316446002000104

Conforti, M., Froio, R., Matteucci, G., Buttafuoco, G., 2015. Visible and near infrared spectroscopy for predicting texture in forest soil: an application in southern Italy. IForest - Biogeosciences For. 8, 339–347. https://doi.org/10.3832/ifor1221-007

Congalton, R.G., 1991. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sens. Environ. 37, 35–46. https://doi.org/10.1016/0034-4257(91)90048-B

Costa, A.C.S. da, Bigham, J.M., Rhoton, F.E., Traina, S.J., 1999. Quantification and characterization of Maghemite in soils derived from volcanic rocks in southern Brazil. Clays Clay Miner. 47, 466–473. https://doi.org/10.1346/CCMN.1999.0470408

Curcio, D., Ciraolo, G., D'Asaro, F., Minacapilli, M., 2013. Prediction of soil texture distributions using VNIR-SWIR reflectance spectroscopy. Procedia Environ. Sci. 19, 494–503. https://doi.org/10.1016/j.proenv.2013.06.056

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

Dhawale, N.M., Adamchuk, V.I., Prasher, S.O., Viscarra Rossel, R.A., 2021. Evaluating the precision and accuracy of proximal soil Vis–NIR sensors for estimating soil organic matter and texture. Soil Syst. 5, 48. https://doi.org/10.3390/soilsystems5030048

Gee, G.W., Bauder, J.W., 1986. Particle-size analysis, in: Methods of Soil Analysis: Part 1 - Physical and Mineralogical Methods. Soil Science Society of America, American Society of Agronomy, pp. 383–411. https://doi.org/10.2136/sssabookser5.1.2ed.c15

Gholizadeh, A., Žižala, D., Saberioon, M., Borůvka, L., 2018. Soil organic carbon and texture retrieving and mapping using proximal, airborne and Sentinel-2 spectral imaging. Remote Sens. Environ. 218, 89–103. https://doi.org/10.1016/j.rse.2018.09.015

Greve, M.H., Kheir, R.B., Greve, M.B., Bøcher, P.K., 2012. Quantifying the ability of environmental parameters to predict soil texture fractions using regression-tree model with GIS and LIDAR data: The case study of Denmark. Ecol. Indic. 18, 1–10. https://doi.org/10.1016/j.ecolind.2011.10.006

Groenendyk, D.G., Ferré, T.P.A., Thorp, K.R., Rice, A.K., 2015. Hydrologic-process-based soil texture classifications for improved visualization of landscape function. PLOS ONE 10, e0131299. https://doi.org/10.1371/journal.pone.0131299

Heggemann, T., Welp, G., Amelung, W., Angst, G., Franz, S.O., Koszinski, S., Schmidt, K., Pätzold, S., 2017. Proximal gamma-ray spectrometry for site-independent in situ prediction of soil texture on ten heterogeneous fields in Germany using support vector machines. Soil Tillage Res. 168, 99–109. https://doi.org/10.1016/j.still.2016.10.008

Kagiliery, J., Chakraborty, S., Acree, A., Weindorf, D.C., Brevik, E.C., Jelinski, N.A., Li, B., Jordan, C., 2019. Rapid quantification of lignite sulfur content: combining optical and X-ray approaches. Int. J. Coal Geol. 216, 103336. https://doi.org/10.1016/j.coal.2019.103336

Kämpf, N., Marques, J.J., Curi, N., 2012. Mineralogia de solos brasileiros, in: Pedologia: Fundamentos. SBCS, Viçosa, MG, pp. 81–145.

Köppen, W., 1936. Das geographische System der Klimate, in: Köppen, W., Geiger, R. (Eds.), Handbuch Der Klimatologie, 1. Gebrüder Borntrager, Berlin, pp. 1-44 part C.

Kottek, M., Grieser, J., Beck, C., Rudolf, B., Rubel, F., 2006. World map of the Köppen-Geiger climate classification updated. Meteorol. Z. 15, 259–263. https://doi.org/10.1127/0941-2948/2006/0130

Kuhn, M., 2008. Building predictive models in R using the caret package. J. Stat. Softw. 28, 1–26. https://doi.org/10.18637/jss.v028.i05

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33, 159–174.

Lazaar, A., Pradhan, B., Naiji, Z., Gourfi, A., Hammouti, K.E., Andich, K., Monir, A., 2021. The manifestation of VIS-NIRS spectroscopy data to predict and map soil texture in the Triffa plain (Morocco). Kuwait J. Sci. 48. https://doi.org/10.48129/kjs.v48i1.8012

Mahmood, H.S., Hoogmoed, W.B., Henten, E.J., 2012. Sensor data fusion to predict multiple soil properties. Precis. Agric. 13, 628–645. https://doi.org/10.1007/s11119-012-9280-7

Parahyba, R.D.B.V., Araújo, M.D.S.B.D., Almeida, B.G.D., Rolim Neto, F.C., Sampaio, E.V.S.B., Caldas, A.M., 2019. Water retention capacity in Arenosols and Ferralsols in a semiarid area in the state of Bahia, Brazil. An. Acad. Bras. Ciênc. 91, e20181031. https://doi.org/10.1590/0001-3765201920181031

Phogat, V.K., Tomar, V.S., Dahyia, R., 2015. Soil Physical Properties, in: Soil Science: An Introduction. Indian Society of Soil Science, New Delhi, pp. 135–171.

