University of São Paulo
"Luiz de Queiroz" College of Agriculture

The Brazilian soils from a spectral library: from fundamental to applications

Ariane Francine da Silveira Paiva

Thesis presented to obtain the degree of Doctor in Science. Area: Soil and Plant Nutrition

Piracicaba
2022

Ariane Francine da Silveira Paiva
Environmental Manager

The Brazilian soils from a spectral library perspective: from fundamental to applications

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Advisor:
Prof. Dr. **JOSÉ ALEXANDRE MELO DEMATTÊ**

Thesis presented to obtain the degree of Doctor in Science. Area: Soil and Plant Nutrition

Piracicaba
2022

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA – DIBD/ESALQ/USP

Paiva, Ariane Francine da Silveira

The Brazilian soils from a spectral library: from fundamental to applications / Ariane Francine da Silveira Paiva. - - versão revisada de acordo com a resolução CoPGr 6018 de 2011. - - Piracicaba, 2022.

159 p.

Tese (Doutorado) - - USP / Escola Superior de Agricultura "Luiz de Queiroz".

1. Sensoriamento próximo 2. Biblioteca espectral 3. Espectroscopia 4. Pedometria I. Título

AGRADECIMENTOS

Agradeço a Deus pelo dom da vida, por me sustentar em todos os momentos e por tantas graças concedidas.

À minha família, em especial minha mãe, Alba, que sempre foi meu porto seguro, por confiar e acreditar em mim quando eu mesma não acreditei. Ao meu esposo, Aulísio, por ser meu Tabor e dividir comigo o peso da cruz. Ao meu filho, Rafael, que é luz aos meus dias e motivação para seguir. Ao filho que carrego no ventro, Filipe, que mesmo tão pequeno já traz tanta alegria à minha vida. E ao meu pai, Adilson Gil da Silveira (*in memorian*) e meu padrasto, Sebastião Chiarinelli (*in memorian*), por todos os ensinamentos e carinho.

Ao meu orientador Professor Doutor José Alexandre Melo Demattê pela oportunidade do doutorado e por toda a paciência ao longo desses anos.

Às agências de fomento Capes, CNPq e, em especial, à Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP, processo n° 2016-26176-6) pelos recursos aplicados na minha formação.

À ESALQ/USP por me acolher durante quase dez anos de formação e por me fornecer a estrutura necessária para desenvolver o doutorado.

Ao grupo Geocis (https://esalqgeocis.wixsite.com/geocis) pelo apoio e amizade, nos momentos de trabalho e de descontração. De modo especial, aos pós-doutorandos André Carnieletto Dotto, Raul Roberto Poppiel e Rodnei Rizzo, por todas as contribuições e apoio, fundamentais para a execução desse trabalho.

A todos os pesquisadores que colocaboraram com a construção da Biblioteca doando amostras.

Aos laboratórios de análise de solos que participaram do ProBASE: Sial Solo Análises Laboratoriais, Solocria Laboratório Agropecuário, Laboratório Terra Brasileira, Laboratório Exata, Safrar Análises Agrícolas, Timac Agro Paraguay, Laboratório

Agropecuário Agronômico, Laboratório Solos e Plantas, Laboratório Agroanálise, Laboratório de análise de solos e folha do Instituto Federal do Sudeste de MG, Labras Ambientais e Agrícolas, Laboratório 3Rlab, Laboratório GEAAP, Laboratório Agronômico, DMLab, Solos Laboratório de Análise, Laboratório Água Limpa, Ribersolo, Agrolab Minas, Laboratório Santa Rita, Laboratório Agrolab, Laboratório Agrozap, e tanto colaboraram no desenvolvimento desse trabalho.

Aos técnicos de T.I., Sérgio e Luciano, por todo o trabalho na construção do site: (https://besbbr.com.br).

CONTENTS

RESUMO

Os solos brasileiros de uma biblioteca espectral: do fundamental às aplicações

O solo é um recurso natural fundamental para o equilíbrio da vida no planeta e para o desenvolvimento econômico, no entanto a maior parte das áreas agrícolas do mundo não é totalmente conhecida, o conhecimento das propriedades físico-químicas do solo e sua variabilidade espacial constituem a essência para seu uso sustentável, planejamento e manejo adequado, visando maior produtividade e conservação. Nosso objetivo com esse trabalho foi montar uma Biblioteca Espectral de Solos do Brasil (BESB), avaliar a variabilidade, caracterizar e discriminar os solos a partir dos espectros, além de estudar como uma biblioteca espectral de solo podem auxiliar na rotina de laboratórios tradicionais de solo. Buscamos também, desenvolver um algoritmo que permita a predição de atributos de solos via espectros e a disponibilização esse algoritmo de predição na internet. As amostras de solo foram obtidas pela doação por pesquisadores e laboratórios de todo o Brasil, passaram por processo de determinação espectroscópica na faixa entre 350 a 2500 nm (vis-NIR-SWIR) e algumas delas passaram por análise na faixa de 2500 a 25000nm (MIR) e fluorescência de raio X (XRF). A BESB permite extrair e associar a informação espectral inerente com as variáveis geográficas e ambientais. Com o desenvolvimento da BESB, foi possível: I) demonstrar o potencial de uma biblioteca espectral para o manejo de solos tropicais, e II) relacionar a refletância espectral do solo com as regiões e estados, biomas, geologia, classes de solos e vegetação. Esse estudou provou que as informações espectrais podem ser utilizadas para caracterizar o solo e sua variação e diversidade no solo brasileiro, além de termos conseguido montar um banco de dados com padrões de solos via espectro e modelos de predição de atributos do solo que pode ser acessado pela comunidade e com isso implementar a sua utilização em futuros levantamentos de solos. Foi possível também identificar como o uso de análises espectrais pode ser inserido na rotina dos laboratórios de análise de solo, mostrando o potencial dos laboratórios híbridos.

Palavras-chave: Detecção espectral, Sensoriamento próximo, Espectroscopia, Pedometria

ABSTRACT

The Brazilian soils from a spectral library: from fundamental to applications

Soil is a fundamental natural resource for the balance of life on the planet and for economic development, however most agricultural areas in the world are not fully known, the knowledge of the physical-chemical properties of the soil and its spatial variability constitute the essence for its sustainable use, planning and adequate management, aiming at greater productivity and conservation. Our objective with this work was to set up a Brazilian Soil Spectral Library (BSSL), evaluate the variability, characterize and discriminate soils from the spectra, and study how a soil spectral library can help in the routine of traditional soil laboratories . We also seek to develop an algorithm that allows the prediction of soil attributes via spectra and the availability of this prediction algorithm on the internet. Soil samples were obtained by donation by researchers and laboratories throughout Brazil, underwent a spectroscopic determination process in the range between 350 and 2500 nm (vis-NIR-SWIR) and some of them underwent analysis in the range of 2500 to 25000 nm (MIR) and X-ray fluorescence (XRF). BSSL allows extracting and associating inherent spectral information with geographic and environmental variables. With the development of BSSL, it was possible: I) to demonstrate the potential of a spectral library for the management of tropical soils, and II) to relate the spectral reflectance of the soil with regions and states, biomes, geology, soil classes and vegetation. This study proved that spectral information can be used to characterize the soil and its variation and diversity in the Brazilian soil, in addition to having managed to assemble a database with soil patterns via spectrum and soil attribute prediction models that can be accessed by the community and thus implement its use in future soil surveys. It was also possible to identify how the use of spectral analysis can be inserted into the routine of soil analysis laboratories, showing the potential of hybrid laboratories.

Keywords: Spectral detection, Proximal sensing, Spectroscopy, Pedometrics

1. INITIAL CONSIDERATIONS

Soil is a fundamental natural resource for human activities. The balance of life on the planet and world economic development depends on the quality of this basic resource (Demattê et al., 2019). Therefore, knowledge about their properties and their spatial variability are the first step to plan actions for sustainable use and adequate management (Wall and Nielsen, 2012).

Only in Brazil there are about 66 million ha with strong agriculture activity and remains 11% possible to use (IBGE, 2019). Annually more than 600 million soil samples can be analyzed worldwide (Demattê et al., 2019). These analyzes are carried out by the traditional method of soil analysis that uses chemical products and takes a long time (Viscarra Rossel et al., 2016). Due to the negative impacts of these analyses, it is necessary to seek alternatives that require less time, are cheaper and have less environmental impact. In this context, the analysis of soils via spectroscopy is a viable alternative, since soils have been studied in terms of their spectral behavior, with proven efficiency (Zheng and Schreier, 1988; Ben-Dor and Banin, 1995; Mulder et al., 2011; Luce et al., 2014).

Spectroscopy is a proximal detection technique based on the detection of electromagnetic radiation reflected by the ground. To study the spectral behavior of soils, the spectral range that corresponds to 400 - 700 nm (visible - Vis), 700 - 1100 nm (near infrared - NIR) and 1100-2500 nm (short waves near infrared - SWIR) is commonly used can be obtained by proximal sensors in the field or laboratory.

Electromagnetic radiation reflected from the ground has been studied since Bower and Hanks (1965). Since then, detection data have shown a strong relationship with several soil attributes, such as organic matter, texture, mineral composition, heavy metal content and others (Nocita et al., 2014). The technique allows the simultaneous characterization of soil attributes with the advantage of being a non-destructive method and allowing in situ soil observation (Viscarra Rossel et al., 2006).

Researchers around the world have been looking for a global communication system based on the so-called Soil Spectral Libraries. (SSLs). The first publication using a SSL with global samples was presented by Stoner and Baugarnder (1981), followed by Brown et al. (2006), and Viscarra Rossel et al. (2016), with participation of 92 countries. Other regionals initiatives came along such as the ICRAF-ISRIC Soil VNIR Spectral Library (Garrity and Bindraban, 2004), the LUCAS framework (Land Use/Cover Area Frame Survey; http://eusoils.jrc.ec.europa.eu/projects/Lucas) (Orgiazzi et al., 2018) with data from 23 countries in Europe (Stevens et al., 2013), and the GEOCRADLE with samples from nine countries in Balkan, Middle East, North and central Africa (Tziolas et al., 2019; Shepherd and Walsh, 2002; Summerauer et al., 2021). In addition, numerous countries have also developed local SSLs, such as the Brazilian Soil Spectral Library (BSSL) (Demattê et. al., 2018, Bellinaso et al., 2010), and those from the Czech Republic (Brodsky et al., 2011), France (Gogé et al., 2012), Denmark (Knadel et al., 2012), Mozambique (Cambule et al., 2012), Spain (Bas et al., 2013), Australia (Viscarra Rossel and Webster, 2011), China (Shi et al., 2014; Ji et al., 2016; Liu et al., 2018), USA (Condit, 1970; Wijewardane et al., 2018), New Zealand (Baldock et al., 2019), and Tajikistan (Hergarten et al. 2013).

We expect the Brazilian Soil Spectral Library (BSSL) to demonstrate relationships between inherent spectral information and geographic and environmental variables. This dataset can lead to new approaches in soil detection using near and remote sensing. Thus, the aim of this study is to present the BSSL and its relationship with soil properties and other characteristics that cover most Brazilian soils. In addition to the development of models that allow the prediction of soil attributes via spectra and availability of this prediction algorithm on the internet.

One of our goals was also to bring spectroscopy closer to soil analysis laboratories, disseminating the technique and understanding how the technique can help in the routine of laboratories, reducing the time and cost of soil analysis, in addition to understanding the limitations of the use of spectroscopy.

1.1.THESIS DEVELOPMENT

Samples from all over Brazil were used, aiming to obtain the largest possible number of samples and representativeness within the states. For the preparation of the Spectral Library, the soil samples used in the study were analyzed for physical and chemical determination using traditional methods.

1.1.1. Obtaining samples

Samples and spectral data were obtained in different ways and sources, as follows:

a) Existing spectral data obtained by the Remote Sensing laboratory of the Department of Soil Science, ESALQ-USP. At the beginning of the project, in February 2017, it was a bank with approximately ten thousand samples;

b) Analysis of new samples obtained by the laboratory. About ten thousand new samples were analyzed;

c) Donor data. These are researchers and professors from the most diverse regions of the country who have provided their data to be part of the library;

d) Samples received from soil analysis laboratories participating in the ProBASE (Brazilian Soil Analysis Program via Spectroscopy), created in 2018. We received around 7200 samples from 36 laboratories.

All material was organized and evaluated for results and methodologies. A screening was performed to select the main results. The spectral data were associated with the respective existing soil analyzes (chemical, granulometric, and others).

1.1.2. Soil spectroscopy analysis

For the vis-NIR-SWIR analysis, the soil samples were dried at 45 °C for 48 hours, ground, sieved with 2 mm mesh, and homogeneously distributed in petri dishes prior the measurement of the spectra in the 380-2500 nm range. The spectral data were acquired using the Fieldspec 3 spectroradiometer (Analytical Spectral Devices, ASD, Boulder, CO). The sampling interval was 1 nm, reporting 2151 channels. The light source was provided by two external 50-W halogen lamps. These lamps were positioned at 35 cm from the sample (non-collimated rays and a zenithal angle of 30°) with an angle of 90° between them. The sensor is calibrated using a white Spectralon plate, representing a 100% reflectance standard (reflectance factor 1.0).

For spectral analysis in the mid-infrared (MIR), the soil samples were ground and passed through 100 mesh. Reflectance spectra were obtained with the Alpha Sample Compartment RT-DLaTGS ZnSe (Bruker Optik GmbH) equipped with an accessory for acquiring diffuse reflectance (DRIFT). The sensor has a HeNe laser positioned inside the equipment and a calibration pattern for each wavelength. It has a KBr beam, allowing a high amplitude of the incident radiance to penetrate the sample. Spectra were acquired between (4000 and 600 cm^{-1}; 2,500 – 20,000 nm), with a spectral resolution of 5 cm^{-1} and 32 scans per minute per spectra. A gold reference plate was used as standard, and the sensor was calibrated every four measurements.

The X-ray fluorescence (XRF) analysis was carried out in a portable X-ray fluorescence spectrometer (pXRF) Olympus Delta Professional (Olympus Corporation, Waltham, MA, USA), with two excitation modes (EM). The first EM employed 40 keV, 91.1 µA, and is equipped with a 2 mm aluminum filter, which is most suitable to quantify the following elements: vanadium, chromium, iron, cobalt, nickel, copper, zinc, tungsten, mercury, arsenic, lead, bismuth, rubidium, uranium, strontium, zirconium, yttrium, aurum, thorium, niobium and, molybdenum and secondarily: titanium and manganese. The second EM employed 10 keV and 80.5 μA, that improves the signal of light elements (mainly magnesium, aluminum, and silicon) and quantifies the following elements: magnesium, aluminum, silicon, phosphorus, sulfur, chlorine, calcium, titanium, and manganese. About 5-15 g of sample was placed in a polyethylene bag (4 µm of thickness and 5 cm of width) and submitted to analysis in a platform with protection for X-ray's emission. The pXRF Delta Professional is furnished with a 50 keV silver X-ray anode and a silicon drift detector, with 2048 channels.

REFERENCES

Baldock, J. A., McNally, S. R., Beare, M. H., Curtin, D., & Hawke, B. (2019). Predicting soil carbon saturation deficit and related properties of New Zealand soils using infrared spectroscopy. *Soil Research*, *57*(8), 835-844.

Bas, M. V., Meléndez-Pastor, I., Navarro-Pedreño, J., Gómez, I., Mataix-Solera, J., & Hernández, E. (2013, April). Saline soils spectral library as a tool for digital soil mapping. In *EGU General Assembly Conference Abstracts* (pp. EGU2013-9738).

Bellinaso, H., Demattê, J. A. M., & Romeiro, S. A. (2010). Soil spectral library and its use in soil classification. *Revista Brasileira de Ciência do Solo*, *34*(3), 861-870.

Ben-Dor, E., & Banin, A. (1995). Near infrared analysis (NIRA) as a method to simultaneously evaluate spectral featureless constituents in soils. *Soil Science*, *159*(4), 259-270.

Bower, S. A., & Hanks, R. J. (1965). Reflection of radiant energy from soils. *Soil Sci*, *100*, 130-138.

Brodský, L., Klement, A., Penížek, V., Kodešová, R., & Borůvka, L. (2011). Building soil spectral library of the Czech soils for quantitative digital soil mapping. Soil and water research, 6(4), 165-172.

Brown, D. J., Shepherd, K. D., Walsh, M. G., Mays, M. D., & Reinsch, T. G. (2006). Global soil characterization with VNIR diffuse reflectance spectroscopy. *Geoderma*, *132*(3-4), 273-290.

Cambule, A. H., Rossiter, D. G., Stoorvogel, J. J., & Smaling, E. M. A. (2012). Building a near infrared spectral library for soil organic carbon estimation in the Limpopo National Park, Mozambique. *Geoderma*, *183*, 41-48.

Condit, H. R. (1970). The spectral reflectance of American soils. *Photogrammetric Engineering*.

Demattê, J. A. M., Dotto, A. C., Bedin, L. G., Sayão, V. M., & e Souza, A. B. (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.

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

Garrity, D., & Bindraban, P. (2004). A globally distributed soil spectral library visible near infrared diffuse reflectance spectra. *ICRAF (World Agroforestry Centre)/ISRIC (World Soil Information) Spectral Library: Nairobi, Kenya*.

Gogé, F., Joffre, R., Jolivet, C., Ross, I., & Ranjard, L. (2012). Optimization criteria in sample selection step of local regression for quantitative analysis of large soil NIRS database. *Chemometrics and Intelligent Laboratory Systems*, *110*(1), 168-176.

Hergarten, C., Nazarmavloev, F., & Wolfgramm, B. (2013). Building a soil spectral library for Tajikistan comparing local and global modeling approaches. In *3rd Global Workshop on Proximal Soil Sensing. Potsdam Germany: Leibniz-lnstitute for Agricultural EngineeringPotsdam-Bornim* (pp. 265-269).

IBGE. Monitoramento da cobertura e uso da terra do Brasil. URL: https://biblioteca.ibge.gov.br/index.php/bibliotecacatalogo?view=detalhes&id=2101703. (accessed:09.10.21).

Ji, W., Li, S., Chen, S., Shi, Z., Rossel, R. A. V., & Mouazen, A. M. (2016). Prediction of soil attributes using the Chinese soil spectral library and standardized spectra recorded at field conditions. *Soil and Tillage Research*, *155*, 492-500.

Knadel, M., Deng, F., Thomsen, A., & Greve, M. (2012). Development of a Danish national Vis-NIR soil spectral library for soil organic carbon determination. *Digital Soil Assessments and Beyond*, 403-408.

Liu, Y., Shi, Z., Zhang, G., Chen, Y., Li, S., Hong, Y., Shi, T., Wang, J., & Liu, Y. (2018). Application of spectrally derived soil type as ancillary data to improve the estimation of soil organic carbon by using the Chinese soil vis-NIR spectral library. *Remote Sensing*, *10*(11), 1747.

Luce, M. S., Ziadi, N., Zebarth, B. J., Grant, C. A., Tremblay, G. F., & Gregorich, E. G. (2014). Rapid determination of soil organic matter quality indicators using visible near infrared reflectance spectroscopy. *Geoderma*, *232*, 449-458.

Mulder, V. L., De Bruin, S., Schaepman, M. E., & Mayr, T. R. (2011). The use of remote sensing in soil and terrain mapping—A review. *Geoderma*, *162*(1-2), 1-19.

Nocita, M., Stevens, A., Toth, G., Panagos, P., van Wesemael, B., & Montanarella, L. (2014). Prediction of soil organic carbon content by diffuse reflectance spectroscopy using a local partial least square regression approach. *Soil Biology and Biochemistry*, *68*, 337-347.

Shepherd, K. D., & Walsh, M. G. (2002). Development of reflectance spectral libraries for characterization of soil properties. *Soil science society of America journal*, *66*(3), 988-998.

Shi, Z., Wang, Q., Peng, J., Ji, W., Liu, H., Li, X., & Rossel, R. A. V. (2014). Development of a national VNIR soil-spectral library for soil classification and prediction of organic matter concentrations. *Science China Earth Sciences*, *57*(7), 1671-1680.

Stevens, A., Nocita, M., Tóth, G., Montanarella, L., & van Wesemael, B. (2013). Prediction of soil organic carbon at the European scale by visible and near infrared reflectance spectroscopy. *PloS one*, *8*(6), e66409.

Stoner, E. R., & Baumgardner, M. F. (1981). Characteristic variations in reflectance of surface soils. *Soil Science Society of America Journal*, *45*(6), 1161-1165.

Summerauer, L., Baumann, P., Ramirez-Lopez, L., Barthel, M., Bauters, M., Bukombe, B., Reichenbach, M., Boeckx, P., Kearsley, E., Van Oost, K., Vanlauwe, B., Chiragaga, D., Heri-Kazi, A.B., Moonen, P., Sila, A., Shepherd, K., Mujinya, B.B., Van Ranst, E., Baert, G., Doetterl, S., & Six, J. (2021). Filling a key gap: a soil infrared library for central Africa. *Soil Discussions*, 1-28.

Tziolas, N., Tsakiridis, N., Ben-Dor, E., Theocharis, J., & Zalidis, G. (2019). A memory-based learning approach utilizing combined spectral sources and geographical proximity for improved VIS-NIR-SWIR soil properties estimation. *Geoderma*, *340*, 11-24.

Viscarra Rossel, R., & Webster, R. (2012). Predicting soil properties from the Australian soil visible–near infrared spectroscopic database. *European Journal of Soil Science*, *63*(6), 848-860.

Viscarra Rossel, R., Behrens, T., Ben-Dor, E., Brown, D. J., Demattê, J. A. M., Shepherd, K. D., Shi, Z., Stenberg, B., Stevens, A., Adamchuk, V., Aichi, H., Barthes, B. G., Barthomeus, H. M., Bayer, A. D., Bernoux, M., Bottcher, K., Brodský, L., Du, C. W., Chappell, A., Fouad, Y., Genot, V., Gomez, C., Grunwald, S., Gubler, A., Guerrero, C., Hedley, C. B., Knadel, M., Morrás, H. J. M., Nocita, L., Ramirez-Lopez, L., Roudier, P., Rufasto Campos, E. M., Sanborn, P., Sellito, V. M., Sudduth, K. A.,Rawlins, B.G., Walter, C., Winowiecki, L. A., Hong, S. Y., & Ji, W. (2016). A global spectral library to characterize the world's soil. Earth-Science Reviews, 155, 198-230. doi:10.1016/j.earscirev.2016.01.012.

Wall, D. H., & Nielsen, U. N. (2012). Biodiversity and ecosystem services: is it the same below ground. *Nature Education Knowledge*, *3*(12), 8.
Wijewardane, N. K., Ge, Y., Wills, S., & Libohova, Z. (2018). Predicting physical and chemical properties of US soils with a mid-infrared reflectance spectral library. *Soil Science Society of America Journal*, *82*(3), 722-731.

Zheng, F., & Schreier, H. (1988). Quantification of soil patterns and field soil fertility using spectral reflection and digital processing of aerial photographs. *Fertilizer research*, *16*(1), 15-30.

2. THE BRAZILIAN SOIL SPECTRAL LIBRARY (BSSL): A GENERAL VIEW, APPLICATION AND CHALLENGES

ABSTRACT

The present study was developed in a joint partnership with the Brazilian pedometrics community to standardize and evaluate spectra within the 350-2500 nm range of Brazilian soils. The Brazilian Soil Spectral Library (BSSL) began in 1995, creating a protocol to gather soil samples from different locations in Brazil. The BSSL reached 39,284 soil samples from 65 contributors representing 41 institutions from all 26 states. Through the BSSL spectra database, it was possible to estimate important soil attributes, such as clay, sand, soil organic carbon, cation exchange capacity, pH and base saturation, resulting in differences among the multi-scale models taking Brazil (overall), regional and state scale. Statistical analyses showed that six basic patterns of spectral signatures represent the Brazilian soils types and that environmental conditions explain the differences in spectra. This study demonstrates that spectroscopy analyses along with the establishment of soil spectral libraries are a powerful technique for providing information on a national and regional levels. We also developed an interactive online platform showing soil sample locations and their contributors: ´a put together community system´. As soil spectroscopy is considered a fast, simple, accurate and nondestructive analytical procedure, its application may be integrated with wet analysis as an alternative to support the sustainable management of soils.

Key words: spectral sensing; proximal sensing; Vis-NIR-SWIR spectroscopy; pedometrics.

2.1. INTRODUCTION

Soil is a fundamental natural resource for sustaining life in the planet and economic development, since it provides several ecosystem services and is the basic resource for many human activities (Adhikari and Hartemink, 2016; Jónsson and Davíðsdóttir, 2016). Thus, the knowledge about soil physical and chemical properties and their spatial variabilities are the essence for their sustainable use, planning and adequate management, aiming for greater productivity and conservation (Wall and Nielsen, 2012).

Earth has about 150 million km² of land and most of it is not totally known in terms of soil surface composition, which is usually made by traditional wet analysis. Since this technique has been used for more than 100 years, it is considered the most relevant to characterize soil properties. However, this approach suffer from the usage of chemical reagents and time consuming (Viscarra Rossel et al., 2016). Besides, there are still uncertainties and discussion of current methods and its results, which frequently lead to difficulties in the interpretation and misleading communication. These issues took research to seek for other strategies on to optimize and/or assist these previous and important wet methods.

Proximal sensing research community has applied spectroscopy techniques systematically on the last 40 years to reach soil properties with important results (Nocita et al., 2014). The spectral range commonly used to study spectral pattern of soils corresponds to 400 - 700 nm (visible - Vis), 700 - 1,100 nm (near infrared - NIR) and 1,100 - 2,500 nm (shortwave infrared - SWIR) which can be obtained by sensors in the field or laboratory and has been the baseline for optical aerial/satellite remote sensing (reflectance spectroscopy, imaging spectroscopy). In this case, from the surface reflectance of the samples measured in laboratory, it is possible to develop models relating the spectral pattern to some soil characteristics which can be extrapolate to satellite spectral data, making possible to map large areas (Demattê, 2016). Since Bowers and Hanks (1965), soil reflectance has been studied and reached a strong background on its interpretation. During this period, sensing data has showed a strong relationship with several soil attributes, i.e., soil organic carbon (Stevens et al., 2008), texture (Brodský at al., 2010), mineral composition (Viscarra Rossel et al., 2006), and others (Nocita et al., 2014). The technique allows the simultaneous characterization of soil attributes with the advantage of being a non-destructive method of *in situ* observation (Viscarra Rossel et al., 2006).

To make spectral information useful for the soil science community, it is imperative to have reference patterns in a database (Viscarra Rossel and Behrens, 2010), commonly named spectral libraries. A diverse database is fundamental to understand soils spectral behavior and reach its attributes prediction from spectra (Shepherd and Walsh, 2002). After this study, others came along such as Brown et al. (2006) and Viscarra Rossel and McBratney (2008). The ICRAF-ISRIC world soil spectral library (Garrity and Bindraban, 2004), for example, is composed of 785 soil profiles from 58 countries from Africa, Europe, Asia, and the Americas. Viscarra Rossel and Webster (2012) described a large spectral library with ~4,000 soil profiles covering the Australian continent. A spectral library covering the United States (US) has been setting on the Rapid Carbon Assessment (RaCA) project (Soil Survey Staff, 2014) with 144,833 Vis-NIR spectral curves from 32,084 soil profiles. The European spectral library called LUCAS consists of about 20,000 topsoil samples, collected from 23 countries in the European continent, and measured for 13 soil properties in a single laboratory (Stevens et al., 2013). Another important example of soil spectral library (SSL) was the ASTER spectral library (Baldridge et al., 2009), a compilation of 2,400 spectra of soils, rocks, minerals and other related materials. SSL initiatives in other countries include: Brazil (Bellinaso et al., 2010), Czech Republic (Brodský et al., 2011), France (Gogé et al., 2012), Denmark (Knadel et al., 2012), Mozambique (Cambule et al., 2012),and China (Ji et al., 2016; Shi et al., 2014). Finally, a world soil spectral library was constructed for soil organic carbon, soil texture, iron, $CaCO_3$, CEC, and pH with 90 participating countries (Viscarra Rossel et al., 2016). Such collaborative initiatives open many doors for its applicability.

Brazil is the largest country in South America, with an area of ~ 851 million ha, and is the fifth largest in the world. It has ~ 152.5 million ha in agricultural land (18% of total). Soil mapping and pedologic properties characterization with conventional survey and laboratory methods is enormous challenge. An example of this effort is the PronaSolos (Polidoro et al., 2016), a forthcoming national program that aims to provide more detailed mapping of soils in Brazil. Thus, soil sensing and the fusion of spectral data are promising to allow quick acquisition of information for surveying large areas of soils (Grunwald et al., 2015). The State of São Paulo had its first soil spectral Atlas performed by Epiphanio et al. (1992), which was published afterwards by Formaggio et al. (1996). Bellinaso et al. (2010) and Terra et al. (2015) created soil spectral libraries from states of the South Central of Brazil. However, the country still does not have a standardized SSL to integrate the soil research community and support different applications for studying soil resources.

The objective of this study was to present the first integrated SSL and its relationship with soil attributes and other environmental characteristics covering most of the Brazilian soils. The Brazilian Soil Spectral Library (BSSL) allows the exploration of new approaches on proximal and remote spectral sensing. We hypothesize that spectral data relate to geographical and environmental variables.

2.2. MATERIAL AND METHODS

2.2.1. The collaborating system

The BSSL started in 1995 with a collection of soil samples from the Department of Soil Science, Luiz de Queiroz College of Agriculture, University of São Paulo (ESALQ-USP), where spectral reflectance was measured and inserted into the database. The collaboration system of the BSSL is shown in Fig. 1a and the flowchart with data description is in Fig. 1b. A dynamic and interactive online platform showing the Brazilian map with all BSSL data was also created. This online platform facilitates the communication between any user who wants to contact the researchers and use their soil spectra dataset. The interactive map can be accessed at <https://bibliotecaespectral.wixsite.com/esalq>.

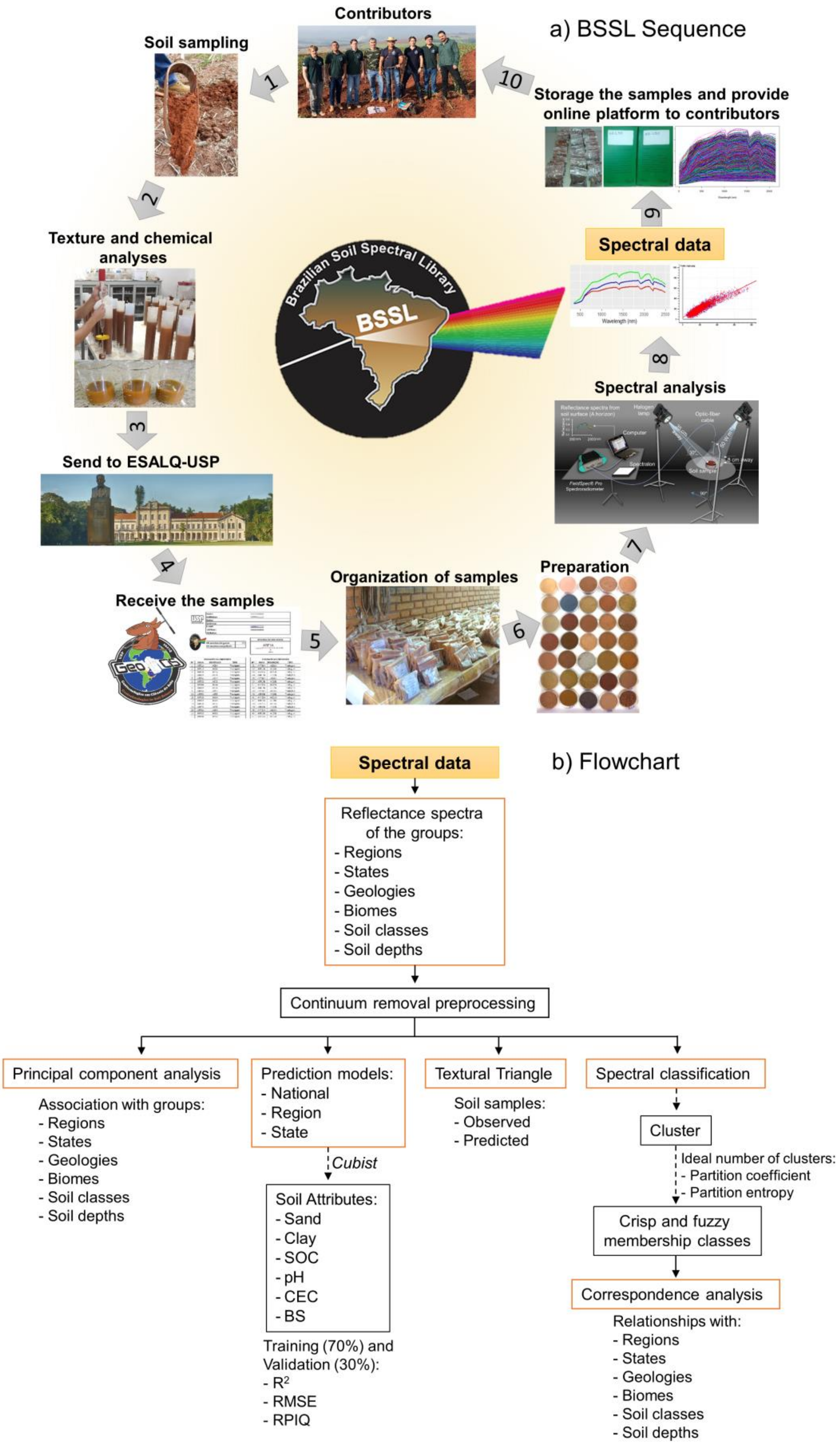


Fig. 1. Methodological sequence for the development of the Brazilian Soil Spectral Library (BSSL) development (a) and flowchart representing the statistical analyses of BSSL(b).

2.2.2. Description of the spectral database

The current spectral library contains 39,284 soil samples from 65 contributors representing 41 institutions. The Brazilian spectral database was constructed combining all the soil samples from the collaborators. Figure 2 shows the maps of Brazil with the region, state, geology, biome, soil class, and sample points. The Brazilian regions are North (N), Northeast (NE), Midwest (MW), Southeast (SE), and South (S) (Fig. 2a). The Brazilian states are Acre (AC), Alagoas (AL), Amapá (AP), Amazonas (AM), Bahia (BA), Ceará (CE), Distrito Federal (DF), Espírito Santo (ES), Goiás (GO), Maranhão (MA), Mato Grosso (MT), Mato Grosso do Sul (MS), Minas Gerais (MG), Pará (PA), Paraíba (PB), Paraná (PR), Pernambuco (PE), Piauí (PI), Rio de Janeiro (RJ), Rio Grande do Norte (RN), Rio Grande do Sul (RS), Rondônia (RO), Roraima (RR), Santa Catarina (SC), São Paulo (SP), Sergipe (SE), and Tocantins (TO) (Fig. 2b). The geology is represented by igneous, metamorphic, and sedimentary rocks (Fig. 2c). The biomes are Amazon, Caatinga, Cerrado, Atlantic Forest, Pampa, and Pantanal (Fig. 2d). Only the most representative soil classes are presented in the Brazilian map, which are Lixisols, Ferralsols, and Arenosols (Fig. 2e). The geographic locations of the soil samples are shown in Fig. 2f. When the information provided by the contributors had no geographical coordinates, the points were allocated at the nearest city.

Most of the samples that compose the database came from the SE and MW regions, with 19,429 and 9,391 samples, 50% and 24% of the samples, respectively (Fig. 3a). São Paulo (SP), Mato Grosso do Sul (MS) and Goiás (GO) states (26,474 samples total) correspond to 68% of all samples (Fig. 3b). Samples are from soils formed mainly in three lithologic groups, igneous and sedimentary with 10,621 and 10,409 samples, respectively 21,030 in total) (Fig. 3c). The most represented biomes are the Atlantic Forest and Cerrado, with 19,248 (53%) and 12,468 (34%), respectively (Fig. 3d). The soil class with most samples is the Ferralsols (22,674 samples, equivalent to 63% of all samples) located mainly in the SE and MW regions (Fig. 3e). Other soil classes represented in the BSSL are Arenosols. Ferralsols and Lixisols represent the two most important soil classes in Brazil, covering about 31.5 and 26.8% of the Brazilian territory, respectively (Santos et al., 2011). These two classes represent 86% of all samples in the database (31,551 samples) (Fig. 3e). Considering the total database, 79% of the samples present A (0-20 cm), B (40-60 cm), C (80-100 cm), and D (100-120 cm) layers (Fig. 3f). Approximately 43% of all samples have soil organic carbon (SOC), 85% have granulometry, 35% have cation exchange capacity (CEC), 67% have values of pH in water and 72% have base saturation (BS) (BS = [Ca + Mg + K + Na]/CEC x 100) (Donagemma et al., 2011) measurements.

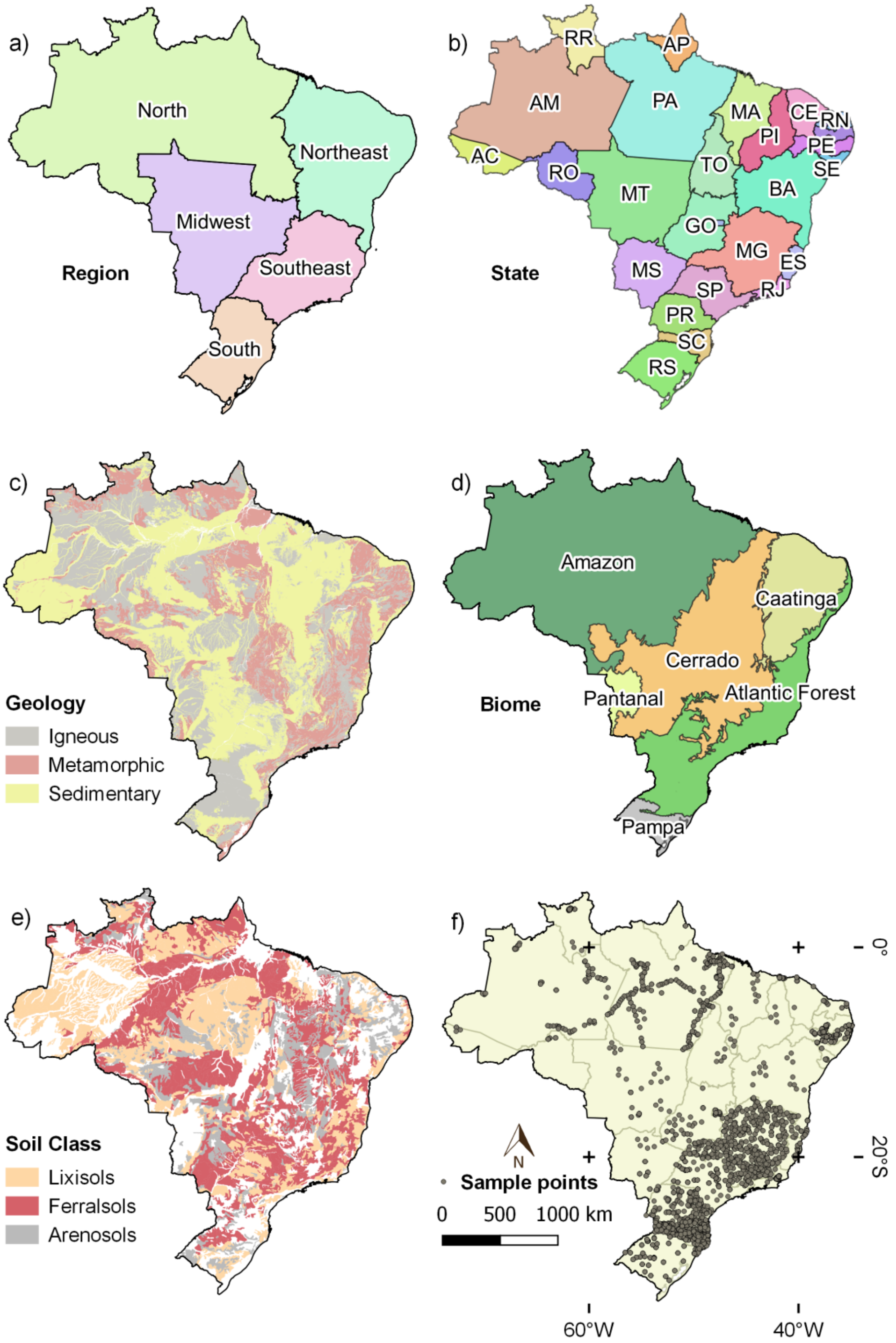


Fig. 2. Maps representing the Brazilian regions (a), states (b), geology (c), biomes (d), main soil classes (e), and sampling locations of the Brazilian Soil Spectral Library (f).

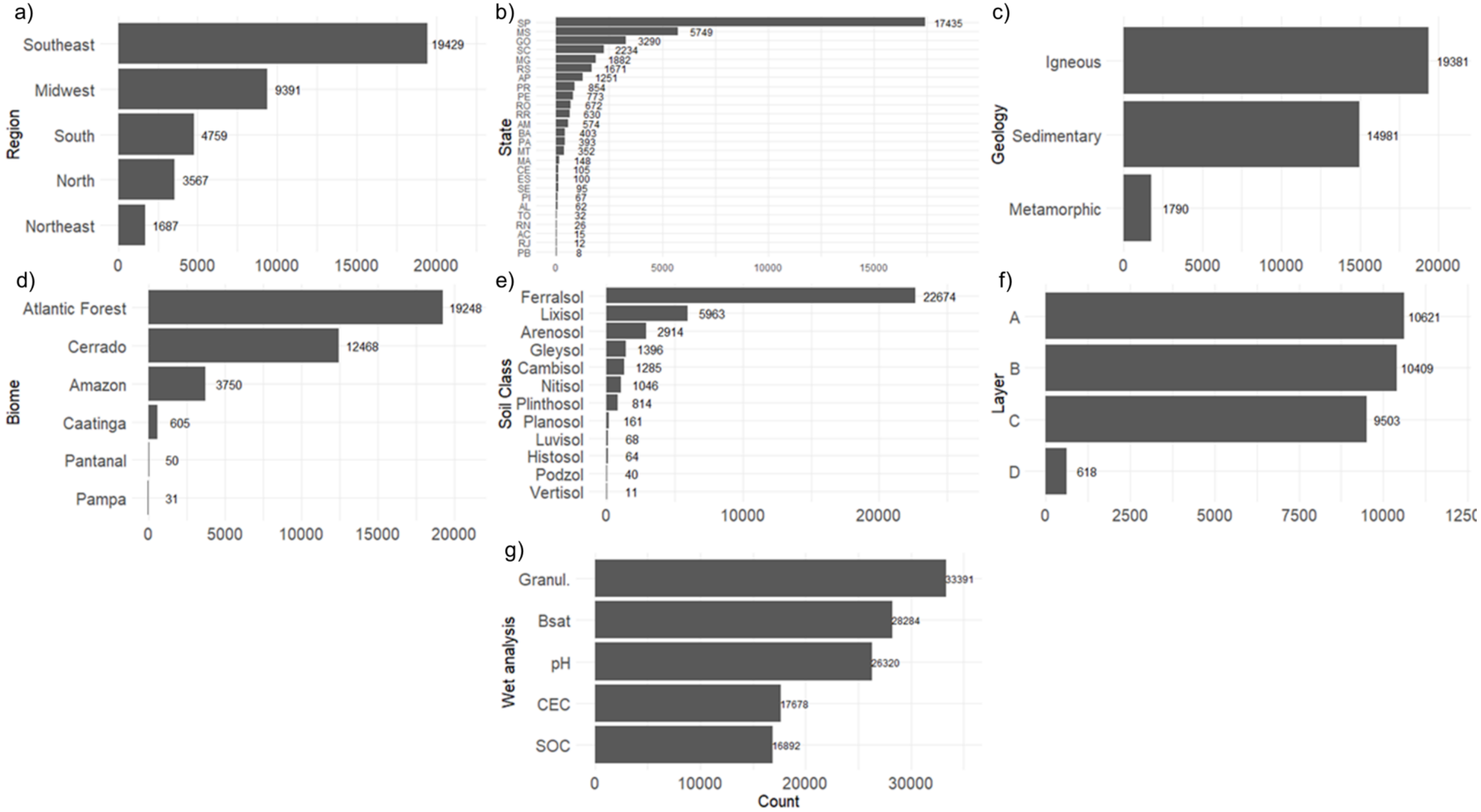


Fig. 3. Distribution of soil samples according to Brazilian regions (a); states (b); geology (c); biomes (d); soil classes (considering layers A and B) (e); soil layers (f); and soil attributes (g). The number of samples vary for each group depending on the available information. Soil classes were defined according to World Reference Base (International Union of Soil Science Working Group WRB, 2015). Soil layers corresponded to A (0-20 cm), B (40-60 cm), C (80-100 cm), and D (100-120 cm).

2.2.3. Spectral data, preprocessing and transformations

All soil samples from the database were previously dried at 45 °C, ground and sieved with 2 mm mesh and then homogeneously distributed in Petri dishes prior the measurement of the spectra. The spectral data were acquired by the Geotechnologies in Soil Science group (GeoCis), São Paulo, Brazil, using the Fieldspec 3 spectroradiometer (Analytical Spectral Devices, ASD, Boulder, CO), which has a spectral range from visible to shortwave infrared (350 – 2,500 nm) and spectral resolution of 1 nm from 350 to 700 nm, 3 nm from 700 to 1,400 nm, and 10 nm from 1,400 to 2,500 nm. The sampling interval of data output is 1 nm reporting 2,151 channels. One of the strengths of the database is that all spectral analyses followed the standardized spectral library analysis protocol.

The spectral sensor, which was used to capture light through a fiber-optic cable, was allocated at 8 cm from the sample surface. The sensor scanned an area of approximately 2 cm^2, and a light source was provided by two external 50-W halogen lamps. These lamps were positioned at a distance of 35 cm from the sample (non-collimated rays and a zenithal angle of 30°) with an angle of 90° between them. A *Spectralon* standard white plate was scanned every 20 minutes during scanning. Two replicates (one involving a 180° turn of the petri dish) were obtained for each sample. Each spectrum was averaged from 100 readings over 10 s. The mean values of two replicates were used for each sample. Ninety-eight percent (98%) of soil spectra were measured in GeoCis Lab, following the protocol proposed by Ben-Dor et al. (2015). Although the other 2% were not measured by the same equipment, protocols for spectra acquirement were strictly followed. Considering that practically all the spectral library was built with the same protocol, performing a calibration transfer function would demand time and resources, while the improvements would most likely be fairly small.

The spectral reflectance was transformed to continuum removal (CR) (Clark and Roush, 1984). This preprocessing removes the continuous features of spectra and is often used to isolate specific absorption features. The CR creates a continuum or hull similar to fitting a rubber band over the original spectrum. The spectrum is normalized by setting the value of the hull to 100% reflection, where the first and last values of the continuum-removed spectrum equal 1. We applied CR preprocessing because of its strength and ability to enhance absorption depths by correcting apparent shifts from wavelength-dependent scattering, highlighting specific absorption bands of a reflectance spectrum(Mutanga et al., 2005).Besides that, the CR preprocessing is capable of providing calibration models with high accuracy.

We used heuristically testing to optimize the clustering procedure, which involved grouping the samples by their reflectance spectra. The reflectance intensity provides important information in the spectral characterization of soils, but in our case, it did not provide good results. Initially, we had clustered the samples based on the reflectance spectra, but the fuzzy performance indicators were not satisfactory. On the other hand, when employing the CR spectra, we not only produced reasonable performance indices but also had results similar to other studies (e.g. Demattê et al., 2016; Terra, Demattê, & Viscarra Rossel, 2018). Although the reflectance intensity corresponds to a large share of spectral variance, there are other information in the spectrum which are extremely important to soils discrimination (e.g. features related to clay minerals at around 1,200, 1,900 and 2,200 nm). The potential of such information should not be underestimated.

2.2.4. Principal component analysis

The CR spectra were analyzed by principal component analysis (PCA) to reduce dimensionality and improve computational efficiency. The data was not standardized to make easier the interpretation of absorption features in continuum spectra. We used both the scores and eigenvectors of PCA to assist in the interpretation of BSSL data.

Geographical and environmental characteristics were associated with the spectral data and the Brazilian spectral samples were separated according to 5 regions, 26 states, 3 geologies, 6 biomes, 11 soil classes and 4 soil depths (Fig. 3). The PCA was used to investigate the associations between groups and spectral data. The PCA correlates the average soil spectral reflectances with regions, states, geology, biomes, soil classes, and layers. Soil classes in the Brazilian Soil Classification System were correlated with the WRB classification (IUSS Working Group WRB, 2015). The BSSL presents a large variation of samples considering layers, surface and subsurface horizons, and complete soil profiles. However, only soil samples that had the following depths were selected for PCA with layers' data: A (0-20 cm), B (40-60 cm), C (80-100 cm), and D (100-120 cm). Considering all samples from the spectral database, 84% were taken collected with auger, 12% from complete profiles and 4% only from the surface layer (0-20 cm).

2.2.5. Spectroscopic modeling of soil attributes

The soil attributes selected for predictive modeling were sand, clay, SOC, pH, CEC, and BS. Several strategies of modeling were performed to predict these attributes. First, national, regional and state models were developed for each attribute, where the national model included the complete database. The datasets for each soil attribute were separated into training and independent validation by a 70:30 split. This separation was carried out using random division, which was able to separate the groups homogeneously. The cubist method (Quinlan, 1992)was applied to train the spectroscopic models. Cubist applies the M5 (Model Tree approach) to grow categorical decision trees to handle continuous classes by placing a multivariate linear model at each leaf. The model building and estimation process were achieved by the *caret* package (Kuhn, 2017) in R (R Core Team, 2018). This package has a set of functions that attempt to streamline the process for creating predictive models. The calibration function was applied to adjust the best fitted model using optimal tuning parameters as follows: cross-validation resampling, committees, and neighbors.

For each soil attribute the performance of the models were assessed by comparing the predicted and observed values based on the independent validation data set. The coefficient of determination (R^2) (Eq. 1), root mean squared error (RMSE) (Eq. 2), and ratio of performance to interquartile distance (RPIQ) (Eq. 3) were assessed to quantify the inaccuracy of the estimates.

$$\mathrm{R}^2 = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\hat{\mathrm{y}}_\mathrm{i} - \bar{\mathrm{y}}_\mathrm{i})^2}{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{y}_\mathrm{i} - \bar{\mathrm{y}}_\mathrm{i})^2} \qquad (1)$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\hat{\mathrm{y}}_\mathrm{i} - \mathrm{y}_\mathrm{i})^2} \qquad (2)$$

$$\mathrm{RPIQ} = \frac{(\mathrm{Q3} - \mathrm{Q1})}{\mathrm{RMSE}} \qquad (3)$$

where ŷ is the predicted value, ȳ is the mean of the observed value, y is the observed values, n is the number of samples with *i* equal to 1, 2, … n, IQ is the difference between the third and first quartiles (Q3 - Q1).

The textural triangle was developed using the reference values for clay, sand and silt, called observed, and using the predicted values for clay and sand obtained by each model, i.e., the predicted values used in the triangle of BSSL were originated from the national model, and the predicted values of S region were originated from the S model. The silt content was calculated by the difference of clay plus sand content. The textural triangle was carried out in R using the *soiltexture* package (Moeys, 2016).

2.2.6. Spectral patterns classification

The classification of spectral patterns was performed aiming to represent Brazilian soils. The reflectance was transformed to CR, which was used to classify spectra into general groups.

The question was: How many classes of spectra were necessary to represent the Brazilian soils? To answer that, we classified the spectra by clustering via similarity measurements. The first three principal component scores were classified by the fuzzy *c*-means algorithm (Bezdek et al., 1984). This approach produces two methods of classification: crisp and fuzzy membership degrees. The first produces the crisp or hard (no-fuzzy) membership degrees of the objects in order to place them into only one discrete cluster. The fuzzy *c*-means technique assigns a fuzzy membership degree to each data point based on its distances to the cluster centers. The fuzzy approach is based on the distance between various input data points (PCA scores). This algorithm assigns a fuzzy membership degree to each data point based on its distances to the cluster centers. The farther from the center of the cluster the smaller the probability of the point being classified in the respective class. The fuzzy membership degrees are continuous and range from 0 to 1. Each sample has a membership in every cluster, where close to 1 indicates a high degree of similarity between the sample and a cluster, while close to 0 implies a low similarity (Bezdek et al., 1984).

Fuzzy clustering requires the user to predefine the number of clusters (*c*), but it is not always possible to know this number in advance. To obtain the ideal number of c-means cluster two validation functions were performed. The two most important validity index functions to determine the optimal number of clusters are as follow: (a) partition coefficient (pC) and(b) partition entropy (pE) (Bezdek et al., 1984). The best performance is achieved when the pC achieves its maximum value or pE obtains its minimum. All analyses and statistical procedures described above were performed by the R programming (R Core Team, 2018). The crisp and fuzzy *c*-means clustering was carried out using the *ppclust* package (Cebeci et al., 2018).

The SSL was clustered by fuzzy c-means and crisp fuzzy techniques to evaluate the potential of spectral data in the discrimination of soil types. By testing the SSL with two different clustering methods, we intended to evaluate if the soil spectral groups were robust and capable of being replicated by different clustering methods. The inter-comparison between c-means and crisp methods was done using Sankey diagram (Schmidt, 2008), which assesses the relative associations between memberships. The diagram represents the relationship between groups by linking the clusters from the c-means technique with the ones from crisp fuzzy analysis. Higher associations between two groups are represented by proportionally larger links, while poor relation between a pair of clusters is represented by thin links.

2.2.7. Correspondence analysis

The associations between geological and environmental characteristics with soil spectral classes defined by cluster analysis were identified through correspondence analysis (CA). This technique is designed specifically for the analysis of categorical variables, and its primary goal is to illustrate the most important relationships among the variables' response categories using a graphical representation (Benzécri, 1992). The concept is similar to PCA but applies to categorical rather than continuous data. It summarizes the associations between the spectral soil classes and the variables (regions, states, geology, biomes, soil classes, and layers) in two-dimensional graphical forms. The CA plots are derived from a table where the rows are the characteristics (e.g. states of Brazil, biome, etc.) and the columns are the six spectral classes. The CA was applied using the *FactoMineR* package (Lê et al., 2008).

2.3. RESULTS AND DISCUSSION

2.3.1. Soil reflectance spectra vs physiographic and soil characteristics

The higher sand content in NE region caused an increase in reflectance compared to other regions (Fig. 4a). On the average, the sand content of NE region is 651 g kg^{-1}, while the average of S region, for example, is 264 g kg^{-1}. Soils from the S region of Brazil, where predominantly formed under the influence of basalt or related to igneous rocks (Fig. 2c), which presented low reflectance values due to iron oxides and opaque minerals (Fig. 4a). This region also has specific temperate climate, which favors the preservation of SOC, and this agree with its lower spectra. In general, in relation to the geology, spectral signatures from igneous rocks are usually rich in calcium and iron and had low reflectance (Fig. 4c), while soils formed in metamorphic parent material showed high reflectance values. In these soils, reflectance features were mainly linked to orthoclase, quartz and plagioclase minerals. The low reflectance values found in soils formed in igneous rocks, such as basalt and diabase with high amounts of iron, were correlated with high clay contents and consequently higher influence of scattering. The soils from the Cerrado biome revealed the lowest spectral reflectances (Fig. 4d). The Atlantic Forest also presented a low reflectance curve. The higher reflectance was found in Caatinga biome (Fig. 4d), which can be explained by the predominance of sandy particles in soils, besides the high temperatures that have accelerate the decomposition of soil organic matter resulting on relatively low SOC. The representation of the spectral curves of each biome is a generalization, because within each of them there is a great complexity of soil types. For instance, the Atlantic Forest biome extends from southern to northern Brazil with different soil types. Among the soil classes, the spectral curve of the Nitisols showed the lowest reflectances (Fig. 4e). This soil originated mainly from mafic rocks such as basalt and diabase, present high amounts of clay, iron oxide and opaque minerals (IUSS Working Group WRB, 2015). The spectral curve of the Histosols presented low reflectance in the visible spectral region due to the high content of SOC. Contrarily, the Podzols had the highest reflectances, followed by the Cambisols rich in quartz (Fig. 4e). The Arenosols, for example, showed greater reflectances in the A layer due to higher sand content in relation to subsurface (Fig. 4f). The average reflectance curves of the four layers (soil depths) (Fig. 4f) indicate the differences in SOC content: from 550 to 900 nm the reflectance increases while SOC decreases. The spectral range from 1,500 to 2,500 nm was influenced by quartz, due to the high reflectance values in all four soil layers. The B and C layers were identical, which is attributed to low variation in mineralogy and texture.

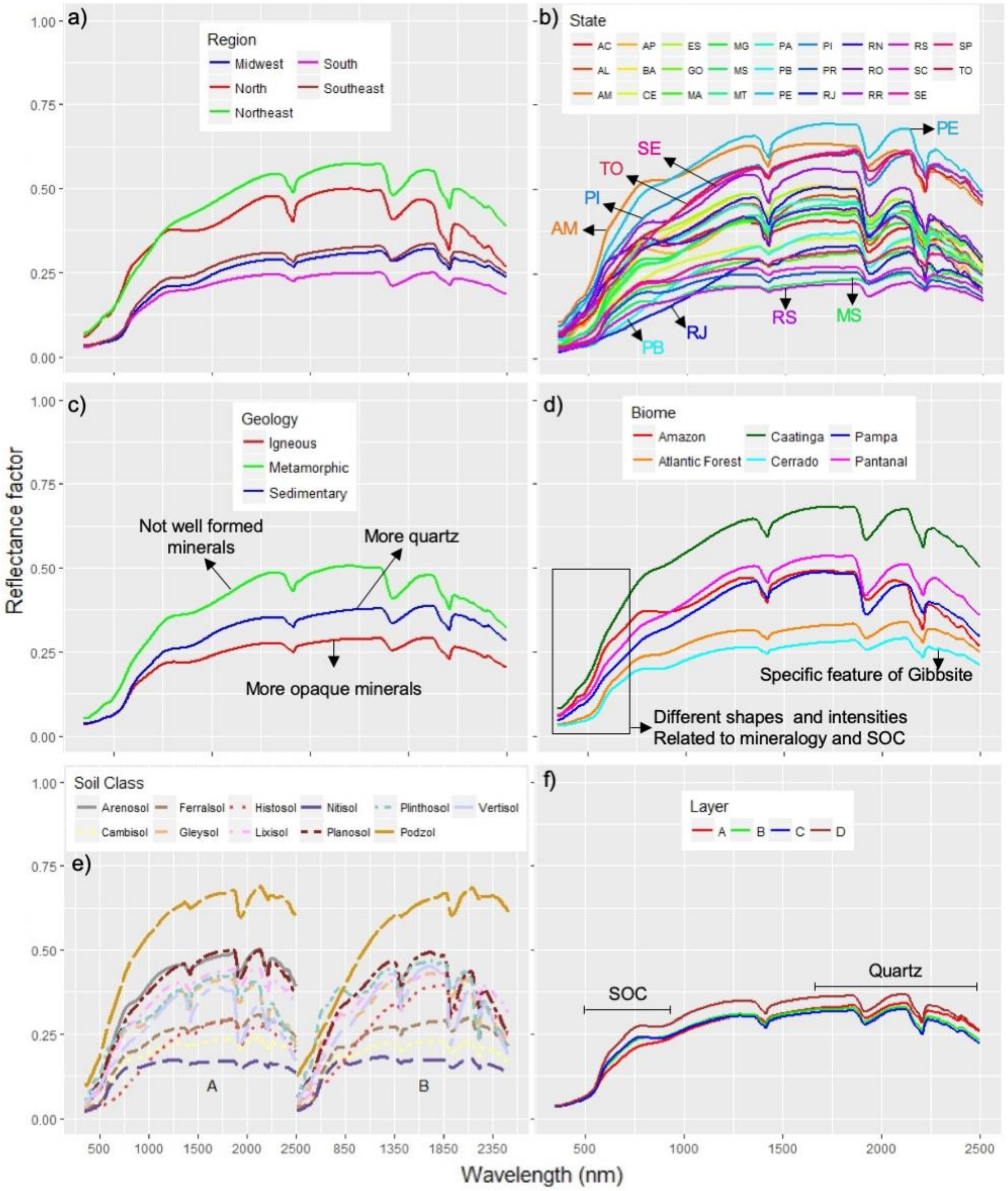


Fig. 4. Soil reflectance spectra averaged according to each region (a); state (b); geology (c); biome (d); soil class (considering layers A and B) (e); and layer (f). The number of samples considered in each group are shown in Fig. 3.

The PCA revealed that the first principal component (PC1) accounts explained 85% of the variance in the data (Fig. 5). The PCA of spectra averaged by region (Fig. 5a) was able to detect that the N, NE and S regions presented different soil spectral patterns. However, soil spectra from the MW and SE were grouped together, showing nearly the same spectral pattern in these regions. Among soils from 26 Brazilian states some showed similar factor loadings (Fig. 5b). For example, the average soil spectra from MT and GO states as well as for AP and

AC; MS and RS; PR, and SC; BA, RN, and ES; MA, and SP; AL, and PA; RJ, and PB grouped together. Soil spectra from AM, PE, PI, TO, SE, and RO states showed different patterns and were not grouped. For the three geological classes, the PCA discriminated them by showing separated data distributions (Fig. 5c). The PCA result for the biomes indicated that Caatinga and Amazon present distinct spectral curves (Fig. 5d). Indeed, these are two important and very distinct environments (Amazon = tropical humid soils; Caatinga = semiarid soils). This finding is corroborated by the result in Fig. 4d, where the Caatinga showed a spectral curve with high reflectances and the Amazon presented higher intensity reflectance in the visible region. This interpretation agrees with spectra by regions (Fig. 5a). However, soils from Pantanal and Pampa presented small differences in spectral pattern similar to the Atlantic Forest and Cerrado (Fig. 5d). The principal component scores discriminated well among Podzols, Plinthosols, Histosols, and Lixisols, using only B layer (40 - 60 cm depth) (Fig. 5e). Conversely, Nitisols, Cambisols, and Ferralsols were grouped relatively close together suggesting similarities spectral pattern. The same arrangement was found among Planosols, Gleysols, and Vertisols, which are soils formed under the influence of hydromorphic conditions with more prolonged water saturation typically exhibiting the Fe^{3+}reduction and high SOC. In contrast Podzols, Plinthosols, Histosols and Lixisols show large differences in the content of SOC, iron oxides and texture. For instance, Podzols have SOC mainly associated with a sandy texture, dominantly quartz, and associated with complexes of Al and Fe. Plinthosols are characterized by high iron oxides and low crystallinity degree in the form of nodules and low content of SOC. Histosols have high content of SOC, and low Fe content. Lixisols have kaolinite dominance, variable texture and low SOC content. Another group is represented by soils without texture gradient such as Nitisols, Cambisols, and Ferralsols. On the other hand, Lixisols with textural gradient or Histosols, with very high SOC were discriminated by principal component scores. The PCA for layers detected that A (surface) and D (subsurface) layers were separated suggesting distinct spectral pattern in these two layers (Fig. 5f). The B and C layer were placed in proximity in the PCA graphs indicating that they were similar in relation to the soil spectra (Fig. 5f).

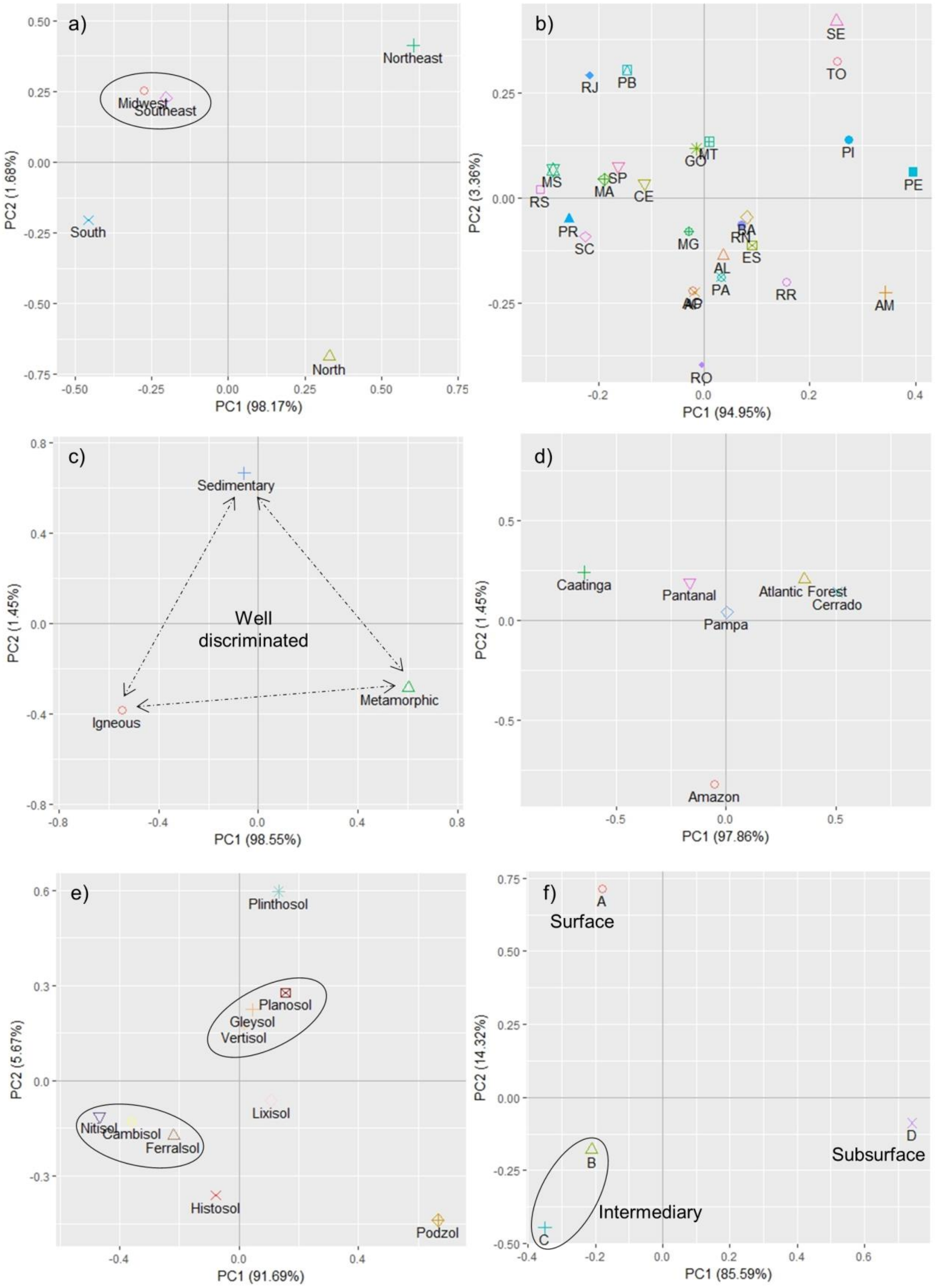


Fig. 5. Principal component scores 1 (PC1) and 2 (PC2) calculated from the average reflectance spectra of each region (a), state (b), geology (c), biome (d), soil class (e), and layer (f).

2.3.2. National, regional, and state prediction of soil attributes

The national model produced R^2 of 0.78 and RMSE of 6.89 g kg^{-1} for SOC prediction in validation model (Table 1). For SOC prediction in validation model, the regional models showed R^2 ranging from 0.58 to 0.84 and RMSE from 2.28 to 9.97 g kg^{-1}. The MW region presented the best results, while the N region the worse. The state that presented the best results was MT with R^2 of 0.89 and RMSE of 1.65 g kg^{-1}. From the 18 states with SOC prediction models, 8 showed a R^2 above 0.80 and only two showed R^2 below 0.32 (AP and RR). The worst R^2 were associated either with a small number of samples or a high variability in SOC.

Among all the soil variables predicted, clay showed the highest coefficient of determination in validation mode at national level with 0.88 and RMSE of 75.93 g kg^{-1} (Table 2). At the regional level, all clay validation models had R^2 higher than 0.71 and the SE and MW regions had R^2 higher than 0.91 and RMSE lesser than 60 g kg^{-1}. At the state level, the best result was found for GO with R^2 of 0.96, RMSE of 52.15 g kg^{-1} and RPIQ of 8.99. From the 21 states with clay prediction, 13 showed R^2 higher than 0.80 and only 3 had R^2 below 0.50.

For sand predictions in validation mode at the national level showed R^2 of 0.87 and RMSE of 103.03 g kg^{-1} (Table 3). At regional scale, the MW region showed the best performing validation sand model with R^2 of 0.94 and the S region the worse, but still with moderate well predictions (R^2 = 0.77 and RMSE = 114.95 g kg^{-1}). At the state level, 20 sand models were generated with GO showing the best results (R^2 = 0.97 and RMSE = 53.94 g kg^{-1}). In 14 states the R^2 were higher than 0.80 and only PE had R^2 below 0.50 for predictive modeling of sand.

At the national level, the validation of the model generated for pH prediction showed a R^2 of 0.54 and RMSE of 0.39 (Table 4). In general, the national pH model was better than the regional ones. Only the NE region presented a higher R^2 (0.65) in validation mode, while the S region showed the smallest R^2 (0.34). The good result for the NE region is related to the well-performing pH models generated for the CE and PE states that belong to this region and had the highest R^2, both with 0.97. From the 16 analyzed states, 9 pH prediction models had poor results with R^2 below 0.50.

The prediction validation of CEC at the national level reached a R^2 of 0.68 and RMSE of 24.02 $cmol_c$ kg^{-1} (Table 5). At the regional level, NE showed the best validation results (R^2 = 0.89 and RMSE = 27.68 $cmol_c$ kg^{-1}) and the S region the worst (R^2 = 0.64 and RMSE = 3.81 $cmol_c$ kg^{-1}). At the state level, MT, RN and SE showed a R^2 above 0.93, and three states showed a R^2 below 0.30 (AL, PA and PE).

Of all soil attributes analyzed in this article, BS presented the poorest result at the national level with R^2 = 0.49 and RMSE = 17.01% in validation mode (Table 6). However, at the regional level only the SE region showed a R^2 below 0.50 (R^2 = 0.49 and RMSE = 16.28%) and the NE showed the best performing BS model (R^2 of 0.79 and RMSE = 13.42%). At the state level, 16 models were generated, of which 14 showed R^2 above 0.65 and AM and GO states showed higher R^2 (0.70 and 0.69, respectively).

Higher errors in the calibration models than in validation are definitely odd, especially when the modeling process uses a machine learning algorithm. In our case, this is related to a limited number of samples for modeling at the state level. The SOC model predictions at the state of AL (Table 1), for example, there were only 32 samples available for calibration and 19 for validation, which resulted in a R^2 of 0.29 (calibration) and 0.43 (validation). Conversely, prediction models for SP state had a total of 8,185 samples, 5,729 for calibration and 2,456 for validation. In this case, the R^2 for calibration and validation were practically the same (0.66). Representative and comprehensive datasets are essential for a robust calibration, because otherwise the errors may be high and results incoherent. The BSSL has been constantly populated with new samples, therefore we believe that soon it will be possible to calibrate good prediction models for all states and regions.

2.3.3. Synthesis of soil attributes prediction

In summary, we computed the best model performances (i.e., highest model fits and lowest errors) by down scaling results (i.e., from national to state levels). For example, for SOC validation models the R^2 were 0.78 - 0.84 - 0.93 (Table 1), for clay 0.88 - 0.94 - 0.96 (Table 2), and for sand 0.87 - 0.94 - 0.95 for national and best performing regional and state models, respectively (Table 3). However, the performance of state-specific soil models differed widely due to multiple factors including sample size, soil variance, and soil-forming factor differences. The model performance for the same soil attribute in different states differed widely (e.g., R^2 of clay varied from 0.42 to 0.96), hindered detailed discussion on several factors which still need further studies. The variability in statistical metrics of soil attributes assessments are not new. Nocita et al. (2014) found several discrepancies in regard to R^2 for the same soil attribute, i.e., SOC, pH, and others. Zeng et al. (2016) indicated that local predictions can be better modeled by understanding the soil development, i.e., parent material, biome and land use. Shepherd and Walsh (2002), obtained R^2 for CEC from 0.6 (national model) to 0.8 (local models). Grunwald et al. (2018) found that upscaled SOC spectral models performed better in terms of R^2 and RPIQ, whereas the downscaled models showed less bias and smaller RMSE in Florida, USA. This study found no universal trend that could explain the scalability of the models, such as spectral variance, soil attribute variance, methods, and environmental characteristics or diversity.

Overall, SOC models with high model fit ($R^2 > 0.85$ in regions CE, MG, MT and SC) coincided with relatively high mean SOC of > 12 g kg^{-1}irrespective of SOC variabilities that were very large (e.g., 55.7 g kg^{-1} standard deviation, SD, in CE) or low (e.g., only 5.0 g kg^{-1}SC in MT). Similar trends were discovered for soil texture models. For example, clay models with high model fit ($R^2 > 0.85$ in regions ES, GO, MG, MS, RO, RS, SC and SP) corresponded with high mean clay content of > 210 g kg^{-1}, though almost all of these models covered a wide range in clay contents with SD values ranging from 130 to 253 g kg^{-1}). Trends among pH, CEC, and BS models were less clear among regions in terms of underlying factors to explain model performance. The model performance in this study showed comparable results with other spectral library studies (Viscarra Rossel et al., 2016) This suggest that soil spectral models for key indicators on Brazilian soils could be developed with similar quality as documented in other soil spectral studies.

Samples from certain states showed high pedological variation (factors and processes of soil formation). States with pedological complexity and/or less sample size may have contributed to lower model fits and higher errors than more homogeneous and/or more densely sampled states. Though high sample size may not necessarily mean better model performance as indicated by models in SP, which performed excellent for clay, sand, and CEC, moderately for SOC, and less well for pH and BS.

In order to confirm the effectiveness of the spectroscopic method to determine clay and sand contents, the textural triangle was produced showing simultaneously the observed values derived from traditional laboratory analysis and the predicted ones by the spectroscopic method (Fig. 6). The textural triangle of the entire BSSL (Fig. 6a) showed that most soil samples were placed in the soil texture classes of sand, loamy sand, sandy loam, sandy clay loam, sandy clay, and clay. The SE region showed similar trends in soil texture when compared to the whole BSSL database because the SE region contributed 52% of the total samples with textural data (Fig. 6b). However, the samples of SE showed lower silt content than the entire BSSL collection. The predicted soil textures for the S region (Fig. 6c) N region (Fig. 6e), and NE region (Fig. 6f) showed more scatter than the MW region (Fig. 6d). In the S region, most soils belong to clay textural class followed by silt clay and clay loam mainly due to the parent material predominantly formed by igneous rocks (Fig. 2c). The samples with the highest silt contents belong to the S region. This is related to the low temperature in this region on soil formation. In the MW region, the prediction model for clay and sand was the most accurate and this is reflected in the predicted samples, which presented the same trend as the observed ones (Fig. 6d). The vast majority of samples present a soil textural class varying from sand to clay.

The percentage of silt in these soils is low. In general, the MW region presents more weathered-leached soils such as Ferralsols, derived from sedimentary and igneous rocks forming loamy sand and clayey soils. Soils from the N region showed a large variation in texture (Fig. 6e) corroborating the lower prediction performance of clay (Table 2) and sand content (Table 3) compared to other models. The NE region present a large amount of sandy soils and for this reason the predicted model had good performance (Fig. 6f). In the NE region, the majority of the soils in this study is formed from sedimentary materials, which is showed by the textural classes with high sand content (Fig. 6f).

The textural variations found in the triangles of each region are due to differences in the geology, climate, and relief. Each region presents textural diversity as there is also a great variation of types of soils (Fig. 2e). It is important to emphasize that the predicted samples had the same tendency of the observed ones. This is a great finding considering the world demand for soil analyses with more than 600 million soil samples processed every year which represents a consumption of about 840 thousand kg of dichromate and ammonium ferrous sulfate and 3 million L of sulfuric acid, just for SOC analysis (Demattê et al., 2019). The effectiveness of soil spectroscopic analysis is justified by the fact that it is fast, simple, accurate, cheap, and most importantly non-polluting method. The possibility of predicting several attributes with just one spectral reading, the easy and rapid data acquisition of large amounts of samples without using environmentally hazardous chemicals are the major advantages of the Vis-NIR-SWIR spectroscopy technique for soil analysis (Minasny and McBratney, 2008; Viscarra Rossel and Behrens, 2010).

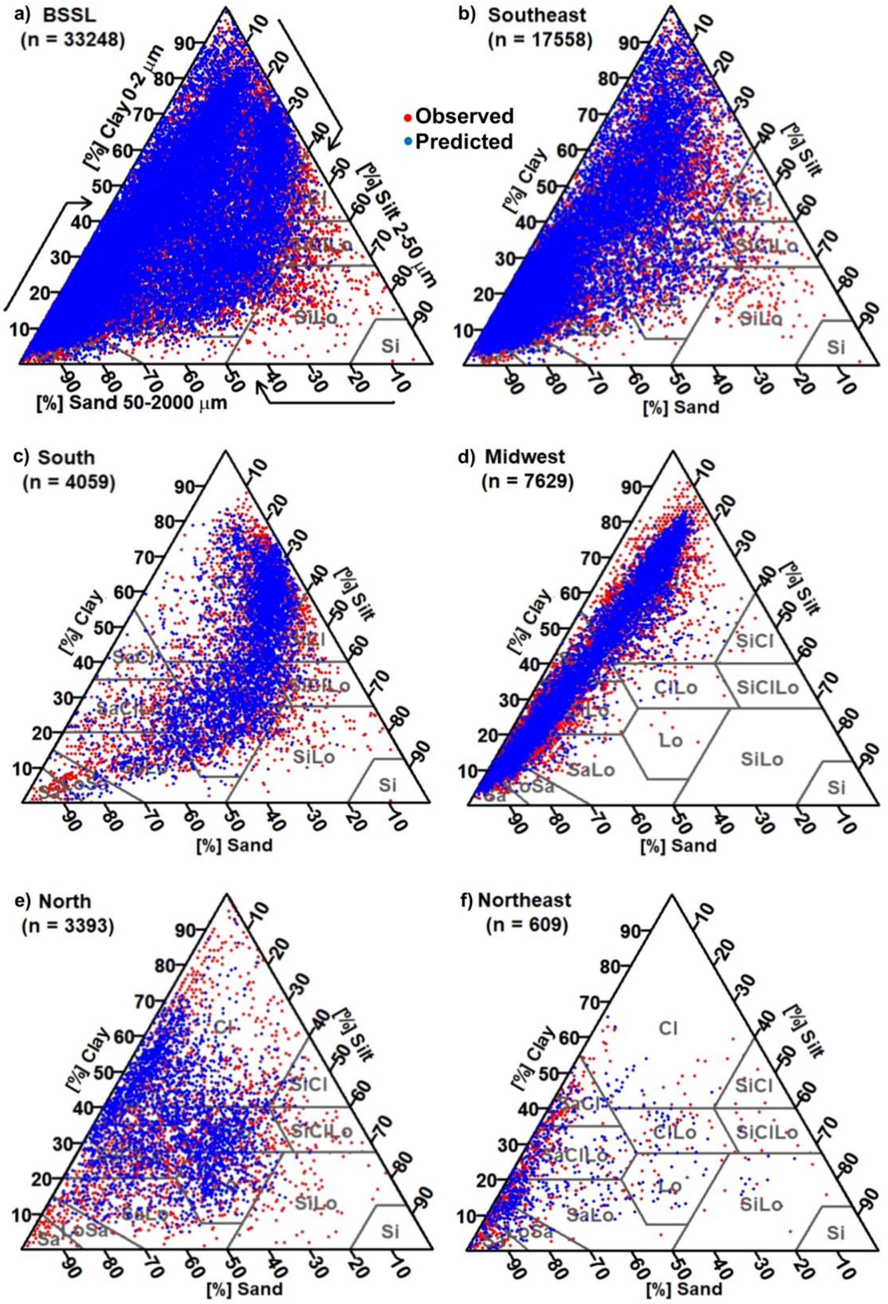


Fig. 6. Soil texture triangle calculated from the entire database (a) and for Brazilian regions, (b) Southeast (SE), (c) South (S), (d) Midwest (MW), (e) North (N), and (f) Northeast (NE) regions. Cl: clay; SiCl: silty clay; SaCl: sandy clay; ClLo: clay loam; SiClLo: silty clay loam; SaClLo: sandy clay loam; Lo: loam; SiLo: silty loam; SaLo: sandy loam; Si: silt; LoSa: loamy sand; and Sa: sand.

Table 1. Cubist model parameters, descriptive statistics, and results of prediction models of Soil Organic Carbon (SOC).

SOC (g kg^{-1})		Descriptive analysis				Observations			Training set			Validation set		
		Mean	SD	Min	Max	Total	Train.	Val.	R^2	RMSE	RPIQ	R^2	RMSE	RPIQ
National		8.9	11.9	0.0	431.1	18076	12653	5423	0.82	5.07	1.38	0.78	6.89	0.94
Regions	South	12.0	15.1	0.0	141.4	1833	1283	550	0.82	6.66	2.69	0.71	8.12	2.05
	Southeast	8.0	5.2	0.0	54.9	9252	6476	2776	0.72	2.82	2.06	0.74	2.75	2.11
	Midwest	8.6	5.3	0.6	57.0	3104	2173	931	0.84	2.10	3.04	0.84	2.28	2.55
	Northeast	13.4	35.2	0.0	431.1	1309	916	393	0.87	14.64	0.69	0.79	9.97	1.15
	North	8.2	6.7	0.0	105.6	2578	1805	773	0.64	4.20	1.66	0.58	4.41	1.80
States	AC	-	-	-	-	-	-	-	-	-	-	-	-	-
	AL	1.5	1.0	0.7	4.1	32	19	13	0.29	0.92	1.49	0.43	0.76	1.22
	AM	8.0	8.1	0.1	105.6	435	304	131	0.78	3.20	1.64	0.72	5.92	1.15
	AP	12.4	6.6	2.0	56.0	817	571	246	0.21	5.70	1.31	0.32	6.09	1.64
	BA	1.8	1.7	0.3	20.2	242	169	73	0.87	0.75	1.64	0.84	0.39	2.31
	CE	40.5	55.7	0.5	310.8	105	73	32	0.98	6.55	5.54	0.93	25.93	0.91
	ES	-	-	-	-	-	-	-	-	-	-	-	-	-
	GO	11.5	6.7	1.7	66.0	618	432	186	0.84	2.70	3.66	0.83	2.73	2.98
	MA	1.5	0.8	0.2	4.0	74	51	23	0.14	0.77	1.25	0.40	0.75	0.96
	MG	11.8	7.6	0.0	59.9	1065	745	320	0.88	2.74	3.81	0.89	2.45	3.85
	MS	7.5	4.3	0.6	32.6	2269	1588	681	0.85	1.73	2.68	0.85	1.72	2.71
	MT	12.8	5.0	4.1	25.6	217	151	66	0.90	1.57	5.38	0.89	1.65	5.11
	PA	6.3	5.3	0.1	38.9	305	213	92	0.61	3.60	1.42	0.69	3.24	1.73
	PB	-	-	-	-	-	-	-	-	-	-	-	-	-
	PE	10.8	8.7	0.0	70.0	773	541	232	0.84	3.45	2.32	0.84	4.02	2.10
	PI	8.5	7.9	0.1	33.3	67	34	33	0.46	5.88	1.13	0.41	6.57	1.70
	PR	-	-	-	-	-	-	-	-	-	-	-	-	-
	RJ	-	-	-	-	-	-	-	-	-	-	-	-	-
	RN	-	-	-	-	-	-	-	-	-	-	-	-	-
	RO	7.7	4.3	1.2	20.9	642	449	193	0.66	2.51	1.85	0.65	2.51	1.62
	RR	1.7	0.9	0.0	8.0	379	265	114	0.17	0.87	1.09	0.24	0.73	1.32
	RS	8.2	15.2	0.0	141.4	1238	866	372	0.81	6.86	0.57	0.76	8.57	0.44
	SC	20.0	11.2	0.2	93.2	595	416	179	0.74	5.92	2.51	0.86	3.89	3.86
	SE	-	-	-	-	-	-	-	-	-	-	-	-	-
	SP	7.5	4.6	0.0	48.0	8185	5729	2456	0.66	2.80	1.90	0.66	2.63	1.99
	TO	-	-	-	-	-	-	-	-	-	-	-	-	-

Table 2. Cubist model parameters, descriptive statistics, and results of prediction models of clay.

Clay (g kg^{-1})		Descriptive analysis		Observations			Training set		Validation set					
		Mean	SD	Min	Max	Total	Train.	Val.	R^2	RMSE	RPIQ	R^2	RMSE	RPIQ
National		327.5	217.6	0.0	986.8	32350	22645	9705	0.88	77.00	4.35	0.88	75.93	4.36
Regions	South	448.6	197.4	0.0	880.0	4059	2841	1218	0.83	82.02	3.91	0.83	81.12	4.20
	Southeast	284.5	201.6	5.0	960.0	17448	12214	5234	0.91	59.89	3.76	0.92	58.82	3.71
	Midwest	352.5	242.4	0.0	910.0	7656	5359	2297	0.94	59.25	7.75	0.94	58.61	7.68
	Northeast	222.4	136.9	3.0	629.0	609	426	183	0.75	70.13	3.06	0.78	59.28	2.99
	North	378.6	190.8	10.0	986.8	2578	1805	773	0.71	102.67	2.62	0.74	96.85	2.48
States	AC	-	-	-	-	-	-	-	-	-	-	-	-	-
	AL	130.9	76.7	30.0	300.0	32	22	10	0.72	45.38	2.20	0.71	38.71	2.07
	AM	315.8	224.0	10.0	986.8	550	385	165	0.85	86.37	2.83	0.80	101.50	2.18
	AP	469.8	172.4	80.0	920.0	432	302	130	0.57	112.74	2.31	0.61	109.45	2.54
	BA	230.2	136.0	11.0	626.0	402	281	121	0.83	56.45	3.60	0.84	52.85	4.01
	CE	256.8	144.4	20.0	534.0	23	16	7	0.49	105.19	1.96	0.42	107.84	1.77
	ES	209.9	130.7	10.0	600.0	100	70	30	0.92	37.60	3.99	0.92	38.53	5.97
	GO	311.6	253.0	20.0	890.0	2148	1504	644	0.97	47.24	9.97	0.96	52.15	8.99
	MA	-	-	-	-	-	-	-	-	-	-	-	-	-
	MG	550.7	244.1	10.0	960.0	1729	1210	519	0.88	83.95	5.10	0.88	87.79	5.01
	MS	372.0	236.7	0.0	910.0	5350	3745	1605	0.94	55.81	7.84	0.93	61.27	7.41
	MT	245.3	182.5	20.0	840.0	158	111	47	0.81	88.04	1.14	0.83	53.89	1.00
	PA	341.4	205.0	15.8	931.4	296	207	89	0.63	124.26	1.93	0.69	116.23	2.38
	PB	-	-	-	-	-	-	-	-	-	-	-	-	-
	PE	268.2	102.7	90.0	490.0	69	48	21	0.67	58.70	3.34	0.63	63.54	2.44
	PI	147.2	148.8	3.0	595.0	66	46	20	0.60	93.72	2.15	0.50	105.55	1.58
	PR	555.7	217.1	0.0	880.0	299	209	90	0.86	83.06	4.33	0.81	88.30	3.45
	RJ	-	-	-	-	-	-	-	-	-	-	-	-	-
	RN	268.2	138.2	98.0	533.0	17	9	8	0.52	100.94	2.16	0.43	98.51	1.49
	RO	451.3	161.1	80.0	880.0	642	449	193	0.89	53.96	4.45	0.89	55.85	4.65
	RR	327.9	129.0	48.0	800.0	626	438	188	0.69	73.49	2.18	0.68	69.38	2.33
	RS	445.6	207.8	0.0	837.0	1642	1149	493	0.84	84.21	4.37	0.85	78.80	4.60
	SC	435.7	181.1	0.0	800.0	2118	1483	635	0.84	72.58	4.00	0.85	70.18	4.13
	SE	-	-	-	-	-	-	-	-	-	-	-	-	-
	SP	255.5	173.2	5.0	910.0	15617	10932	4685	0.90	54.24	3.13	0.90	54.33	3.17
	TO	106.3	124.8	10.0	530.0	32	19	13	0.82	34.39	2.04	0.80	86.52	0.69

Table 3. Cubist model parameters, descriptive statistics, and results of prediction models of sand.

Sand (g kg^{-1})		Descriptive analysis				Observations			Training set			Validation set		
		Mean	SD	Min	Max	Total	Train.	Val.	R^2	RMSE	RPIQ	R^2	RMSE	RPIQ
National		529.7	284.1	0.0	990.0	33481	23437	10044	0.87	102.06	5.17	0.87	103.03	5.08
Regions	South	242.9	213.9	10.0	990.0	4059	2841	1218	0.78	101.82	2.65	0.77	101.54	2.65
	Southeast	606.2	252.4	0.0	970.0	17687	12381	5306	0.90	80.67	4.34	0.90	80.02	4.35
	Midwest	567.8	278.9	0.0	966.0	7656	5359	2297	0.94	66.10	8.05	0.94	68.04	7.90
	Northeast	651.9	237.3	26.0	988.0	682	477	205	0.80	106.18	2.74	0.83	102.00	2.98
	North	363.0	244.2	0.0	980.0	3397	2378	1019	0.79	113.73	4.04	0.78	114.95	3.68
States	AC	-	-	-	-	-	-	-	-	-	-	-	-	-
	AL	796.9	87.9	610.0	940.0	32	19	13	0.40	64.16	1.44	0.38	78.69	1.21
	AM	373.6	197.6	0.0	910.0	551	386	165	0.71	108.92	2.31	0.71	119.88	2.41
	AP	203.2	223.4	0.0	808.0	1250	875	375	0.84	90.47	3.54	0.83	93.95	3.30
	BA	734.6	150.7	96.0	988.0	402	281	121	0.78	69.31	3.13	0.80	71.66	2.51
	CE	434.1	280.9	26.0	977.0	95	67	29	0.88	107.93	5.22	0.90	87.27	5.43
	ES	747.4	136.1	360.0	950.0	100	70	30	0.86	46.28	3.35	0.87	57.95	4.23
	GO	626.9	294.5	36.9	951.0	2148	1504	644	0.96	55.65	10.26	0.97	53.94	10.31
	MA	-	-	-	-	-	-	-	-	-	-	-	-	-
	MG	271.2	215.7	0.0	970.0	1729	1210	519	0.80	98.88	3.20	0.78	102.97	3.25
	MS	542.0	269.5	0.0	966.0	5350	3745	1605	0.94	64.73	8.03	0.95	61.89	8.11
	MT	638.6	237.0	20.0	960.0	158	111	47	0.87	81.81	1.92	0.83	110.60	1.81
	PA	474.8	260.8	13.0	945.0	296	207	89	0.67	152.30	2.83	0.73	134.56	3.09
	PB	-	-	-	-	-	-	-	-	-	-	-	-	-
	PE	467.07	166.3	93.0	897.0	69	48	21	0.21	147.06	1.40	0.30	141.66	1.38
	PI	610.2	358.2	27.0	985.0	67	47	20	0.83	153.65	4.56	0.80	169.41	3.83
	PR	298.6	247.4	10.0	960.0	299	209	90	0.91	74.91	5.07	0.90	72.17	3.50
	RJ	-	-	-	-	-	-	-	-	-	-	-	-	-
	RN	-	-	-	-	-	-	-	-	-	-	-	-	-
	RO	508.2	162.9	100.0	900.0	642	449	193	0.88	54.50	4.40	0.88	59.80	4.52
	RR	446.1	177.7	6.0	910.0	626	438	188	0.55	121.23	1.90	0.58	117.05	1.95
	RS	241.7	229.5	10.0	928.0	1642	1149	493	0.78	111.46	2.39	0.81	99.93	2.43
	SC	235.9	194.2	10.0	990.0	2118	1483	635	0.82	82.92	3.27	0.82	83.12	3.01
	SE	-	-	-	-	-	-	-	-	-	-	-	-	-
	SP	641.9	228.6	0.0	969.0	15856	11099	4757	0.89	76.13	2.89	0.89	75.58	3.02
	TO	846.6	164.9	290.0	980.0	32	22	10	0.91	48.14	1.77	0.88	73.29	1.50

Table 4. Cubist model parameters, descriptive statistics, and results of prediction models of pH.

pH (H_2O)		Descriptive analysis				Observations			Training set			Validation set		
		Mean	SD	Min	Max	Total	Train.	Val.	R^2	RMSE	RPIQ	R^2	RMSE	RPIQ
National		5.4	0.7	0.0	8.9	26163	18314	7849	0.53	0.40	1.70	0.54	0.39	1.66
Regions	South	4.7	0.5	3.8	6.5	328	229	99	0.35	0.44	1.42	0.34	0.47	1.19
	Southeast	5.5	0.6	0.6	8.7	17001	11900	5101	0.49	0.42	1.83	0.49	0.42	1.83
	Midwest	5.4	0.7	0.0	8.5	5947	4162	1785	0.51	0.45	1.82	0.52	0.43	1.84
	Northeast	5.5	1.0	2.8	8.9	732	512	220	0.60	0.69	2.26	0.65	0.63	2.01
	North	4.9	0.6	2.5	7.7	2155	1508	647	0.41	0.47	1.69	0.46	0.45	1.76
States	AC	-	-	-	-	-	-	-	-	-	-	-	-	-
	AL	6.0	0.4	5.4	7.0	32	19	13	0.41	0.26	1.48	0.46	0.06	1.54
	AM	4.6	0.5	3.2	7.1	501	350	151	0.35	0.43	1.73	0.43	0.41	1.41
	AP	-	-	-	-	-	-	-	-	-	-	-	-	-
	BA	5.2	0.8	3.6	8.4	403	282	121	0.57	0.53	1.58	0.58	0.50	1.61
	CE	5.7	1.2	3.0	8.3	33	19	14	0.73	0.62	1.23	0.97	0.45	1.31
	ES	-	-	-	-	-	-	-	-	-	-	-	-	-
	GO	5.5	0.6	4.0	8.2	2050	1435	615	0.41	0.57	2.20	0.41	0.54	2.19
	MA	-	-	-	-	-	-	-	-	-	-	-	-	-
	MG	5.1	0.7	3.0	8.3	1352	946	406	0.50	0.46	1.61	0.55	0.45	1.45
	MS	5.3	0.7	0.0	8.5	3730	2611	1119	0.53	0.42	1.80	0.58	0.42	1.74
	MT	5.6	0.6	3.8	7.4	167	116	51	0.31	0.72	1.69	0.45	0.66	1.79
	PA	4.7	0.7	2.5	7.7	308		93	0.59	0.45	1.42	0.50	0.40	1.88
	PB	-	-	-	-	-	-	-	-	-	-	-	-	-
	PE	5.1	0.9	3.9	8.2	69	41	28	0.62	0.32	1.49	0.97	0.20	0.72
	PI	6.0	1.7	2.8	8.9	67	40	27	0.54	1.25	2.96	0.29	1.38	1.68
	PR	-	-	-	-	-	-	-	-	-	-	-	-	-
	RJ	-	-	-	-	-	-	-	-	-	-	-	-	-
	RN	-	-	-	-	-	-	-	-	-	-	-	-	-
	RO	5.1	0.4	3.9	6.4	672	470	202	0.26	0.59	1.94	0.30	0.54	2.00
	RR	4.9	0.6	3.5	7.6	627	438	189	0.53	0.22	1.38	0.55	0.15	1.32
	RS	5.0	0.5	4.4	6.0	23	16	7	0.43	0.26	1.39	0.35	0.07	0.14
	SC	4.7	0.4	3.8	6.5	305	213	92	0.29	0.62	1.40	0.30	0.57	1.33
	SE	-	-	-	-	-	-	-	-	-	-	-	-	-
	SP	5.5	0.6	0.6	8.1	15547	10882	4665	0.47	0.42	1.90	0.48	0.44	1.89
	TO	-	-	-	-	-	-	-	-	-	-	-	-	-

Table 5. Cubist model parameters, descriptive statistics, and results of prediction models of Cation Exchange Capacity (CEC).