Piikki, K., Söderström, M., Eriksson, J., Muturi John, J., Ireri Muthee, P., Wetterlind, J., Lund, E., 2016. Performance Evaluation of Proximal Sensors for Soil Assessment in Smallholder Farms in Embu County, Kenya. Sensors 16, 1950. https://doi.org/10.3390/s16111950

Pinheiro, H.S.K., Carvalho Junior, W. de, Chagas, C. da S., Anjos, L.H.C. dos, Owens, P.R., 2018. Prediction of topsoil texture through regression trees and multiple linear

regressions. Rev. Bras. Ciênc. Solo 42. https://doi.org/10.1590/18069657rbcs20170167

R Development Core Team, 2018. R: A language and environmental for statistical computing. R Found. Stat. Comput.

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

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

Savitzky, A., Golay, M.J.E., 1964. Smoothing and differentiation of data by simplified least squares procedures. Anal. Chem. 36, 1627–1639. https://doi.org/10.1021/ac60214a047

Schaetzl, R.J., Anderson, S., 2015. Soil: Genesis and Geomorphology, 2nd ed. Cambridge University Press, New York.

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

Silva, S.H.G., Weindorf, D.C., Pinto, L.C., Faria, W.M., Acerbi Junior, F.W., Gomide, L.R., de Mello, J.M., de Pádua Junior, A.L., de Souza, I.A., Teixeira, A.F. dos S., Guilherme, L.R.G., Curi, N., 2020. Soil texture prediction in tropical soils: A portable X-ray fluorescence spectrometry approach. Geoderma 362, 114136. https://doi.org/10.1016/j.geoderma.2019.114136

Soil Survey Staff, 2014. Keys to Soil Taxonomy, 12th ed. USDA, Washington, DC.

Stevens, A., Ramirez-Lopez, L., 2020. An introduction to the prospectr package. R package Vignette R package version 0.2.1.

Stiglitz, R., Mikhailova, E., Post, C., Schlautman, M., Sharp, J., 2017. Using an inexpensive color sensor for rapid assessment of soil organic carbon. Geoderma 286, 98–103. https://doi.org/10.1016/j.geoderma.2016.10.027

Stiglitz, R., Mikhailova, E., Post, C., Schlautman, M., Sharp, J., 2016. Evaluation of an inexpensive sensor to measure soil color. Comput. Electron. Agric. 121, 141–148. https://doi.org/10.1016/j.compag.2015.11.014

Stiglitz, R., Mikhailova, E., Sharp, J., Post, C., Schlautman, M., Gerard, P., Cope, M., 2018. Predicting soil organic carbon and total nitrogen at the farm scale using quantitative color sensor measurements. Agronomy 8, 212. https://doi.org/10.3390/agronomy8100212

Swetha, R.K., Chakraborty, S., 2021. Combination of soil texture with Nix color sensor can improve soil organic carbon prediction. Geoderma 382, 114775. https://doi.org/10.1016/j.geoderma.2020.114775

Tümsavaş, Z., Tekin, Y., Ulusoy, Y., Mouazen, A.M., 2019. Prediction and mapping of soil clay and sand contents using visible and near-infrared spectroscopy. Biosyst. Eng. 177, 90–100. https://doi.org/10.1016/j.biosystemseng.2018.06.008

Wan, M., Qu, M., Hu, W., Li, W., Zhang, C., Cheng, H., Huang, B., 2019. Estimation of soil pH using PXRF spectrometry and Vis-NIR spectroscopy for rapid environmental risk assessment of soil heavy metals. Process Saf. Environ. Prot. 132, 73–81. https://doi.org/10.1016/j.psep.2019.09.025

Wang, S., Li, W., Li, J., Liu, X., 2013. Prediction of Soil Texture Using FT-NIR Spectroscopy and PXRF Spectrometry With Data Fusion: Soil Sci. 178, 626–638. https://doi.org/10.1097/SS.0000000000000026

Weindorf, D.C., Chakraborty, S., 2018. Portable apparatus for soil chemical characterization. US10107770B2.

Weindorf, D.C., Chakraborty, S., 2016. Portable X-ray fluorescence spectrometry analysis of soils, in: Methods of Soil Analysis. Madison: Soil Science Society of America, pp. 1–8. https://doi.org/10.2136/methods-soil.2015.0033

Weindorf, D.C., Chakraborty, S., Herrero, J., Li, B., Castañeda, C., Choudhury, A., 2016. Simultaneous assessment of key properties of arid soil by combined PXRF and Vis-NIR data: Arid soil assessment by PXRF and Vis-NIR. Eur. J. Soil Sci. 67, 173–183. https://doi.org/10.1111/ejss.12320

Weiss, N.A., 2015. Introductory statistics, 10th Revised ed. ed. Pearson, Boston.

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, 316–333. https://doi.org/10.1111/ejss.12875

Zhu, Y., Weindorf, D.C., Zhang, W., 2011. Characterizing soils using a portable X-ray fluorescence spectrometer: 1. Soil texture. Geoderma 167–168, 167–177. https://doi.org/10.1016/j.geoderma.2011.08.010