CEC (cmol$_c$ kg^{-1})		Descriptive analysis				Observations			Training set			Validation set		
		Mean	SD	Min	Max	Total	Train.	Val.	R^2	RMSE	RPIQ	R^2	RMSE	RPI
National		47.7	41.9	0.0	958.0	17433	12203	5230	0.66	25.78	1.46	0.68	24.02	1.5
Regions	South	12.6	5.9	1.3	35.7	631	442	189	0.60	3.72	1.77	0.64	3.81	1.8
	Southeast	54.2	41.5	0.0	778.7	9896	6927	2969	0.75	21.64	1.63	0.79	20.02	1.7
	Midwest	49.9	32.9	0.0	528.4	3974	2782	1192	0.77	16.06	2.35	0.75	17.29	2.1
	Northeast	53.6	88.6	0.0	958.0	682	477	205	0.82	39.44	1.72	0.89	27.68	2.3
	North	23.3	23.9	0.1	248.0	2250	1575	675	0.74	13.16	1.81	0.72	12.40	1.9
States	AC	-	-	-	-	-	-	-	-	-	-	-	-	
	AL	49.7	15.9	27.0	97.0	31	18	13	0.11	20.41	1.10	0.02	10.26	0.7
	AM	22.0	35.8	1.0	248.0	501	351	150	0.91	11.75	0.56	0.84	14.92	0.4
	AP	37.3	16.3	15.1	138.8	432	302	130	0.67	9.70	1.99	0.59	10.17	1.7
	BA	33.5	39.6	0.0	224.0	403	282	121	0.79	19.42	3.08	0.77	19.30	3.0
	CE	25.4	24.5	3.3	117.7	33	23	10	0.70	15.84	0.86	0.74	10.44	1.3
	ES	85.5	28.0	44.7	183.5	100	70	30	0.35	22.94	1.58	0.29	25.23	1.1
	GO	68.3	38.9	3.2	454.8	606	424	182	0.72	23.32	1.71	0.72	17.14	2.3
	MA	-	-	-	-	-	-	-	-	-	-	-	-	
	MG	66.9	45.1	0.5	778.7	1745	1222	524	0.68	26.71	1.92	0.67	24.77	1.8
	MS	49.0	30.2	10.1	528.4	3101	2171	930	0.79	13.99	2.50	0.76	14.85	2.3
	MT	18.5	19.3	0.0	75.0	267	187	80	0.97	3.16	12.04	0.93	5.39	6.8
	PA	15.6	22.7	1.4	183.2	302	211	91	0.45	20.01	0.50	0.26	17.87	0.5
	PB	-	-	-	-	-	-	-	-	-	-	-	-	
	PE	6.8	2.6	2.7	17.3	69	48	21	0.25	1.95	1.48	0.16	2.71	1.3
	PI	199.8	177.4	17.1	958.0	48	24	24	0.53	88.89	2.37	0.53	211.64	0.9
	PR	-	-	-	-	-	-	-	-	-	-	-	-	
	RJ	16.6	8.6	7.4	32.9	12	6	6	0.27	10.41	0.50	0.56	6.94	2.0
	RN	36.6	35.0	1.9	102.2	25	13	12	0.95	7.35	5.40	0.99	5.11	11.4
	RO	29.8	13.1	8.1	102.1	638	447	191	0.69	7.30	1.96	0.65	8.06	1.9
	RR	4.5	2.6	0.1	24.0	377	264	113	0.40	1.91	1.41	0.36	2.64	1.1
	RS	13.5	7.2	1.3	35.7	326	228	98	0.55	5.14	1.66	0.50	4.63	1.7
	SC	11.6	4.0	3.8	24.5	305	214	92	0.63	2.43	2.53	0.65	2.32	2.3
	SE	142.4	142.1	23.0	627.4	65	45	20	0.93	41.99	1.23	0.94	37.78	5.3
	SP	51.0	40.1	0.0	564.0	8039	5627	2412	0.82	18.67	1.50	0.79	17.65	1.5
	TO	-	-	-	-	-	-	-		-	-	-	-	

Table 6. Cubist model parameters, descriptive statistics, and results of prediction models of Base Saturation (BS).

BS (%)		Descriptive analysis				Observations		Training set			Validation set		
		Mean	SD	Min	Max	Train.	Val.	R^2	RMSE	RPIQ	R^2	RMSE	RPIQ
National		39.6	23.6	0.0	100.0	19915	8535	0.50	16.93	2.39	0.49	17.01	2.40
Regions	South	44.4	22.9	1.7	89.7	228	98	0.56	15.26	2.55	0.54	15.51	2.60
	Southeast	44.9	22.4	0.0	100.0	11887	5094	0.48	16.33	2.20	0.49	16.28	2.24
	Midwest	31.4	21.9	1.0	99.0	5183	2221	0.58	14.62	2.28	0.57	14.72	2.32
	Northeast	37.9	29.3	1.4	100.0	445	191	0.81	12.93	3.47	0.79	13.42	3.20
	North	29.8	24.1	0.0	100.0	2172	931	0.70	13.42	2.72	0.69	13.93	2.70
States	AC	-	-	-	-	-	-	-	-	-	-	-	-
	AL	41.4	18.1	16.0	74.0	19	13	0.14	17.16	1.51	0.33	15.40	2.11
	AM	19.0	14.2	1.0	100.0	351	150	0.47	9.29	1.29	0.70	9.91	1.41
	AP	40.7	27.2	0.0	88.0	874	375	0.71	14.75	3.53	0.63	16.63	3.10
	BA	23.1	17.8	1.4	94.4	281	121	0.53	12.44	1.72	0.57	12.33	1.70
	CE	86.0	11.1	52.0	100.0	23	10	0.69	6.76	1.26	0.60	8.07	0.93
	ES	-	-	-	-	-	-	-	-	-	-	-	-
	GO	29.8	24.6	1.0	99.2	1492	640	0.70	13.65	3.01	0.69	13.40	2.87
	MA	-	-	-	-	-	-	-	-	-	-	-	-
	MG	30.2	23.3	0.0	100.0	1222	524	0.57	16.75	2.12	0.55	16.96	2.02
	MS	31.5	20.5	2.0	97.0	3585	1537	0.58	13.73	2.24	0.43	15.98	1.92
	MT	50.3	18.8	9.0	96.0	105	45	0.58	11.89	1.77	0.38	16.01	1.12
	PA	26.8	21.1	0.0	95.5	211	91	0.54	14.81	1.73	0.52	15.00	1.75
	PB	-	-	-	-	-	-	-	-	-	-	-	-
	PE	41.4	22.0	9.0	100.0	48	21	0.59	14.22	1.93	0.28	19.56	1.28
	PI	76.9	20.4	25.0	100.0	47	20	0.57	12.37	1.88	0.37	18.68	1.49
	PR	-	-	-	-	-	-	-	-	-	-	-	-
	RJ	-	-	-	-	-	-	-	-	-	-	-	-
	RN	-	-	-	-	-	-	-	-	-	-	-	-
	RO	15.0	8.6	2.0	63.0	449	193	0.46	6.76	1.63	0.45	6.25	1.28
	RR	38.1	22.8	1.1	100.0	264	113	0.55	16.13	2.10	0.39	17.88	2.07
	RS	44.4	22.9	1.7	89.7	228	98	0.69	13.13	3.02	0.35	18.61	2.11
	SC	-	-	-	-	-	-	-	-	-	-	-	-
	SE	-	-	-	-	-	-	-	-	-	-	-	-
	SP	46.8	21.6	0.0	99.8	10587	4537	0.52	15.36	2.22	0.33	17.69	1.97
	TO	-	-	-	-	-	-	-	-	-	-	-	-

2.3.4. Spectral classification

In order to categorize how many spectral patterns are required to represent Brazilian soils according to the shape of the 39,284 spectral signatures, the first three principal component scores (Fig. 7a) were used as variables in the cluster analysis (Terra et al., 2015). The eigenvectors of PC1 are dominated by positive loadings along the wavelengths, which captured 64% of the total variance. The high positive loadings were found in the visible region that showed the characteristic absorptions for iron oxides (Fig. 7a). The eigenvector of the PC2 (10%) showed high negative loadings near at wavelengths for the characteristic absorptions of 2:1 clay mineral (illite and smectite) and possibly organic matter. The PC3 (7.6%) fluctuated between positive and negative loadings.

In order to reduce the dimensionality of the data, the first three principal component scores (Fig. 7a), were applied to determine the optimal number of clusters. We selected six clusters (classes) because, since the pE was maximized and the pC was minimized when six clusters were obtained (Table 7), which was then selected to represent the most satisfactory cluster for the data. In the crisp clustering, each observation receives membership values of 0 or 1 for each cluster. In the scatter diagram it shows the distribution of all observations colored in the 6 crisp classes (Fig. 7b).

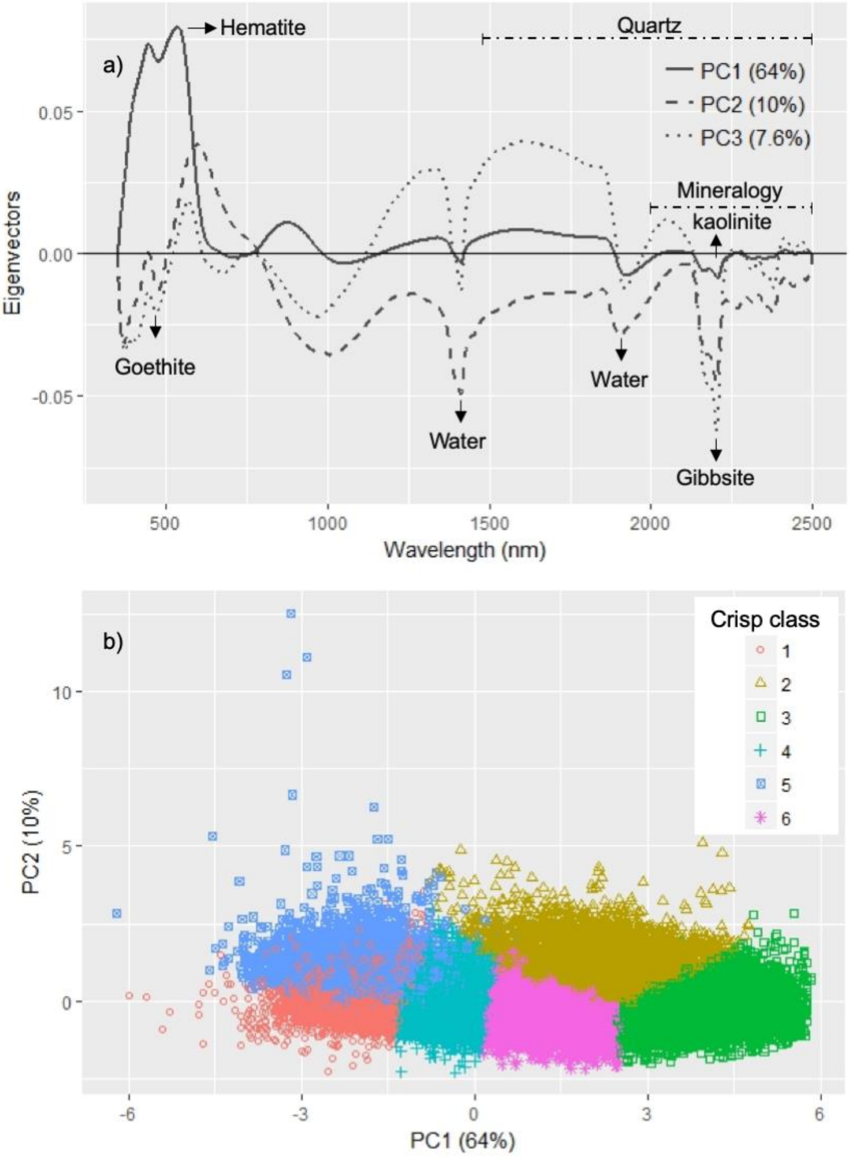


Fig. 7. Principal components eigenvectors of PC 1, 2, and 3, (a) and crisp fuzzy-*c*-means classification, considering six groups (b). Principal components analysis was performed with the continuum removed spectra. Sampling points clustering was based on PC scores.

Table 7. Fuzzy validation indices for the optimum number of clusters, the partition entropy (pE), and the partition coefficient (pC).

Number of clusters*	pE	pC
3	0.57	0.68
4	0.63	0.68
5	0.71	0.65
6	**1.06**	**0.48**
7	0.86	0.61
8	0.91	0.60
9	0.97	0.58
10	1.04	0.55
11	1.11	0.53
12	1.05	0.56

* In bold is the optimal number of clusters.

The average CR spectra of 6 classes is presented in Fig. 8. The average spectrum of classes 1, 4, and 5 (Fig. 8a, g, i) were characterized by absorptions representative of soils with abundant iron oxides (400 to 600 nm), while the classes 2, 3, and 6 (Fig. 8c, e, l) showed relevant absorptions in the NIR-SWIR regions. Although fairly similar, class 4 may be differentiated from 1 and 5 classes by the absorption features at 500 and 900 nm. The spectrum from class 4 (Fig. 8g) presented features less pronounced than the other two (Fig. 8a, i). These features were related to the crystal field electronic effect of hematite mineral and consequently to the ferric ion (Fe^{+3}) observed in such iron oxides. The interaction between electromagnetic energy and hematite results in electronic transitions, creating the absorption features centered at 530 and 885 nm. Spectra from classes 1 and 5 can be distinguished from each other by the CR reflectance factor at the SWIR-1 range (1,000-1,800 nm), whereas class 5 presented a lower CR factor. Fuzzy class 2 can be distinguished from the others by the lower CR factor of features centered at 1,200, 1,900 and 2,200 nm. Finally, class 3 spectrum has high CR factor between 350-750 nm, which is related to low content of iron oxides in soils.

In the fuzzy-*c*-means clustering, each data point can belong to more than one cluster. The probability of each soil sample being classified in the fuzzy membership class 1 is shown in Fig. 8b. In the center of the fuzzy membership class 1 are the samples with high probability of pertaining to this class (red color). The same analogy applies to the other five classes. These findings suggest that six types of spectra represent the whole population of Brazilian soils. The six classes of spectra were discriminated according to the spectral pattern of soils, which is directly linked to intrinsic heterogeneous characteristics, where by contents of SOC, iron oxides, mineralogy of the clay fraction, particle size distribution, and moisture, are the ones that most influence the spectral responses.

Stoner and Baumgardner (1981) found five spectral classes in a large database of the U.S. and Brazil. The authors suggested that five soil spectral reflectance curves could be distinguished as sharing in common certain differentiating characteristics concerning mainly the organic matter and iron oxide contents. One of the classes was detected because it had its origin in Brazil (Paraná state). Formaggio et al. (1996) identified four patterns of spectral curves according to the shape and intensity of the parameters in one state (São Paulo) in Brazil. Viscarra Rossel et al. (2016), used a global spectral library to characterize the world's soil and found six classes of spectra. The authors also stated that grouping the spectra into more homogeneous spectral classes can improved the modeling by removing bias in the predictions. Terra et al. (2018) found 6 different patterns of soil spectra based on differences in reflectance intensity and absorption features caused by weathering intensification, which enabled to distinguish soil samples regarding similarity of particle size distribution, mineralogy, and some chemical properties.

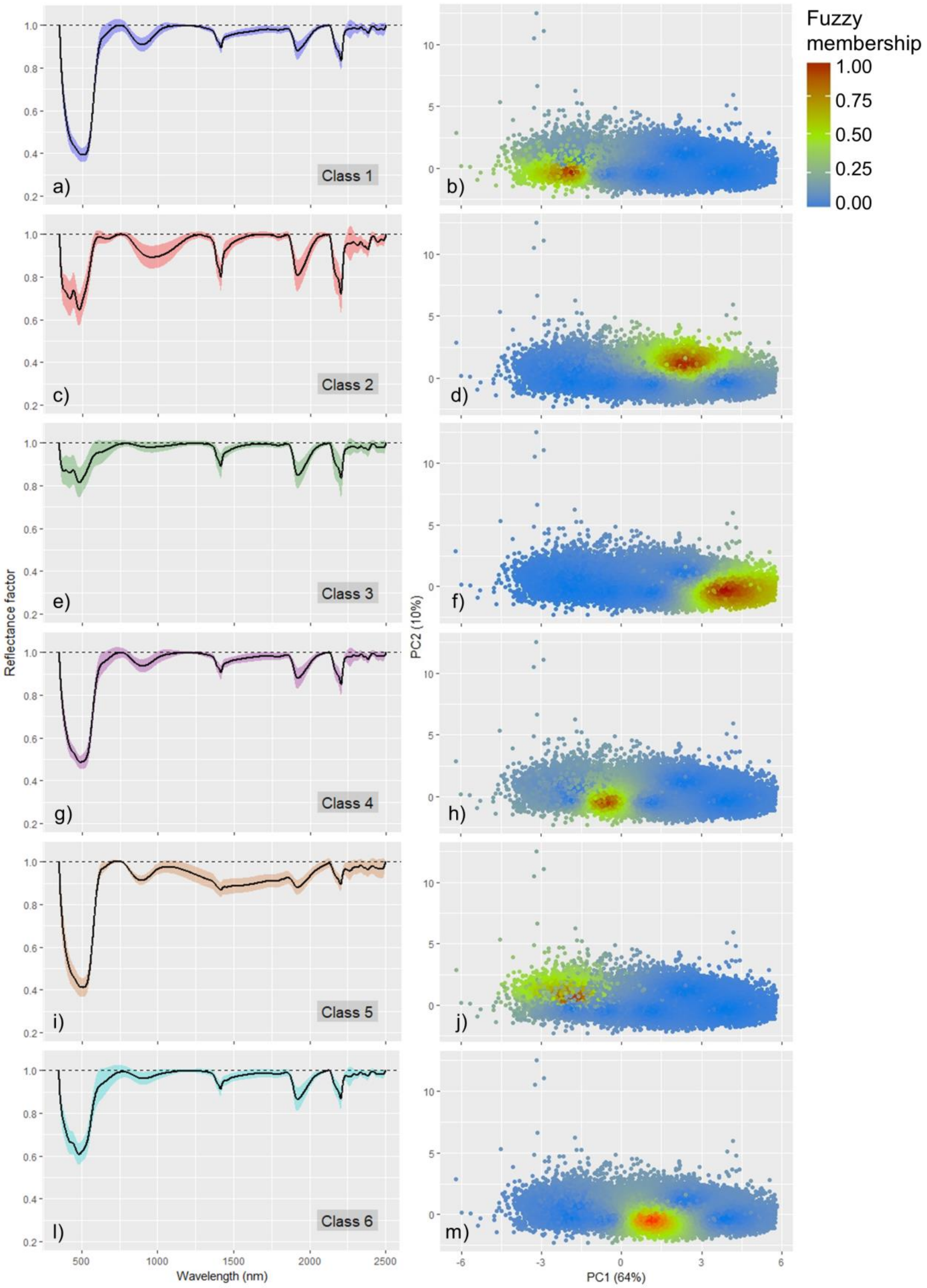


Fig. 8. The average continuum removed spectrum of each fuzzy cluster (a, c, e, g, i, and l) and fuzzy membership values for the 6 clusters (b, d, f, h, j, and m).

From the six classes defined by crisp fuzzy (Crisp-1 to Crisp-6), Crisp-4 presents the largest number of samples (9,893 samples), closely followed by Crisp-1 (9,530 samples) (Fig. 9). Most of the samples in fuzzy c-means classes (FMD-1 to FMD-6) were correctly assigned to the correspondent crisp classes. More than half of FMD-1 samples were classified as Crisp-1 (Fig. 9), another part was misclassified as Crisp-4 and Crisp-5, while few ones were defined as Crisp-2, Crisp-3, and Crisp-6. This suggests that Crisp-1, 4, and 5 are related to each other, which is corroborated by the similarity in CR spectra of these classes (Fig. 8a, g, i). In FMD-2, the dominant misclassified classes were Crisp-3 and Crisp-6 (Fig. 9). Most of FMD-3 individuals were correctly assigned as Crisp-3, with few samples misclassified as Crisp-2 and Crisp-6. FMD-4 showed higher misclassification with Crisp-1, followed by Crisp-6. As expected, FMD-5 was mostly misclassified as Crisp-1 and Crisp-4, confirming the correlation between their spectral pattern. Finally, FMD-6 was mainly misclassified as Crisp-4, demonstrating the similarity between the spectra of these classes (Fig. 8g, l).

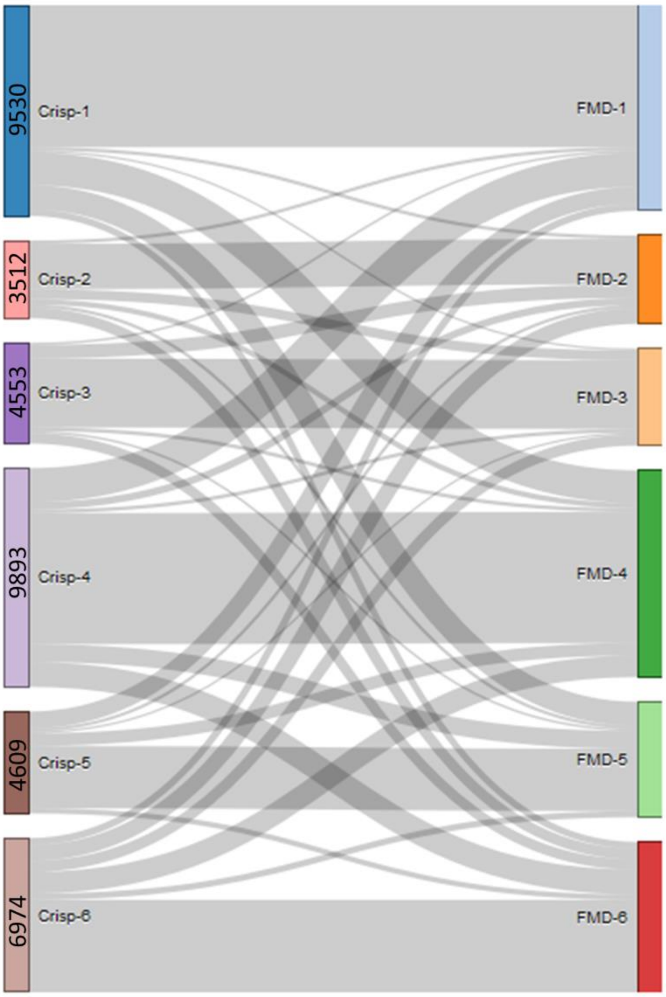


Fig. 9. Sankey diagram showing the relative associations between crisp and Fuzzy Membership Degree (FMD) for each class.

2.3.5. Correspondence Analysis

The CA analysis showed that spectral classes 1 and 5 were correlated with the MW region, while class 4 was similar with region SE, class 6 with region S, class 2 with region N, and class 3 with region NE (Fig. 10a). The spectral classes 5, 1, 4, and 6 resemble MS, SP, PR, GO, MA, RS, RJ, MT, and PB states, that is the points are very close in the simultaneous plot of row and column coordinates (Fig. 10b). The spectral class 3 showed proximity with AL, RN, AM, ES, BA, and PE with most of them from regions N and NE. The spectral class 2 showed some similarity with AP, RO, RR, PA, and AC states. The sedimentary rocks were highly associated with spectral class 1, metamorphic rocks were correlated with classes 2 and 3, and igneous rocks with classes 5 and 6 (Fig. 10c). For the CA between spectral classes and biomes (Fig. 10d), the classes 1, 4, and 6 were related with Atlantic Forest biome, class 5 with Cerrado, class 3 with Pampa, Pantanal and Caatinga, and class 2 with Amazon. For the CA of spectral and soil classes only profile samples that have layer B collected were used. The Gleysols and Plinthosols classes were associated with spectral class 2, Ferralsol is highly associated to classes 1 and 4 (Fig. 10e). Nitosols and Lixisols were associated with class 6, and Cambisols with class 3. Histosols, Arenosols, Podzols, Planosols, and Vertisols were not associated with any particular spectral class but it is worth mentioning that they were closer to class 3. The CA of spectral classes and soil layers showed that classes 6 and 3 were correlated with A layer (Fig. 10f). The B layer showed some association with spectral classes 2, 4, and 5 (Fig. 10f), which is corroborated by the fact that spectral classes 4 and 5 showed similarity in CR spectrum (Fig. 8g, i). The C layer was strongly associated with class 1. It is also important to mention that the spectral class 4 was right in the middle of layers A and B, and spectral class 5 was in the middle of layers B and C. The D layer did not present direct associations with any of the spectral classes (Fig. 10f), indicating that the D layer presented contradictory spectral pattern considering the 6 spectral classes.

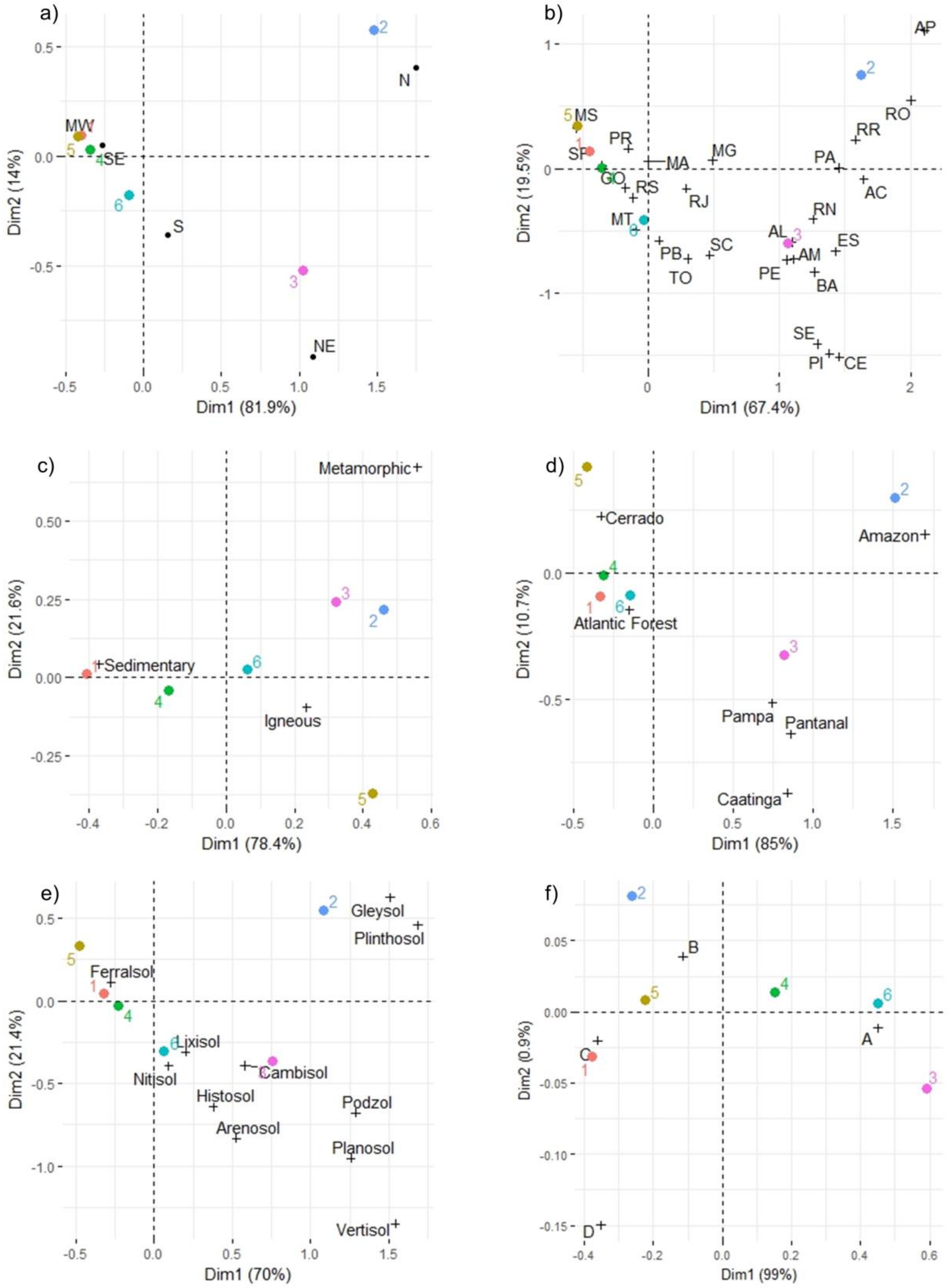


Fig. 10. Ordination diagrams from the correspondence analysis (CA) between the 6 spectral classes and Brazilian regions (a), states (b), geology (c), biomes (d), soil classes (only profile samples that have layer B were used) (e), and layers (f).

2.4. CONCLUSIONS

The BSSL provided strong evidence to be a useful tool to estimate soil attributes such as clay, sand, SOC, CEC, pH, and BS, with variable results. There were differences among models considering national (for all Brazil), regional and state scales. The results were coherent for clay, sand, SOC, and CEC. The attributes with low content in soils are more prone to show high inaccuracy and need further evaluations, such as chemical ones (Ca, K, P, others). Cluster analysis showed that Brazil has six classes of spectral signatures among the studied whole population and there were clear differences among spectra developed in different geographic (i.e., states) and environmental locations (i.e., geology). The results endorse the importance and relevance of spectral libraries for soil evaluation in support of the quantification of soil quality and classification. The large spectral database, will be enhanced at scales that meet the users' needs for soil mapping in Brazil. The use of sensors and geotechnologies, due to their rapid and low cost analysis, allows a higher sample throughput and denser sampling. Both can generate data for soil survey and mapping that will assist in the sustainable management of agriculture and forest systems. We believe that soil spectroscopy in Brazil is on the right track, because it is a fast, simple, accurate and most importantly non-pollutant method. In addition, a strong collaborative network and infrastructure has been formed that supports the expansion of soil spectral soil mapping. With the approval of Brazil's National Soils Program (PronaSolos), which is promising in relation to the mapping of Brazilian soils, we hope that techniques such as soil spectroscopy can be applied.

Besides the importance of spectral standardization, spectral libraries such as the BSLL must be accompanied by chemical and physical characterization of soil attributes. Depending on the volume of soil samples, the standard procedures applied to spectral measurements can be complex and time-consuming. Despite the spectral variation can also be an issue to be faced, the incorporation of soil chemical and physical data is crucial. The spectral acquisition is no more an obstacle to the organization of spectral libraries but the collection of reliable and consistent soil chemical and physical data poses challenges. Soil spectral information is dependent on wet analysis. In some cases, the soil sample collection and wet chemistry and physical analyses were conceived before the BSSL initiative. Therefore, standardized collection and analyses of samples were not possible in all circumstances. We agreed that it can create a margin of error in both calibration and validation processes, but the novelty of this database and the difficulty to gather information must be taken into consideration. Furthermore, having a SSL that represent Brazilian soils is just as important as defining the degree of uncertainty. Therefore, in the first stage of the SSL’s development we decided to include samples from many wet chemistry and physical laboratories.

The results are a first step towards the establishment of the BSSL. The database is being continuously increased with new information, consequently increasing its representativeness along the Brazilian territory. Improvements in the frame-work have been conduced, including computational routines implementing sophisticated statistical procedures, which will reduce the uncertainties in the calibration procedure. Among them, data will be filtered and standardized with wavelets, which will help to account for the inconsistencies in sample preparation, different measurement protocols and instruments that were used. In parallel, different machine learning algorithms have been evaluated, aiming to define the most suitable data mining method for our dataset. These mining procedures account for local relationships in the data providing to the models a wide usability at different spatial scales (local, regional and national). We also developed an interactive online platform to disseminate the use of spectroscopy in soil science, and to interact with the database administrators. The soil dataset and their contributors can be accessed at <https://bibliotecaespectral.wixsite.com/esalq>. By increasing the number of users, the data available and knowledge will also increase and consequently the BSSL will be constantly improved to represent the variability of Brazilian soils. We have to keep in mind that the importance of spectra is not only concentrated on chemical agriculture information (i.e., Ca, Mg, K, others). From one measurement, we can achieve chemical, physical and mineralogical information, which all are important for soil mapping and agriculture as well. Finally, in our vision, wet analysis is an important method and now have the great opportunity to merge knowledge (and aggregate information) with proximal sensing, to evolve on soil analysis to a new generation and its benefits.

ACKNOWLEDGEMENTS

We would like to thank the São Paulo Research Foundation (FAPESP) for the financial support for first author (Project grant n. 2014/22262-0), second author (Project grant n. 2017/03207-6) and third author (Project grant n. 2016/26176-6) and the National Council for Scientific and Technological Development (CNPq). Also, to thank the Geotechnologies on Soil Science group - GeoCis (*esalqgeocis.wixsite.com/english*) and to everybody that directly or indirectly assisted on publishing this study.

REFERENCES

Adhikari, K., Hartemink, A.E., 2016. Linking soils to ecosystem services — A global review. Geoderma 262, 101–111. https://doi.org/10.1016/J.GEODERMA.2015.08.009

Baldridge, A.M., Hook, S.J., Grove, C.I., Rivera, G., 2009. The ASTER spectral library version 2.0. Remote Sens. Environ. 113, 711–715. https://doi.org/10.1016/J.RSE.2008.11.007

Bellinaso, H., Demattê, J.A.M., Romeiro, S.A., 2010. Soil Spectral Library and Its Use in Soil Classification. R. Bras. Ci. Solo 34, 861–870. https://doi.org/10.1590/S0100-06832010000300027

Benzécri, J.P., 1992. Correspondence Analysis Handbook. Marcel Dekker, New York, NY.

Bezdek, J.C., Ehrlich, R., Full, W., 1984. FCM: The fuzzy c-means clustering algorithm. Comput. Geosci. 10, 191–203. https://doi.org/10.1016/0098-3004(84)90020-7

Bowers, S.A., Hanks, R.J., 1965. Reflection of Radiant Energy from Soils. Soil Sci. 100, 130–138.

Brodský, L., Klement, A., Penížek, V., Kodešová, R., Borůvka, L., 2011. Building soil spectral library of the Czech soils for quantitative digital soil mapping. Soil Water Res 6, 165–172.

Brown, D.J., Shepherd, K.D., Walsh, M.G., Dewayne Mays, M., Reinsch, T.G., 2006. Global soil characterization with VNIR diffuse reflectance spectroscopy. Geoderma 132, 273–290. https://doi.org/10.1016/j.geoderma.2005.04.025

Cambule, A.H., Rossiter, D.G., Stoorvogel, J.J., Smaling, E.M.A., 2012. Building a near infrared spectral library for soil organic carbon estimation in the Limpopo National Park, Mozambique. Geoderma 183–184, 41–48. https://doi.org/10.1016/j.geoderma.2012.03.011

Cebeci, Z., Yildiz, F., Kavlak, A.T., Cebeci, C., Onder, H., 2018. ppclust: Probabilistic and Possibilistic Cluster Analysis. R Packag. version 0.1.1.

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

Demattê, J.A.M., 2016. From Profile Morphometrics to Digital Soil Mapping, in: Digital Soil Morphometrics. Springer International Publishing, pp. 383–399. https://doi.org/10.1007/978-3-319-28295-4_24

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

Demattê, J.A.M., Oliveira, J. de C., Tavares, T.R., Lopez, L.R., Terra, F. da S., Araújo, S.R., Fongaro, C.T., Maia, S.M.F., Mello, F.F. de C., Rizzo, R., Vicente, S., de Melo Bortolleto, M.A., Cerqueira, P.H.R., 2016. Soil chemical alteration due to slaughterhouse waste application as identified by spectral reflectance in São Paulo State, Brazil: an environmental monitoring useful tool. Environ. Earth Sci. 75. https://doi.org/10.1007/s12665-016-6042-2

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

Epiphanio, J.C.N., Formaggio, A.R., Valeriano, M.D.M., Oliveira, J.B., 1992. Comportamento espectral de solos do Estado de São Paulo. Instituto Nacional de Pesquisas Espaciais, São José dos Campos, São Paulo.

Formaggio, A.R., Epiphanio, J.C.N., Valeriano, M.M., Oliveira, J.B., 1996. Comportamento espectral (450-2.450 nm) de solos tropicais de São Paulo. Rev. Bras. Ciência do Solo 20, 467–474.

Garrity, D., Bindraban, P., 2004. A Globally Distributed Soil Spectral Library Visible Near Infrared Diffuse Reflectance Spectra. ICRAF (World Agroforestry Centre)/ISRIC (World Soil Information) Spectral Library, Nairobi, Kenya.

Gogé, F., Joffre, R., Jolivet, C., Ross, I., Ranjard, L., 2012. Optimization criteria in sample selection step of local regression for quantitative analysis of large soil NIRS database. Chemom. Intell. Lab. Syst. 110, 168–176. https://doi.org/10.1016/J.CHEMOLAB.2011.11.003

Grunwald S., G.M. Vasques, R.G. Rivero. 2015. Fusion of soil and remote sensing data to model soil properties. In: Sparks, D.L. (Ed.), Advances in Agronomy, Vol. 131, pp. 1–109.

Grunwald S., C. Yu, X. Xiong. 2018. Transferability and scalability of total soil carbon prediction models in Florida, USA. Pedosphere J. 28(6): 856-872.

IUSS Working Group WRB, 2015. World Reference Base for Soil Resources 2014, update 2015. International soil classification system for naming soils and creating legends for soil maps. World Soil Resources Reports No. 106. FAO, Rome.

Ji, W., Li, S., Chen, S., Shi, Z., Viscarra Rossel, R.A., Mouazen, A.M., 2016. Prediction of soil attributes using the Chinese soil spectral library and standardized spectra recorded at field conditions. Soil Tillage Res. 155. https://doi.org/10.1016/j.still.2015.06.004

Jónsson, J.Ö.G., Davíðsdóttir, B., 2016. Classification and valuation of soil ecosystem services. Agric. Syst. 145, 24–38. https://doi.org/10.1016/j.agsy.2016.02.010

Knadel, M., Deng, F., Thomsen, A., Greve, M., 2012. Development of a Danish national Vis–NIR soil spectral library for soil organic carbon determination, in: Minasny, B., Malone, B.P., McBratney, A.B. (Eds.), Digital Soil Assessments and Beyond: Proceedings of the 5th Global Workshop on Digital Soil Mapping. Sydney, Australia, pp. 403–408.

Kuhn, M. et al., 2017. caret: Classification and Regression Training. R package version 6.0-73.

Lê, S., Josse, J., Husson, F., 2008. FactoMineR: An R Package for Multivariate Analysis. J. Stat. Softw. 25, 1–18. https://doi.org/10.18637/jss.v025.i01

Minasny, B., McBratney, A.B., 2008. Regression rules as a tool for predicting soil properties from infrared reflectance spectroscopy. Chemom. Intell. Lab. Syst. 94, 72–79. https://doi.org/10.1016/j.chemolab.2008.06.003

Moeys, J., 2016. soiltexture: Functions for Soil Texture Plot, Classification and Transformation. R package version 1.4.1.

Mutanga, O.M.C., Skidmore, A.K., Kumar, L., Ferwerda, J., 2005. Estimating tropical pasture quality at canopy level using band depth analysis with continuum removal in the visible domain. Int. J. Remote Sens. 26, 1093–1108. https://doi.org/10.1080/01431160512331326738

Nocita, M., Stevens, A., Toth, G., Panagos, P., van Wesemael, B., Montanarella, L., 2014. Prediction of soil organic carbon content by diffuse reflectance spectroscopy using a local partial least square regression approach. Soil Biol. Biochem. 68, 337–347. https://doi.org/10.1016/j.soilbio.2013.10.022

Polidoro, J.C., Mendonça-Santos, M.D.L., Lumbreras, J.F., Coelho, M.R., Carvalho Filho, A. De, Motta, P.E.F. Da, Carvalho Junior, W. De, Araujo Filho, J.C. De, Curcio, G.R., Correia, J.R., Martins, E.D.S., Spera, S.T., Oliveira, S.R.D.M., Bolfe, E.L., Manzatto, C. V., Tosto, S.G., Venturieri, A., Sa, I.B., Oliveira, V.A. De, Shinzato, E., Anjos, L.H.C. Dos, Valladares, G.S., Ribeiro, J.L., Medeiros, P.S.C. De, Moreira, F.M.D.S., Silva, L.S.L., Sequinatto, L., Aglio, M.L.D., Dart, R.D.O., 2016. Programa Nacional de Solos do Brasil (PronaSolos), 1st ed. Embrapa Solos, Rio de Janeiro, RJ.

Quinlan, J., 1992. Learning with continuous classes, in: Adams, A., Sterling, L. (Eds.), Proceedings AI’92, 5th Australian Conference on Artificial Intelligence.World Scientific. Singapure, pp. 343–348.

R Core Team, 2018. R: A language and environment for statistical computing.

Schmidt, M., 2008. The Sankey Diagram in Energy and Material Flow Management. J. Ind. Ecol. 12, 82–94. https://doi.org/10.1111/j.1530-9290.2008.00004.x

Shepherd, K.D., Walsh, M.G., 2002. & SOIL & PLANT ANALYSIS Development of Reflectance Spectral Libraries for Characterization of Soil Properties 988–998.

Shi, Z., Wang, Q., Peng, J., Ji, W., Liu, H., Li, X., Viscarra Rossel, R.A., 2014. Development of a national VNIR soil-spectral library for soil classification and prediction of organic matter concentrations. Sci. China Earth Sci. 57, 1671–1680. https://doi.org/10.1007/s11430-013-4808-x

Soil Survey Staff, 2014. Soil taxonomy: A basic system of soil classification for making and interpreting soil surveys, Twelfth Ed. ed. Natural Resources Conservation Service. U.S. Department of Agriculture Handbook.

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

Stevens, A., van Wesemael, B., Bartholomeus, H., Rosillon, D., Tychon, B., Ben-Dor, E., 2008. Laboratory, field and airborne spectroscopy for monitoring organic carbon content in agricultural soils. Geoderma 144, 395–404. https://doi.org/10.1016/J.GEODERMA.2007.12.009

Stoner, E.R., Baumgardner, M.F., 1981. Characteristic Variations in Reflectance of Surface Soils. Soil Sci. Soc. Am. J. 45, 1161. https://doi.org/10.2136/sssaj1981.03615995004500060031x

Terra, F.S., Demattê, J.A.M., Viscarra Rossel, R.A., 2018a. Proximal spectral sensing in pedological assessments: vis–NIR spectra for soil classification based on weathering and pedogenesis. Geoderma 318, 123–136. https://doi.org/10.1016/j.geoderma.2017.10.053

Terra, F.S., Demattê, J.A.M., Viscarra Rossel, R.A., 2018b. Proximal spectral sensing in pedological assessments: vis–NIR spectra for soil classification based on weathering and pedogenesis. Geoderma 318, 123–136. https://doi.org/10.1016/J.GEODERMA.2017.10.053

Terra, F.S., Demattê, J.A.M., Viscarra Rossel, R.A., 2015. Spectral libraries for quantitative analyses of tropical Brazilian soils: Comparing vis–NIR and mid-IR reflectance data. Geoderma 255–256, 81–93. https://doi.org/10.1016/j.geoderma.2015.04.017

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

Viscarra Rossel, R.A., Behrens, T., Ben-Dor, E., Brown, D.J., Demattê, J.A.M., Shepherd, K.D., Shi, Z., Stenberg, B., Stevens, A., Adamchuk, V., Aïchi, H., Barthès, B.G., Bartholomeus, H.M., Bayer, A.D., Bernoux, M., Böttcher, K., Brodský, L., Du, C.W., Chappell, A., Fouad, Y., Genot, V., Gomez, C., Grunwald, S., Gubler, A., Guerrero, C., Hedley, C.B., Knadel, M., Morrás, H.J.M., Nocita, M., Ramirez-Lopez, L., Roudier, P., Campos, E.M.R., Sanborn, P., Sellitto, V.M., Sudduth, K.A., Rawlins, B.G., Walter, C., Winowiecki, L.A., Hong, S.Y., Ji, W., 2016. A global spectral library to characterize the world’s soil. Earth-Science Rev. 155, 198–230. https://doi.org/10.1016/j.earscirev.2016.01.012

Viscarra Rossel, R.A., McBratney, A.B., 2008. Diffuse Reflectance Spectroscopy as a Tool for Digital Soil Mapping, in: Digital Soil Mapping with Limited Data. Springer Netherlands, Dordrecht, pp. 165–172. https://doi.org/10.1007/978-1-4020-8592-5_13

Viscarra Rossel, R.A., Webster, R., 2012. Predicting soil properties from the Australian soil visible-near infrared spectroscopic database. Eur. J. Soil Sci. 63, 848–860. https://doi.org/10.1111/j.1365-2389.2012.01495.x

Viscarra Rossel, R.A.A., McGlynn, R.N.N., McBratney, A.B.B., 2006. Determining the composition of mineral-organic mixes using UV–vis–NIR diffuse reflectance spectroscopy. Geoderma 137, 70–82. https://doi.org/10.1016/j.geoderma.2006.07.004

Wall, D.H., Nielsen, U.N., 2012. Biodiversity and Ecosystem Services: Is It the Same Below Ground? Nat. Educ. Knowl. 3, 8.

Zeng, R., Zhao, Y.-G., Li, D.-C., Wu, D.-W., Wei, C.-L., Zhang, G.-L., 2016. Selection of "Local" Models for Prediction of Soil Organic Matter Using a Regional Soil Vis-NIR Spectral Library. Soil Sci. 181, 13–19. https://doi.org/10.1097/SS.0000000000000132

3. THE BRAZILIAN PROGRAM OF SOIL ANALYSIS VIA SPECTROSCOPY (PROBASE): BRINGING WET LABORATORIES TO UNDERSTAND NEW TECHNOLOGIES

ABSTRACT

Soil analysis by spectroscopy has been extensively studied. However, the applications, potential and limitations of the technique are still not well known by end-users, such as commercial laboratories. The present work deals with bringing together spectroscopy and the commercial community (wet laboratories) to perform practical dynamics and understand how spectroscopy works. A total of 35 laboratories sent soil samples for this study. They received a protocol with instructions and sent soil samples (200 per laboratory with a total of 7200-ProBASE dataset) with their respective analysis to a centralized spectroscopy laboratory (University of São Paulo). Samples were measured for visible-Near-Short-wave-infrared, vis-NIR-SWIR (400-2500 nm), Middle-infrared, MIR (3000-25000 nm) and X-ray fluorescence (XRF). We also used in this exercise the Brazilian Soil Spectral Library (BSSL) model (vis-Nir-SWIR) constructed with 49753 samples (BSSL-dataset). We performed three strategies with vis-NIR-SWIR as follows: a) applied a model using the ProBASE-dataset in each laboratory dataset (200 samples per laboratory); b) applied a model using the entire ProBASE-dataset (about 7200 samples); c) applied the BSSL-dataset model to the entire ProBASE-dataset. Afterwards, we modelized using other spectral ranges for comparison. There was a great variation of results which can be summarized as follows: a) all techniques are environmentally friendly, b) vis-NIR-SWIR and XRF are the faster; c) XRF had the best quantification for calcium; d) MIR had the best performance for clay and sand contents; e) vis-NIR-SWIR had for organic matter, Mg, base sum and CEC; f) the three spectral range had good performance for sand and clay; g) chemical elements (Ca, K, Mg, pH) models vary in accordance with the regional population used. In our case, the best models were first from the local population (laboratory samples), going to the ProBASE-dataset (regional) and after using a national model (BSSL). Results indicate that all spectral ranges are useful to be part of a wet laboratory dynamics, where each one has advantages and limitations, but can be complementary. This indicates that it is feasible to create a hybrid laboratory. The dynamics was very well received by the commercial community (95% responded positively for the initiative) which now understands the system and can make better decisions.

Key words: soil analysis, quantification, commercial spectroscopy, sensors

3.1. INTRODUCTION

Brazil has about 66 million ha with strong agriculture activity and more 11% possible to use (IBGE, 2019). These are numbers from a single country that gathers a large extension of the world’s agricultural soils. The optimization of the productivity requires enough soil data to support its management, which exponentially increases the demand for soil analysis. It is estimated about 600.000 millions soil analyses around the world (Demattê et al., 2019).

Molin and Tavares (2019) indicated that for mapping soil attributes would be at least 1 sample per ha in precision agriculture. To perform all these soil analyses, we have the production of several pollutants residues. For example, for organic carbon determination by wet-combustion, it is used about 0,196 g of ammonium sulphate ($Cr_2O_7^{2-}$), 1,2 g of hexahydrated iron and 5 ml of sulphuric acid (H_2SO_4) per sample (Demattê et al., 2019). These will take us to a total of 840.000 kg of ammonium and iron dichromate and sulfate, 3.000.000 of liters of sulfuric acid per year (Demattê et al., 2019). In the case of dry-combustion, there is no use of chemical reagents, but the costs for reaching the content by it is much higher. Generally, the methods employed by wet laboratories (that employ chemistry in the liquid phase for soil analyses) are non environmental friendly, expensive and time consuming (Soriano-Disla, 2014; Nocita et al., 2015) Thus, the question is if the traditional soil analysis is sustainable.

Due to the indicated issues (costs, use of residues and crescent demand), new techniques have been extensively studied, such as proximal sensing (PS). In the past century, important results have been raised, but the technique has not evolved due to the lack of equipment and strong statistical packages. This restricted the number of researchers from around the world to study the subject. In 1995, Ben-dor and Banin published a pioneering study regarding NIRs (700–1100 nm) to quantify soil attributes. Thus, soil properties prediction increased exponentially as observed in a great review made by Soriano-Disla et al. (2014). The visible (400–700 nm), NIR (700-1100 nm), and SWIR (1100–2500 nm) spectral regions is the main range used, due to their response to numerous soil attributes and the fact that most of the commercial equipment operate in these spectral ranges.

As the use of vis-NIR-SWIR grew (Angelopoulou et al., 2020), other spectral ranges began to be studied. The Middle-infrared (MIR) range responses with higher sensibility to smaller changes in the soil condition due to fundamental vibrations and it improved the predictions for several soil attributes (Hutengs et al., 2019). Several studies indicated that its results are better than vis-NIR-SWIR (Viscarra-Rossel et al., 2006; Soriano-Disla et al., 2014; Terra et al., 2015). An unfavorable point of MIR range is the time-consuming soil sample preparation for the spectral reading (Helfenstein et al., 2021). Also, in recent years studies have emerged in the X-ray range, by the use of X-ray fluorescence technique (XRF), with emphasis on portable instruments (pXRF) (Weindorf et al., 2014). The X-ray energy reaches the atoms of elements present in the soil sample and this interaction at the atomic level emits the fluorescence energy. (Weindorf et al., 2014) This energy is specific for each element and proportional to its content, allowing a measurement of chemical composition of soil sample. (Weindorf et al., 2014) This region has been used to estimate indirectly several soil attributes with accurate results (Stockmann et al., 2016; Zhang and Hartemink, 2019, Tavares et al., 2020).

Many studies have pointed to the use of PS as an alternative because it is a fast, low-cost, environmentally friendly and allows the simultaneous analysis of several elements (Brown et al., 2006; Stenberg et al., 2010; Nocita et al., 2015). In the academic world, the use of spectroscopy applied to soils is already a well-established subject, but not clear on how to put it in practice for commercial approach. For example, what is the population of samples necessary to training the models? The dataset should be from a continent, a country, state, or farm? Regional or local? How can this technique be implemented in a commercial wet laboratory? As the results reached the community, wet laboratories started to be concerned with the technique. Would this technique substitute the wet soil analysis? Will it be the end of the traditional soil laboratories?

A concern of laboratory managers is if spectroscopy makes traditional analyzes obsolete or substituted. On the other hand, Demattê et al. (2019) pointed out, spectroscopy is dependent on traditional methods and the low quality of wet chemistry analysis propagates the error to spectroscopy estimation. Demattê et al. (2019) indicated that the future of soil analysis is the denominated hybrid laboratory. This strategy shows that the sensors will be part of the system and will work in conjunction with wet chemistry. In 2017, FAO created the GLOSOLAN (Global Soil Laboratory Network), the global group of traditional soil laboratories, which have as objectives to stimulate and standardize the use of spectroscopy in soil analysis around the world. Although, there are still important issues to be considered on the communication between final customers (farmers), suppliers (sensor industry), and service providers (wet laboratories). If this communication does not be solved the technology will have difficulties to go forward, despite the advances in the research field.

Thus, this work is an innovative initiative where we bring together several traditional wet soil laboratories to participate in a joint dynamic. Users should compare their wet laboratory analysis with a sensor modeling, while we indicate the advantages and limitations on the reached performance in the role of routine commercial work. We used three proximal sensors, i.e., vis-NIR-SWIR, MIR and pXRF. The process was called the ´Brazilian Program for Soil Analysis by Spectroscopy´ (ProBASE), which aims to give knowledge support to all parts evolved (users, labs, and spectral commercial companies), and indicate usefulness of PS in the traditional laboratory. We hope that laboratories will be able to understand and create their own data set and achieve important results in quantifying soil properties as to go forward on a hybrid laboratory construction.

3.2. MATERIALS AND METHODS

3.2.1. The program and laboratorial analyses

The ProBASE (https://esalqgeocis.wixsite.com/geocis/probase) started in 2018. We contacted 34 commercial laboratories from Brazil and 2 from Paraguay. All of them are in the soil analysis market, have an ISO document of excellence and market their services directly to farmers or agricultural consultants. The number of participating laboratories by Brazilian region were as follows: Minas Gerais (MG) with 12, São Paulo (SP) with 7, Paraná (PR) with 5, Goiás (GO) and Mato Grosso do Sul (MS) with 3, and Bahia (BA), Maranhão (MA), Mato Grosso (MT), Tocantins (TO) with one laboratory each.

The soil samples were analyzed by the participating laboratories using the traditional physical and chemical methods according to Teixeira et al. (2017) and Van Raij et al. (2001). They determined the soil particle size distribution (clay, silt and sand contents) (pipette or densimeter method), soil organic carbon (SOC) (Walkey-Black method with determination by colorimetry), pH in water and the following exchangeable/available elements (macronutrients and soil acidity): Ca^{2+}, Mg^{2+} and Al^{3+} (KCl extraction), P and K^{+} (Mehlich-1 or anion exchange resin) and $H^{+}+Al^{3+($NaOH extraction). Thus, the organic matter content (OM, Eq. 1), sum of bases (SB, Eq. 2), cation exchange capacity (CEC, Eq. 3), base saturation (V%, Eq. 4) and Al saturation (m%, Eq. 5) were calculated.

$$OM= SOC*1.724 \qquad (1),$$

$$SB=Ca^{2+}+Mg^{2+}+K^{+} \qquad (2),$$

$$CEC=SB+H^{+}+Al^{3+} \qquad (3),$$

$$V\%=(SB\div CEC)x100 \qquad (4),$$

$$m\%= (Al^{3+} \div (SB+Al^{3+}))x100 \qquad (5).$$

The following nutrients also were determined: Fe, Mn, Cu, Zn, S and B (DTPA extraction).

Among these 36 laboratories, 35 sent samples that were used in this study. The laboratories were instructed to choose 200 soil samples, based on clay content, in order to obtain samples distributed in several soil texture classes, from their dataset about 7,000 soil samples were selected and sent to the Luiz de Queiroz College of Agriculture, University of São Paulo (ESALQ/USP), in the Geotechnologies in Soil Science Group laboratory (Geocis, https://esalqgeocis.wixsite.com/english). However, the number of samples available for each attribute varied, because there were not all laboratory analyses available for all 7,000 samples. The flowchart with the data description is shown in figure 1.

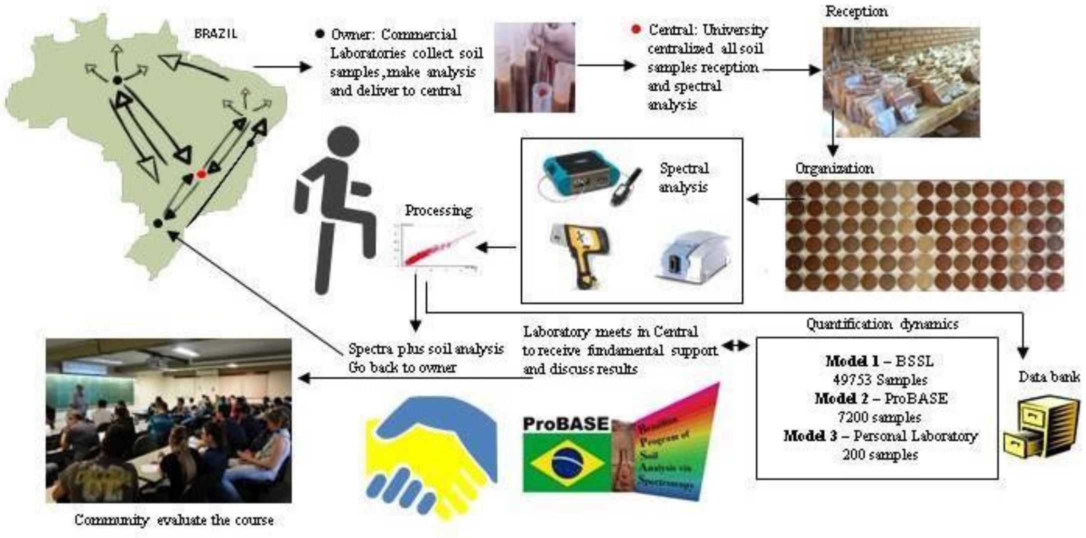

BRAZIL
● Owner: Commercial Laboratories collect soil samples ,make analysis and deliver to central
● Central: University centralized all soil samples reception and spectral analysis
Reception
Organization
Spectral analysis
Processing
Quantification dynamics
Data bank
Model 1 – BSSL
49753 Samples
Model 2 – ProBASE
7200 samples
Model 3 – Personal Laboratory
200 samples
Spectra plus soil analysis
Go back to owner
Laboratory meets in Central to receive fundamental support and discuss results
ProBASE
Brazilian Program of Soil Analysis via Spectroscopy
Community evaluate the course

Figure 1. Flowchart of the exercise with wet laboratories.

3.2.2. Spectral measurements

The soil samples received in ESALQ/USP, were organized per laboratory, and cataloged. All the samples delivered by the 35 laboratories were analyzed in the vis-NIR-SWIR range. However in the MIR range and pXRF, only 50 samples from 8 laboratories were analyzed, in order to compare these three sensors. In this way, a total of 400 soil samples was chosen for pXRX and MIR analysis, maintaining the proportion between the textural classes.

The commercial laboratories provided 50 g of soil samples passed through a 2 mm sieve and submitted to spectral acquisition using a FieldSpec Pro 3 spectroradiometer (Analytical Spectral Devices, Boulder, Colo.) under laboratory conditions. This instrument has a spectral resolution of 1 nm from 350 to 1,000 nm,and of 10 nm from 1,000 to 2500 nm. After, the spectral data were resampled to 1nm by the equipment, totalizing 2151 bands. The data acquisition and the geometry of the setup are described in Demattê et al. (2019). Approximately 95 cm^3 of soil samples were placed inside petri dishes and leveled to reduce surface roughness. Four measurements were obtained per sample by rotating the dish every 90° to obtain a better representativeness of surface area, and an average spectrum was calculated for each sample. A standard Spectralon® white plate was scanned every 20 minutes during calibration.

For MIR analysis, the samples were ground to obtain particles smaller than 100 mesh. Reflectance spectra were obtained with the Alpha Sample Compartment RT-DLaTGS ZnSe (Bruker Optik GmbH) equipped with an accessory for acquiring diffuse reflectance (DRIFT) according to Terra et al. (2015). The sensor has a HeNe laser positioned inside the equipment and a calibration pattern for each wavelength. It has a KBr beam, allowing a high amplitude of the incident radiance to penetrate the sample. Spectra were acquired between 4000 and 400 cm^{-1} (2500 – 25000 nm), with a spectral resolution of 2 cm^{-1} and 32 scans to obtain an average spectra. The reflectance values from 500 to 400 cm^{-1} (20000–25000 nm) were removed due to the low signal to noise ratios. A gold reference plate was used as standard, and the sensor was calibrated every four measurements.

The XRF analysis was carried out in a portable X-ray fluorescence spectrometer (pXRF) Olympus Delta Professional (Olympus Corporation, Waltham, MA, USA), with two excitation modes, according to Rosin et al. (2021). The first excitation mode employed 40 keV, 91.1 µA, and is equipped with a 2 mm aluminum filter and was used to quantify the following elements: vanadium, chromium, iron, cobalt, nickel, copper, zinc, tungsten, mercury, arsenic, lead, bismuth, rubidium, uranium, strontium, zirconium, yttrium, aurum, thorium, niobium The second excitation mode employed 10 keV and 80.5 µA, and was used to quantify the following elements: magnesium, aluminum, silicon, phosphorus, sulfur, chlorine, calcium, titanium, and manganese. About 15 g of sample was placed in a polyethylene bag (20 µm of thickness) and submitted to analysis in a platform with protection for X-ray’s emission. The pXRF Delta Professional is furnished with a 50 keV silver X-ray anode and a silicon drift detector, with 2048 channels. The elements were quantified by the factory calibration called Geochem mode.

3.2.3. Modeling

We first only evaluated the vis-NIR-SWIR range for attributes modelling. This quantification process took three strategies of population set as follows: a) an individual laboratory model to quantify each attribute (Lab dataset, about 200 samples); b) a model using samples of all ProBASE laboratories (ProBASE dataset, about 7,000 samples), c) using the Brazilian Soil Spectral Library database (Demattê et al., 2019), (BSSL database, 49,753 samples). The BSSL details can be seen in Demattê et al. (2019). This strategy was to compare the differences between different populations in dimension and regional occurrence. Complementary, in order to compare the Vis-NIR MIR and XRF for soil attributes prediction, we developed models with these three sensors using only the dataset with 400 selected samples.

The statisticals procedures, such as data split and algorithms for quantification, were the same for all these approaches.

The raw Vis-NIR and MIR spectra and the elemental content from XRF were used as input for the models. The elemental contents below the limit of detection were considered as zero. The data set for each soil attribute was divided into a training and validation set, defined by simple random sampling, where 70% of samples was used for calibration and 30% for validation of models. The Cubist algorithm (Quinlan, 1992) provided in *Cubist* R package (Kuhn and Quinlan, 2021) was chosen for soil attributes prediction, based on previous tests (unpublished results) and literature background.

The Cubist frequently showed better results than other algorithms for use in soil spectroscopy (Chen et al., 2021; Moura-Bueno et al., 2020; Silva et al., 2019). Cubist is a rule-based algorithm that derives from Quinlan's M5 model (Quinlan, 1992). Cubist creates a tree structure with the provided data and collapses paths through the tree to create rules using boosting training (Khaledian and Miller, 2020). Then, differently from other decision trees models, the Cubist uses multiple linear regression models at each node instead of the average (Khaledian and Miller, 2020). The algorithm uses a boosting-like method called committees, which controls the number of model's trees that are sequentially created and the most k common neighbors or instances to build the trees (Silvero et al., 2021; Khaledian and Miller, 2020). We used the default configuration of the Cubist R package for hyperparameters selection. Thus, three possibilities of the number of committees (1, 10 and 20) and neighbors (0, 5, and 9) were tested. The root mean square error (RMSE) of the calibration dataset was used to select the best combination of hyperparameters.

A total of 21 soil attributes , as follows: sand, silt, clay, OM, pH in water, Ca$^{2+,}$ Mg^{2+} K^{+}, Al^{3+}, H^{+}+Al^{3+}, SB, CEC, V%, m%, P, Fe, Mn, Cu, Zn, S and B. The models were evaluated by the coefficient of determination (R^{2}), root mean squared error (RMSE), and ratio of performance to interquartile distance (RPIQ). were evaluated to quantify the inaccuracy of the estimates and quantify the bias. Finally, we compared the results between regional and continental population and sensors efficiency.

3.2.4. Empirical identification of the granulometric analysis through spectral data

We compared the spectral signatures in Vis-NIR of the samples with two standard spectral signatures (one sandy and one clayey), in order to verify if the use of spectroscopy can assist in the identification of errors in soil analysis. The qualitative evaluation of spectra were performed according to Demattê et al. (2014) that considered the reflectance intense in the SWIR region a key to determine samples by texture. In this way, a sample considered as clayey or sandy by wet method and presented the spectral signature discrepant from the standard of these textural classes were considered outlier and reanalyzed in a other wet laboratory for evaluation.

3.2.5. Errors and hits by K and P content ranges

Often the user does not need the exact content of an element but knowing the class that the soil fits in already helps in their activities. Thus we compared the K+ and P contents from wet laboratories with the contents predicted by Vis-NIR, MIR and XRF, using the selected dataset with 400 samples.

The predicted the contents were divided into the following five ranges and compared with the traditional laboratory data also divided into the same ranges. The limits of ranges was defined according to the IAC (1996). In this way, for K, the ranges were divided as very low (0.0-0.7), low (0.8-1.5), medium (1.6-3.0), high (3.1-6.0), very high (>6.0 mmolc.dm^3). For P, the ranges were divided as very low (0-6), low (7-15), medium (16-40), high (41-80), very high (>80 mg.dm^3).

3.2.6. Prediction models for clay testing and the outliers influence

In order to verify the influence of outliers in clay prediction by Vis-NIR, the ProBASE dataset was used to develop models using: a) all dataset, including outliers found by qualitative analysis described in section 2.4 (6595 samples), (b) without the outliers detected (6408 samples), (c) with the outliers checked and reanalysed (6595 samples), (d) using only the outliers (182 samples) and (e) a using only the that was identified as outlier, however after the reanalyze (182 samples). The influence of the outliers was availed by R^{2,} RMSE and RPIQ.

3.3. RESULTS

3.3.1. Laboratories and soil analysis

Regarding spectral models using the vis-NIR-SWIR range (Table 1), we found only 4 laboratories that had R^2 lower than 0.5 and RPIQ than 2 for clay prediction. 25 laboratories reached values of R^2 equal to or greater than 0.70 for this property. 5 laboratories showed R^2 lower than 0.5 for sand prediction. For OM 12 laboratories reached R^2 greater than 0.7 and RMSE lower than 8.1 g.kg^{-1}.

Table 1. Prediction models for clay, sand, and organic matter in general and by laboratory.

	Clay $g.kg^{-1}$						Sand $g.kg^{-1}$						Organic matter $g.kg^{-1}$					
	Training set			Validation set			Training set			Validation set			Training set			Validation set		
	R^2	RMSE	RPIQ	R^2	RMSE	RPIQ	R^2	RMSE	RPIQ	R^2	RMSE	RPIQ	R^2	RMSE	RPIQ	R^2	RMSE	RPIQ
General	0.8	82.2	3.9	0.7	110.1	3.0	0.8	89.6	3.9	0.7	110.1	3.0	0.7	7.9	1.9	0.6	9.5	1.6
1	0.8	81.1	3.6	0.6	146.2	2.5	0.8	90.0	3.8	0.7	155.2	2.5	0.7	9.2	1.7	0.5	10.7	1.9
2	0.9	64.5	5.1	0.7	119.6	3.0	0.9	69.1	5.8	0.6	147.4	2.5	0.8	3.7	3.2	0.4	5.0	1.9
3	0.8	47.6	3.1	0.4	76.1	1.7	0.8	59.8	2.5	0.2	99.9	1.2	0.8	3.6	2.9	0.8	3.7	3.2
4	0.9	46.8	7.5	0.7	87.1	3.3	0.9	57.0	6.1	0.8	87.0	3.7	0.9	2.8	4.1	0.5	4.4	2.2
5	0.9	39.7	4.8	0.7	66.7	2.4	0.9	38.4	5.1	0.7	70.4	2.5	0.7	3.1	2.3	0.4	4.6	1.6
6	0.8	38.9	3.2	0.6	60.4	2.2	0.8	43.7	2.9	0.6	68.2	2.2	0.6	6.1	1.3	0.1	9.5	0.8
7	0.7	77.9	3.7	0.6	90.0	3.2	0.7	92.8	3.8	0.6	112.0	3.1	0.4	7.0	2.0	0.4	7.3	1.9
8	1.0	27.6	4.8	0.9	33.4	3.7	1.0	30.4	4.5	0.9	42.2	4.2	0.9	3.0	4.0	0.7	5.4	2.3
9	1.0	45.9	10.0	0.9	88.2	4.5	1.0	59.3	10.0	0.9	100.7	6.0	0.9	4.6	3.5	0.7	10.9	1.8
10	0.7	55.4	2.6	0.7	83.0	2.3	0.8	61.9	3.3	0.6	90.6	1.9	0.7	3.8	2.7	0.4	5.4	1.7
11	0.9	39.4	8.3	0.9	54.2	6.2	0.9	48.5	8.7	0.9	71.4	6.0	0.9	2.4	4.4	0.8	2.9	3.0
12	0.9	43.3	6.4	0.8	78.1	4.2	0.9	58.9	6.1	0.8	101.9	4.1	0.8	3.7	3.7	0.7	5.3	2.9
13	0.8	63.0	2.9	0.5	100.3	1.9	0.7	81.0	2.5	0.5	122.5	2.0	0.6	2.4	2.1	0.6	3.0	1.6
14	0.9	49.8	4.5	0.7	96.7	2.1	0.9	65.9	3.8	0.7	114.5	2.4	0.8	5.2	1.8	0.1	6.5	1.6
15	0.8	65.7	3.4	0.7	91.8	2.5	0.8	68.4	3.8	0.7	99.7	2.4	0.7	8.0	2.1	0.4	10.5	1.8
16	0.9	59.6	5.4	0.7	96.2	3.3	0.9	77.0	6.9	0.7	132.9	3.7	0.5	5.9	1.9	0.3	7.4	1.5
17	0.9	31.8	3.3	0.8	49.6	2.8	0.9	37.3	3.2	0.7	61.4	2.2	0.8	2.9	2.1	0.5	4.3	1.5
18	1.0	27.2	21.3	0.8	108.8	4.7	1.0	31.4	23.4	0.9	103.8	6.2	1.0	2.2	8.5	0.7	5.0	2.8
19	0.9	37.4	5.7	0.8	58.1	2.5	0.9	70.7	3.7	0.7	99.0	2.4	0.6	4.6	1.9	0.2	5.5	1.6
20	1.0	41.4	9.9	0.9	69.6	6.6	1.0	49.9	10.0	0.9	83.6	6.1	0.9	3.3	5.6	0.8	5.5	3.4
21	0.6	62.3	2.3	0.4	70.3	1.6	0.7	70.5	2.1	0.4	79.3	1.5	0.9	3.5	3.5	0.6	6.0	1.9
22	0.9	42.5	7.6	0.9	59.4	5.3	0.9	68.8	6.5	0.9	91.5	5.4	0.8	3.8	3.0	0.7	4.8	3.0
23	0.9	53.3	5.6	0.8	81.5	4.4	0.9	64.6	6.2	0.8	99.8	4.3	0.9	4.4	5.7	0.6	8.0	2.8
24	0.8	78.1	3.1	0.7	111.3	2.4	0.8	122.4	1.9	0.5	174.9	1.9	0.6	13.4	1.4	0.6	13.7	1.1
25	0.8	55.3	3.8	0.4	86.4	1.7	0.8	77.5	3.9	0.5	115.5	1.8	0.7	5.6	2.4	0.1	7.9	1.4
26	1.0	59.3	7.9	0.8	120.5	3.7	0.9	69.3	7.4	0.8	119.8	4.3	0.9	4.3	4.3	0.7	8.1	2.1
27	0.9	62.2	5.3	0.8	92.0	3.3	0.9	72.2	4.6	0.8	105.3	3.2	0.7	5.1	2.4	0.3	7.5	1.5
28	0.8	77.8	3.1	0.6	102.0	2.4	0.9	115.7	5.2	0.6	166.6	3.5	0.8	5.1	3.4	0.5	6.8	2.0
30	0.8	60.1	2.5	0.5	72.5	2.0	0.7	71.7	2.1	0.4	86.1	1.8	0.7	5.8	2.4	0.3	8.0	1.5
31	0.9	49.3	6.1	0.8	65.0	2.9	0.9	80.8	4.8	0.8	107.9	3.0	0.4	8.9	1.7	0.3	10.9	1.5
32	-	-	-	-	-	-	-	-	-	-	-	-	0.8	7.9	2.7	0.5	11.4	1.9
33	-	-	-	-	-	-	-	-	-	-	-	-	0.8	0.5	2.2	0.3	0.9	1.5
34	0.9	62.8	7.1	0.8	112.0	4.6	0.9	70.7	6.6	0.8	135.5	4.9	0.6	9.7	2.0	0.6	8.7	2.2
35	1.0	33.6	8.8	0.9	47.6	6.7	0.9	58.0	6.6	0.9	65.8	5.9	0.5	4.4	1.9	0.7	4.3	2.6
36	0.9	49.3	6.1	0.8	65.0	2.9	0.9	80.8	4.8	0.8	107.9	3.0	0.4	8.9	1.7	0.3	10.9	1.5

Table 2. Prediction models for pH, P and Al in general and by laboratory.

	pH						P						Al mmolc.kg^{-1}					
	Training set			Validation set			Training set			Validation set			Training set			Validation set		
	R^2	RMSE	RPIQ	R^2	RMSE	RPIQ	R^2	RMSE	RPIQ	R^2	RMSE	RPIQ	R^2	RMSE	RPIQ	R^2	RMSE	RPIQ
General	0.4	0.5	1.7	0.3	0.6	1.5	0.1	52.9	0.4	0.2	31.7	0.6	0.4	4.4	0.5	0.1	4.9	0.4
1	0.6	0.4	1.7	0.4	0.5	1.2	0.8	25.1	2.9	0.2	62.9	0.8	0.8	1.9	0.5	0.2	3.1	0.3
2	0.5	0.4	1.9	0.3	0.5	1.9	0.6	16.1	0.3	0.0	19.7	0.4	0.8	1.5	2.1	0.1	3.2	0.9
3	0.8	0.3	2.4	0.4	0.6	1.4	0.9	9.8	0.4	0.0	36.1	0.2	0.8	3.1	4.5	0.6	4.9	3.1
4	0.7	0.3	2.3	0.3	0.4	1.8	0.7	16.7	1.8	0.3	25.7	1.1	0.6	8.0	0.2	0.1	6.4	0.3
5	0.5	0.4	2.2	0.0	0.4	1.2	0.4	25.5	1.1	0.2	22.6	1.5	0.8	1.4	0.2	0.0	1.1	0.3
6	0.6	0.2	2.0	0.2	0.3	1.1	0.7	25.6	3.1	0.0	34.9	0.7	0.9	0.8	2.4	0.2	1.5	1.4
7	0.5	0.4	1.7	0.3	0.5	1.3	0.6	24.4	1.3	0.3	28.3	1.0	0.5	1.6	0.0	0.1	1.5	0.0
8	0.9	1.0	5.0	0.8	0.1	3.6	0.9	7.1	4.5	0.9	15.4	2.2	0.9	1.4	2.2	0.7	2.5	1.3
9	0.2	0.4	1.6	0.0	1.0	0.7	0.0	45.0	0.3	0.0	37.5	0.4	0.9	3.2	1.7	0.2	5.8	0.9
10	0.0	0.4	1.1	0.0	0.5	1.4	0.5	12.2	1.3	0.3	87.8	0.1	0.7	0.8	0.0	0.0	2.9	0.3
11	0.9	0.2	6.8	0.7	0.3	3.5	0.9	7.6	1.2	0.4	7.8	1.2	0.9	1.1	5.2	0.6	2.4	2.6
12	0.6	0.3	2.1	0.1	0.4	1.3	0.8	7.8	2.2	0.1	13.5	1.2	0.8	1.2	2.1	0.0	1.7	0.0
13	0.4	0.4	1.7	0.4	0.5	1.8	0.9	26.4	0.6	0.1	55.3	0.3	0.6	3.1	1.0	0.1	3.6	1.1
14	0.5	0.3	1.6	0.1	0.5	1.1	0.9	26.3	0.7	0.0	36.1	0.4	0.9	0.9	0.0	0.0	2.2	0.0
15	0.6	0.3	2.1	0.2	0.7	1.6	0.8	10.4	1.2	0.1	16.3	1.1	0.1	2.1	0.9	0.2	14.0	0.2
16	0.3	0.4	1.3	0.1	0.4	1.2	0.5	11.2	1.1	0.1	16.7	1.2	0.6	2.4	0.0	0.0	1.4	0.2
17	0.9	0.2	3.5	0.5	0.4	2.3	0.0	6.6	1.1	0.0	2.1	1.2	0.7	2.3	2.6	0.6	3.4	1.8
18	0.6	0.3	2.0	0.4	0.4	2.1	0.5	8.3	1.7	0.2	7.5	1.1	0.9	1.1	0.6	0.5	1.8	0.4
19	0.4	0.4	2.0	0.1	0.6	1.4	0.5	4.6	1.1	0.0	4.6	1.2	0.6	2.2	1.3	0.1	2.4	1.5
20	0.7	0.4	2.7	0.3	0.7	1.7	0.8	8.9	3.3	0.3	16.5	1.1	0.6	2.1	1.8	0.5	4.0	0.7
21	0.4	0.5	1.6	0.1	0.6	1.2	0.0	49.2	0.4	0.0	40.8	0.3	0.5	2.7	2.0	0.0	4.3	1.1
22	0.4	0.4	1.7	0.1	0.5	1.1	0.0	56.4	0.7	0.0	57.9	0.9	0.0	1.9	0.0	0.0	1.1	0.0
23	0.5	0.3	1.8	0.1	0.5	1.3	0.2	28.9	0.9	0.0	26.9	1.0	0.5	1.1	0.4	0.1	1.7	0.5
24	-	-	-	-	-	-	0.9	214.6	0.0	0.0	32.4	0.1	0.3	12.6	0.9	0.2	24.3	0.4
25	0.4	0.7	1.7	0.1	0.8	1.2	0.8	14.9	1.3	0.2	54.3	0.5	0.3	3.6	1.0	0.1	5.2	1.1
26	0.6	0.3	2.1	0.5	0.3	1.5	0.7	17.8	1.5	0.1	24.8	1.0	0.9	0.9	0.0	0.3	1.4	0.0
27	0.5	0.3	2.0	0.1	0.4	1.3	0.1	27.8	0.4	0.0	14.5	0.7	0.8	1.7	0.6	0.0	0.9	0.4
28	0.7	0.4	2.7	0.4	0.6	1.6	0.8	43.3	1.2	0.2	52.7	0.6	0.6	1.4	0.7	0.1	3.1	0.3
30	0.0	0.6	1.6	0.0	0.6	1.2	0.0	19.1	1.3	0.0	17.7	1.0	0.2	3.5	0.9	0.0	2.8	0.5
31	-	-	-	-	-	-	0.8	42.2	1.2	0.1	65.4	1.1	0.8	2.2	0.9	0.1	2.4	0.4
32	-	-	-	-	-	-	0.6	54.9	0.7	0.0	44.0	0.9	0.8	1.7	0.6	0.0	2.9	0.7
33	0.0	0.5	1.2	0.0	0.6	1.3	0.0	55.5	0.3	0.0	23.0	0.7	0.9	4.7	0.5	0.0	7.0	0.6
34	-	-	-	-	-	-	0.0	38.3	0.5	0.0	31.4	0.6	0.7	1.8	1.1	0.0	1.9	1.0
35	-	-	-	-	-	-	0.0	32.5	0.4	0.0	58.5	0.2	0.5	2.4	1.3	0.1	2.9	1.0
36	-	-	-	-	-	-	0.8	42.2	1.2	0.1	65.4	1.1	0.6	4.5	0.4	0.0	9.0	0.3

Table 3. Prediction models for Ca, Mg and K in general and by laboratory.

	Ca mmolc.kg^{-1}						Mg mmolc.kg^{-1}						K mmolc.kg^{-1}					
	Training set			Validation set			Training set			Validation set			Training set			Validation set		
	R^2	RMSE	RPIQ	R^2	RMSE	RPIQ	R^2	RMSE	RPIQ	R^2	RMSE	RPIQ	R^2	RMSE	RPIQ	R^2	RMSE	RPIQ
General	0.8	21.8	1.2	0.7	32.4	0.8	0.8	6.6	1.2	0.4	11.9	0.7	0.4	1.9	1.3	0.2	2.3	1.0
1	0.9	58.3	4.2	0.7	109.5	2.7	0.9	15.1	4.0	0.7	30.6	2.6	0.8	2.8	2.6	0.4	4.6	1.4
2	0.7	13.4	1.8	0.2	16.4	1.0	0.8	2.3	3.1	0.3	4.0	2.0	0.8	0.6	2.8	0.3	1.1	1.8
3	0.7	16.2	1.3	0.2	18.9	0.8	0.8	3.6	1.8	0.6	6.9	0.5	0.5	1.4	1.0	0.2	1.5	1.0
4	0.5	8.6	1.7	0.1	10.5	1.1	0.5	3.2	1.8	0.1	4.5	1.0	0.1	1.3	1.5	0.0	1.2	1.2
5	0.0	21.4	0.0	0.0	19.4	0.0	0.5	3.3	1.2	0.3	4.0	1.0	0.7	0.6	2.0	0.1	1.0	0.9
6	0.7	4.9	2.1	0.4	7.7	1.6	0.7	3.0	2.4	0.3	4.2	1.5	0.4	1.0	1.0	0.1	1.3	1.5
7	0.4	11.8	1.6	0.4	11.2	1.9	0.7	3.2	2.0	0.4	3.5	1.6	0.4	0.9	1.7	0.5	0.8	2.0
8	0.7	4.7	2.1	0.3	5.8	1.2	0.6	2.3	1.3	0.3	2.6	1.3	0.6	0.5	1.7	0.2	0.8	1.4
9	0.9	10.4	3.9	0.5	20.8	1.7	0.5	6.9	1.4	0.5	4.5	1.8	0.2	3.7	0.9	0.0	4.4	0.5
10	0.7	4.6	2.4	0.3	7.2	1.8	0.6	2.3	2.2	0.2	2.8	1.4	0.0	1.1	1.4	0.0	2.3	0.9
11	0.9	6.0	3.4	0.6	6.9	2.8	0.9	1.6	5.7	0.7	3.0	3.1	0.6	0.6	1.6	0.0	0.7	0.9
12	0.8	5.2	1.9	0.3	8.0	1.5	0.8	1.8	2.2	0.3	2.6	1.7	0.5	0.4	1.6	0.2	0.6	1.0
13	0.9	9.3	1.8	0.8	17.3	1.0	0.5	3.7	1.4	0.2	2.9	1.6	0.5	1.2	1.6	0.0	3.3	0.4
14	0.5	18.9	1.4	0.5	23.8	1.6	0.0	5.6	1.1	0.2	6.9	1.8	0.5	1.6	1.6	0.3	1.9	1.5
15	0.3	20.9	1.0	0.4	21.8	1.4	0.6	6.2	1.3	0.1	9.5	1.4	0.6	2.0	1.2	0.2	2.7	1.0
16	0.6	5.8	2.1	0.3	6.2	1.8	0.7	3.1	2.4	0.3	4.3	1.6	0.5	1.1	1.7	0.1	1.3	1.4
17	0.7	8.1	2.3	0.5	10.8	2.2	0.9	2.0	2.5	0.6	2.9	2.4	0.7	1.0	1.6	0.2	1.6	1.3
18	0.8	8.2	2.6	0.4	13.6	1.8	0.8	2.1	2.7	0.4	3.3	1.7	0.6	1.7	1.7	0.3	1.8	0.5
19	0.6	7.0	2.3	0.1	10.7	1.4	0.8	2.7	2.6	0.5	4.5	1.7	0.7	0.8	1.8	0.2	1.0	1.4
20	0.8	14.4	2.9	0.7	16.1	3.0	0.7	2.0	1.7	0.4	3.2	1.7	0.8	1.1	2.8	0.5	1.8	1.8
21	0.9	14.5	1.1	0.0	43.1	0.5	0.2	3.1	1.5	0.0	3.9	1.4	0.5	1.4	1.7	0.0	2.3	1.1
22	0.6	11.0	1.8	0.3	16.7	1.8	0.4	3.8	1.6	0.4	4.9	1.7	0.5	1.6	1.7	0.0	3.5	1.1
23	0.7	8.7	2.3	0.4	13.2	1.3	0.8	2.9	3.2	0.7	4.8	1.5	0.7	0.7	2.3	0.5	1.0	1.2
24	0.9	15.5	1.2	0.3	31.7	0.6	0.9	5.7	1.3	0.6	9.4	0.9	0.3	1.5	1.0	0.0	5.5	0.3
25	0.4	15.5	1.8	0.1	17.6	1.8	0.6	4.2	1.9	0.2	5.8	1.5	0.6	1.3	1.9	0.1	2.1	1.5
26	0.8	11.2	4.2	0.7	18.3	2.7	0.8	4.3	2.4	0.6	4.5	2.5	0.8	1.2	3.2	0.5	1.9	1.9
27	0.6	36.8	0.7	0.0	38.2	0.6	0.4	11.7	0.8	0.0	10.5	0.8	0.0	2.1	1.3	0.1	1.9	1.1
28	0.6	28.9	1.7	0.1	59.6	0.8	0.9	5.9	4.1	0.6	10.2	1.6	0.8	1.4	4.2	0.6	2.1	1.9
30	0.1	16.5	1.2	0.0	18.8	1.1	0.4	4.1	2.0	0.0	5.3	1.3	0.4	1.1	1.5	0.0	1.7	1.1
31	0.4	4.7	1.5	0.0	14.2	0.5	0.2	3.2	1.3	0.0	4.6	0.7	0.6	1.0	1.0	0.1	1.2	0.7
32	0.6	31.8	0.8	0.0	35.3	0.8	0.8	3.9	2.0	0.2	6.2	1.2	0.6	1.6	1.9	0.3	2.4	1.0
33	0.5	9.6	1.9	0.3	11.3	1.8	0.2	4.6	1.5	0.0	5.1	1.4	0.8	0.2	2.3	0.3	0.3	1.6
34	0.6	24.3	2.0	0.1	30.3	1.3	0.6	7.0	1.8	0.4	7.4	1.6	0.0	3.5	0.9	0.1	2.3	1.1
35	0.3	13.9	1.3	0.1	32.3	0.5	0.3	7.7	1.2	0.1	13.6	0.7	0.1	1.2	1.1	0.1	1.5	0.9
36	0.7	15.7	2.2	0.4	15.8	2.2	0.7	6.8	1.8	0.6	7.9	1.8	0.5	1.5	1.3	0.1	2.1	1.2

Table 4. Prediction models for CEC, base saturation, and Al saturation in general and by laboratory.

	CEC mmolc.kg-1						Base saturation %						Al saturation %					
	Training set			Validation set			Training set			Validation set			Training set			Validation set		
	R^2	RMSE	RPIQ	R^2	RMSE	RPIQ	R^2	RMSE	RPIQ	R^2	RMSE	RPIQ	R^2	RMSE	RPIQ	R^2	RMSE	RPIQ
General	0.8	28.4	1.2	0.8	37.9	0.8	0.5	15.0	1.9	0.3	17.8	1.6	0.5	13.0	0.6	0.2	16.5	0.5
1	0.9	70.2	4.4	0.7	140.9	2.6	0.7	9.3	1.8	0.6	14.3	1.9	0.6	8.4	0.1	0.3	14.7	0.1
2	0.6	20.3	1.4	0.2	19.9	1.1	0.6	15.2	2.4	0.2	23.3	1.9	0.7	9.9	2.3	0.2	17.9	1.0
3	0.9	8.9	2.0	0.4	22.4	0.4	0.9	11.1	4.4	0.5	19.8	2.2	0.9	12.1	5.7	0.7	17.7	3.7
4	0.5	11.3	1.6	0.4	11.2	1.4	0.7	11.3	2.4	0.3	14.1	1.5	0.9	29.0	0.2	0.0	42.2	0.2
5	0.8	6.3	2.6	0.2	12.5	1.4	0.8	7.9	2.9	0.3	10.6	1.2	0.9	6.6	0.4	0.0	8.1	0.2
6	0.6	8.3	1.9	0.4	8.3	1.4	0.7	8.2	2.1	0.3	13.6	1.6	0.8	3.8	0.0	0.2	7.9	1.1
7	0.7	11.3	2.2	0.4	14.3	1.9	0.6	11.1	1.5	0.4	14.9	1.6	0.5	11.2	0.0	0.2	11.8	0.0
8	0.8	6.1	2.0	0.5	7.4	1.3	0.7	6.0	2.7	0.6	7.5	1.9	0.8	5.1	1.6	0.6	7.2	1.2
9	0.8	17.7	2.9	0.6	23.9	2.3	0.6	11.1	2.7	0.3	15.3	1.6	0.6	9.9	1.5	0.2	13.2	0.9
10	0.5	7.9	1.8	0.3	9.9	1.5	0.1	10.4	1.4	0.0	15.7	1.1	0.8	3.1	0.0	0.0	12.3	0.4
11	0.8	7.1	3.4	0.7	7.1	3.3	0.9	6.6	4.9	0.6	12.4	2.7	1.0	5.9	5.8	0.5	19.7	1.8
12	0.8	6.0	2.2	0.3	9.7	1.6	0.6	10.4	1.7	0.0	14.2	1.1	0.9	6.8	1.3	0.0	10.2	0.0
13	0.7	17.9	1.1	0.2	26.1	0.8	0.5	16.3	2.2	0.1	23.9	1.8	0.5	19.1	1.1	0.0	29.6	1.0
14	0.5	21.5	1.7	0.5	29.1	1.8	0.3	14.8	1.4	0.2	16.6	1.9	0.5	9.6	0.0	0.0	11.3	0.0
15	0.3	27.2	1.0	0.1	36.5	1.2	0.3	13.7	1.9	0.1	18.7	1.6	0.7	5.9	0.8	0.1	11.9	0.6
16	0.7	7.4	2.6	0.4	9.1	1.9	-	-	-	-	-	-	0.8	1.5	0.0	0.3	5.9	0.1
17	0.7	9.2	2.0	0.5	11.8	1.9	0.8	6.0	3.8	0.5	10.3	2.2	0.9	3.9	2.8	0.6	6.8	1.7
18	0.8	9.6	2.9	0.5	15.7	1.9	0.5	10.5	1.6	0.2	13.8	1.6	0.8	5.3	0.3	0.1	7.3	0.3
19	0.8	7.7	2.5	0.5	12.0	1.8	0.3	13.5	2.0	0.1	18.1	1.5	0.4	12.8	1.0	0.2	9.7	1.1
20	0.9	11.3	5.2	0.7	16.9	3.2	0.8	10.6	3.4	0.5	14.8	2.5	0.8	9.5	3.0	0.3	17.8	1.6
21	0.6	16.7	1.1	0.0	14.7	1.7	0.3	18.3	1.8	0.0	23.5	1.5	0.3	17.2	1.9	0.1	23.3	1.2
22	0.5	14.9	1.8	0.4	21.4	1.8	0.1	14.7	1.4	0.1	14.6	1.6	0.9	4.0	0.0	0.0	5.5	0.0
23	0.8	10.4	2.9	0.5	17.5	1.5	0.4	11.2	1.6	0.0	15.7	1.5	0.0	7.7	0.2	0.0	11.0	0.3
24	0.8	21.8	1.6	0.7	29.7	1.2	0.8	16.4	3.2	0.3	22.3	2.1	0.6	25.1	2.6	0.3	28.6	2.3
25	0.6	16.3	1.8	0.2	19.1	1.3	0.3	17.9	1.6	0.2	20.7	1.7	0.0	13.5	0.7	0.1	18.0	1.2
26	0.9	12.9	4.0	0.7	21.0	2.7	0.7	8.5	2.6	0.6	10.1	1.9	0.7	4.7	0.0	0.7	3.7	0.0
27	0.7	60.3	0.6	0.0	51.2	0.7	0.0	15.2	1.3	0.0	14.2	1.1	0.7	5.6	0.4	0.0	3.0	0.1
28	0.7	30.0	3.1	0.2	65.8	1.0	0.8	11.9	3.4	0.4	20.0	2.1	0.9	9.0	0.6	0.2	20.4	0.4
30	0.4	16.0	1.5	0.1	21.0	1.2	0.3	16.7	1.8	0.0	20.3	1.1	0.3	11.9	0.9	0.0	12.9	0.3
31	0.5	6.5	1.8	0.1	16.9	0.5	0.3	11.9	1.5	0.0	16.3	1.4	0.8	9.8	0.8	0.2	11.5	0.4
32	0.7	30.7	1.2	0.1	33.7	1.1	0.4	17.0	1.8	0.0	21.7	1.4	0.8	8.3	0.7	0.0	14.3	0.8
33	0.4	10.8	1.8	0.4	10.2	1.9	-	-	-	-	-	-	0.8	10.6	0.5	0.0	15.6	0.5
34	0.6	30.0	1.9	0.3	30.1	1.8	0.1	12.8	1.5	0.0	15.7	1.2	0.4	3.5	1.3	0.0	6.6	0.7
35	0.6	17.1	1.7	0.2	42.0	0.7	0.1	16.4	1.4	0.0	18.6	1.1	0.6	13.2	1.2	0.1	13.0	1.1
36	0.7	20.2	2.1	0.4	24.3	2.0	0.4	17.6	1.8	0.2	16.7	1.4	0.6	16.8	0.5	0.1	18.3	0.6

The pH prediction at vis-NIR-SWIR (Table 2) did not present satisfactory results. Only 2 laboratories achieved R^2 above 0.7 and RPIQ reached 3.5 in the validation model. The general pH model was better than the laboratory ones. For P (Table 2), only one laboratory reached R^2 of 0.9 and RPIQ of 2.2. Most of the laboratories and for the general model, the results were with R^2 below 0.3. For Al, only one laboratory was able to reach R^2 of 0.7, with the majority having R^2 below 0.5.

The general Ca model (Table 3) reached R^2 of 0.7, with low RPIQ (0.8). In addition, three laboratories achieved R^2 of 0.7 and RPIQ above 2.7. However, most laboratories (27) obtained R^2 lower than 0.5. For Mg validation models (table 3) there were 3 laboratories that reached R^2 greater than 0.7, while the general model was 0.4 and RPIQ of 0.7. For K (Table 3), no laboratory achieved R^2 greater than 0.7, only one laboratory achieved R^2 of 0.6 and RPIQ of 1.9, while the general model also had R^2 of 0.2.

For CEC (Table 4) the general model had R^2 of 0.8, but the RPIQ was only 0.8. Four laboratories achieved R^2 greater than 0.7 and RPIQ greater than 2.6. For BS (table 4) the best results achieved were 4 laboratories with R^2 of 0.6 and RPIQ from 1.9. However, most laboratories had R^2 below 0.4. For Al saturation (table 4) 1 laboratory achieved R^2 of 0.7 and RPIQ of 3.7, while most laboratories had R^2 below 0.3.

3.3.2. Soil analysis by different spectral ranges

We analyzed soil properties using the vis-NIR-SWIR and MIR (Figure 2). General models usually reported better results for sand and clay using these three ranges. Results reached R^2 between 0.7 and 0.9 and RPIQ between 2.9 and 4.9 (Figure 2). We obtained important validation results from the ProBASE vis-NIR-SWIR bank (Figure 2a, b) for sand, clay, Ca, base sum, and CEC, all with R^2 above 0.7. For sand and clay the RPIQ was above 3.0. However, Ca, base sum and CEC presented low RPIQ. The other elements showed R^2 below 0.6, so they are not reliable. The validation values using the models generated with the BSSL database were slightly below the models using the ProBASE database (figure 2a, 2b). The difference for sand, clay, and CEC was only 0.1 less for the R^2 values.

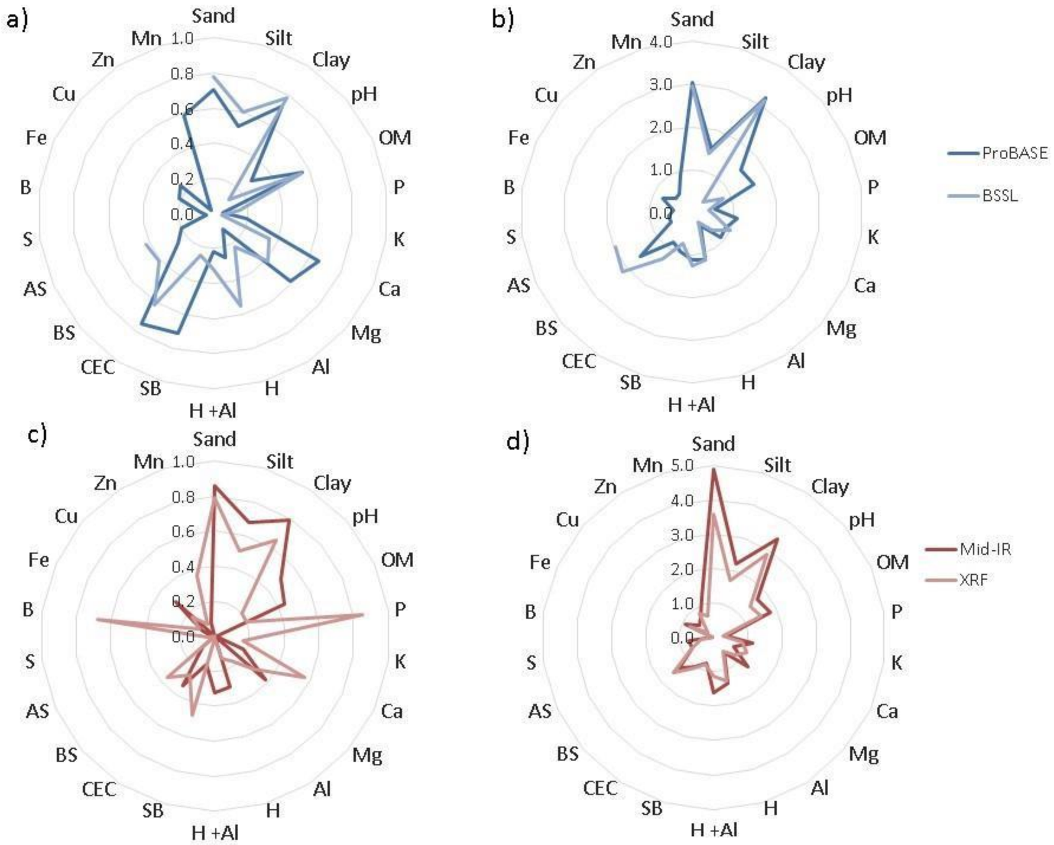


Figure 2. Comparison of ProBASE, BSSL, MIR and XRF models: a) R^2 of the models calibrated with the ProBASE and BSSL database; b) RPIQ of models calibrated with the ProBASE and BSSL database; c) R^2 of the models calibrated with the MIR and XRF database; b) RPIQ of models calibrated with the MIR and XRF database.

3.3.3. Number of laboratories per R^2 range

We analyzed how many laboratories were in each R^2 range for the models using each spectral range (Table 5). The models with the best results were sand and clay for the three spectral bands. For the validation set in the vis-NIR-SWIR range, most laboratories presented R^2 between 0.5 and 0.8 for sand, silt, clay, and OM. For all other elements, R^2 was below 0.5. On the training set, most laboratories had greater R^2 than 0.8 (sand, clay, Cu, and Zn) and between 0.5 and 0.8 for all other elements.

In the MIR range (Table 5), the elements that showed the best results for the training set were sand, silt, clay, and OM. For pH, CEC, and H + Al, most laboratories showed R^2 between 0.5 and 0.8, while for the other elements and laboratories presented R^2 below 0.5. The validation set, in turn, only sand reached R^2 greater than 0.8. Clay and CEC showed R^2 between 0.5 and 0.8, while the other elements were below 0.5.

The XRF range in turn in the training set the elements most laboratories achieved R^2 greater than 0.8 for clay and CEC (Table 5). Most laboratories had R^2 below 0.5 for OM, Ca, H, H + Al, while for pH and base saturation all laboratories had R^2 below 0.5. For the validation set, all laboratories had R^2 below 0.5 for OM, K, Al, and H + Al. Most of them had R^2 below 0.5 for silt, pH, Ca, Mg, H, base sum, CEC, and base saturation. For clay, most laboratories reached R^2 between 0.5 and 0.8, and only for sand had R^2 above 0.8 (Table 5).

Table 5. Percentage of laboratories per R^2 range.

	vis-NIR-SWIR						MIR						XRF					
	Training set			Validation set			Training set			Validation set			Training set			Validation set		
Element	< 0.5	0.5 - 0.8	> 0.8	< 0.5	0.5 - 0.8	> 0.8	< 0.5	0.5 - 0.8	> 0.8	< 0.5	0.5 - 0.8	> 0.8	< 0.5	0.5 - 0.8	> 0.8	< 0.5	0.5 - 0.8	> 0.8
Sand	0	21	79	12	58	30	0	0	100	17	33	50	0	50	50	25	25	50
Silt	15	45	39	42	48	9	0	17	83	33	67	0	25	75	0	75	25	0
Clay	0	18	82	12	48	39	0	17	83	17	50	33	25	25	50	25	50	25
OM	6	53	41	44	41	15	25	25	50	50	25	25	60	20	20	100	0	0
pH	45	41	14	90	7	3	25	50	25	88	13	0	100	0	0	83	0	17
Ca	29	51	20	77	23	0	50	38	13	75	25	0	50	33	17	83	0	17
Mg	31	43	26	71	29	0	50	25	25	75	25	0	40	40	20	80	20	0
K	43	49	9	94	6	0	63	38	0	100	0	0	50	50	0	100	0	0
Al	20	43	37	86	14	0	86	14	0	100	0	0	0	100	0	100	0	0
H	36	36	27	82	15	3	40	40	20	100	0	0	50	33	17	83	17	0
H +Al	30	48	21	91	6	3	25	50	25	88	13	0	50	33	17	100	0	0
SB	17	57	26	80	17	3	38	25	38	63	38	0	40	20	40	60	40	0
CEC	11	66	23	80	20	0	13	50	38	38	50	13	0	33	67	67	17	17
BS	52	39	9	79	21	0	50	25	25	100	0	0	100	0	0	83	17	0
AS	23	46	31	86	14	0	-	-	-	-	-	-	-	-	-	-	-	-
P	34	40	26	97	0	3	-	-	-	-	-	-	-	-	-	-	-	-
Fe	29	42	29	77	23	0	-	-	-	-	-	-	-	-	-	-	-	-
Mn	19	54	27	69	31	0	-	-	-	-	-	-	-	-	-	-	-	-
Cu	13	26	61	61	39	0	-	-	-	-	-	-	-	-	-	-	-	-
Zn	32	26	42	84	16	0	-	-	-	-	-	-	-	-	-	-	-	-
S	42	46	12	88	12	0	-	-	-	-	-	-	-	-	-	-	-	-
B	56	30	15	96	0	4	-	-	-	-	-	-	-	-	-	-	-	-

3.3.4. Spectroscopy on the evaluation of wet laboratory quality analysis

Based on the ProBASE dataset, we found 185 soil samples with high RMSE. These were considered as outliers or suspect and could be: a) real information or b) laboratory wet analysis error. For this, we first plotted the different texture patterns extracted from the basic dataset (Fig. 3 a). These spectra present the major spectral signature (reflectance intensity, shapes, and behavior), strongly studied in Brazil (Demattê et al., 2019). After we made the average of the spectra for clay content classes using the 185 samples (Fig. 3 b). Finally, the 185 soil samples were remade in the wet analysis and then we plotted again the textures (Fig. 3 c). The comparison of the spectral pattern of the suspect samples was not consistent with the patterns. The spectral mean of the samples classified as sand were the only ones that presented a coherent result, as they showed greater reflectance. However, the clay samples showed high reflectance which agrees with the sandy spectra. The texture class that showed the lowest reflectance was the loamy sand, inconsistent with literature and to the patterns (Fig 3 a). In turn, the samples considered suspect outliers when reclassified according to the granulometric reanalysis (Fig 3 c) presented a gradual decrease corresponding to the increase in the clay content and decrease in the sand, confirming that the initial analyzes were wrong and the use of spectroscopy pointed out the issue. In the case of the average spectrum of the textural classes of the ProBASE data set, it is possible to verify that when the database is large, these errors are diluted (the ProBASE data set has 6595 granulometric analysis), so the 185 samples with incorrect granulometric analysis were not enough to alter the result of the medium spectra.

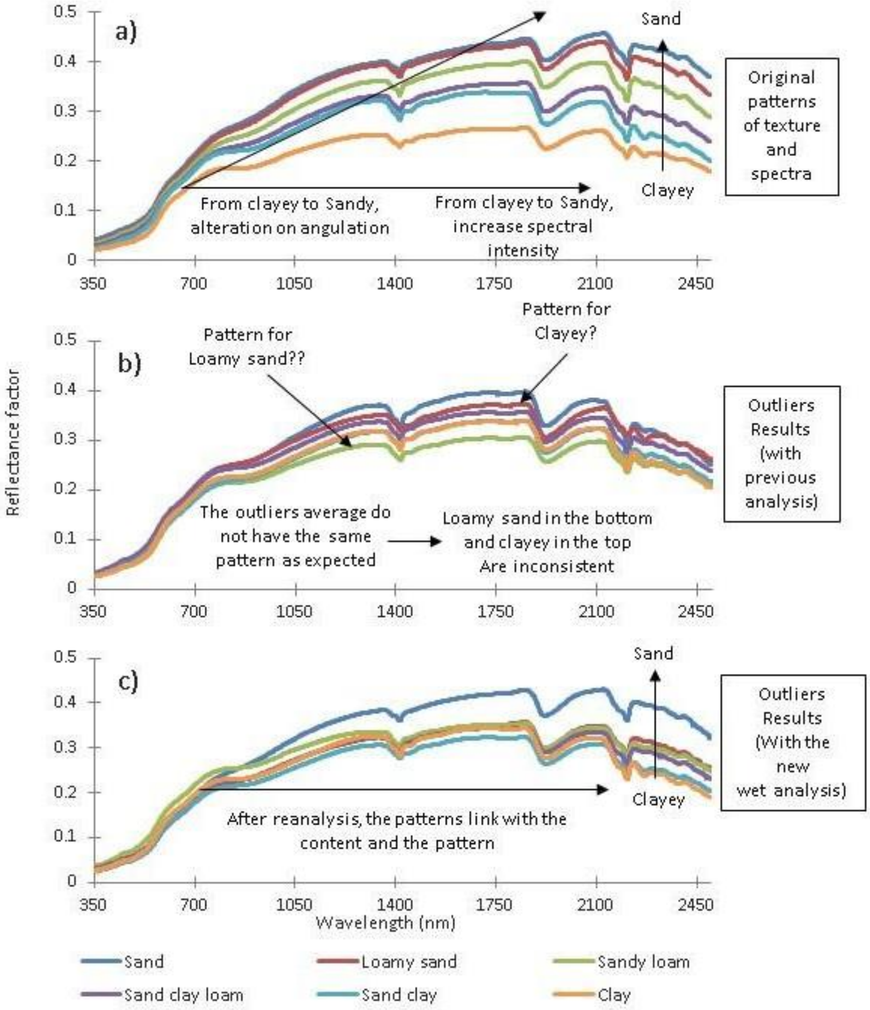


Figure 3. Average spectra by textural class. a) ProBASE data set, b) 185 outlier samples with original analysis, c) 185 outlier samples with reanalysis.

Figure 4 shows four sample cases in which the spectral signature analysis allowed the identification of errors in the quantification of sand and clay. Sample (a) (Fig 4a) was classified as clayey (607 g.kg^{-1} of clay) by the laboratory wet analysis. However, analyzing the spectral signature, we observed high reflectance across the spectral range of vis-NIR-SWIR, which is characteristic of sandy soils. In the same graph spectral signatures patterns of sandy soil (in red) and clay soil (in green) are plotted with the evaluated spectra (in blue). As we observe, the specific sample has a signature typical of sandy soil. We reanalyzed the soil sample in a wet laboratory, and the new analysis showed that, indeed, it had 106 g.kg^{-1} confirming the visual analysis of the spectrum.

The same happened with sample (b) (fig. 4b) where the initial analysis indicated a content of 673 g.kg^{-1} of clay, while the spectrum does not match with the pattern. Reanalysis determined a content of 213 g.kg^{-1} of clay, confirming that the initial analysis was wrong. In this way, we were able to confirm that the spectral analysis can assist in the identification of sand and clay quantification errors or be used as a quality method.

In turn, samples c and d (fig 4c and 4d, respectively) show low reflectance across the spectral range of vis-NIR-SWIR, so it has characteristics of clay soil, but the initial analyzes determined clay contents of 92 g.kg^{-1} for sample c and 175 g.kg^{-1} for sample d. The results of the reanalysis coincided with the spectral characteristics of both samples, 625 g.kg^{-1} (sample c) and 569 g.kg^{-1} of clay (sample d).

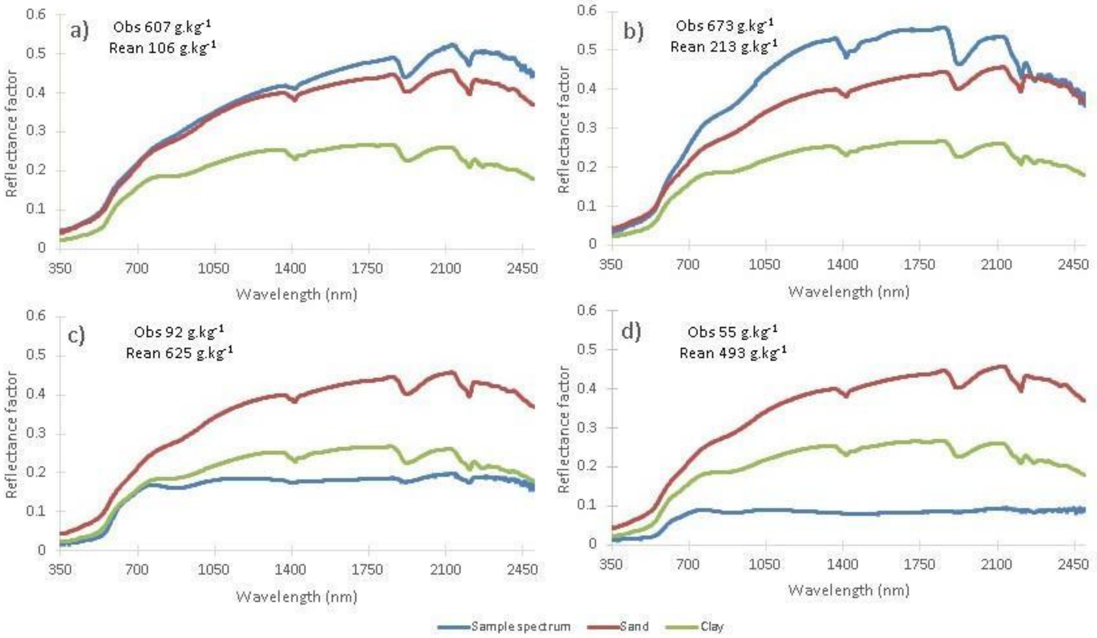


Figure 4. Four examples of case study indicating the detection of wet soil analysis error on clay content. Blue spectra are the original spectra of the soil sample. The green and red lines are the average patterns from the ProBASE dataset for clayey and sandy samples. Each figure compares the measurement with the patterns, as indicated by the previous analysis (detected as an outlier) with the reanalysis.

3.3.5. Impact of quality analysis in spectroscopic models

The model with only outlier-reanalysis (Table 6) was the one that presented the best results both the training (R^2 of 0.96, RMSE of 39.4 and RPIQ of 7.1) and validation set (R^2 of 0.89, RMSE of 72.5 and RPIQ of 4.8). The worst result was presented by the model with the original outliers, which ratifies the laboratory´s first wet analysis as error. Thus, going back to the original models, without the outliers and with the outliers-reanalysis, both increased the results.

Table 6. Prediction models for clay testing the influence of outliers.

Clay g.kg^{-1}	Descriptive Analysis				Observations		Training set				Validation set		
	Mean	SV	Min	Max	Total	Train.	Val.	R^2	RMSE	RPIQ	R^2	RMSE	RPIQ
General model with outliers	394.3	195.2	2.0	950.0	6595	4616	1979	0.79	89.59	3.89	0.69	110.08	3.02
General model without outliers	393.5	194.2	18.0	950.0	6408	4487	1921	0.82	82.20	3.78	0.76	76.07	3.43
General model with reanalysis	394.7	194.5	18.0	950.0	6595	4616	1979	0.81	84.69	3.57	0.74	100.83	3.24
Model with only outliers	423.7	226.6	2.0	925.0	182	127	55	0.45	178.19	2.26	0.01	217.87	1.42
Model with outlier’s reanalysis	439.5	200.3	58.0	838.0	182	127	55	0.96	39.43	7.13	0.89	72.49	4.78

3.3.6. Application on fertilization

We evaluated the application approach for fertilization. For K, the spectral range showed more correct hits for vis-NIR-SWIR, while for P, the XRF was better and very close to MIR (Table 7). Regarding the K contents, the range with medium contents presented more hights than wrongs in the three different spectral ranges studied. And the high range had more hits than wrongs only in the vis-NIR-SWIR range. For P, the low range had more hights than wrongs using the bands of vis-NIR-SWIR and XRF. While the range with more accuracy than wrongs in the MIR spectral range was the range with medium content. In general, samples that were not classified in the correct range of contents were in ranges close to below or just above the correct content.

Table 7. Wrongs and hights by K and P content ranges, comparing the traditional soil analysis and the spectral model.

Element	Range	vis-NIR-SWIR		MIR		XRF	
		Wrongs	Hights	Wrongs	Hights	Wrongs	Hights
K	Very low	734	368	62	9	66	10
	Low	935	850	48	30	39	39
	Medium	759	1482	33	61	42	54
	High	743	747	50	24	53	22
	Very high	364	161	20	4	22	3
P	Very low	1382	584	52	14	44	21
	Low	888	1011	51	46	47	51
	Medium	798	772	22	49	34	41
	High	530	36	29	7	26	21
	Very high	292	5	19	11	19	11

3.4. DISCUSSION

3.4.1. Laboratories and soil analysis

The BSSL models gave slightly better results than the general ProBASE data. The best results were from some specific laboratories (8, 11, 20, 22 and 35). These results are due to the quality of traditional laboratory analysis. A good laboratory analysis results in good spectral models. Demattê et al. (2019) indicates that the results of determination of attributes through chemical analyzes with low quality led to high errors in the spectral models. Therefore, high R^2 and RPIQ and low RMSE of these five laboratories may indicate high quality in their chemical analyzes. There is a lack of works that verify the impact of analyzes carried out by different laboratories in the construction of prediction models from soil spectra. The differences of results between laboratories can be related with several factors such as, the quality of its analysis, the geology, climate, and biome.

Best results were for clay, sand, CEC, and OM, which agrees with literature. On the other hand, chemistry is mostly indicated as to reach low data, but we see good results for Aluminum. From the three great macro nutrients, we saw the sequence Ca, Mg, K where decrease the results. pH is repeating the poor results with literature. Thus, we see that, despite chemistry being difficult to reach, there is light to continue in these elements. Results indicate that their success is related to good soil analysis and the population and/or region. Also, it was observed that some laboratories presented great results for macro elements as others poor. This is also a clear indication regarding the importance of local model population and soil analysis quality that interfere in the model.

3.4.2. Soil analysis by different spectral ranges

Regarding the models using different spectral ranges for sand and clay prediction, the MIR range (fig. 2 c and d) presented slightly better results than the vis-NIR-SWIR and XRF in agreement with Gholizadeh et al. (2013). Terra et al. (2015) also modeled various soil elements using the bands of vis-NIR-SWIR and MIR and obtained the best modeling results for sand and clay (R^2> 0.85 and RPIQ> 3.88), where this range was the best. This is due to the strong interaction between the MIR radiation and the mineral particles in the soil that make up the sand and clay fractions by the fundamental vibration process, producing more spectral characteristics when compared to vis-NIR-SWIR (Janik et al., 2007).

For OM prediction, the vis-NIR-SWIR range showed the best results (Fig. 2a and b). In general, studies show that the MIR range presents better results for predicting OM (Janik et al., 2007; Terra et al. 2015), as the absorption resources associated with various organic functional groups can be identified. The reasons why we achieved different results may be related to the size and variability of the dataset. The vis-NIR-SWIR model was built with 6962 samples, while the MIR was built with 340 samples. And for P prediction, the best results were obtained using the XRF range (fig. 2c and d) with R^2 >0.8, results like that obtained by Kaniu et al. (2012) who obtained an R^2 >0.9 for P analysis using an XRF spectrometer. In terms of practicality, vis-NIR-SWIR is easier and quicker but fails in the best R^2 for most elements. MIR is great due to the fundamental bands but fails on velocity and sample preparation. XRF is not so fast but gives the total element detailed peaks. In summary, we cannot say one is better than the other, but the choice is related to the user´s necessity or having all one can complement the information of the other.

3.4.3. Detection of inconsistency on wet soil analysis by spectroscopy patterns

The medium spectrum of the textural classes of the ProBASE dataset (Fig 3a) presented a gradual decrease as the content of clay increased and the sand decreased. This characteristic was well demonstrated by several qualitative analysis studies of the spectra (Sorensen and Dalsgaard, 2005; Demattê and Terra, 2014). When we plot the samples of the database sent by all laboratories, we observed that some spectra were not consistent with the clay content. The samples with greater error were taken out and compared with the ProBASE pattern. The qualitative spectral analysis showed a completely different pattern from the laboratory analysis. As the soil sample was reanalyzed, the value of clay changed and matched with spectral pattern. These findings show that many of the current works creating spectroscopy models can be compromised. It should have to have a controlling factor before going into the model. This can explain lots of low values in models or the great variances detected in reviews such as in Soriano Disla et al. (2014) and Nocita et al. (2015).

3.4.4. Impact of quality analysis in spectroscopic models

Using the whole population for clay estimation using the vis-NIR-SWIR, we reached a 0.69 R^2. The visual analysis plus the error indicated suspect samples which did not match with the pattern of the determined clay content. While we took out these outliers R^2 went up to 0.76 (what was expected!). Only using the outliers, we had 0.01. After we reanalyzed the clay content of these outliers and inserted again in the complete population and reached 0.89. This indicates that the first model was incorrect due to laboratory errors. Thus, it must be taken in consideration that results vary according to soil wet source. Indeed Cantarella et al. (2006) mentioned soil analysis variation in commercial laboratories. In fact, even elements with worse R^2 can have the same issue and must be addressed by future works.

3.4.5. Application on fertilization

Results for potassium and phosphorus classification vary. The best class reached by all sensors was the medium. All the others had about 50% wrongs and hights. These results indicate that not only the number of the spectral soil analysis is inadequate but also the class in which it reaches. For users in the field, the operational costs are extremely high with fertilizers and confidence is important. On the other hand, works such as Demattê et al. (2019) indicate that 4 laboratories reached completely different results regarding some chemical analysis, i.e., for phosphorus and potassium. If results from laboratories lie on variances, can explain the poor results in spectroscopy and thus on the classification of its content. The strategy to classify a given element inserting in a classification can led to wrongs and thus, caution is important. The main point is to reach the best fixed result and not try to insert into a simple class. These results agree with Jin et al. (2020), where the authors studied various treatments and processing and found pretty good results for quantification. On the other hand, they state that *´a trustworthy model that can be used across industries for NIR quantification is difficult to build´*. For instance, this is possible in research in controlled situations, but not for a whole community.

3.4.6. Users experience in the dynamics

At the end of the modelling results were presented to participants, to see the results of their wet-laboratory and spectral readings. An example is shown in table 5. We observe the number of the laboratory and its result on R^2 on the quantification of the indicated soil component. With this, users could see if their soil wet analysis were good or not. Results vary from lab to lab, but for certain elements they are all constant. For sand and clay almost, all laboratories remain over 0.7 of R^2. Laboratories under 0.5 could need reanalysis and evaluate their results.

Finally, users responded to topics regarding their feeling of the program dynamics (Figure 5). Observe that before the course, 47% have knowledge over note 5, and only 13% over 8. After the course, 62% reached over 8 and all participants over 5. In question 3, 93% considered it important to understand the dynamics of soil spectroscopy in a wet laboratory. These results indicate that the community is not well prepared to receive nor accept the use of the technique. This is only possible when users understand what they are using, advantages and limitations. Wet soil analysis laboratories are the strongest community that link with the agriculture users. Overcome this community and go direct to farmers is not a good strategy to put in practice spectroscopy. It is necessary to first, bring together this community, show how dynamics are with sensors and create the hybrid laboratory. This will go forward and pave the bridge until farmers. In the next step, sensor based commercial communities will have more strength to go by themselves and insert the technique.

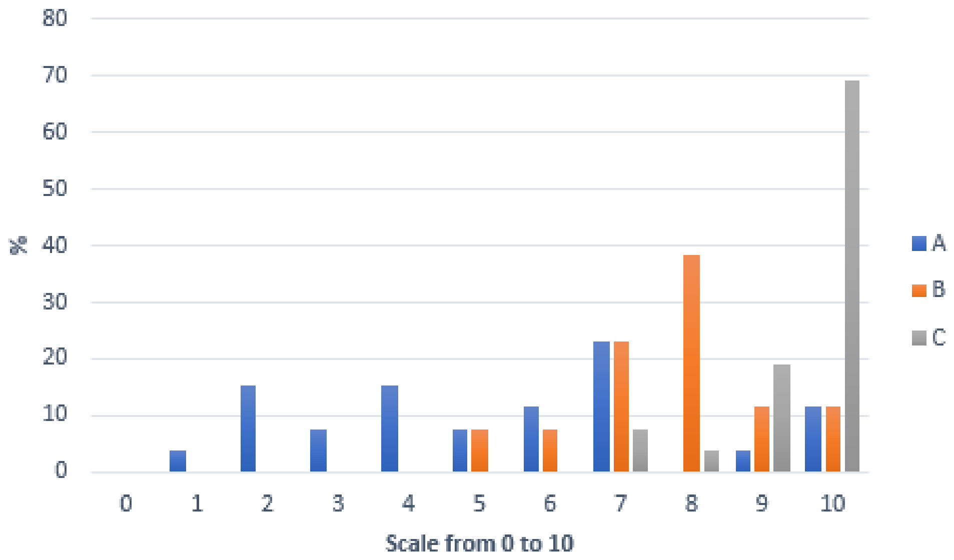


Figure 5. Questionnaire regarding the program realized by the participants. The questions answered are the following: A) What was your level of knowledge about the use of a sensor in soil analysis?; B) How far do you think you have reached in understanding the topic?; C) How important is ProBASE for wet soil analysis laboratories to understand advantages and limitations in the use of sensors?

3.5. CONCLUSION AND FINAL CONSIDERATION

Models vary on quantification of soil properties regarding the population of samples. In general, local samples reached better results than general ones. The best and consistent results were for clay, sand, organic matter, and cation exchangeable capacity. For chemical elements the best was for calcium, magnesium and finally potassium. The results for these elements varied according to each laboratory going from great to poor results. Micronutrients had very poor results. The best ones were for Cu, Mn and Fe reaching about 30% of laboratories with R^2 between 0.5 to 0.8.

Each wet laboratory has its dynamics in which errors are possible and impact directly on spectral analysis. Spectral patterns were able to identify wet laboratory errors regarding granulometric data. As this was corrected, models became better. This indicates the impact of the wet analysis quality on the spectral models. Thus, spectroscopy is not only indicated to quantify an element, but to evaluate a soil analysis.

There is not the best spectral range. Vis-NIR-SWIR, MIR and XRF each have advantages and limitations, where the user is the one to choose which can solve its objective. Although, it is observed that all spectral ranges give important directions on soil analysis, and if possible, should be used together.

The application of models to reach fertilizer (potassium and phosphorus in this case) was not suitable and does not give confidence to deliver for users. More research should be addressed to make this for wider application.

Models vary on results from better to lower on the sequence, local to regional to country. Spectroscopy method cannot substitute soil wet laboratory analysis, but can give important clues on the process, i.e., previous analysis and soil quality control. With this, the sample density can increase as the technique brings confidence to users. On the other hand, users still do not understand the technique, and thus it is important to create courses where they understand the advantages and limitations of the technique. This is the best path to put to work this great technique.

ACKNOWLEDGEMENTS

We are grateful to Geotechnologies in Soil Science Group (GeoCiS/ESALQ-USP; http://esalqgeocis.wixsite.com/english) for team support. This work was supported by the São Paulo Research Foundation (FAPESP) (grant numbers 2014/22262-0, 2016/26176-6, and 2020/04306-0). We thank the laboratories participating in ProBASE: Sial Solo Análises Laboratoriais, Solocria Laboratório Agropecuário, Laboratório Terra Brasileira, Laboratório Exata, Safrar Análises Agrícolas, Timac Agro Paraguay, Laboratório Agropecuário Agronômico, Laboratório Solos e Plantas, Laboratório Agroanálise, Laboratório de análise de solos e folha do Instituto Federal do Sudeste de MG, Labras Ambientais e Agrícolas, Laboratório 3Rlab, Laboratório GEAAP, Laboratório Agronômico, DMLab, Solos Laboratório de Análise, Laboratório Água Limpa, Ribersolo, Agrolab Minas, Laboratório Santa Rita, Laboratório Agrolab, Laboratório Agrozap.

REFERENCES

Angelopoulou, T., Balafoutis, A., Zalidis, G., Bochtis, D., 2020. From Laboratory to Proximal Sensing Spectroscopy for Soil Organic Carbon Estimation—A Review. Sustainability. 12(2), 443. https://doi.org/10.3390/su12020443

Ben-Dor, E., Banin, A., 1995. Near infrared analysis (NIRA) as a method to simultaneously evaluate spectral featureless constituents in soils. Soil Science. 159(4), 259-270.

Brown, D.J., Shepherd, K.D., Walsh, M.G., Mays, M.D., Reinsch, T.G., 2006. Global soil characterization with VNIR diffuse reflectance spectroscopy. Geoderma. 132(3-4), 273-290. https://doi.org/10.1016/j.geoderma.2005.04.025

Cantarella, H., Quaggio, J.A, van Raij, B., Abreu, M.F., 2006. Variability of soil analysis of commercial laboratories: implications for lime and fertilizer recommendations. Commun. Soil Sci. Plant Analysis. 37, 2213–2225.

Chen, S., Xu, H., Xu, D., Ji, W., Li, S., Yang, M., Hu, B., Zhou, Y., Wang, N., Arrouays, D., Shi, Z., 2021. Evaluating validation strategies on the performance of soil property prediction from regional to continental spectral data. Geoderma. 400, 115159. https://doi.org/10.1016/j.geoderma.2021.115159

Demattê, J.A.M., Terra, F.S., 2014. Spectral pedology: a new perspective on evaluation of soils along pedogenetic alterations. Geoderma. 217, 190-200. https://doi.org/10.1016/j.geoderma.2013.11.012

Demattê, J.A.M., Bellinaso, H., Romero, D.J., Fongaro, C.T., 2014. Morphological Interpretation of Reflectance Spectrum (MIRS) using libraries looking towards soil classification. Scientia Agricola, 71(6), 509-520. https://doi.org/10.1590/0103-9016-2013-0365

Demattê, J.A.M., Bellinaso, H., Araújo, S.R., Rizzo, R., Souza, A.B., 2016. Spectral regionalization of tropical soils in the estimation of soil attributes. Revista Ciência Agronômica. 47, 589-598. https://doi.org/10.5935/1806-6690.20160071

Demattê, J.A.M., Ramirez-Lopez, L., Rizzo, R., Nanni, M.R., Fiorio, P.R., Fongaro, C.T., Medeiros Neto, L.G., Safanelli, J.L., Barro, P.P.S., 2016. Remote sensing from ground to space platforms associated with terrain attributes as a hybrid strategy on the development of a pedological map. Remote Sensing. 8(10), 826. https://doi.org/10.3390/rs8100826

Demattê, J.A.M., Dotto, A.C., Bedin, L.G., Sayão, V.M., Souza, A.B., 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

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

Gholizadeh, A., Borůvka, L., Saberioon, M., Vašát, R., 2013. Visible, near-infrared, and middle-infrared spectroscopy applications for soil assessment with emphasis on soil organic matter content and quality: State-of-the-art and key issues. Applied spectroscopy. 67(12), 1349-1362. https://doi.org/10.1366/13-07288

Helfenstein, A., Baumann, P., Viscarra-Rossel, R., Gubler, A., Oechslin, S, Six, J., 2021. Quantifying soil carbon in temperate peatlands using a mid-IR soil spectral library. SOIL. 7, 193–215, https://doi.org/10.5194/soil-7-193-2021, 2021.

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

IBGE. Monitoramento da cobertura e uso da terra do Brasil. URL: https://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=2101703. (accessed:09.10.21).

Janik, L.J., Skjemstad, J.O., Shepherd, K.D., Spouncer, L.R., 2007. The prediction of soil carbon fractions using mid-infrared-partial least square analysis. Soil Research. 45(2), 73-81. https://doi.org/10.1071/SR06083

Jin, X., Li, S., Zhang, W., Zhu, J., Sun, J., 2020. Prediction of soil-available potassium content with visible near-infrared ray spectroscopy of different pretreatment transformations by the boosting algorithms. Applied Sciences. 10(4), 1520. https://doi.org/10.3390/app10041520

Khaledian, Y., Miller, B.A., 2020. Selecting appropriate machine learning methods for digital soil mapping. Applied Mathematical Modelling, 81, 401-418. https://doi.org/10.1016/j.apm.2019.12.016

Kaniu, M.I., Angeyo, K.H., Mwala, A.K., Mwangi, F.K., 2012. Energy dispersive X-ray fluorescence and scattering assessment of soil quality via partial least squares and artificial neural networks analytical modeling approaches. Talanta. 98, 236-240. https://doi.org/10.1016/j.talanta.2012.06.081

Kuhn, M., Quinlan, J,R., (John R., 2021. Cubist: Rule- And Instance-Based Regression

Modeling. R package version 0.3.0.

Molin, J.P., Tavares, T.R., 2019. Sensor systems for mapping soil fertility attributes: Challenges, advances, and perspectives in brazilian tropical soils. Engenharia Agrícola. 39, 126-147. http://dx.doi.org/10.1590/1809-4430-Eng.Agric.v39nep126-147/2019

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

Nocita, M., Stevens A., van Wesemael, B., Aitkenhead, M., Bachmann M., Barthès, B., Ben-Dor, E., Brown, D.J., Clairotte, M., Csorba, A., Dardenne, P., Demattê, J.A.M, Genot, V., Guerrero, C., Knadel,, M., Montanarella, L., Noon, C., Ramirez-Lopez, L., Robertson, J., Sakai, H., Soriano-Disla, J.M., Shepherd, K.D., Stenberg, B., Towett, E.K., Vargas, R., Wetterlind, J., 2015. Soil spectroscopy: An alternative to wet chemistry for soil monitoring. In Advances in agronomy. 132, 139-159. https://doi.org/10.1016/bs.agron.2015.02.002

Quinlan, J.R. (John R., 1992. Learning with continuous classes, in: 5th Australian Joint Conference on Artificial Intelligence. pp. 343–348. https://doi.org/10.1142/9789814536271

Rosin, N.A., Demattê, J.A.M.., Leite, M.C.A., Carvalho, H.W.P., Costa, A.C., Greschuk, L T., Curi, N., Silva, S.H.G., 2021. The fundamental of the effects of water, organic matter, and iron forms on the pXRF information in soil analyses.CATENA, 105868. https://doi.org/10.1016/j.catena.2021.105868

Silva, E.B., Giasson, E., Dotto, A.C., Caten, A.T., Demattê, J.A.M., Bacic, I.L.Z., Veiga, M.D., 2019. A regional legacy soil dataset for prediction of sand and clay content with VIS-NIR-SWIR, in southern Brazil. Revista Brasileira de Ciência Do Solo. 43, e0180174. https://doi.org/10.1590/18069657rbcs20180174

Silvero, N.E.Q., Demattê, J.A.M., Amorim, M.T A., Santos, N.V., Rizzo, R., Safanelli, J.L., Poppiel, R.R., Mendes, W.S., Bonfatti, B.R., 2021. Soil variability and quantification based on Sentinel-2 and Landsat-8 bare soil images: A comparison. Remote Sensing of Environment, 252, 112117. https://doi.org/10.1016/j.rse.2020.112117

Sorensen, L.K., Dalsgaard, S., 2005. Determination of clay and other soil properties by near infrared spectroscopy. Soil Science Society of America Journal. 69(1), 159-167. https://doi.org/10.2136/sssaj2005.0159

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

Stenberg, B., Viscarra-Rossel R.A., Mouazen, A.M., Wetterlind, J., 2010. Visible and near infrared spectroscopy in soil science. Advances in agronomy. 107, 163-215. https://doi.org/10.1016/S0065-2113(10)07005-7

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

Tavares, T.R., Molin, J.P., Nunes, L.C., Alves, E.E.N., Melquiades, F.L., Carvalho, H.W.P., Mouazen, A.M., 2020. Effect of X-ray tube configuration on measurement of key soil fertility attributes with XRF. Remote Sensing. 12(6), 963. https://doi.org/10.3390/rs12060963

Teixeira, P.C., Donagemma, G.K., Fontana, A., Teixeira, W.G., 2017. Manual de métodos de análise de solo, 3a edição. ed. Embrapa Solos, Brasilia, DF.

Terra, F.S., Demattê, J.A.M, Viscarra-Rossel, R.A., 2015. Spectral libraries for quantitative analyses of tropical Brazilian soils: Comparing vis–NIR and mid-IR reflectance data. Geoderma. 255, 81-93. https://doi.org/10.1016/j.geoderma.2015.04.017

Viscarra-Rossel, R.A., Walvoort, D.J.J., McBratney, A.B., Janik, L.J., Skjemstad, J.O., 2006. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma. 131(1-2), 59-75. https://doi.org/10.1016/j.geoderma.2005.03.007

Weindorf, D. C., Bakr, N., Zhu, Y., 2014. Advances in portable X-ray fluorescence (PXRF) for environmental, pedological, and agronomic applications. Advances in agronomy, 128, 1-45. https://doi.org/10.1016/B978-0-12-802139-2.00001-9

Zhang, Y., Hartemink, A.E., 2019. Soil horizon delineation using vis-NIR and pXRF data. Catena. 180, 298-308. https://doi.org/10.1016/j.catena.2019.05.001

4. BRIDGING THE GAP BETWEEN SOIL SPECTROSCOPY AND TRADITIONAL LABORATORY: INSIGHTS FOR ROUTINE IMPLEMENTATION

ABSTRACT

We need innovative solutions to help laboratories to rapidly characterize the soils and adopt internal quality controls with lower costs and ecosystem impacts. Soil spectroscopy emerged as an alternative to wet chemistry. However, soil laboratories still do not widely use this technology in their routines, mainly due to the lack of: i) standards and protocols, ii) spectral libraries, iii) capacity in spectral methods, and iv) professionals with expertise in chemometrics. Therefore, we aimed to provide basic guidelines for the use of soil spectra in laboratory routines regarding internal quality control and samples selection for analysis and prediction. We used 350-2500nm spectra and unsupervised random forests to compute proximity matrices and cluster spectrally similar soil samples for outlier detection using an adjusted method for skewed distribution, which allowed us to establish an internal quality control. Then, we used the most important wavelengths from unsupervised random forest to order soil samples by laboratory and subset them into different set sizes for training and testing prediction models for clay, sand, organic matter cation exchange capacity and base saturation. Our results indicated that internal quality control based on soil spectra and unsupervised analysis can be implemented for laboratory routines. Spectra-based soil samples selection and attributes predictions can reduce the need for laboratory analysis by half. Soil spectroscopy can become an important alternative to improve traditional analytical results, save time, reduce costs, and mitigate environmental impact with acceptable accuracy.

Keywords: Soil Science; Pedometrics; Unsupervised learning; Laboratory routines.

4.1. INTRODUCTION

The global soil resources, which produce 95% of all food, are under great stress (Amundson et al., 2015). They are essential to achieve land-related Sustainable Development Goals (SDG), including food (SDG 2), water (SDG 6), climate (SDG 13) and biodiversity (SDG 15) (Bouma, 2020). Reaching these goals requires soil mapping and land use planning, which exponentially increases the number of soil analyses performed every year. However, limited funding or laboratory resources in many countries cause lack of soil data, resulting in poor modelling and significant uncertainty (Wadoux et al., 2021). This situation limits soil management and policy having a direct impact on the SDG.

Traditional analytical methods, which provide valuable soil data over time, are costly, time-consuming, and not environmentally friendly (Viscarra Rossel and McBratney, 1998). The quality of results from traditional laboratories is susceptible to systematic and random errors that frequently depend on themselves and/or the analytical techniques used, leading to outliers (Leeuwen et al., 2021). These errors were observed in a soil test conducted with 100 laboratories around the world using standard procedures (Hartmann and Suvannang, 2019), which showed a disturbing variation in the quality and consistency of measurement of soil attributes.

A soil attribute outlier is a value that is significantly distant (dissimilar) from the population of samples. Outliers of soil attributes may most often be due to measurement errors from classical analytical methods (Demattê et al., 2019)—where the extensive sequence of steps may introduce errors from different sources like subsampling, reagents, rounding off readings, instrumental, among others (Mountier et al., 1966). Therefore, repeated measurements on soil samples are required to quantify the error of the laboratory, as part of costly laboratory proficiency testing programs (Leeuwen et al., 2021). Now, we need innovative approaches to help laboratories to rapidly characterize the soil conditions and adopt internal quality controls with lower costs and ecosystem impacts.

Soil spectroscopy is a fast, cost-effective and environmentally-friendly technique that emerged as an alternative to wet chemistry for estimating soil attributes (Nocita et al., 2015). The 350-2500 nm reflectance spectroscopy was the most studied spectral range for pedometrics due to its functionality and suitable results (Dematte et al., 2019; McBratney et al., 2019; Stevens et al., 2013; Viscarra Rossel et al., 2016). If spectroscopy has proven to be suitable for pedometrics, why do soil laboratories still not widely use this technology in their routines? Moreover, what would be the applications of spectroscopy in commercial soil laboratories? Could spectroscopy have any utility other than predicting soil attributes? Which and how many samples should be selected for analysis and prediction? What would be the guidelines for laboratory technicians? How could a breakthrough in soil analysis improve our understanding in soils? These are some of the questions that we tried to answer in this study.

Recently, the Global Soil Laboratory Network—GLOSOLAN (FAO, 2021) launched its initiative on soil spectroscopy and recognized the potential of visible to infrared spectra for soil analysis, monitoring and mapping, due to the fact that spectral signatures respond to soil mineral and organic composition. GLOSOLAN also highlighted the constraints that still hamper the wider uptake of this technology were the lack of: i) standards and protocols, ii) spectral libraries, and iii) capacity of traditional soil laboratories in spectral methods. Souza et al. (2016) emphasized that data analysis requires professionals with expertise in chemometrics, and it also hinders the implementation of spectroscopy as a routine soil analysis technique.

Despite GLOSOLAN's initiative, there are no currently global standard operating procedures for the use of spectra in laboratory routines. This was evidenced through a search with the keywords "soil spectroscopy" and "laboratory routine" in the Scopus database (Figure 1), showing that almost all articles published between 1995 and 2020 did not address the implementation of the technique in the laboratory routine. However, soil spectroscopy received more attention for laboratory routines from 2020.

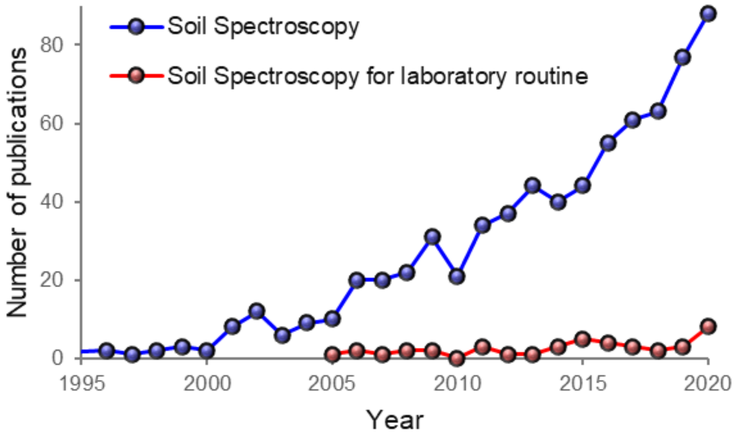


Figure 1. Number of publications searched in April 2021 from the Scopus database using the keywords "soil spectroscopy" and "laboratory routine".

Unsupervised analysis, such as clustering and principal component analysis, have gained popularity for pedometrics (Dematte et al., 2019; McBratney et al., 2019; Stevens et al., 2013; Viscarra Rossel et al., 2016). The unsupervised random forest (Ciss, 2015a) is another analysis capable of discovering relevant patterns in the data (Breiman, 2001), which has been little used for pedometrics. Perhaps, one reason for this is the belief that random forests only perform supervised analysis (Afanador et al., 2016). Regarding variations in spectra, Romero et al. (2018) observed low differences between sensor measurements mainly caused by geometry and equipment variation. Aware of this, Ben Dor et al. (2015) determined a protocol to standardize measurements between sensors, taking soil spectroscopy a step ahead of traditional methods (Demattê et al., 2019).

We assumed that: i) the soil spectra have a degree of association with soil attributes that is statistically significant to group samples into clusters with homogeneous soil attributes values; ii) the triplicate acquisition of soil spectra is less prone to measurement error than traditional analytical methods; iii) unsupervised analysis of soil spectra can be implemented for internal quality control in laboratories; iv) spectra can be used to select representative samples and reduce the need for traditional analysis. Therefore, we aimed to provide basic guidelines for the use of soil spectra in laboratory routines regarding internal quality control and samples selection for analysis and prediction.

4.2. MATERIALS AND METHODS

4.2.1. Soil samples acquisition

The soil samples were collected with an auger at 0-20 cm depth by 36 commercial laboratories from Brazil and Paraguay that participated in the Brazilian Program of Soil

Analysis via Spectroscopy—ProBASE (https://esalqgeocis.wixsite.com/geocis/probase) at the University of São Paulo, Brazil. These laboratories regularly participate in Brazilian proficiency soil-testing program of the Agronomic Institute of Campinas (IAC) and Embrapa. ProBASE is a training course on fundamentals and applications of soil reflectance spectroscopy for students, researchers, consultants and company directors and officers. In the first edition of the training in 2019, each participating commercial laboratory sent us about 200 soil samples for this study, totaling nearly 7,200 samples. The tropical soils sampled were mainly used for sugarcane, soybean, and corn agriculture.

4.2.2. Soil data acquisition

4.2.2.1. Traditional soil laboratory analysis

Each commercial soil laboratory air-dried, grounded and sieved to 2-mm mesh the soil samples for physical and chemical determination using traditional methods according to Teixeira et al. (2017). They determined the amount of clay and sand in g kg^{-1} by the densimeter method, organic matter (OM) in g kg^{-1} by the Colorimetric method, cation exchange capacity [CEC=Ca^{2+}+Mg^{2+}+K^{+}+H^{+}+Al^{3+}] in mmol$_c$ kg^{-1} and base saturation [V%=((Ca^{2+}+Mg^{2+}+K^{+})x100)÷CEC] in percentage.

4.2.2.2. Soil spectroscopy analysis

The commercial laboratories provided 50 g of soil fractions smaller than 2 mm that were conditioned by us in Petri dishes to measure the 350-2500 nm spectra at 1 nm resolution in laboratory, using the Fieldspec 3 spectroradiometer (Analytical Spectral Devices, ASD, Boulder, CO). We obtained three spectra—each one averaged from 100 scans— by soil sample from different rotation positions and we averaged them to get the final spectrum. The equipments were arranged according to the long light geometry from Brazilian protocols (Romero et al., 2018), where a sensor placed perpendicularly 8 cm away from the sample surface captured the light reflected from 2 cm^2. We corrected the splices of the final spectrum positioned at 1000 and 1800 nm by linear interpolation of 10 bands using the prospectr package (Stevens and Ramirez-Lopez, 2013) in the R software (R Core Team, 2018).

4.2.2.3. Correlation analysis

We performed a correlation analysis by laboratory using the Pearson's coefficient to measure the relationship between the soil spectra and traditional soil attributes. The strength of the correlation provided information about the ability of the spectra to produce statistically significant ($p < 0.01$) groups with relatively homogeneous soil attribute values.

4.2.3. Quality assessment of physical-chemical analytical results

4.2.3.1. Spectral clustering of soil samples

To decrease computational time of this step, we resampled the spectra from a resolution of 1 nm to 10 nm by simple interpolation using the prospectr package in R. After that, we used the resampled spectra and randomUniformForest (Ciss, 2015b) package in R for: i) unsupervised learning and computing proximity matrices, ii) dimension reduction, iii) clustering and iv) variable importance. We repeated (looped) all procedures separately for each commercial laboratory—we used the code (number) of the laboratory as a variable for stratification.

For the first step we used the unsupervised mode of Random Uniform Forests (RUF) algorithms (Ciss, 2015a)—that follows the Breiman (2001) formulation—to find complex patterns in resampled spectra and provide proximities measures between the soil samples. Proximity was based on the frequency (number of times divided by number of trees) that pairs of soil samples were in the same terminal nodes (Ciss, 2015a).

In the second step, we applied a classical (metric) multidimensional scaling (Gower, 1966) to the proximity matrices to reduce their dimensions to two principal coordinates that were enough to achieve the best possible separation between points.

For the third step, we used the first two principal coordinates of the proximity distances to calculate the Euclidean (straight-line) distances (Pythagoras theorem) between points and find the nearest neighbor distance—since similar things are near to each other— to cluster the soil samples using the K-means algorithm (Ciss, 2015b). We adjusted the optimal number of clusters (limited to ten) by the gap statistic (Tibshirani et al., 2001). Therefore, we obtained spectral clusters of soil samples for each laboratory that we used for outlier detection in the next section.

In the fourth step we converted the unsupervised RUF models into supervised ones—we obtained the clusters labels and sent them to the supervised RUF classifier for learning with the data— to compute global variable importance (Ciss, 2015b). The importance showed which wavelengths were influential for defining the RUF proximity measures.

4.2.4. Outlier detection

We used the spectral clusters obtained in the previous step to identify—separately within each one— the outliers of the soil attributes that were determined using classical analytical methods by the commercial laboratories.

Since the soil data were frequently not normally distributed within each group according to the Shapiro-Wilk's test, we used an adjusted method for skewed distribution proposed by Hubert and Vandervieren (2008) as part of the robustbase package for R, to detect outliers without making any assumption about the distribution of the data.

The method used the interquartile range [IQR = $Q_3 – Q_1$] and the medcouple–MC, a robust measure of skewness introduced by Brys et al. (2004)— to calculate the lower [$Q_1 - 1.5 \text{ x } e^{(-4 \text{ x MC})}$ x IQR] and upper [$Q_3 + 1.5 \text{ x } e^{(3 \text{ x MC})}$ x IQR] whiskers when MC ≥ 0 (right skewed), or the lower [$Q_1 – 1.5 \text{ x } e^{(-3 \text{ x MC})}$ x IQR] and upper [$Q_3 + 1.5 \text{ x } e^{(4 \text{ x MC})}$ x IQR] whiskers for MC < 0 (left skewed), as the boundaries of the data distribution. The Q_1 and Q_3 are the 1st (25%) and 3rd (75%) quartiles, and *e* is the exponential model used to define the whiskers. Any values that fell outside of these thresholds (far away from the rest of the data) were flagged as outliers.

4.2.5. Samples selections for laboratorial analysis

4.2.5.1. Selection of training and testing sets

We split samples based on spectra into subsets with data distribution close to that of the full sample set of each laboratory to reduce the variance between them and improve prediction results (Chen et al., 2021; Khaledian and Miller, 2020). Firstly, we removed the outliers detected in the previous step and ordered the soil samples by the mean spectral reflectance from the most important wavelengths—global variable importance higher than 70%. Then, we split the whole data into test and train sets by (systematically) selecting the K-th sample of the sequence—e.g., we selected the first sample (50%) to train and the second one (50%) to test, or the first two samples (67%) to train and the third one (33%) to test the algorithm. Therefore, we used samples split into 50%-50%, 67%-33%, 75%-25%, 80%-20% and 87%-13% for training and testing the prediction algorithm and evaluating the effect of the set size on the model performance for each soil attribute and laboratory. This approach can indicate the ability of a given laboratory of soil analysis to reduce the number of samples and the associated costs required for traditional determinations.

4.2.5.2. Soil attributes modelling from spectra

We selected Cubist algorithm (Quinlan, 1992) provided in Cubist R package (Kuhn and Quinlan, 2021) based on tests (unpublished results) and literature (Chen et al., 2021; Moura-Bueno et al., 2020; Silva et al., 2019) that frequently showed better results for Cubist than other algorithms for soil spectral modeling. Therefore, we used Cubist—a machine learning algorithm that fits the final prediction by a linear regression—to train regression models—using 100 committees for stable predictions and other default parameters—between raw spectra—described in section 2.2.2— and soil attributes for each commercial laboratory.

Then, we used the trained Cubist models to make prediction in the testing set and calculate the following metrics for each laboratory: i) root mean square error (RMSE) to assess how close the predicted values were to the observed ones; ii) relative RMSE [rRMSE (%) = (mean RMSE/(*Y*_max – *Y*_min))x100], where *Y* is the observed value, to compare RMSE with different units; iii) coefficient of determination (R^2) to indicate the proportion of the variance in the target variable that the model explains; and iv) ratio of the performance to interquartile distance (RPIQ = IQR/RMSE) to evaluate the spread of the dataset to the models accuracy (Khaledian and Miller, 2020). Predictive models with higher performances usually have lower RMSE (or rRMSE) and higher R^2 and RPIQ values.

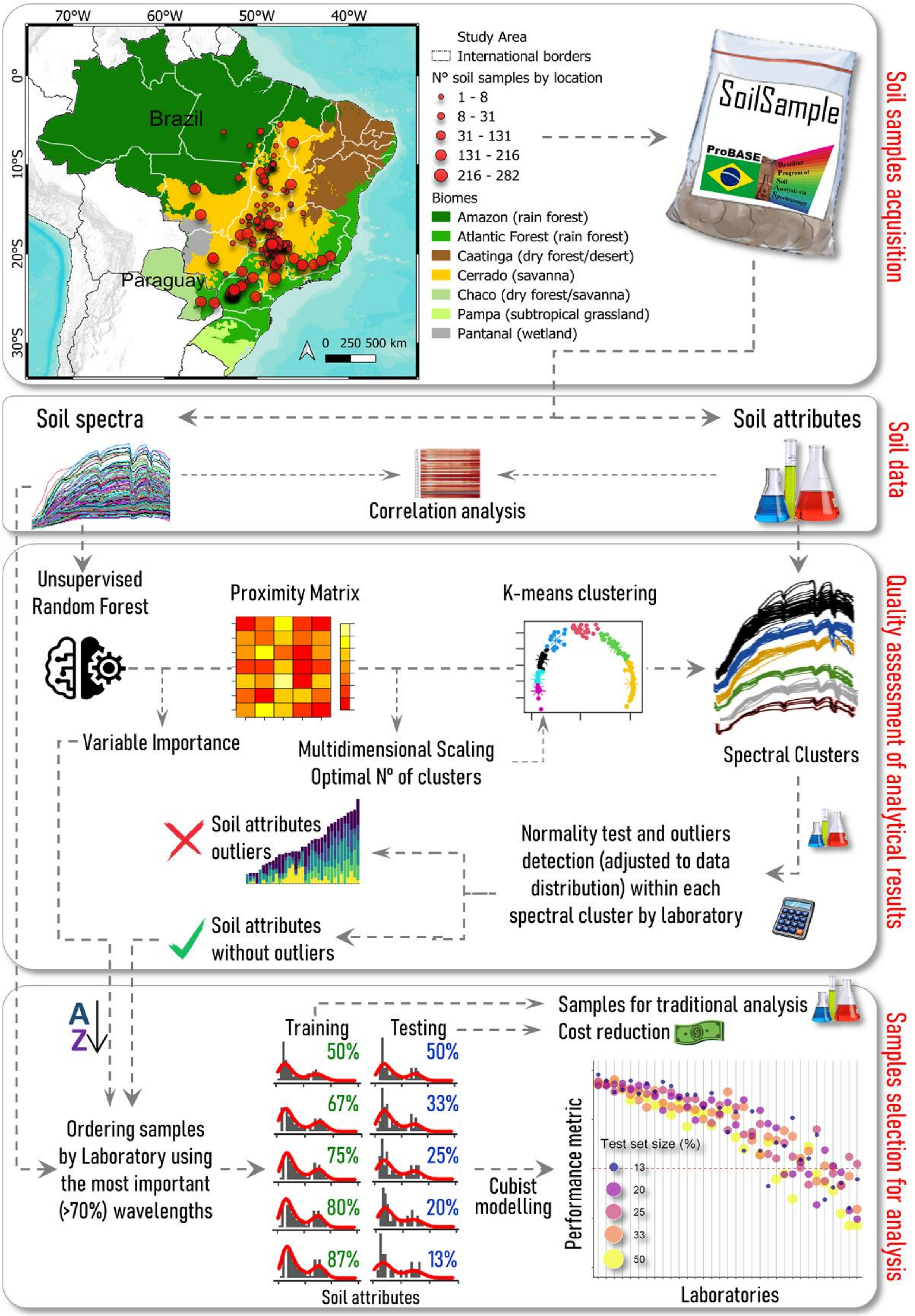


Figure 2. Methodological flowchart and location of the soil samples acquired through the Brazilian Program of Soil Analysis via Spectroscopy (ProBASE) that were used in this study.

4.3. Results

4.3.1. Relationship between soil spectra and traditional attributes

We performed about 180 correlation analyses between 36 labs and 5 soil attributes, and for this reason we described and discussed only average values. The averaged 350-2500 nm soil spectra (Figure 3) showed the presence of absorption features of iron oxides—mainly hematite and goethite—, kaolinite, gibbsite and 2:1 clay mineral in the soil fractions smaller than 2 mm. These minerals are typical in the studied tropical soils—mainly Ferralsols— which developed in flattened or smoothed reliefs with good drainage conditions that facilitated their agricultural mechanization.

The soil spectra (Figures 3, 4 and 5) had significant ($p < 0.01$) Pearson's correlation with clay ($–0.82 < r < –0.04$), sand ($–0.09 < r < 0.82$), OM ($–0.68 < r < 0.12$), CEC ($–0.69 < r < 0.13$) and V% ($–0.56 < r < 0.31$). Higher amounts of clay and OM in the soil—normally associated with higher levels CEC and V%— produced higher absorption of the electromagnetic radiation and negative correlation values. Conversely, higher amounts of sand fractions increased the reflectance factor of the soil and caused positive correlation values.

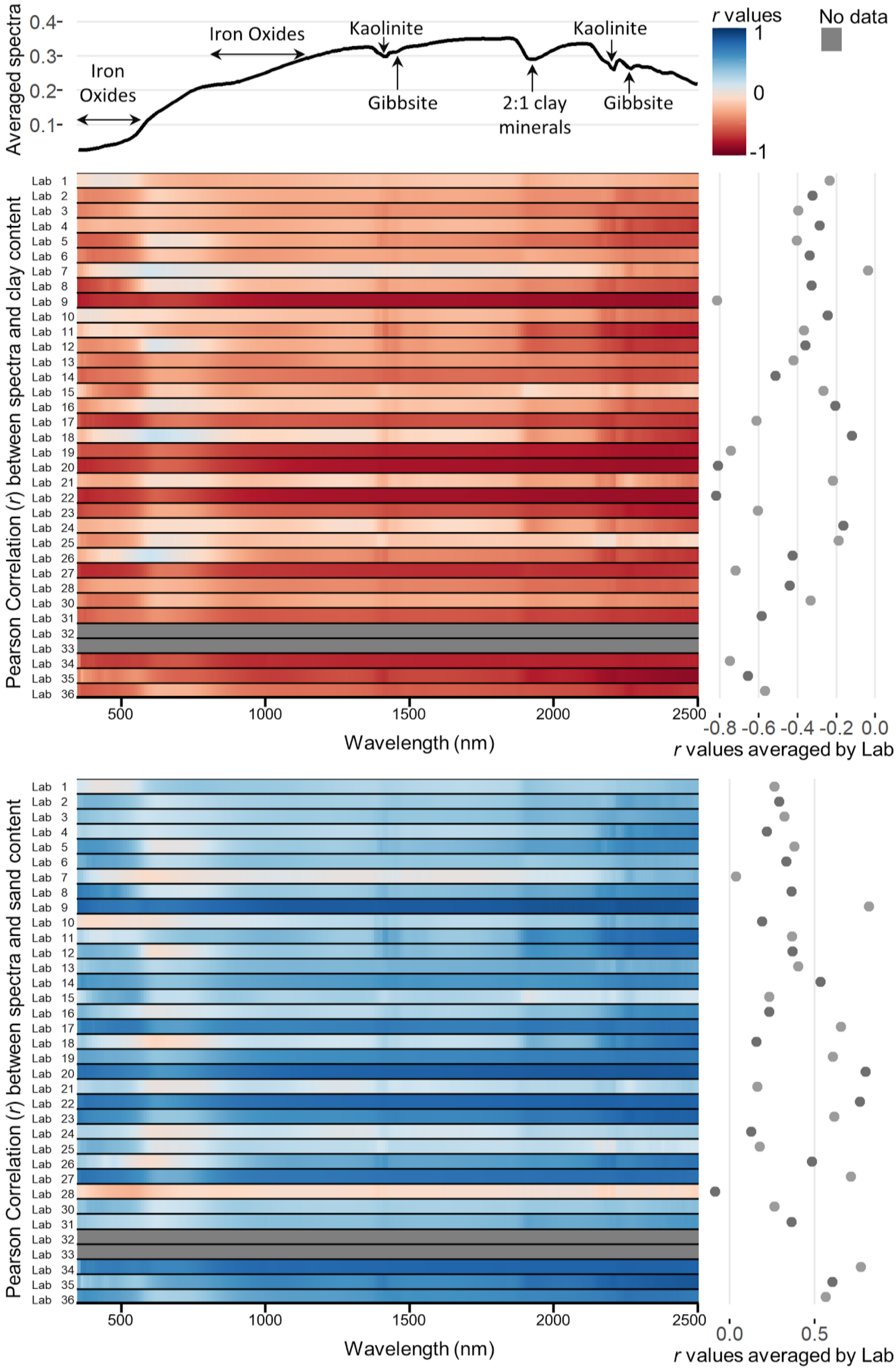


Figure 3. Pearson correlation values of raw spectra with clay (top panel) and sand (bottom panel) contents displayed by laboratory.

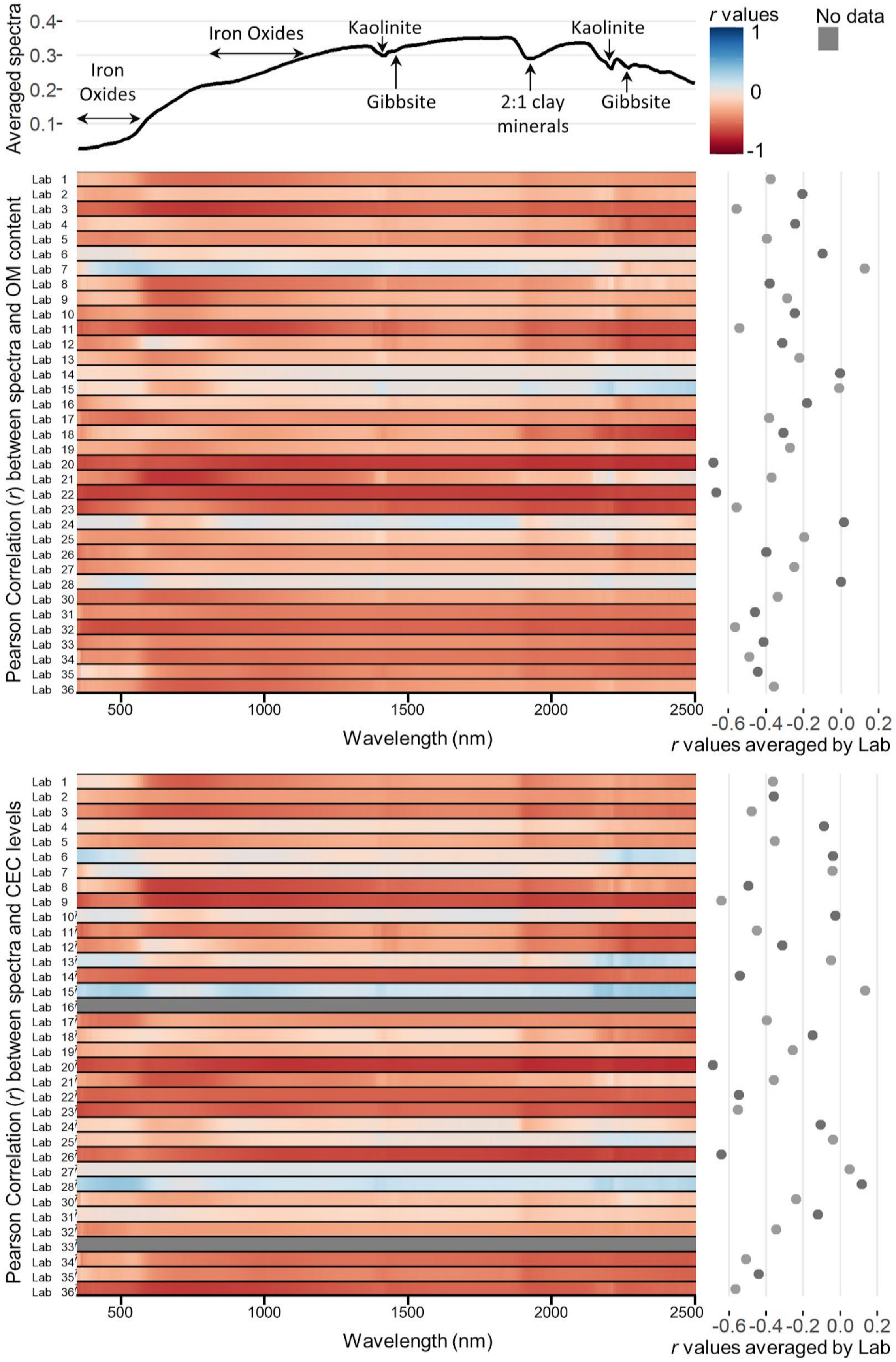


Figure 4. Pearson correlation values of raw spectra with organic matter (top panel) and cation exchange capacity (bottom panel) levels displayed by laboratory.

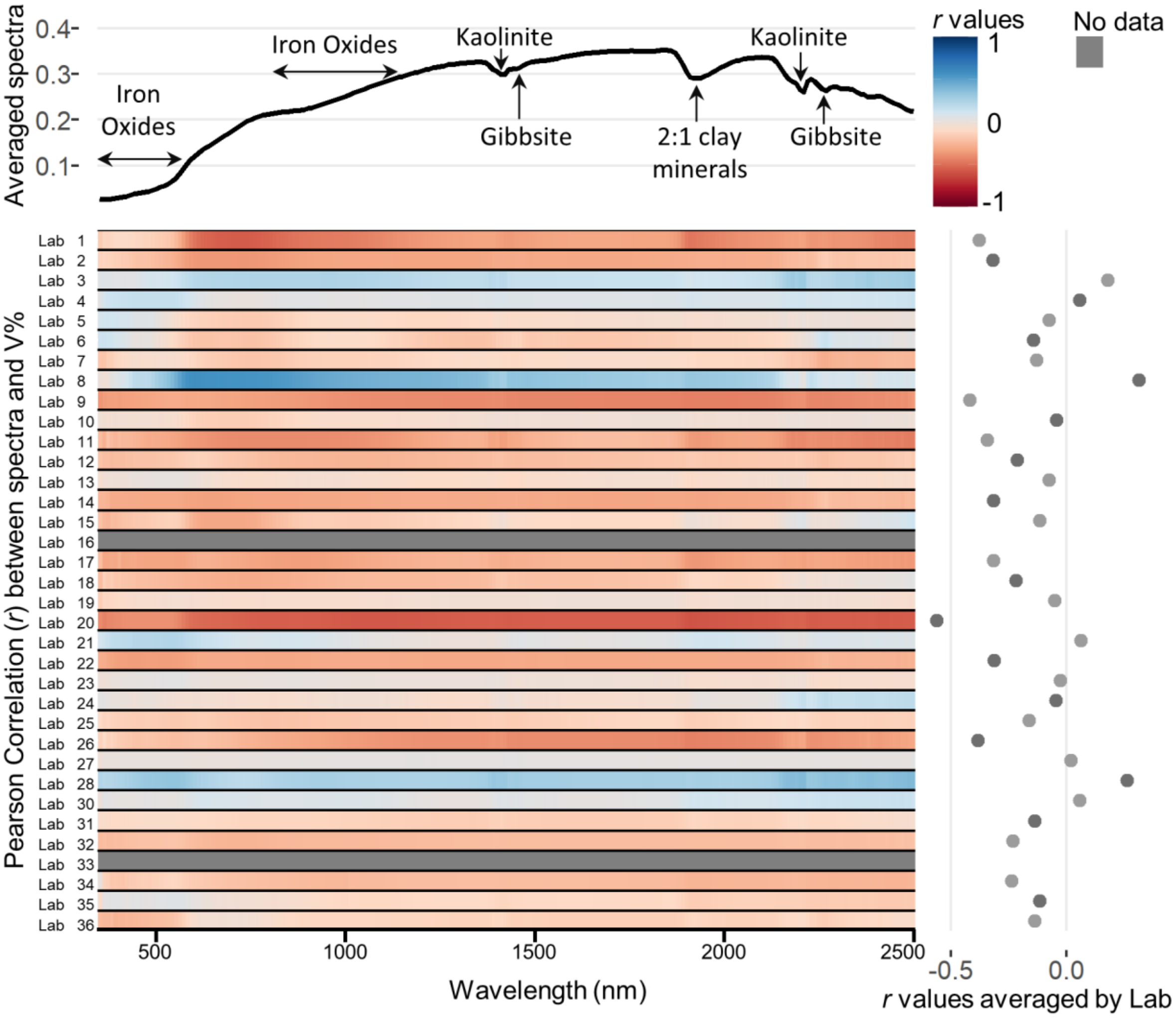


Figure 5. Pearson correlation values of raw spectra with base saturation levels displayed by laboratory.

4.3.2. Quality in analytical methods

Overall, the variable importance based unsupervised RUF spectral classification (Figure 6) showed strong interaction effects between the variables (> 50% of variance explained per wavelength). The ranges near 500, 700, 1000–2100 and 2300 nm were the most important (> 70%) for proximity measures between soil samples. These influential wavelengths—matching the spectral ranges linearly correlated with the studied soil attributes (Figures 3–5)— guided the proximity measures—reduced to two principal coordinates— and, therefore, the formation of the spectral clusters (left panel in Figure 7). It suggested that soil samples were clustered into homogeneous groups with similar values of clay, sand, OM, CEC, and V%, allowing a more precise outlier analysis within each cluster for these soil attributes. Figure 7 (left panel) shows an example for the laboratory number 35 with ten spectral clusters formed by grouping from 4 to 43 spectra with very close (similar) spectral patterns.

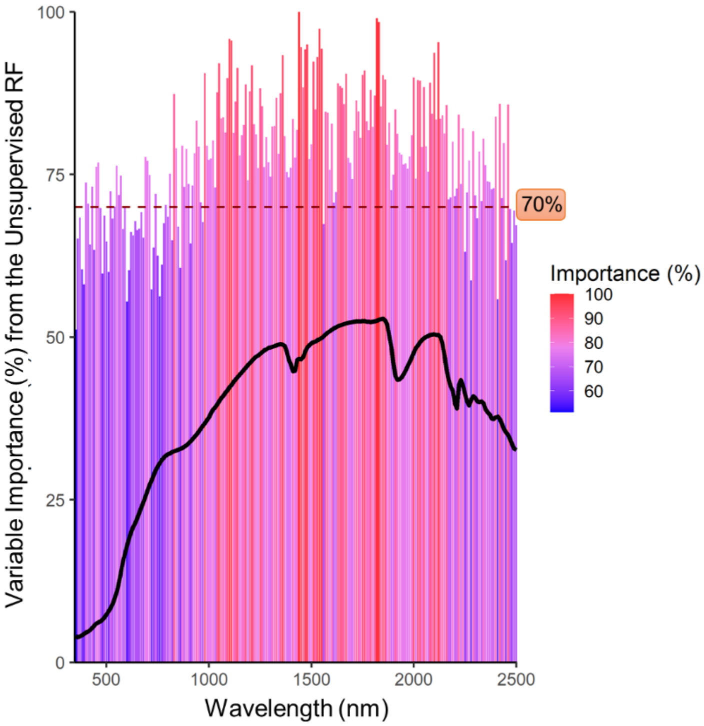


Figure 6. Global variable importance measures (%) based on interactions computed for the unsupervised RUF spectral classification of soil samples.

The outlier analysis used in this study (right panel in Figure 7) precisely adjusted the lower and upper thresholds for distribution of soil data within each spectral cluster and flagged values as outliers. Outlier detection was performed about 1,080 times—36 labs with 5 soil attributes and near 6 spectral clusters by lab. Figure 7 shows an example of how this method worked for clay contents determined by the laboratory number (code) 35, where values far away from the rest of the data were flagged as TRUE (outlier) and their respective spectra were marked in red color. Therefore, soil samples spectrally near to each other (Figure 7)—with spectra correlated with soil attributes data (Figures 3–5)— detected samples with possible error of measurement in classical soil analytical methods. Unfortunately, due to limited resources, we were unable to reanalyze the samples identified as outliers by soil analytical methods to confirm these results.

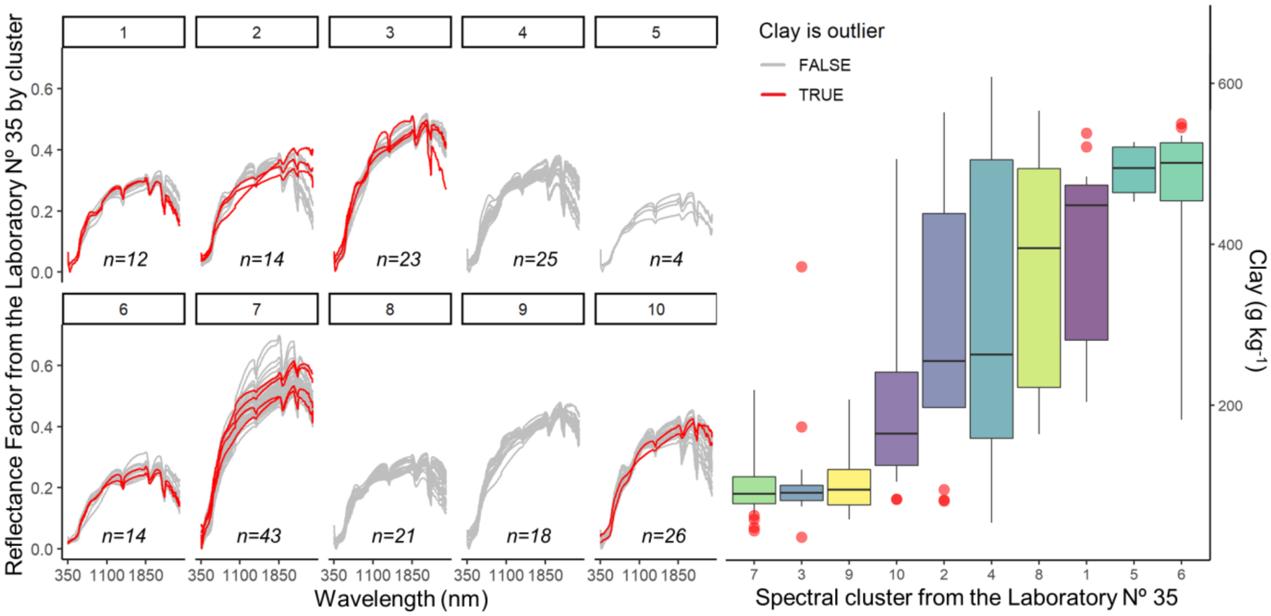


Figure 7. Outlier detection method performed for clay content determined by the laboratory number 35, displaying the spectral cluster (left panel) and boxplot adjusted for skewed distribution (right panel) with outliers highlighted in red color. *n*: number of spectra grouped within a specific cluster.

After we detected and flagged outliers in wet chemistry data within the clusters for each laboratory, we found an increasing total number of outliers detected—sum of all laboratories— for sand (255 ≈ 3.64%) < V% (265 ≈ 3.79%) < CEC (303 ≈ 4.33%) < clay (320 ≈ 4.57%) < OM (324 ≈ 4.63%)—Figure 8. It suggested that sand had the lowest laboratory inaccuracies and OM the largest ones. The total number of outliers for a specific laboratory—sum of all soil attributes— varied from 21 (≈10.5% for lab 4) to 69 (≈34.5% for lab 35), suggesting that laboratory 35 had nearly three times more global inaccuracy than laboratory 4. The top six laboratories with the highest quality (lowest number of outliers or inaccuracy) for clay were the codes 8, 34, 21, 4, 1 and 36, for sand were the codes 28, 36, 31, 10, 8 and 16, for OM were the codes 4, 11, 26, 12, 10 and 28, for CEC were the codes 16, 8, 12, 6, 19 and 10 and for V% were the codes 16, 19, 24, 34, 20 and 9 (Figure 8). The study laboratories used the same analytical methods and obtained different numbers of outliers for the same soil attributes. It means they had different levels of accuracy or laboratory measurement error.

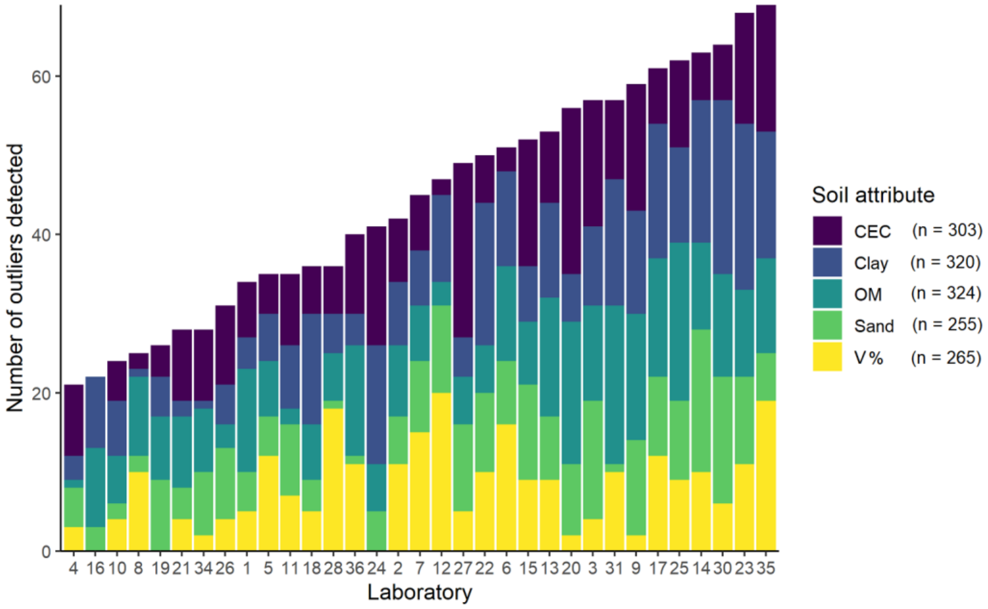


Figure 8. Total number of outliers detected for each soil attribute determined using analytical methods grouped by laboratory. *n*: total number of outliers detected for a specific soil attribute.

4.3.3. Sample selection for determination and prediction

Systematic sampling of samples based on spectra by laboratory resulted in 900 training and 900 testing sets of different sizes—36 labs with 5 soil attributes and 5 set sizes— with soil attribute values equally distributed among the subsets. To save space, we selected as an example the clay content determined by laboratories for different training and testing set sizes and displayed its histograms in Figures 9 and 10. These histograms showed soil data with different distribution among laboratories, usually due to soil spatial variability between the different study locations (Figure 2). We emphasize that no soil attribute values were used to split the sample sets, and that our approach yielded promising results for spectral library applications.

Therefore, the soil samples that a laboratory receives from its customer (users) could have their spectra acquired to support the selection of which samples will have their attribute values determined using wet chemistry, and which of them its values predicted using spectra-based models. The size of the training set will result from previous studies that the laboratory will need to perform using existing data from your laboratory.

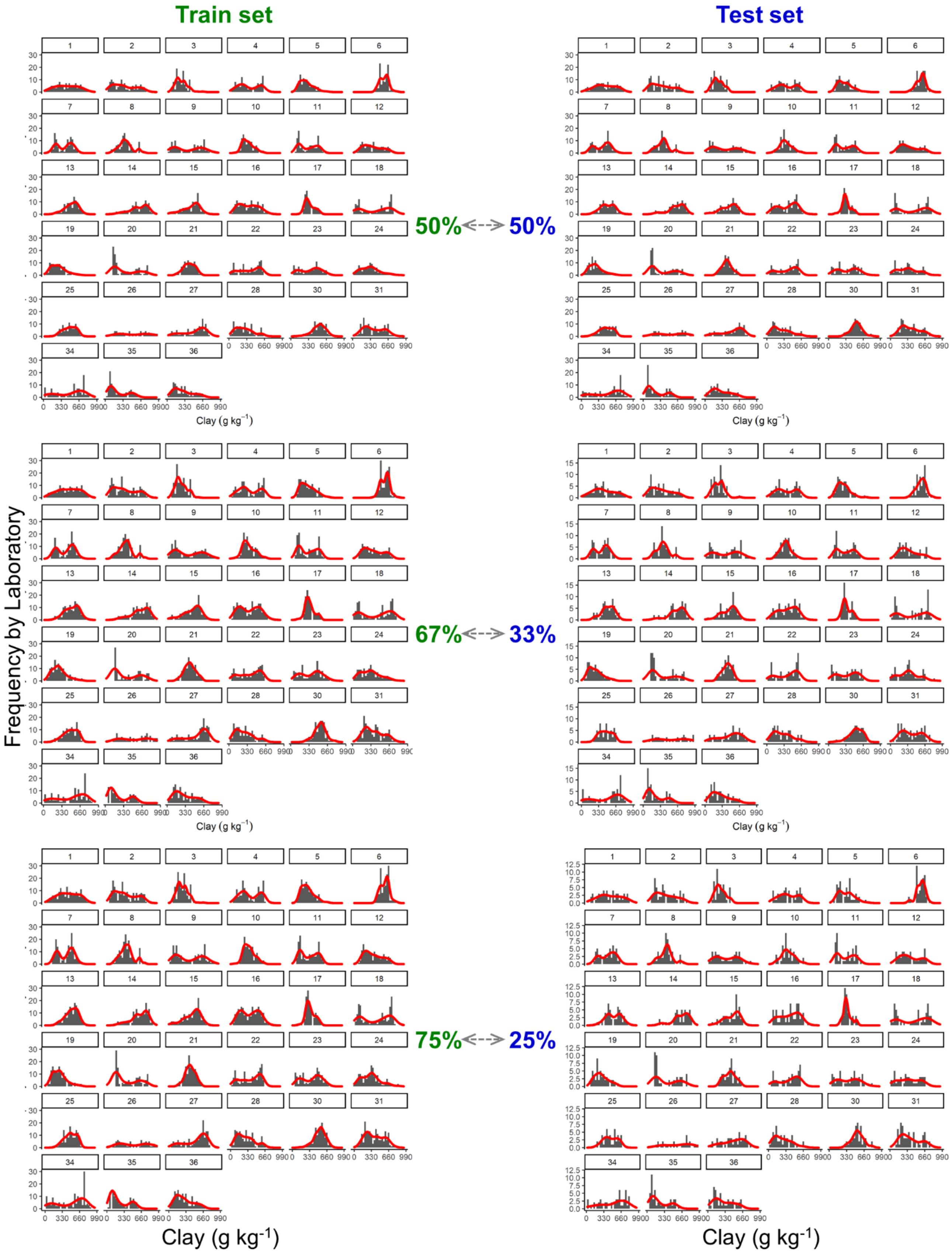


Figure 9. Histogram of clay content determined by laboratories for different training (left panel) and testing (right panel) set sizes.

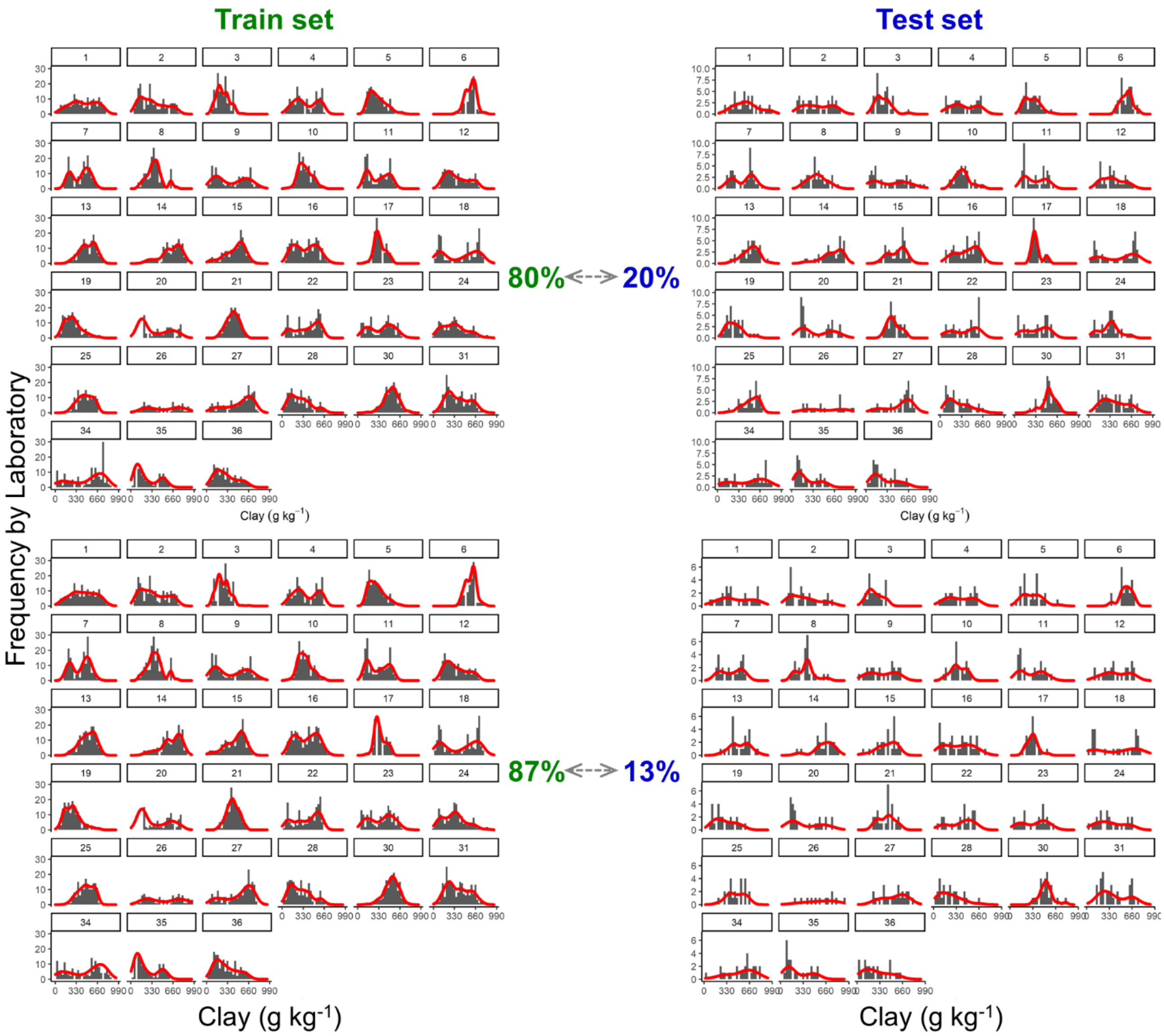


Figure 10. Histogram of clay content determined by laboratories for different training (left panel) and testing (right panel) set sizes.

4.3.4. Soil predictions

Since we trained and tested nearly 900 prediction models—36 labs with 5 soil attributes and 5 set sizes— using Cubist algorithm, only the mean results were described and discussed. Overall, we obtained highly variable performance results between laboratories for each soil attribute from the testing sets (Figures 11-13), with R^2 values ranging from near 1 to 0 (Figure 11), rRMSE from 28 to 2% (Figure 12) and RPIQ from 9 to 1 (Figure 13). The predictions in the validation sets showed mean values of i) R^2 between 0.71 and 0.28 for clay > sand > CEC > OM > V%, ii) rRMSE between 12.76 and 19.65 for clay < sand < OM < CEC < V%, and iii) RPIQ between 3.61 and 1.69 for clay > sand > CEC > OM > V%. It means that clay and sand often had higher average performance levels than CEC, OM, and V%.

We found that smaller test sets (size = 13%), and therefore, larger training sets (size = 87%), promoted moderately higher performances for only 30% of the laboratories (Figures 11-13). We also observed an overall weak relationship of test set size with R^2 ($r \approx –0.18$), rRMSE ($r \approx 0.19$), and RPIQ ($r \approx –0.13$) for all labs. These results suggested that larger training sets did not always guarantee better model fit for our conditions.

The laboratories that outperformed the average in soil predictions using 50% of the samples (for testing) comprised 27% for clay ($R^2 > 0.71$, rMRSE < 12.76%, RPIQ > 3.61), 30% for sand ($R^2 > 0.69$, rMRSE < 12.67%, RPIQ > 3.47), 17% for OM ($R^2 > 0.53$, rMRSE < 13.79%, RPIQ > 2.21), 33% for CEC ($R^2 > 0.54$, rMRSE < 14.31%, RPIQ > 2.36), and 21% for CEC ($R^2 > 0.28$, rMRSE < 19.65%, RPIQ > 1.69). These levels of performance suggested that near 25% of the laboratories had a higher potential than others to reduce the number of soil samples for training by at least 50%, while reducing laboratory costs related to physical-chemical analytical methods.

Conversely, about 31% of the laboratories had difficulties in predicting clay, sand, OM, and CEC, while 45% of them were almost unable to predict V% for all test set sizes. We found that R^2 and RPIQ values from prediction models presented direct correlation with standard deviation ($0.15 < r < 0.71$) and indirect with kurtosis ($–0.61 < r < –0.29$) from observed soil attributes, but also slightly increased rRMSE ($r \approx 0.10$). It suggested that lower prediction performances were related to the smaller standard deviations and larger effects of kurtosis within the training and testing sets from each laboratory—see the example of the distribution of clay data in Figures 9 and 10. However, when we analyzed the kurtosis within the spectral clusters from each laboratory, we found mean correlation values of kurtosis with R^2 and RPIQ varying between –0.08 to 0.03, and it suggested no effect of kurtosis. For all cases, skewness had a stronger effect only for R^2 ($r \approx –0.17$) and rRMSE ($r \approx 0.42$) from V% predictions.

Finally, when we tried to model soil attributes within each spectral cluster, we obtained worse prediction performances (unpublished results) likely due to the limited number of samples remaining (after the splitting process) for training and testing the Cubist algorithms. For instance, considering that each laboratory had nearly 200 samples and an average of 6 spectral clusters, we had about 33 samples for splitting into training and testing sets. This low number of samples probably failed to correctly represent the soil variations and led to worse results than those we obtained using the full sample set from each laboratory.

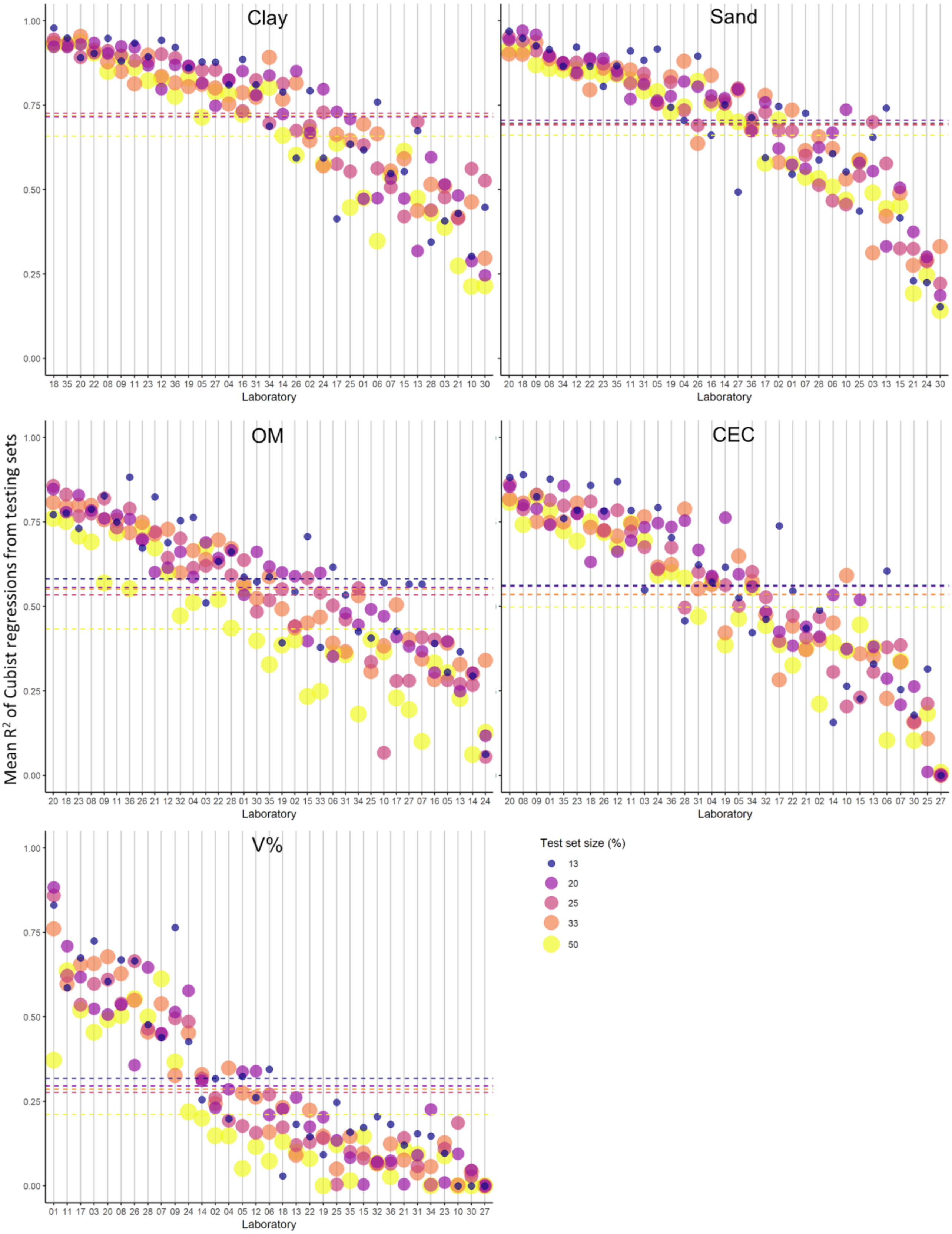


Figure 11. Coefficient of determination (R^2) of the soil attributes predicted by Cubist from different testing set sizes grouped by laboratory. OM: organic matter; CEC: cation exchange capacity; V%: base saturation. Dashed lines represent mean values for each test set size.

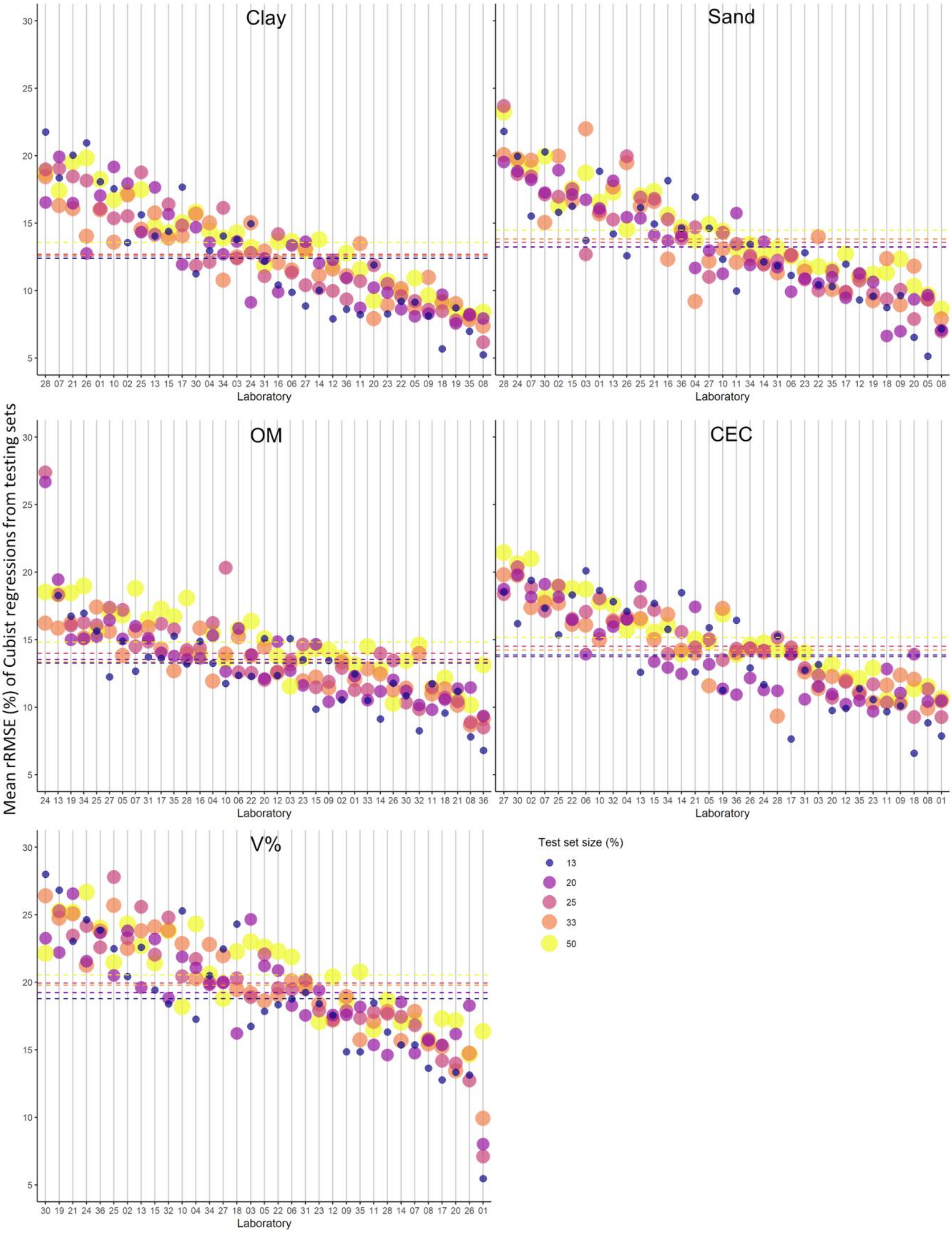


Figure 12. Relative root mean square error (rRMSE, %) of the soil attributes predicted by Cubist from different testing set sizes grouped by laboratory. OM: organic matter; CEC: cation exchange capacity; V%: base saturation. Dashed lines represent mean values for each test set size.

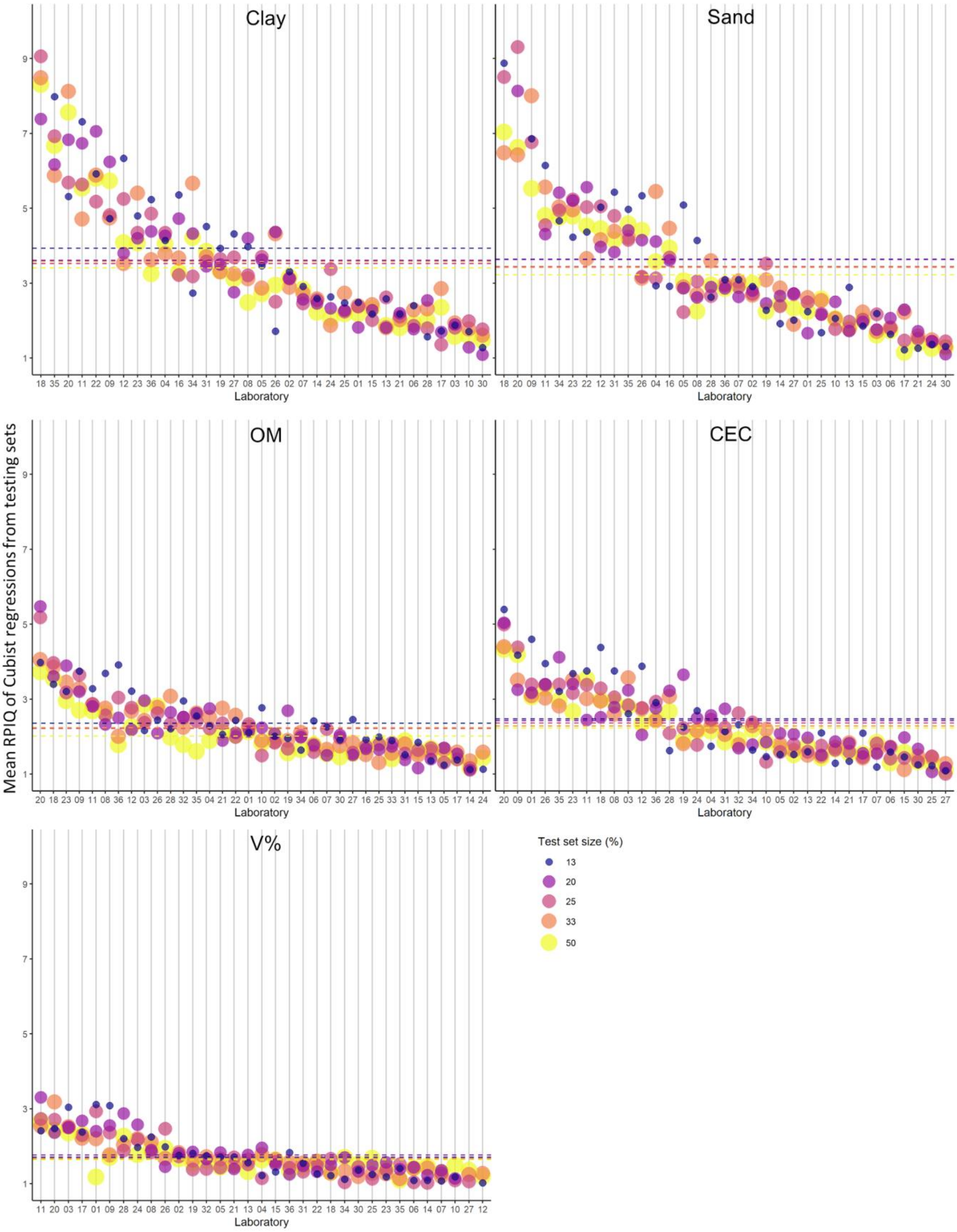


Figure 13. Ratio of the performance to interquartile distance (RPIQ) of the soil attributes predicted by Cubist from different testing set sizes grouped by laboratory. OM: organic matter; CEC: cation exchange capacity; V%: base saturation. Dashed lines represent mean values for each test set size.

4.4. Discussion

4.4.1. Relationship between soil spectra and traditional attributes

The mineralogical composition of the studied soils originated from their parent materials that were exposed for long-term to strong weathering conditions and resulted in intensive leaching of silica from soil fractions (Schaefer et al., 2008). The minerals with spectral responses from tropical soils (Figures 3–5) occur frequently in the clay fraction in the form of iron oxides–that reduces the reflectance near 500 and 900 nm due to electronic transitions—, kaolinite, gibbsite and 2:1 clay mineral—that lower the reflectance near 1400, 1900 and 2300 nm due to molecular vibrations involving stretching and bending (Clark et al., 1990; Guimaraes et al., 2021). Highly reflective minerals as quartz are present in the sand fractions (Figures 3–5) increasing the soil reflectance from near 1750 nm (Lacerda et al., 2016).

The negative correlations of raw spectra with OM can be related to the absorption effect of organic compounds across 350-2500 nm range, where correlations near 500 and 700 nm are probably related to soil color (Demattê et al., 2003; Nawar et al., 2016). The minerals in the clay fraction and the OM—both with spectral response— usually generate electrical charges that retains ions in the soil—such as Ca^{2+}, Mg^{2+}, K^{+} and Al^{3+} without spectral response—causing and indirect correlation of soil spectra with CEC and V% (Santana et al., 2018). Nevertheless, the Pearson’s correlation values near to zero for some laboratories (Figures 3–5) can be caused by complex relationships—not always considered linear— between spectral data and soil attributes (Santana et al., 2018) or measurement error in wet chemistry soil data (Leeuwen et al., 2021).

Similar descriptions between 350–2500 nm spectra and soil attributes (Figures 3–5) were reported by the Brazilian Soil Spectral Library (Dematte et al., 2019) for soils from Cerrado and Atlantic Forest biomes (Figure 2). The Soil spectral library of Piauí State in Brazil (Mendes et al., 2021) also reported similar relationships, where the soil clay and organic carbon contents had negative correlations and sand had positive correlation values with 350–2500 nm spectra. Moura-Bueno et al. (2020) found that the first principal component from 350–2500 nm soil spectra had inverse correlation with organic carbon (r = −0.75) and clay (r = −0.63) contents, and positive correlation with sand (r = 0.45) content from subtropical Brazilian soils. In addition, Stevens et al. (2013) reported principal components form 350–2500 nm spectra strongly correlated with clay (r = −0.45), CEC (r = −0.5), organic carbon (r = −0.55) and weaker with sand (r = 0.17) content.

4.4.2. Quality in analytical methods

Despite RUF is an important unsupervised method, a keywords search in Scopus database—TITLE-ABS-KEY("unsupervised random forest" OR "random uniform forest" AND "chemometrics" OR "spectroscopy")— revealed that it has been little applied to chemometrics in spectroscopy, possibly due to the belief that random forest can only be used for supervised analysis. In our study, the unsupervised RUF confidently classified similar soil samples as belonging to the same spectral class, resulting in high values of proximity compressed to lower dimensional space for clustering. That is because during the unsupervised RUF classification, spectrally similar soil samples followed the same routes along the different decision trees and ended up in the same terminal node (Afanador et al., 2016).

Highly predictive wavelengths were the most important for proximity measures (Figure 6) because they explained more the variance between samples, and it helped to understand the structure of the spectral clusters (Ciss, 2015b). Similar influential wavelengths were also reported by Moura-Bueno et al. (2020), who reduced 350–2500 nm spectra to three principal components and found eigenvectors related to iron oxides (near 450, 600–800 nm), OM (near 800–900 and 1850 nm), and clay minerals (near 1400 and 1900 nm). Mendes et al. (2021) reported that the 1000–2500 nm spectral range had high importance to estimate soil attributes such as pH, sand, clay, and organic carbon. The first the principal components from the Brazilian Soil Spectral Library (Dematte et al., 2019) were related to iron oxides and OM (near 500 and 900 nm), kaolinite, gibbsite, quartz and 2:1 clay mineral (near 1000–2100 and 2300 nm). Similar wavelengths—to those found in our study for soil clustering— at near 400–750, 1000–1250 and 2300 nm were also important to predict soil clay, sand, organic carbon and CEC in the global spectral library (Viscarra Rossel et al., 2016) due to their intrinsic relationship with soil constituents.

Ramirez-Lopez et al. (2013a) and Zeng et al. (2021) showed that the soil 350–2500 nm similarity is directly related to the soil compositional similarity for soil organic carbon, pH, CEC, clay, silt, and sand content. Similarly, Ramirez-Lopez et al. (2013b) used a regional (Brazilian) and global 350–2500 nm soil library to test a spectrum-based learner by: i) computing their principal components, ii) measuring the Mahalanobis distances between samples, and iii) training prediction models based on nearest neighbors (similarity). They reported accurate predictions for soil clay, organic carbon and Ca^{2+} contents attributed to the use of local models based on spectral distances calculated from principal components. That method is comparable to our approach, as we also performed a local analysis based on spectral distance measurements between samples, however, applied to outlier detection. Therefore, the nearer the samples are spectrally between 350–2500 nm, the more similar they can be in their pedological composition. It means that for samples within a spectral cluster, the variability of a specific soil attribute (Figure 7) can be partly explained by the variability of the spectral similarity measures (Ramirez-Lopez et al., 2013a, 2013b; Zeng et al., 2021). Likewise, Araújo et al. (2014) divided a global soil dataset—nearly 7,000 samples— into spectrally (mineralogically) similar clusters and it improved the prediction performance of clay and OM contents.

The performance of 50 commercial soil laboratories in Brazil was assessed by Agbenin and Cantarella (2011) using Z-scores—based on two-sided probability distribution— to detect outliers for each soil attribute where organic carbon (Walkley-Black method) had the largest outliers and Ca^{2+} and Mg^{2+} (Mehlich I method) the smallest ones, similar to our results. Nevertheless, Z-score is not robust enough for a skewed data distribution (Hubert and Vandervieren, 2008). Demattê et al. (2019) evaluated the analytical quality of soil attributes from Brazilian laboratories and observed the lowest number of penalized samples by laboratories (sum of all laboratories) for CEC, sand, and H^{+}+Al^{3+}, like our findings.

Souza et al. (2016) used 1000–2500 nm spectra to predict soil organic carbon (Walkey-Black method) from 111 laboratories and detected about 10% outliers in the data using methods such as extreme leverage and unmodelled residuals. Santana et al. (2018) considered soil samples as outliers if the predicted values—using 350–2500 nm spectra— presented residues exceeding ±3×RMSE or if standard deviation was higher than a threshold for soil attributes. They excluded from the full samples set between 2.3 and 3.1% of outliers for OM < sum of bases (SB) < Sand < CEC—when modeled by random forest— and between 5.6 and 15.6% of outliers for Sand < Clay < SB < CEC < OM—when modeled by partial least squares. However, all these methods—to detect outliers— assume different statistical distributions and allow for different types of error.

During the inter-laboratory comparison for Asia (Suvannang and Hartmann, 2019), the 16 participating laboratories had on average 14% and 10% of outliers—detected by standard deviation thresholds— for organic carbon (Walkey-Black method) and K^{+} (Ammonium Acetate method), respectively. The inter-comparison of 16 Latin American laboratories (Guerrero and Bertsch, 2020) resulted in the lowest average outliers for Mg^{2+} (≈2%), followed by organic carbon (≈4%), Ca^{2+} (≈5%) and K^{+} (≈13%). Both studies related the occurrence of outliers possibly to i) the lack of quality control inside the laboratories, and ii) the lack of sufficient professional training and qualification of the staff for soil analysis.

The study of Anas et al. (2016) demonstrated that the quality of soil laboratories cannot be equivalent even when they used the same analytical method for soil attributes determination. Whenever a specific soil attribute is measured using multiple methods, each of them has its own bias (Leeuwen et al., 2021). The main factors causing measurement error are: i) sample or subsample handling and preparation methods, ii) the analyst, iii) changing laboratory conditions—e.g., humidity, temperature—, and iv) complex analytical methods (Libohova et al., 2019; Viscarra Rossel and McBratney, 1998).

Therefore, to improve quality in general, an inter-laboratory study in analytical chemistry (Hund et al., 2000) recommend that laboratories should implement their own quality assurance systems—as the one proposed in this study, which is faster and cheaper— instead of trying to acquire expensive instrumentation that does not guarantee high quality results. These insights are relevant for initiatives such as Global Soil Laboratory Network—GLOSOLAN (FAO, 2021).

4.4.3. Sample selection for determination and prediction

Our findings add to the few studies on soil spectral modelling that have examined distinct test set sizes from commercial laboratories. Other studies (Dematte et al., 2019; Gogé et al., 2014; Moura-Bueno et al., 2020; Ramirez-Lopez et al., 2013a; Santana et al., 2018; Silva et al., 2019; Stevens et al., 2013; Viscarra Rossel et al., 2016)— where they randomly divided the data set using only a single proportion (e.g., 70:30, 75:25)— also reported that the distribution of soil attributes (e.g., clay, sand, OM, SB, CEC) values were not normally distributed, mainly because many samples were collected from different soils.

The effects of subsetting strategies on soil spectral modelling were assessed by Clingensmith et al. (2019), where the systematic sampling based on soil spectral data improved model performance—for clay, sand, organic carbon, total nitrogen, and Fe available— compared with other sampling strategies. The study also highlighted the importance of systematic sampling—from the full set of samples— using spectra for areas with i) greater spectral variation and spatial heterogeneity, ii) larger geographic extent, and iii) sub-optimally sampled data. Likewise, Nawar and Mouazen (2018) compared three sample selection methods and different subsetting sizes for soil organic carbon prediction using 350–2500 nm spectra, and found the systematic sampling—based on spectral similarity matrices— provided uniform distribution of data between subsets and improved accuracy and robustness of prediction models. Zeng et al. (2021) selected similar soil samples using spectral data and improved the predictive accuracy of soil organic carbon, pH, CEC, clay, silt, and sand content. Therefore, data split based on spectra reduced the training set size, and, thus, the costs with soil analysis, while improving model performance.

4.4.4. Soil predictions

Our soil predictions from raw spectra yielded performance values similar to those obtained from soil spectral libraries (Dematte et al., 2019; Stevens et al., 2013; Viscarra Rossel et al., 2016) for sand, clay, organic carbon, CEC, and V% using Cubist and several spectral modelling strategies—e.g., preprocessing, stratification, splitting and auxiliary predictors. We also had a high number of laboratories that outperformed the results from the soil spectral library of Piauí in Brazil (Mendes et al., 2021) and other studies on spectral modelling (Moura-Bueno et al., 2020; Santana et al., 2018; Silva et al., 2019). The prediction of clay and sand content outperformed estimation of other soil attributes because they reflect the direct influence of soil mineralogy on soil spectral patterns within the 350–2500 nm range (Lacerda et al., 2016).

Kuang and Mouazen (2011) and Wetterlind and Stenberg (2010) found the standard deviation and range of observed values of soil attributes—that explained the variability in sets from 70 to 205 samples— influenced the performance of prediction models for clay, silt, sand, organic carbon, and pH. The authors also described that a larger standard deviation and wider range of soil attributes resulted in larger R^2 and RPD (ratio of performance to deviation) values, but also larger RMSE values. Larasati et al. (2018) reported that the effect of kurtosis level of the data significantly affected the performance of the artificial neural network predictive model. Such misestimation often occurs when soil samples are under-represented in the tails of the soil attribute data distribution (Stevens et al., 2013). Furthermore, the measurement error of laboratories—using traditional analytical methods— may have a larger impact on soil modelling than the acquisition and processing of spectral data (Horst-Heinen et al., 2021).

Gogé et al. (2014) and Lucà et al. (2017) collected between 144 and 216 soil samples to assess the influence of calibration set size and found that soil predictions—for clay, organic carbon, CEC, iron and $CaCO_3$ based on spectra— became accurate when at least 40–72 samples were used to train the algorithms, whereas Araújo et al. (2014) reported model improvements with smaller subsets when using a large dataset (≈7,000 samples). These studies support our choice of using the full set of samples (≈200) from each laboratory—instead of smaller sets (≈33) from spectral clusters— to model soil attributes. When the sample set is small, the model would perform better using all samples, while for large sample sets, clustering samples would improve model performance.

Therefore, studies on soil spectral modelling—as the present work— have potential to be used in laboratory routines, and may reach even more significant levels of applicability and robustness with their popularization (Souza et al., 2016). The application of proximal soil sensing as a cost-effective tool in routine analysis is hence essential for the adoption and implementation of precision agriculture (Nocita et al., 2015; Viscarra Rossel and McBratney, 1998).

4.4.5. Guidelines for the use of spectra in laboratory routines

Most studies based on soil spectra do not focus on bridging the gap between soil spectroscopy and their real implementation in laborory routines. With the development of the present work, we attempted to provide guidelines about how spectroscopy could be implemented in the routine of soil laboratories. Therefore, users may use the following recommended guidelines.

4.4.5.1. Internal quality control

After spectra acquisition, the soil laboratory can perform a set of procedures for continuous monitoring of their internal quality in order to verify whether the analytical results are reliable enough to be released. The quality control can reach more significant levels of robustness as the spectral library grows daily. Steps:

- Spectra acquisition from soil samples using standard protocols, e.g., Brazilian protocols (Romero et al., 2018), and splicing correction;
- Clustering soil samples, collected from a plot or group of plots of a farm, based on their spectral proximity values. Laboratories could construct a reference library containing samples with soil attributes data and spectra determined with high confidence—e.g., using replicates. This (local or regional) library could be coupled with the samples for clustering, where the library could be used as reference for outlier detection within the spectral clusters in routine analysis.
- Detection of outlying analytical results within each spectral cluster using a method (auto) adjusted to the data distribution;
- Discard outlying values and resubmit soil sample for analytical determination;
- Repeat clustering and outlier detection procedures until the desired levels of quality are obtained.

4.4.5.2. Samples selection for analysis and prediction

When the soil samples arrive at the laboratory, the analyst does not know any information about the soils. Then, the specialist needs to determine which and how many samples should be submitted to analytical methods and which and how many to predict. Steps:

- Spectra acquisition from soil samples using standard protocols, e.g., Brazilian protocols (Romero et al., 2018), and splicing correction;
- Performing unsupervised spectral classification of soil samples using Random Uniform Forest and computing proximity matrices and wavelength importance;
- Ordering soil samples based on the mean reflectance from the most important wavelengths or their spectral proximity values.
- Systematic sampling of samples (after ordering) for splitting into training and testing sets. The optimal size of the subsets should be explored by the laboratory using their existing database for each region served (e.g., agricultural zone, municipality, or geographical region).
- Submitting soil samples from the training set to analytical methods.
- Obtaining the model performance of the training set (with analytical results) for each soil attribute. First, systematically split the whole training set in two smaller subsets for training and testing the prediction algorithms, and then, calculate the performance metrics for the model. It will provide a reference about the accuracy of the attribute predictions that will be made in the following steps using the whole testing set (for which there is no analytical results).
- Predicting soil attributes for the testing set by applying the fitted prediction model.

4.4.6. Future recommendations

For future studies we recommend to: i) re-analyze samples detected as outliers using soil analytical methods, ii) study different soil attributes, iii) apply different spectral preprocessing methods, iv) investigate different methods for data stratification (e.g., environmental variables such as soil class, biomes, parent material, etc.), v) test the proposed approach using data obtained within the medium-infrared spectral range.

4.5. CONCLUSIONS

Internal quality control based on soil spectra and unsupervised analysis can be implemented for laboratory routines. Spectra-based soil samples selection and attributes predictions can reduce the need for laboratory analysis by half. More importantly, the use of spectra in laboratory routines can benefit countries even if they have limited—or great— funding or laboratory resources. Soil spectroscopy is complementary to traditional laboratory methods and can become an important alternative to improve their results, save time, reduce costs, and mitigate environmental impact with acceptable accuracy.

ACKNOWLEDGMENTS

We thank the laboratories that participated in the Brazilian Program of Soil Analysis via Spectroscopy—ProBASE (https://esalqgeocis.wixsite.com/geocis/probase) at the University of São Paulo, Brazil, for providing samples and soil data. We also thank the GeoCiS research group (https://esalqgeocis.wixsite.com/english) for the technical support.

REFERENCES

Afanador, N.L., Smolinska, A., Tran, T.N., Blanchet, L., 2016. Unsupervised random forest: a tutorial with case studies. J. Chemom. 30, 232–241. https://doi.org/https://doi.org/10.1002/cem.2790

Agbenin, J.O., Cantarella, H., 2011. Performance of commercial soil laboratories in a proficiency test program in Brazil. Accredit. Qual. Assur. 16, 553. https://doi.org/10.1007/s00769-011-0814-x

Amundson, R., Berhe, A.A., Hopmans, J.W., Olson, C., Sztein, A.E., Sparks, D.L., 2015. Soil and human security in the 21st century. Science (80-.). 348, 1261071–1261071. https://doi.org/10.1126/science.1261071

Anas, M.S., Ahmed, Y.A., Yusuf, S., Yusuf, J.A., 2016. Assessment of Laboratory Performance Evaluation in Determining Al, Fe, and N Content on Some Soil Samples Based on Soil Analytical Method using Youden Plot and Ranking Test. J. Nucl. Energy Sci. Power Gener. Technol. 05. https://doi.org/10.4172/2325-9809.1000151

Araújo, S.R., Wetterlind, J., Demattê, J.A.M., Stenberg, B., 2014. Improving the prediction performance of a large tropical vis-NIR spectroscopic soil library from Brazil by clustering into smaller subsets or use of data mining calibration techniques. Eur. J. Soil Sci. 65, 718–729. https://doi.org/https://doi.org/10.1111/ejss.12165

Ben Dor, E., Ong, C., Lau, I.C., 2015. Reflectance measurements of soils in the laboratory: Standards and protocols. Geoderma 245–246, 112–124. https://doi.org/http://dx.doi.org/10.1016/j.geoderma.2015.01.002

Bouma, J., 2020. Soil security as a roadmap focusing soil contributions on sustainable development agendas. Soil Secur. 1, 100001. https://doi.org/10.1016/j.soisec.2020.100001

Breiman, L., 2001. Random forests 45, 5–32. https://doi.org/10.1023/A:1010933404324

Brys, G., Hubert, M., Struyf, A., 2004. A Robust Measure of Skewness. J. Comput. Graph. Stat. 13, 996–1017. https://doi.org/10.1198/106186004X12632

Chen, S., Xu, H., Xu, D., Ji, W., Li, S., Yang, M., Hu, B., Zhou, Y., Wang, N., Arrouays, D., Shi, Z., 2021. Evaluating validation strategies on the performance of soil property prediction from regional to continental spectral data. Geoderma 400, 115159. https://doi.org/https://doi.org/10.1016/j.geoderma.2021.115159

Ciss, S., 2015a. Random Uniform Forests.

Ciss, S., 2015b. randomUniformForest: random Uniform Forests for Classification, Regression and Unsupervised Learning.

Clark, R.N., King, T.V. V, Klejwa, M., Swayze, G.A., Vergo, N., 1990. High spectral resolution reflectance spectroscopy of minerals. J. Geophys. Res. Solid Earth 95, 12653–12680. https://doi.org/10.1029/JB095iB08p12653

Clingensmith, C.M., Grunwald, S., Wani, S.P., 2019. Evaluation of calibration subsetting and new chemometric methods on the spectral prediction of key soil properties in a data-limited environment. Eur. J. Soil Sci. 70, 107–126. https://doi.org/10.1111/ejss.12753

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

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

Demattê, J.A.M., Epiphanio, J.C.N., Formaggio, A.R., 2003. Influência da matéria orgânica e de formas de ferro na reflectância de solos tropicais. Bragantia 62, 451–464. https://doi.org/10.1590/S0006-87052003000300012

FAO, 2021. Spectroscopy [WWW Document]. Spectroscopy. URL http://www.fao.org/global-soil-partnership/glosolan/soil-analysis/dry-chemistry-spectroscopy/en/ (accessed 10.9.21).

Gogé, F., Gomez, C., Jolivet, C., Joffre, R., 2014. Which strategy is best to predict soil properties of a local site from a national Vis–NIR database? Geoderma 213, 1–9. https://doi.org/10.1016/j.geoderma.2013.07.016

Gower, J.C., 1966. Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis. Biometrika 53, 325. https://doi.org/10.2307/2333639

Guerrero, A., Bertsch, F., 2020. Primer informe del ejercicio de intercomparación de la red latinoamericana de laboratorios de suelos (LATSOLAN). FAO, Rome, Italy. https://doi.org/10.4060/ca9251es

Guimaraes, C.C.B., A. M. Demattê, J., Carlos de Azevedo, A., Simão Diniz Dalmolin, R., ten Caten, A., Sayão, V.M., Cipriano da Silva, R., Poppiel, R.R., Mendes, W. de S., Urbina Salazar, D.F., Barros e Souza, A., 2021. Soil weathering behavior assessed by combined spectral ranges: Insights into aggregate analysis. Geoderma 402, 115154. https://doi.org/https://doi.org/10.1016/j.geoderma.2021.115154

Hartmann, C., Suvannang, N., 2019. Global Soil Laboratory Assessment. Rome, Italy.

Horst-Heinen, T.Z., Dalmolin, R.S.D., Samuel-Rosa, A., Grunwald, S., 2021. The interplay among analytical method, preprocessing, and modeling on soil organic carbon Vis-NIR-SWIR predictions, in: EGU General Assembly 2021. EGU General Assembly 2021, online, pp. EGU21-7851. https://doi.org/10.5194/egusphere-egu21-7851

Hubert, M., Vandervieren, E., 2008. An adjusted boxplot for skewed distributions. Comput. Stat. Data Anal. 52, 5186–5201. https://doi.org/https://doi.org/10.1016/j.csda.2007.11.008

Hund, E., Massart, D.L., Smeyers-Verbeke, J., 2000. Inter-laboratory studies in analytical chemistry. Anal. Chim. Acta 423, 145–165. https://doi.org/10.1016/S0003-2670(00)01115-6

Khaledian, Y., Miller, B.A., 2020. Selecting appropriate machine learning methods for digital soil mapping. Appl. Math. Model. 81, 401–418. https://doi.org/10.1016/j.apm.2019.12.016

Kuang, B., Mouazen, A.M., 2011. Calibration of visible and near infrared spectroscopy for soil analysis at the field scale on three European farms. Eur. J. Soil Sci. 62, 629–636. https://doi.org/10.1111/j.1365-2389.2011.01358.x

Kuhn, M., Quinlan, J.R. (John R., 2021. Cubist: Rule- And Instance-Based Regression Modeling. R package version 0.3.0.

Lacerda, M., Demattê, J., Sato, M., Fongaro, C., Gallo, B., Souza, A., 2016. Tropical Texture Determination by Proximal Sensing Using a Regional Spectral Library and Its Relationship with Soil Classification. Remote Sens. 8, 701. https://doi.org/10.3390/rs8090701

Larasati, A., Dwiastutik, A., Ramadhanti, D., Mahardika, A., 2018. The effect of Kurtosis on the accuracy of artificial neural network predictive model. MATEC Web Conf. 204, 02018. https://doi.org/10.1051/matecconf/201820402018

Leeuwen, C.C.E., Mulder, V.L., Batjes, N.H., Heuvelink, G.B.M., 2021. Statistical modelling of measurement error in wet chemistry soil data. Eur. J. Soil Sci. n/a, ejss.13137. https://doi.org/10.1111/ejss.13137

Libohova, Z., Seybold, C., Adhikari, K., Wills, S., Beaudette, D., Peaslee, S., Lindbo, D., Owens, P.R., 2019. The anatomy of uncertainty for soil pH measurements and predictions: Implications for modellers and practitioners. Eur. J. Soil Sci. 70, 185–199. https://doi.org/https://doi.org/10.1111/ejss.12770

Lucà, F., Conforti, M., Castrignanò, A., Matteucci, G., Buttafuoco, G., 2017. Effect of calibration set size on prediction at local scale of soil carbon by Vis-NIR spectroscopy. Geoderma 288, 175–183. https://doi.org/10.1016/j.geoderma.2016.11.015

McBratney, A., de Gruijter, J., Bryce, A., 2019. Pedometrics timeline. Geoderma 338, 568–575. https://doi.org/10.1016/j.geoderma.2018.11.048

Mendes, W. de S., Boechat, C.L., Gualberto, A.V.S., Barbosa, R.S., Silva, Y.J.A.B. da, Saraiva, P.C., Sena, A.F.S. de, Duarte, L. de S.L., 2021. Soil spectral library of Piauí State using machine learning for laboratory analysis in Northeastern Brazil. Rev. Bras. Ciência do Solo 45. https://doi.org/10.36783/18069657rbcs20200115

Mountier, N.S., Griggs, J.L., Oomen, G.A.C., 1966. Sources of error in advisory soil tests. New Zeal. J. Agric. Res. 9, 328–338. https://doi.org/10.1080/00288233.1966.10420784

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

Nawar, S., Buddenbaum, H., Hill, J., Kozak, J., Mouazen, A.M., 2016. Estimating the soil clay content and organic matter by means of different calibration methods of vis-NIR diffuse reflectance spectroscopy. Soil Tillage Res. 155, 510–522. https://doi.org/https://doi.org/10.1016/j.still.2015.07.021

Nawar, S., Mouazen, A.M., 2018. Optimal sample selection for measurement of soil organic carbon using on-line vis-NIR spectroscopy. Comput. Electron. Agric. 151, 469–477. https://doi.org/10.1016/j.compag.2018.06.042

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

Quinlan, J.R. (John R., 1992. Learning with continuous classes, in: 5th Australian Joint Conference on Artificial Intelligence. pp. 343–348.

R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Ramirez-Lopez, L., Behrens, T., Schmidt, K., Rossel, R.A.V., Demattê, J.A.M., Scholten, T., 2013a. Distance and similarity-search metrics for use with soil vis–NIR spectra. Geoderma 199, 43–53. https://doi.org/10.1016/j.geoderma.2012.08.035

Ramirez-Lopez, L., Behrens, T., Schmidt, K., Stevens, A., Demattê, J.A.M., Scholten, T., 2013b. The spectrum-based learner: A new local approach for modeling soil vis–NIR spectra of complex datasets. Geoderma 195–196, 268–279. https://doi.org/https://doi.org/10.1016/j.geoderma.2012.12.014

Romero, D.J., Ben-Dor, E., Demattê, J.A.M., Souza, A.B. e, Vicente, L.E., Tavares, T.R., Martello, M., Strabeli, T.F., da Silva Barros, P.P., Fiorio, P.R., Gallo, B.C., Sato, M.V., Eitelwein, M.T., 2018. Internal soil standard method for the Brazilian soil spectral library: Performance and proximate analysis. Geoderma 312, 95–103. https://doi.org/10.1016/j.geoderma.2017.09.014

Santana, F.B., de Souza, A.M., Poppi, R.J., 2018. Visible and near infrared spectroscopy coupled to random forest to quantify some soil quality parameters. Spectrochim. Acta Part A Mol. Biomol. Spectrosc. 191, 454–462. https://doi.org/https://doi.org/10.1016/j.saa.2017.10.052

Schaefer, C.E.G.R., Fabris, J.D., Ker, J.C., 2008. Minerals in the clay fraction of Brazilian Latosols (Oxisols): a review. Clay Miner. 43, 137–154. https://doi.org/10.1180/claymin.2008.043.1.11

Silva, E.B., Giasson, É., Dotto, A.C., Caten, A. ten, Demattê, J.A.M., Bacic, I.L.Z., Veiga, M. da, 2019. A Regional Legacy Soil Dataset for Prediction of Sand and Clay Content with Vis-Nir-Swir, in Southern Brazil. Rev. Bras. Ciência do Solo 43. https://doi.org/10.1590/18069657rbcs20180174

Souza, A.M., Filgueiras, P.R., Coelho, M.R., Fontana, A., Winkler, T.C.B., Valderrama, P., Poppi, R.J., 2016. Validation of the near Infrared Spectroscopy Method for Determining Soil Organic Carbon by Employing a Proficiency Assay for Fertility Laboratories. J. Near Infrared Spectrosc. 24, 293–303. https://doi.org/10.1255/jnirs.1219

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

Stevens, A., Ramirez-Lopez, L., 2013. prospectr: Processing and sample selection for vis-NIR spectral data.

Suvannang, N., Hartmann, C., 2019. First Inter-laboratory Comparison Report of the Regional Soil Laboratory Network for Asia (SEALNET). FAO, Rome, Italy.

Teixeira, P.C., Donagemma, G.K., Fontana, A., Teixeira, W.G., 2017. Manual de métodos de análise de solo, 3a edição. ed. Embrapa Solos, Brasilia, DF.

Tibshirani, R., Walther, G., Hastie, T., 2001. Estimating the number of clusters in a data set via the gap statistic. J. R. Stat. Soc. Ser. B (Statistical Methodol. 63, 411–423. https://doi.org/https://doi.org/10.1111/1467-9868.00293

Viscarra Rossel, R.A., Behrens, T., Ben-Dor, E., Brown, D.J., Demattê, J.A.M., Shepherd, K.D., Shi, Z., Stenberg, B., Stevens, A., Adamchuk, V., Aïchi, H., Barthès, B.G., Bartholomeus, H.M., Bayer, A.D., Bernoux, M., Böttcher, K., Brodský, L., Du, C.W., Chappell, A., Fouad, Y., Genot, V., Gomez, C., Grunwald, S., Gubler, A., Guerrero, C., Hedley, C.B., Knadel, M., Morrás, H.J.M., Nocita, M., Ramirez-Lopez, L., Roudier, P., Campos, E.M.R., Sanborn, P., Sellitto, V.M., Sudduth, K.A., Rawlins, B.G., Walter, C., Winowiecki, L.A., Hong, S.Y., Ji, W., 2016. A global spectral library to characterize the world’s soil. Earth-Science Rev. 155, 198–230. https://doi.org/http://dx.doi.org/10.1016/j.earscirev.2016.01.012

Viscarra Rossel, R.A., McBratney, A.B., 1998. Soil chemical analytical accuracy and costs: implications from precision agriculture. Aust. J. Exp. Agric. 38, 765. https://doi.org/10.1071/EA97158

Wadoux, A.M.J.-C., Heuvelink, G.B.M., Lark, R.M., Lagacherie, P., Bouma, J., Mulder, V.L., Libohova, Z., Yang, L., McBratney, A.B., 2021. Ten challenges for the future of pedometrics. Geoderma 401, 115155. https://doi.org/10.1016/j.geoderma.2021.115155

Wetterlind, J., Stenberg, B., 2010. Near-infrared spectroscopy for within-field soil characterization: small local calibrations compared with national libraries spiked with local samples. Eur. J. Soil Sci. 61, 823–843. https://doi.org/10.1111/j.1365-2389.2010.01283.x

Zeng, R., Zhang, J.P., Cai, K., Gao, W.C., Pan, W.J., Jiang, C.Y., Zhang, P.Y., Wu, B.W., Wang, C.H., Jin, X.Y., Li, D.C., 2021. How similar is "similar," or what is the best measure of soil spectral and physiochemical similarity? PLoS One 16, e0247028. https://doi.org/10.1371/journal.pone.0247028

5. THE BRAZILIAN SOIL SPECTRAL SERVICE (BRAZ3): AN EASY SYSTEM FOR WORLD WIDE COMMUNITY

ABSTRACT

We are in the Soil Spectral Libraries (SSLs) era but most of the soil information available to land users (farmers, researchers, and policy-makers) is still obtained by traditional soil analysis, which is becoming non-environmentally friendly and cannot go along with the world's demand. However, soil scientists have been using spectroscopy techniques to obtain massive reliable soil data, their usage is still limited to the academic arena. In order to provide fast and credible soil data, the online Brazilian Soil Spectral Service (Braz3) for estimating soil properties was developed using the Brazilian Soil Spectral Library (BSSL). BSSL has 49,753 and 4,751 soil spectra from the Visible–Near–Short Wave Infrared (vis–NIR–SWIR; 350–2,500 nm) and the Mid-Infrared (MIR; 2,500–20,000 nm), respectively. The platform provides an interactive way to find researchers and data, and to predict several soil properties at the Brazilian territory from which we focus on clay and Soil Organic Carbon (SOC) contents estimation. The system was tested by 500 Brazilian users as well as users invited from 65 countries. The users accessed the online platform (besbbr.com.br) and followed the instructions, which basically consisted of uploading the soil spectra in a defined format. After a few minutes, they automatically received their results (predictions of SOC and clay content) by email. The Braz3 provided good results for the Brazilian users and variables for other countries. When overseas users made their Local Models (LMs, local dataset), the result was better. Clay content was the best predicted property in all cases with $R^2 > 0.5$. MIR generally presented higher R^2 values than vis–NIR–SWIR for both soil properties. The online system showed to be a reliable service that works properly. Users experienced how spectra can easily deliver soil analysis in almost a real-time situation. Afterwards, different models were developed outside the system. We merged the BSSL with the soil spectra from other countries and generated a World Soil Spectral Library (WSSL) and a global model. We observed that LMs were the best followed by the WSSL and the BSSL. This trend suggests a need to keep expanding online SSLs, in particular a WSSL, but with a dynamic system in which users can choose the population to run the model and seek for the best local one (for their own purpose: level, scale). Receiving soil analysis as a service, without making and testing models, can improve the interest of end-users for this powerful technique.

Keywords: Spectral service, soil analysis, spectral library, spectroscopy, environment, soil quality, precision agriculture.

5.1. INTRODUCTION

Soil is an important component of the environment as it offers crucial services such as food production, water cleaning, and carbon sequestration (Lal et al., 2021). To achieve sustainable use of soil resources, it is crucial that the world's soil community interact and seek reliable methods for obtaining soil information. So far, traditional soil laboratory analysis has been the most common way to obtain soil data, but it is non-environmentally friendly and becomes expensive when large amounts of soil samples need to be analyzed (Viscarra Rossel and Mcbratney, 1998). This is especially important in developing countries, where farmers either do not carry out soil analysis due to high costs or they are unaware of its importance and work in the 'darkness' related to soil management. Despite the disadvantages, the traditional laboratory analysis is, and will continue to be, the most suitable way to obtain soil data. However, alternatives approches such as soil spectroscopy have proved as a suitable way to optimize soil analysis and disseminate its use to all interested parties.

The interaction of soil components with electromagnetic energy, which is the basis of soil spectroscopy, is well-documented in the literature (Viscarra Rossel et al., 2009, Soriano Disla et al., 2014, Nocita et al., 2013). Understanding of these interactions provided researchers with the tools to search for a global communication system based on the so-called Soil Spectral Libraries (SSLs). The first publication using a SSL with global samples was presented by Stoner and Baugarnder (1981), followed by Brown et al. (2006), and Viscarra Rossel et al. (2016), with participation of 92 countries. Other regionals initiatives came along such as the ICRAF-ISRIC Soil VNIR Spectral Library (Garrity and Bindraban, 2004), the LUCAS framework (Land Use/Cover Area Frame Survey; http://eusoils.jrc.ec.europa.eu/projects/Lucas) (Orgiazzi et al., 2018) with data from 23 countries in Europe (Stevens et al., 2013), and the GEOCRADLE with samples from nine countries in Balkan, Middle East, North and central Africa (Tziolas et al., 2019; Shepherd and Walsh, 2002; Summerauer et al., 2021). In addition, numerous countries have also developed local SSLs, such as the Brazilian Soil Spectral Library (BSSL) (Demattê et. al., 2018, Bellinaso et al., 2010), and those from the Czech Republic (Brodsky et al., 2011), France (Gogé et al., 2012), Denmark (Knadel et al., 2012), Mozambique (Cambule et al., 2012), Spain (Bas et al., 2013), Australia (Viscarra Rossel and Webster, 2011), China (Shi et al., 2014; Ji et al., 2016; Liu et al., 2018), USA (Condit, 1970; Wijewardane et al., 2018), New Zealand (Baldock et al., 2019), and Tajikistan (Hergarten et al. 2013).

These SSLs are important initiatives but all of them were only available in scientific journals. Other initiatives made spectra available to users, but who are the spectra users? This basic data still needs complex processing, which is certainly accessible to researchers but not to the general public. To make an analogy, it would be like when satellite imagery was available for free for the first time. Most users were unable to use them due to pre-processing issues (e.g., atmospheric correction and georeferencing) (Bunting, 2017). This shortage was eliminated, when images were made available already pre-processed and georeferenced. Nowadays, we have an analogous situation, where many SSLs are available worldwide but end-users (farmers) cannot see their importance. As a first step on a learning curve, why not start delivering the product directly to users? This would not interrupt research on this topic, although rather will boost it.

Taking this into account, in this study, we present a free web-based online platform for soil properties prediction using Visible–Near–Short Wave Infrared (vis–NIR–SWIR) and Mid-Infrared (MIR) spectral ranges, and as background the BSSL. The system was tested by users from Brazil and 65 other countries. After understanding the valuable use of an online system, we evaluated other modeling with locals and a world-wide dataset for comparison with the BSSL.

We expected that the model works better for Brazil than for the other countries as the World Soil Spectral Library (WSSL) brings light to all cases, but local ones would be the best. Since this is the first initiative of this kind, the main idea is pedagogic, with the objective that users start to understand how soil spectra can assist their needs. This initiative is not free from uncertainties and has its limitations; however, we hope to bring to light a new and fast generation of online communication in soil analysis.

5.2. MATERIAL AND METHODS

5.2.1. The Brazilian Soil Spectral Service (Braz3) construction

We developed an online platform denominated The Brazilian Soil Spectral Service (Braz3) with support of the Geotechnologies in Soil Science Group (GEOCIS, https://esalqgeocis.wixsite.com/english) Laboratory at the Luiz de Queiroz College of Agriculture (ESALQ), University of São Paulo (USP). Braz3 is divided into three complementary modules: data localization, soil data visualization, and soil processing and quantification (Figure 1a). In the data locations module, the user visualizes the number of samples by state and identifies the authors and partner institutions. The soil data visualization module shows soil spectra in the Vis, NIR, SWIR, and MIR bands filtered by classifications, orders, groups, layers, and textures.

All spectra and models are kept inside the system and are not disclosed. Scripts for modeling are in the backend of the system so the user only needs to choose the desired properties to quantify. The web platform, which can be accessed in http://besbbr.com.br/, provides several services inserted in the results section.

This system was prepared only for the BSSL as an experimental phase. The predictive models were prepared in R software (R Development Core Team, Vienna, Austria, 2020) and then inserted into the web interface created in JavaScript. The baseline of the system functioning is illustrated in Figure 1.

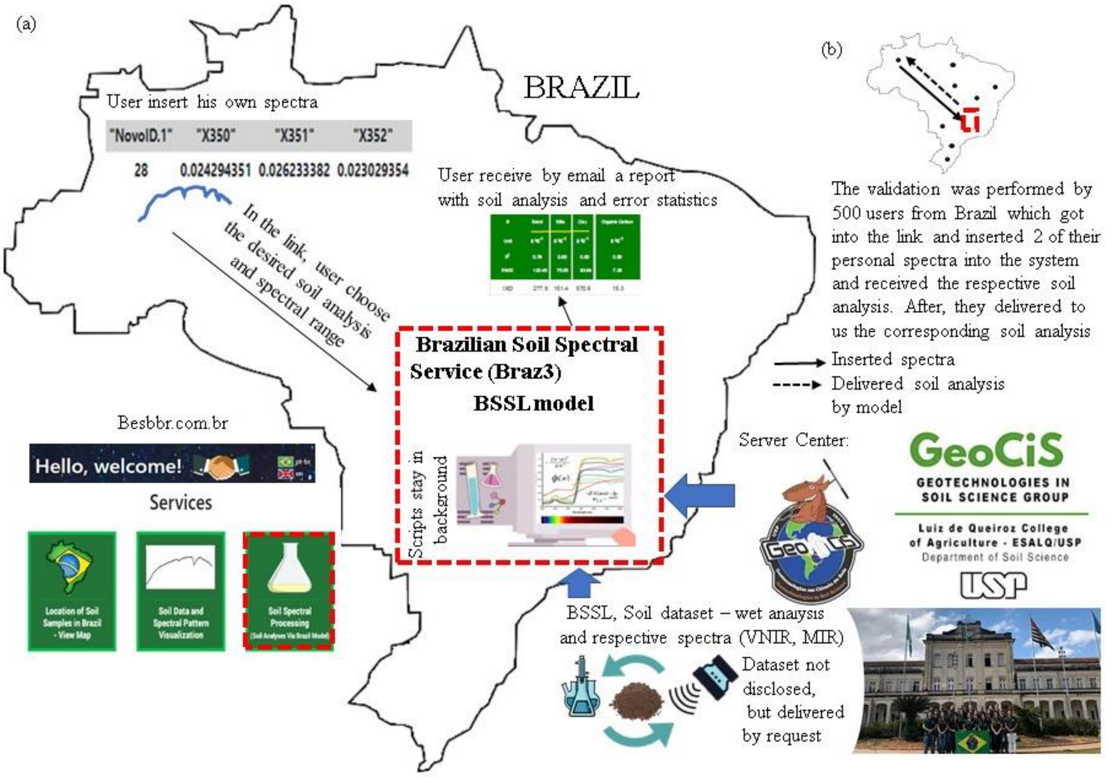

(a)
BRAZIL
User insert his own spectra
"NovoID.1" "X350" "X351" "X352"
28 0.024294351 0.026233382 0.023029354
In the link, user choose the desired soil analysis and spectral range
User receive by email a report with soil analysis and error statistics
Brazilian Soil Spectral Service (Braz3)
BSSL model
Scripts stay in background
Besbbr.com.br
Hello, welcome!
Services
Location of Soil Samples in Brazil - View Map
Soil Data and Spectral Pattern Visualization
Soil Spectral Processing (Soil Analyses Via Brazil Model)
(b)
The validation was performed by 500 users from Brazil which got into the link and inserted 2 of their personal spectra into the system and received the respective soil analysis. After, they delivered to us the corresponding soil analysis
Inserted spectra
Delivered soil analysis by model
Server Center:
GeoCiS
GEOTECHNOLOGIES IN SOIL SCIENCE GROUP
Luiz de Queiroz College of Agriculture - ESALQ/USP
Department of Soil Science
USP
BSSL, Soil dataset – wet anaysis and respective spectra (VNIR, MIR)
Dataset not disclosed, but delivered by request

Figure 1. (a) How the system functions; (b) how the system was tested.

The web server was created using the Apache software (The Apache Software Foundation, Wakefield, MA, USA, 2021) and PHP programming language (Tasneem and Ammar, 2012). Apache is an open-source Hypertext Transfer Protocol (HTTP) server project with the goal of providing a secure, efficient, and extensible web server on HTTP standards. PHP is a fast and flexible scripting language, mainly used in web development (Mitchell, 2016).

The predictive models were carried out in the form of a loadable module in Apache, allowing the server to interpret the R scripts. The prediction of the soil attributes is a task that requires high computational resources (Padarian et al., 2020) and to meet this requirement, we employed tools for organizing the spectral data sent by users in queues and distributing them with other computers. The First-In-First-Out (FIFO) (Manohar and Appaiah, 2017) model was selected and implemented on the server, allowing for a dynamic queue data structure that allows removal and insertion of processing on the server. A high-performance processing cluster was created and thus made it possible to distribute the processes with low-cost computers in the R environment. For the web server, a workstation was acquired with 2 XEON 5120T processor hardware with 14 cores each and a video card with 4000 GPUs, which are essential for the application of the predictive models.

5.2.2. Soil dataset construction

As mentioned, the Braz3 is a service based on the soil dataset from the BSSL (Demattê et al., 2019). Using these data, we constructed the platform with vis–NIR–SWIR, resulting in 49,753 soil samples donated by 81 collaborators, representing 69 institutions from all over the country (https://bibliotecaespectral.wixsite.com/english/lista-de-cedentes). The BSSL in the MIR range comprises 4,951 soil samples.

The BSSL contains laboratory analysis for contents of sand, silt, clay, SOC, calcium, magnesium, potassium, phosphorus, pH, sodium, Fe_2O_3, TiO_2, and MnO as well as respective spectra for vis–NIR–SWIR and MIR regions. In this paper, we focused on clay and SOC. For the vis–NIR–SWIR analysis, the soil samples were dried at 45 °C for 48 hours, ground, sieved with a 2 mm mesh, and homogeneously distributed in petri dishes prior the measurement of the spectra in the 400–2500 nm range (Demattê et al., 2019). The spectral data were acquired using the Fieldspec 3 spectroradiometer (Analytical Spectral Devices, ASD, Boulder, CO, USA). The sampling interval was 1 nm, reporting 2151 channels. The light source was provided by two external 50-W halogen lamps, which were positioned at 35 cm from the sample (non-collimated rays and a zenithal angle of 30°) with an angle of 90° between them. The sensor is calibrated using a white Spectralon plate (Lab-sphere, North Sutton, NH, USA) representing a 100% reflectance standard (reflectance factor 1.0).

For spectral analysis in the MIR, the soil samples were ground and passed through 100 mesh. Reflectance spectra were obtained with the Alpha Sample Compartment RT-DLaTGS ZnSe (Bruker Optik GmbH, Ettlingen, Germany) equipped with an accessory for acquiring Diffuse Reflectance Infrared Fourier Transform (DRIFT). The sensor has a HeNe laser positioned inside the equipment and a calibration pattern for each wavelength. It has a KBr beam allowing a high amplitude of the incident radiance to penetrate the sample. Spectra were acquired between 4000 to 600 cm^{-1} (2,500 – 20,000 nm) with a spectral resolution of 5 cm^{-1} and 32 scans per minute per spectra. A gold reference plate was used as standard, and the sensor was calibrated every four measurements.

5.2.3. Data pre-processing and modeling

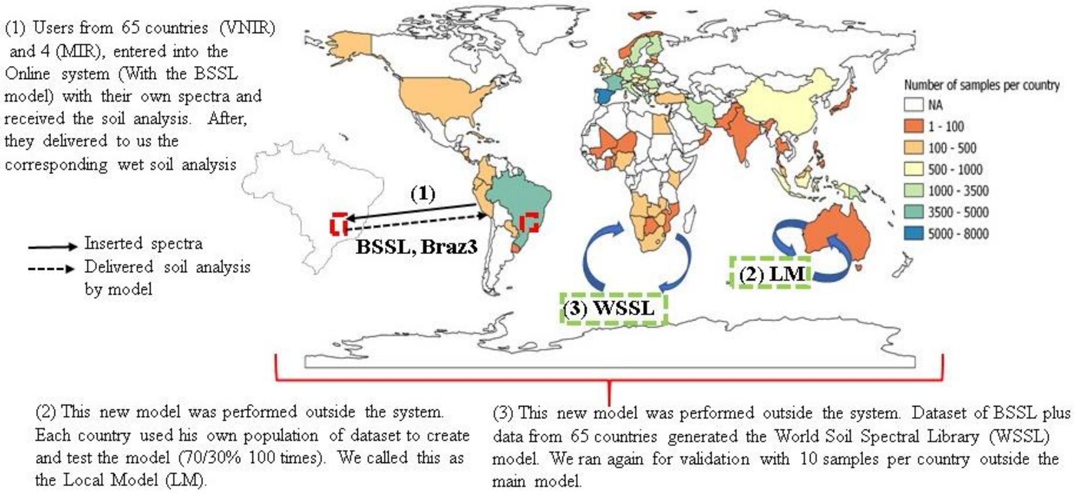
Different pre-processing methods were tested for the vis–NIR–SWIR range and those that presented the best results for each soil property were selected. The Standard Normal Variable (SNV) and Continuum Removal (CR) pre-processing techniques were used to pre-process the spectra for modeling of clay and SOC, respectively. All calculations were performed using the *prospectr* package in R (Stevens and Ramirez-Lopes, 2020). In order to minimize the influence of noise, the ranges from 350 to 420 nm and from 2480 to 2500 nm were removed, and we used the range from 420 to 2480 nm. Finally, we resampled the spectra at a resolution of 10 nm to reduce spectral multicollinearity and the processing time of the site, and to improve the modeling efficiency for this large dataset (Hong-Yan et al., 2009; Zhang et al., 2012; Zhang and Huang, 2019). For the MIR spectral range, the Savitzky-Golay first derivative (SGD) with a first-order polynomial and a window size of 9 nm and SNV were applied.

The datasets for each soil property were randomly split into a calibration (training;70%) and an independent validation (testing; 30%) datasets. The complete BSSL dataset was used to calibrate spectroscopy models using the cubist machine learning algorithm (Quinlan, 1992). Cubist is a rules-based algorithm that applies the M5 (ModelTree) approach to create categorical decision trees to deal with continuous classes. The trees are produced by the algorithm through rules that use boost training. Reinforcement training is based on converting weak learners into strong learners, in addition to giving stronger learners more weight (Khaledian and Miller, 2020). The final model is regulated by a set of nodes along the tree and two hyperparameters (committees and neighbors), which improve the model's performance. The model construction and estimation process were performed by the *caret* package in R (Kuhn, 2020), which has a set of functions that seek to simplify the process of creating predictive models. The criteria used to select the optimal models were Coefficient of Determination (R^2), Ratio of Performance to InterQuartile distance (RPIQ), and Root Mean Square Error (RMSE).

The online system was tested by users with their own spectra, as by 500 Brazilian (two spectra per user, total of 1000 spectra) for the vis–NIR–SWIR and 200 samples for MIR and received the soil analysis results. The real analysis coming from the pattern of the wet chemistry in the laboratory was later delivered to our team, which evaluated the observed and determined data. Users were also invited to make criticism regarding the system for a better improvement.

5.2.4. Exploring other modeling with the dataset

To go forward on possible future activities, we created new insights to evaluate our dataset. This part of the work was not inside the online system, but directly on a computer. We created four types of models for comparison: (a) the BSSL was tested using 28,255 soil samples for vis–NIR–SWIR from 65 countries and 3,488 samples from 4 countries for MIR (in the MIR range there were 8240 samples: 391 from Australia, 170 from Iran, and 2728 from the USA). We entered these spectra in the BSSL model and observed results, (b) we created for each of the 65 countries Local Models (LMs) only with their spectral population and predicted the same soil properties (c) finally, we merged the BSSL with the spectra from the other 65 countries and generated a WSSL. The processing was the same as the one used for BSSL, as it performed by random data split of 70% for model calibration (training) and 30% for validation (testing) purposes and used the same processing as for the BSSL. Finally, we compared the results from the BSSL tested by other countries and compared with the developed LMs; BSSL and WSSL models compared with the same 65 countries. This made it possible to evaluate the differences between global, national, and local datasets on the quantification of soil properties. These dynamics are shown in Figure 2.


(1) Users from 65 countries (VNIR) and 4 (MIR), entered into the Online system (With the BSSL model) with their own spectra and received the soil analysis. After, they delivered to us the corresponding wet soil analysis

(2) This new model was performed outside the system. Each country used his own population of dataset to create and test the model (70/30% 100 times). We called this as the Local Model (LM).

(3) This new model was performed outside the system. Dataset of BSSL plus data from 65 countries generated the World Soil Spectral Library (WSSL) model. We ran again for validation with 10 samples per country outside the main model.

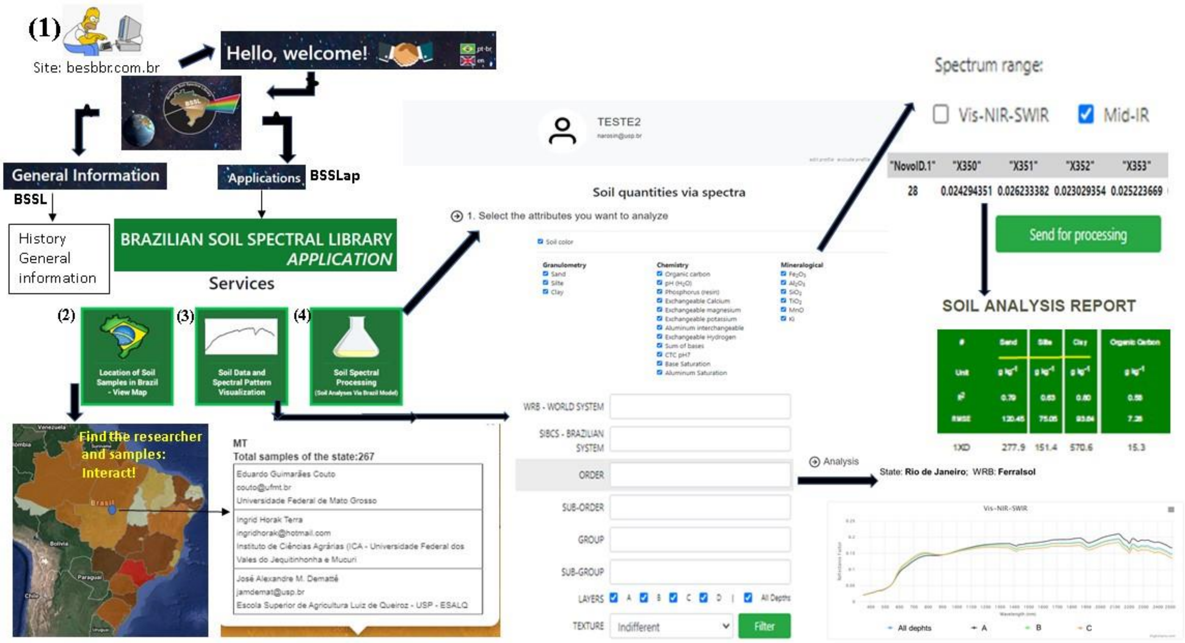
Figure 2. World participants. Exploiting different populations and models to quantify soil properties (1) spectra from 65 countries were tested into the BSSL model, (2) Local model, (3) World Soil Spectral Model.

5.3. RESULTS

5.3.1. Online interaction experience

The web site is available on "*besbbr.com.br*" and brings together spectral information in vis-NIR-SWIR and MIR ranges (Figure 3). This is a user-friendly interface in order to provide a great experience for users. The web is prepared for: a) end users who want the soil analysis, b) for researchers and employees who want to test and evaluate their models, c) for students interested to learn, d) for soil scientists, to test and have new insights, e) for pedologists and soil scientists in all levels to see the soil spectral signature (pattern).

The web site presents the following sequence (Figure 3). Entering the system user makes his registration. Afterwards, he can go to the general information of how the Brazilian Soil Spectral Library was developed and or go directly to BSSL-applied services. The web offers three services as follows (Figure 3): entering in align (2) the user can find the owners of spectral data and its personal information. With this, users can get in contact directly to the owner and ask for this dataset to initiate joint collaboration. Also, the user will see where to find background users on spectroscopy. This idea was performed to stimulate users that want to interact and create new groups. The interactive map of contributors allows one to search for specific institutions or researchers and to visualize in which Brazilian State each one has spectral data. This allows the interaction between researchers and encourages spectral data sharing and partnerships; In align (3) users can have several examples of soil types patterns. He can ask for a specific soil classification, for example a Ferralsol, and the system will show an average of all ferralsols in the dataset or from a specific state going on filtering. In addition to soil classification, users can see patterns of different soil depths and specific soil properties. For example, he can ask for samples in vis-NIR-SWIR of sandy soils at surface depth and in a specific state. The results of the search are the average spectra of samples. Depending on the number of soil spectra falling under specific criteria, the process may take some time. As an example, we selected the SP state, the vis-NIR-SWIR spectra, the sandy textural class from the first layer (A), with no indications for soil classification that takes about 3 minutes.



MT
Total samples of the state:267

Eduardo Guimarães Couto
couto@ufmt.br
Universidade Federal de Mato Grosso

Ingrid Horak Terra
ingridhorak@hotmail.com
Instituto de Ciências Agrárias (ICA - Universidade Federal dos Vales do Jequitinhonha e Mucuri

José Alexandre M. Demattê
jamdemat@usp.br
Escola Superior de Agricultura Luiz de Queiroz - USP - ESALQ

Soil quantities via spectra

1. Select the attributes you want to analyze

☑ Soil color

Granulometry
☑ Sand
☑ Silte
☑ Clay

Chemistry
☑ Organic carbon
☑ pH (H_2O)
☑ Phosphorus (resin)
☑ Exchangeable Calcium
☑ Exchangeable magnesium
☑ Exchangeable potassium
☑ Aluminum interchangeable
☑ Exchangeable Hydrogen
☑ Sum of bases
☑ CTC pH7
☑ Base Saturation
☑ Aluminum Saturation

Mineralogical
☑ Fe_2O_3
☑ Al_2O_3
☑ SiO_2
☑ TiO_2
☑ MnO
☑ Ki

WRB - WORLD SYSTEM
SIBCS - BRAZILIAN SYSTEM
ORDER
SUB-ORDER
GROUP
SUB-GROUP
LAYERS ☑ A ☑ B ☑ C ☑ D | ☑ All Depths
TEXTURE Indifferent
Filter

☐ Vis-NIR-SWIR ☑ Mid-IR

"NovoID.1"	"X350"	"X351"	"X352"	"X353"
28	0.024294351	0.026233382	0.023029354	0.025223669

SOIL ANALYSIS REPORT

#	Sand	Silte	Clay	Organic Carbon
Unit	g kg⁻¹	g kg⁻¹	g kg⁻¹	g kg⁻¹
R²	0.79	0.63	0.80	0.58
RMSE	120.45	75.05	93.84	7.28
1XD	277.9	151.4	570.6	15.3

State: **Rio de Janeiro**; WRB: **Ferralsol**

Figure 3. General flux of the site and its usefulness for users: (1) enter the system, link; (2) service to find the researcher and spectral data; (3) observation of soil classification spectral patterns; (4) quantification of soil properties

Entering in align (4) we have the soil Braz3. The major idea is that the user has spectral data and wants to make a soil analysis in any part of the world. To have access to the prediction module, the user must register in the platform indicating his/her email address to receive the results. After that, the user must log into his/her account in the platform to: 1) download the template model that must be used to organize the soil spectra, 2) upload the file (csv format) with the soil spectra, 3) select the attributes to be predicted, and 4) send the data for processing in the platform. Thus, the user uploads the spectra (regarding on the same spectral wavelengths of the system which is indicated), chooses if by vis-NIR-SWIR or MIR spectral range, chooses the desired soil properties and runs the processing. The system has all scripts on the background which are not disclosed on the web but are presented in this paper. After about fifteen minutes, depending on the filtering and number of samples chosen, the user will receive a report by email. If the system does not respond it is because there is no data in the chosen filter. In the report, will appear all soil analysis of the specific spectra, the method, reference, and statistics. The user also has the option to give feedback online sending the wet soil analysis of his spectra so the system will be uploaded and recharged after an aderency evaluation.

5.3.2. The Brazilian Soil Spectral Service quantification

Figure 4 presents the differences between clay and SOC quantification in the vis-NIR-SWIR using different population models such as the BSSL, the WSSL and the LM. For clay, the predicted content behavior of all models were very similar with observed distribution, with 90% of the population ranging mainly from 150 to 400 g.kg^{-1}. For SOC, the predicted values distribution for WSSL and LM were in accordance with observed values, with 90% of the population ranging mainly from 10 to 180 g.kg^{-1}. On the other hand, the BSSL model underestimated the SOC predicted values.

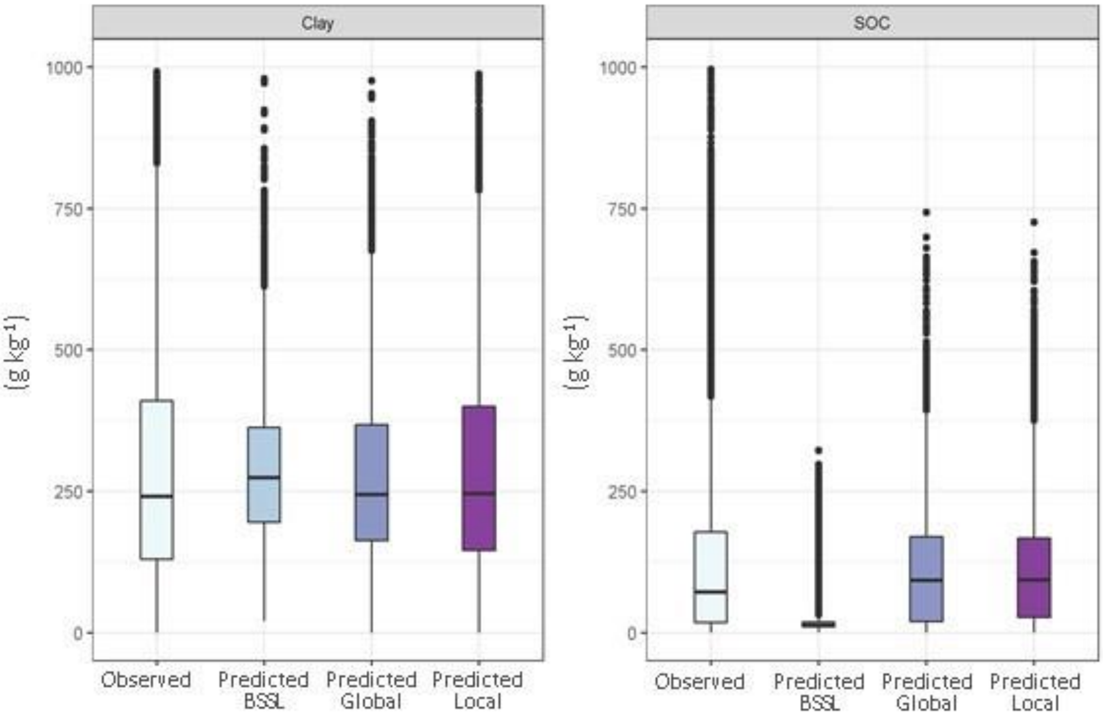


Figure 4. Boxplot Clay and SOC vis-NIR-SWIR

Differences between quantification in the MIR range using different population models, such as BSSL, WSSL and LM (Figure 5) for clay, the predicted content behavior of the WSSL and LM models was very similar to the observed distribution, with 90% of the population ranging mainly from 150 to 400 g.kg^{-1}, but the BSSL model underestimated the contents, with most of the population ranging between 240 and 280 g.kg^{-1}. For SOC, the distribution of predicted values for WSSL and LM agreed with the observed values, with 90% of the population ranging mainly from 2 to 20 g.kg^{-1}. The BSSL model underestimated the predicted SOC values, but the difference was small.

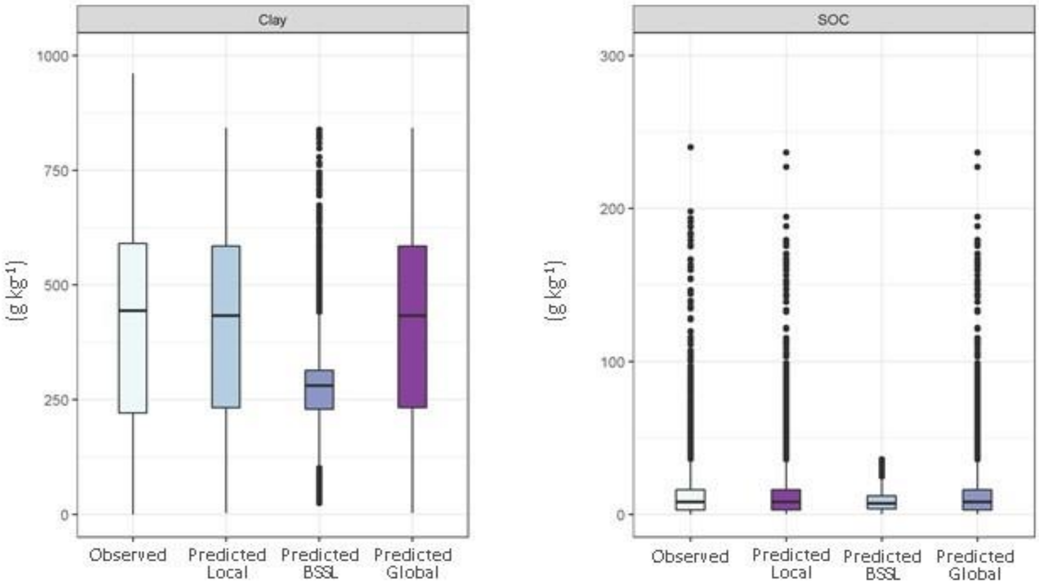


Figure 5. Box plot using MIR dataset per country: (a) Clay and (b) SOC and R^2 validation (c) Clay and (d) SOC by country.

5.3.3. Predictions in function to the population

Figures 6 and 7 presents the results of the vis-NIR-SWIR R^2 model using the different populations per country for clay and SOC respectively, as Figure 8 indicates the RMSE. The local models showed better results, the world model came close behind and BSSL in general had the lowest results, the best result, as expected was for Brazil. In some cases, such as Jamaica and Japan, BSSL presented good results for both clay and SOC.

In relation to RMSE for clay and SOC, BSSL presents the biggest errors, followed by global and LM presents the smallest errors for both elements. However, for SOC, some countries such as Czech Republic, Ireland and Italy stand out for presenting high RMSE for LM.

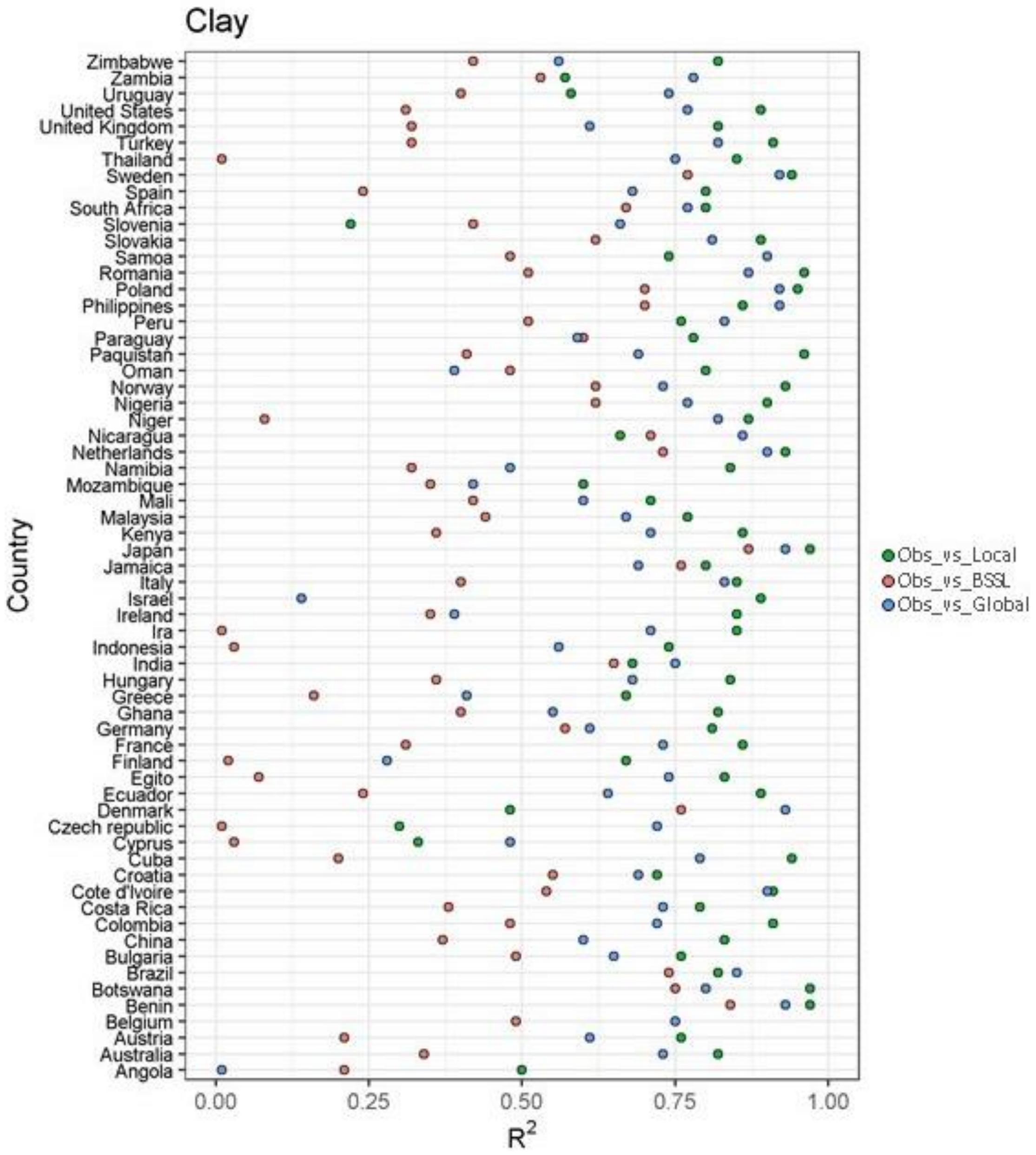


Figure 6. R^2 for clay per country in the vis-NIR-SWIR

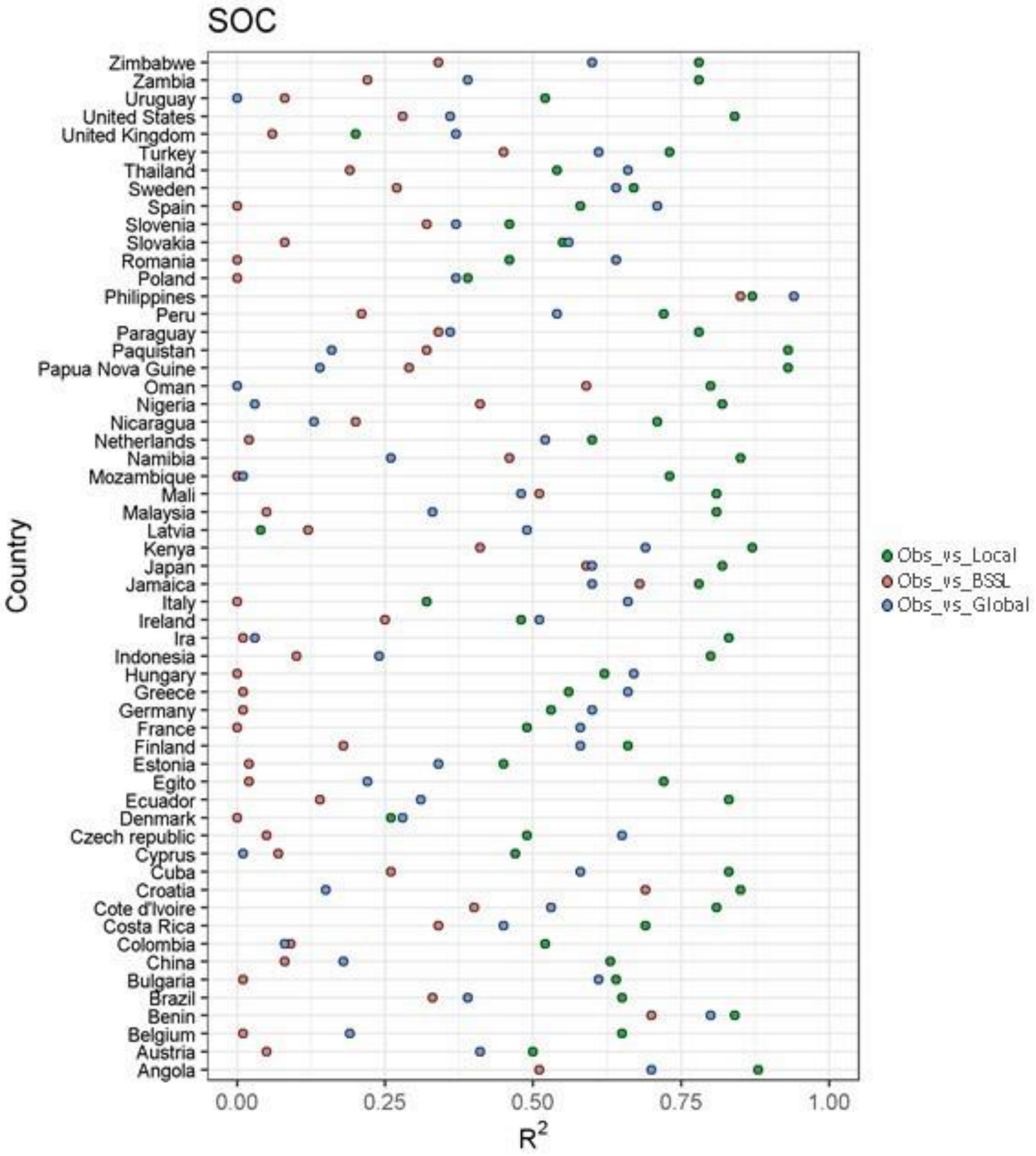


Figure 7. R^2 for SOC per country in the vis-NIR-SWIR.

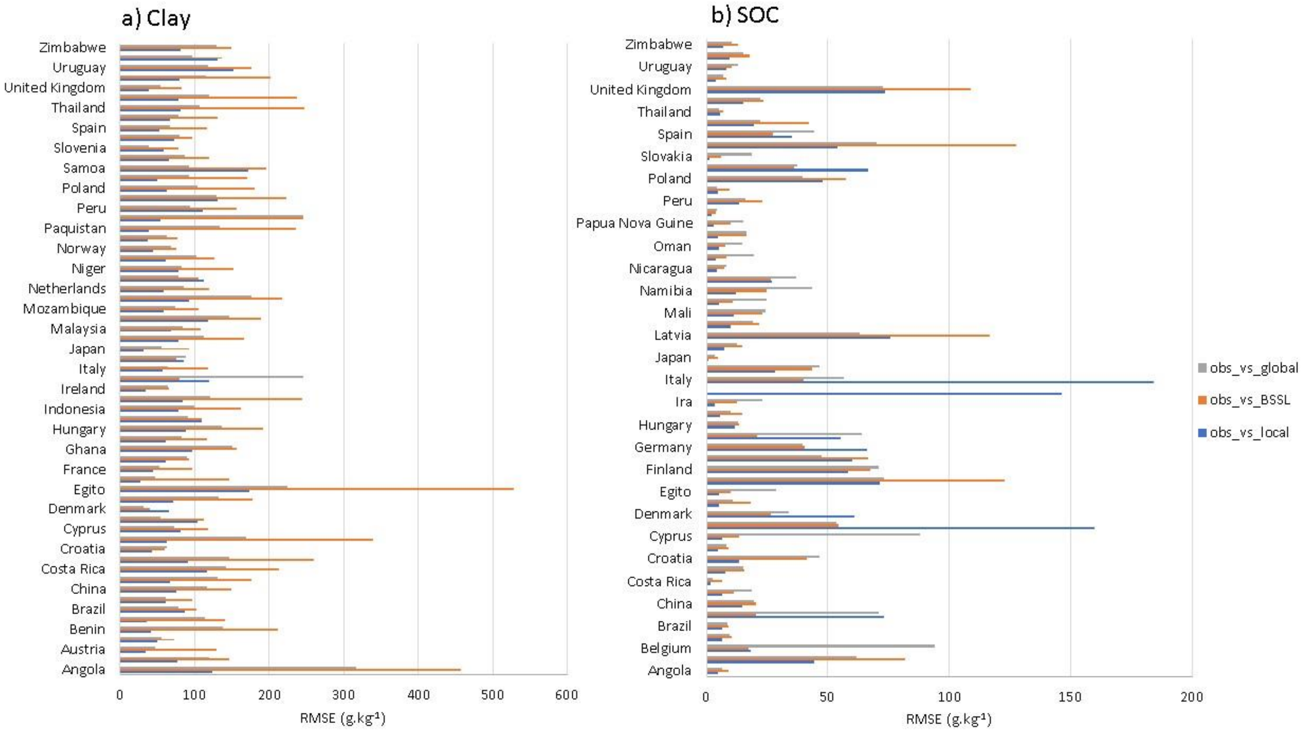


Figure 8. RMSE for clay and SOC per country in the vis-NIR-SWIR.

The R^2 results (Figure 9) for clay followed the pattern of the results obtained for vis-NIR-SWIR, but for SOC they were higher even using BSSL. The RMSE (Figure 10) Global and LM were similar for all countries for both clay and SOC. While the BSSL was the largest among the three, but much smaller than the RMSE obtained by the vis-NIR-SWIR models.

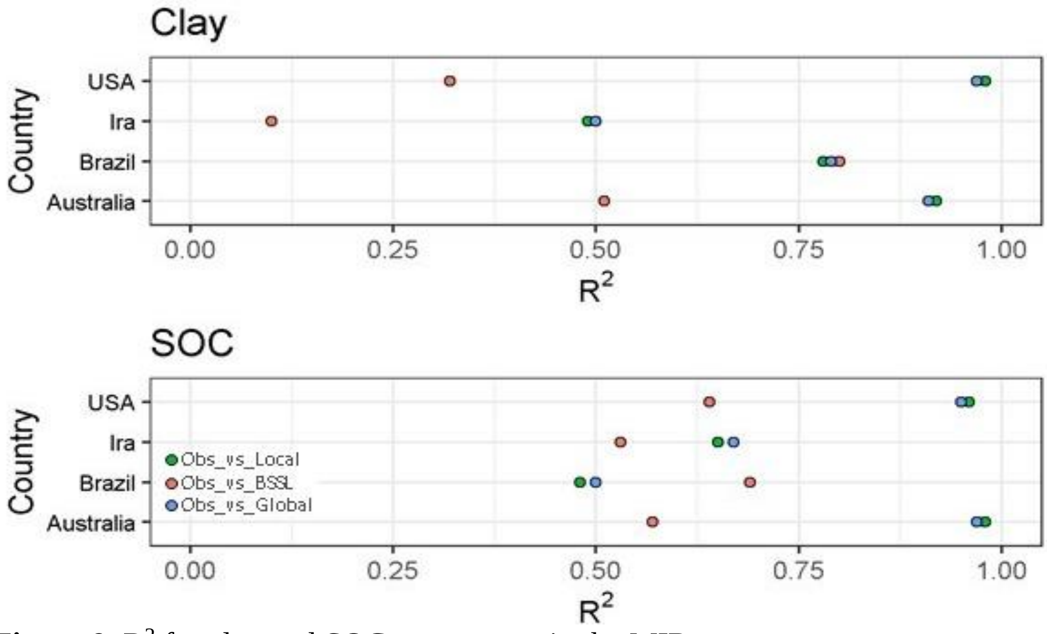


Figure 9. R^2 for clay and SOC per country in the MIR.

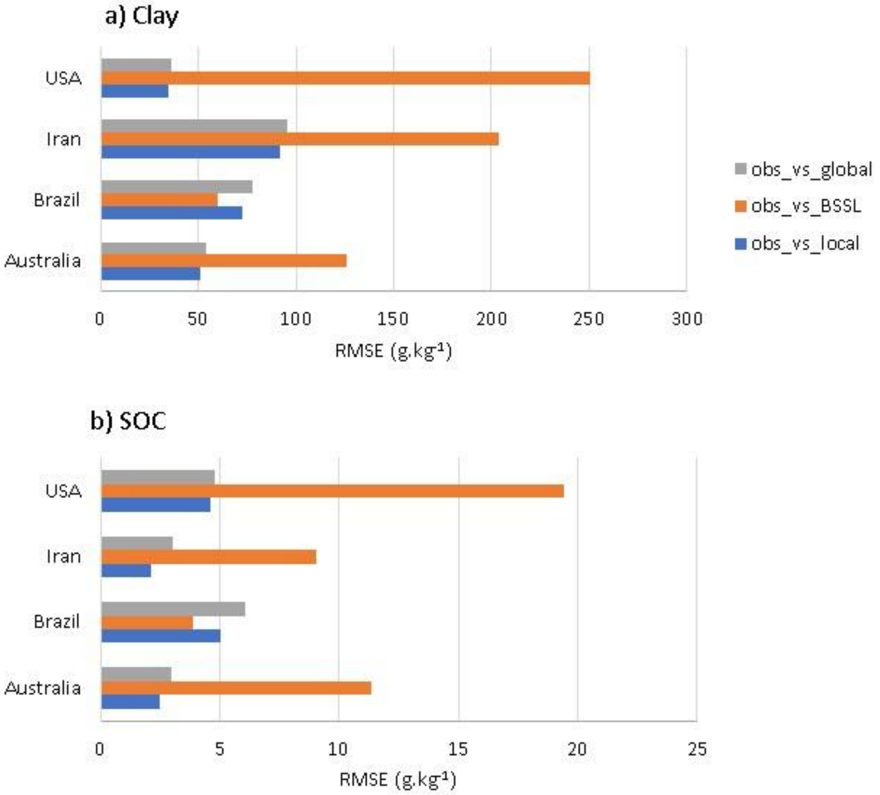


Figure 10. RMSE for clay and SOC per country in the MIR.

The use of a continental country such as Brazil has advantages due to its soil and biomes. Table 1 shows that using the BSSL model 24 countries reached over 0.5 of R^2 for clay. When applied in African countries from 16, 7 had the same score. In fact, many countries from Africa have very similar soils as Brazil. Asia has completely different soils and indeed from 11 only 3 had better results. When we created the WSSL results increased. We had a total of 54 countries over 0.5 against 24 using BSSL. All results increased using WSSL. Going on the other side, when we tested the local models, results increased more where 59 countries were over 0.5.

Table 1. Number of countries with variable R^2 for clay using vis-NIR-SWIR.

Model	R^2	Total	Africa	Asia	Europe	North America	Oceania	South America
LM Clay	0 - 0.3	1	0	0	1	0	0	0
	0.3 -0.5	3	0	0	3	0	0	0
	0.5 -0.7	8	3	1	2	1	0	1
	0.7 - 0.8	11	1	2	4	1	1	2
	> 0.8	40	11	9	13	3	1	3
BSSL Clay	0 - 0.3	14	3	3	6	1	0	1
	0.3 -0.5	25	6	5	8	2	2	2
	0.5 -0.7	12	4	1	5	0	0	2
	0.7 - 0.8	9	1	1	4	2	0	1
	> 0.8	3	2	1	0	0	0	0
WSSL Clay	0 - 0.3	3	1	1	1	0	0	0
	0.3 -0.5	6	2	1	3	0	0	0
	0.5 -0.7	18	3	4	8	1	0	2
	0.7 - 0.8	18	5	3	4	3	1	2
	> 0.8	18	4	3	7	1	1	2

For SOC results were worse when using BSSL (Table 2) but maintained the trend of better the WSSL and the best the local models. This is due to the great variability of SOC over all regions of the world and clay is more stable.

Table 2. Number of countries with variable R^2 for SOC using vis-NIR-SWIR.

Model	R^2	Total	Africa	Asia	Europe	North America	Oceania	South America
LM SOC	0 - 0.3	3	0	0	3	0	0	0
	0.3 -0.5	9	0	0	9	0	0	0
	0.5 -0.7	17	0	2	11	1	0	3
	0.7 - 0.8	9	4	1	0	4	0	0
	> 0.8	19	7	7	1	2	1	1
BSSL SOC	0 - 0.3	38	3	5	22	3	1	4
	0.3 -0.5	11	5	2	1	1	0	2
	0.5 -0.7	6	2	2	1	1	0	0
	0.7 - 0.8	1	0	0	1	0	0	0
	> 0.8	1	0	1	0	0	0	0
WSSL SOC	0 - 0.3	17	4	5	4	1	1	2
	0.3 -0.5	14	2	1	6	2	0	3
	0.5 -0.7	22	3	3	13	2	0	1
	0.7 - 0.8	2	1	0	1	0	0	0
	> 0.8	2	1	1	0	0	0	0

5.4. DISCUSSION

5.4.1. The web service vantages and limitation

Soil spectroscopy is a non friendly discipline for the major community and is mostly understood by researchers. In fact, this field has gathered hundreds of papers in the last 60 years (Nocita et al., 2015, Soriano Disla et al., 2014). Despite this strong background, the technique did not advance to end users. This impacted on that all soil analysis continues to be performed by wet chemistry in the last 130 years. There is no doubt regarding the importance of the wet method, but it is also a fact that the demand for soil analysis is increasing, and the use of chemistry is not sustainable (Tuğrul, 2019). Many papers until now already proved the efficiency of soil spectroscopy (Stenberg et al., 2010) but could not deliver a service to facilitate end users. This online platform makes the service simple. The main limitation for the user is still on how to make modeling. This work provides a service simple for any user. The user only inserts the spectra and receives the soil analysis. The great advantage of this system is that it is quick, environmentally friendly and provides the ability to find background users and thus, create new groups on spectral sensing. The system also brings papers and fundamental explanations so students and researchers can create new and better systems.

Results are encouraging but this was not the only matter of this paper, since literature already said this to us. We really intended to show the idea on how to put in practice a soil service. Otherwise, we will continue on publishing, and nothing reaches the community. This dynamic of online processing permits a fast and interesting prediction of several soil attributes, serving as parameter for spectral analyses and encouraging the dissemination and use of spectral techniques. The platform also brings papers and explanations regarding the spectral ranges and its fundamentals.

This service also indicated how to see spectral patterns from the laboratory. Soil spectral libraries can be applied for many purposes, including modeling of soil attributes, soil survey, classification and mapping, and monitoring, by extracting the baseline electromagnetic properties of soils, which can be compared with pattern samples. Soil spectra can also be used for communication among researchers (soil classification has several systems, but spectra are the same!), and development of field, aerial, and space sensors, among others.

To understand the usefulness of soil SL, consider the following example: the interested parts (farmers or researchers) could send their soil samples to a central spectral library (e.g., a national or global) where they would be scanned and the spectral curves stored, or they could send already acquired soil spectral curves that compose their local spectral libraries. Local SL can be explored for personal interests (e.g., soil monitoring), and feed global SL, growing a global repository. Once having a global SL, spectral curves from a profile of an unknown type could be compared with other spectra from the global SL and a preliminary soil classification or the SOC or clay content, could be estimated. The ideal scale (global, continental, regional, local or farm) for a soil spectral library application has had much inquiry, and the general result is that the spatial scale of coverage and application depends.

The main limitation we observed is that the dataset must be robust and normalized for spectra and for soil analysis. Quality is crucial for great results, but this is difficult to reach with old dataset. Thus, it is imperative to start using protocols on soil spectral acquisition and a link with soil wet analysis quality. The link between these two communities is the basis of the success of this task.

5.4.2. Brazilian dataset and validation

For the Brazilian vis-NIR-SWIR dataset, clay content presented good results using the platform, with R^2 0.75. SOC was lower in the validation contrast than the model and reached 0.45. SOC spectral estimation has been a challenge in Brazilian agricultural areas because of the low content. For MIR results were significantly better where for clay BSSL reached 0.8 and

0.7 for clay and SOC respectively and in agreement with Terra et al. (2015). Thus, the BSSL online platform can be used as an important service for soil analysis, taking in account the level of accuracy and the element that the user desires.

5.4.3. Overseas Users for BSSL

Still under the online platform, several countries inserted spectra to see if their Local samples would fit into a BSSL model. We observed that for clay, 3 countries from Africa should over 0.7 of R^2, only two from Asia, 4 from Europe and 2 from North America. Despite that, in Europe there are still 5 countries in the R^2 range of 0.5-0.7. For SOC results were much less expressive, which agrees that SOC is more dependent with other factors such as biomes, land use and others. In the case of clay, results indicate that the BSSL model does not have expressive results for all countries for good for some and moderate to others or very poor. It is a first indication that Local models are better, but the choice of the model can still be with the user, knowing the limitations.

5.4.4. Exploiting new modeling with the dataset

Spectral libraries can have several approaches and levels, a farm (Ramirez-Lopez et al., 2019), a region (Rizzo et al., 2020), a country (Shi et al., 2014), a continent (Orgiazzi et al., 2018) or the world (Viscarra Rossel et al., 2016). The present paper played different approaches to understand the population of dataset and results around the world. This part of the manuscript was performed outside the platform.

We observed that Local Population was clearly better at quantifying clay and SOC in almost all cases and continents. The users Local models preserved the main characteristics of their soils, parent materials, biomes and other information which spectra carries. This comes in agreement with Brown (2007) for whom from global to local, the second would be the best. This occurred for clay and SOC in all continents. We had only a few cases in Europe where results were pretty low (R^2 range 0.3 to 0.5). When we use these datasets associated with the BSSL and reach a World Soil Spectral Library (WSSL), results were better than BSSL alone, but not greater than the LM. This indicates that LM in fact is better, and a worldwide approach is the best to reach more users and in agreement with Wetterlind and Stenberg (2010). A single country such as BSSL, should not be the best approach. On the other hand, it must be stated that the Spectral Libraries Era is already happening, and the user will be the one to choose which SL to use. This comes in agreement with findings of Debaene et al. (2014). The authors found little significant increase in prediction capacity of soil attributes with use of an entire data set, watching an increase on the R^2 of 0.63 to 0.72 for SOC and R^2 of 0.71 to 0.73 for clay. Thus, Continental SL can be improved to better quantify local situations. Indeed Araújo et al. (2014) when analysing spectra of 7,172 tropical soil samples. They found that separating the global dataset into more mineralogically uniform clusters improved predictive performance of clay content regardless of the geographical origin, showing that probably physically based, soil-related stratification criteria in libraries offer better results.

The WSSL of clay compared with the local, comes in agreement with Genot et al. (2011). These authors built a methodological framework for the use of NIR spectroscopy on a local and global scale by spectral treatment and regression methods. In addition, evaluated the ability of NIR spectroscopy to predict Total Organic Carbon (R^2= 0.91 local and R^2= 0.70 global), and clay (R^2 = 0.64 local and R^2 = 0.61 global) above several soils' conditions.

When we analysed MIR, the same situation was observed on increasing R^2 on the sequence LM - WSSSL - BSSL. This is directly related to the MIR fundamental spectral ranges bond specific which gives better results.

5.5. CONCLUSION AND FINAL CONSIDERATION

The online soil spectroscopy dynamics presented a great communication between worldwide users and delivered important information. The system can be applied for several approaches such as research, farming, wet laboratories, industries, create startups, teach students, pedologists, soil mapping, and others. The system is easy, and users do not have to do modeling and only insert the spectra and receive soil analysis with error and statistical information. Also, the user has the capacity to find the owners of spectra, get in contact and make partnerships as requested data. Directly on the platform, users can see the behavior of spectral patterns for soils with different textures, SOC, and many properties, despite soil depth and classification. Finally, the user inserted spectra in the web and received online in its e-mail the report with soil analysis.

In the case of clay content, the BSSL model presented great results for the Brazilian community and can be used as a first measurement. SOC had a good model but lower validation. MIR presented better results than vis-NIR-SWIR.

When other countries used the BSSL platform, results were variable. In conclusion, an overseas user can use the BSSL but be aware of the R^2 and results under this paper.

After we played outside the platform with data and observed that Local Model libraries presented expressively greater results in all continents mainly for clay. A World Soil Spectral Library could purchase important information, but not better than the LM. A specific Country spectral library, such as BSSL, was not able to support nor achieve the same results as previous models. On the other hand, some countries presented good data, which indicate that in future, the chosen spectral library service can be done by the user.

Developing regional, continental, or global libraries does not exclude embracing a local or physically based library, and the user will decide according to their necessity. It is a matter of the objective of the user as he is aware of the errors inside the population he is using. Also, it will be a matter of the user's country structure. And if his country is not inside a global library but inside a nearby country? In the Era of spectral libraries, users can choose the service that best attends.

We should mention our experience on what could be the issues on interference at modeling, as follows: (a) consistency of spectral acquisition of each user, (b) consistency of soil wet analysis, (c) Soil formation, mineralogy, and biome. These should be addressed in future works to achieve the best results

We strongly advise the necessity to go forward and construct a worldwide system, but that inside, brings the option to the user to choose the population he wants to create the model and afterwards receive the result.

ACKLOWDEGMENTS

We are grateful to Geotechnologies in Soil Science Group (GeoCiS/ESALQ-USP; http://esalqgeocis.wixsite.com/english) for team support. The authors had support from São Paulo Research Foundation (FAPESP) (grant numbers 2014/22262-0, 2016/26176-6, and 2020/04306-0).

REFERENCES

Apache Software Foundation, 2021. Acesso em: 20 de julho de 2021. Disponível em: https://www.apache.org/.

Araújo, S. R., Wetterlind, J., Demattê, J. A. M., & Stenberg, B. (2014). Improving the prediction performance of a large tropical vis-NIR spectroscopic soil library from B razil by clustering into smaller subsets or use of data mining calibration techniques. European Journal of Soil Science, 65(5), 718-729. doi:10.1111/ejss.12165.

Bas, M. V., Meléndez-Pastor, I., Navarro-Pedreño, J., Gómez, I., Mataix-Solera, J., & Hernández, E. (2013, April). Saline soils spectral library as a tool for digital soil mapping. In EGU General Assembly Conference Abstracts (pp. EGU2013-9738).

Bellinaso, H., Demattê, J. A. M., & Romeiro, S. A. (2010). Soil spectral library and its use in soil classification. Revista Brasileira de Ciência do Solo, 34(3), 861-870. doi:10.1590/s0100-06832010000300027

Brodský, L., Klement, A., Penížek, V., Kodešová, R., & Borůvka, L. (2011). Building soil spectral library of the Czech soils for quantitative digital soil mapping. Soil and water research, 6(4), 165-172.

Brown, D. J., Shepherd, K. D., Walsh, M. G., Mays, M. D., & Reinsch, T. G. (2006). Global soil characterization with VNIR diffuse reflectance spectroscopy. Geoderma, 132(3-4), 273-290. doi:10.1016/j.geoderma.2005.04.025

Brown, D. J. (2007). Using a global VNIR soil-spectral library for local soil characterization and landscape modeling in a 2nd-order Uganda watershed. Geoderma, 140(4), 444-453. doi:10.1016/j.geoderma.2007.04.021

"Browse Books." PHP Team Development | Guide Books. https://dl.acm.org/doi/book/10.5555/2829069.

Bunting, P. (2017). Pre-processing of remotely sensed imagery. In The Roles of Remote Sensing in Nature Conservation (pp. 39-63). Springer, Cham.

Cambule, A. H., Rossiter, D. G., Stoorvogel, J. J., & Smaling, E. M. A. (2012). Building a near infrared spectral library for soil organic carbon estimation in the Limpopo National Park, Mozambique. Geoderma, 183, 41-48. doi:10.1016/j.geoderma.2012.03.011

Debaene, G., Niedźwiecki, J., Pecio, A., & Żurek, A. (2014). Effect of the number of calibration samples on the prediction of several soil properties at the farm-scale. Geoderma, 214, 114-125. doi:10.1016/j.geoderma.2013.09.022

Demattê, J. A. M., & Garcia, G. J. (1999). Alteration of soil properties through a weathering sequence as evaluated by spectral reflectance. Soil Science Society of America Journal, 63(2), 327-342. doi:10.2136/sssaj1999.03615995006300020010x

Demattê, J. A. M., Fongaro, C. T., Rizzo, R., & Safanelli, J. L. (2018). Geospatial Soil Sensing System (GEOS3): A powerful data mining procedure to retrieve soil spectral reflectance from satellite images. Remote Sensing of Environment, 212, 161-175. doi:10.1016/j.rse.2018.04.047

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

Donagema, G. K., Campos, D. B. V. B., Calderano, S. B., Teixeira, W. G., & Viana, J. H. N. (2011). Manual de métodos de análise de solo [Manual of soil analysis methods]. Rio de Janeiro (RJ): Empresa Brasileira de Pesquisa Agropecuária. Portuguese.

Fontán, J. M., Calvache, S., López-Bellido, R. J., & López-Bellido, L. (2010). Soil carbon measurement in clods and sieved samples in a Mediterranean Vertisol by Visible and Near-Infrared Reflectance Spectroscopy. Geoderma, 156(3-4), 93-98. doi:10.1016/j.geoderma.2010.02.001.

Garrity, D., & Bindraban, P. (2004). A globally distributed soil spectral library visible near infrared diffuse reflectance spectra. ICRAF (World Agroforestry Centre)/ISRIC (World Soil Information) Spectral Library: Nairobi, Kenya.

Ge, Y., Morgan, C. L., Grunwald, S., Brown, D. J., & Sarkhot, D. V. (2011). Comparison of soil reflectance spectra and calibration models obtained using multiple spectrometers. Geoderma, 161(3-4), 202-211. doi:10.1016/j.geoderma.2010.12.020.

Genot, V., Colinet, G., Bock, L., Vanvyve, D., Reusen, Y., & Dardenne, P. (2011). Near infrared reflectance spectroscopy for estimating soil characteristics valuable in the diagnosis of soil fertility. Journal of Near Infrared Spectroscopy, 19(2), 117-138.

Gogé, F., Joffre, R., Jolivet, C., Ross, I., & Ranjard, L. (2012). Optimization criteria in sample selection step of local regression for quantitative analysis of large soil NIRS database. Chemometrics and Intelligent Laboratory Systems, 110(1), 168-176. doi:10.1016/j.chemolab.2011.11.003.

Hong-Yan, R. E. N., Zhuang, D. F., Singh, A. N., Jian-Jun, P. A. N., Dong-Sheng, Q. I. U., & Run-He, S. H. I. (2009). Estimation of As and Cu contamination in agricultural soils around a mining area by reflectance spectroscopy: A case study. Pedosphere, 19(6), 719-726. doi.org/10.1016/S1002-0160(09)60167-3

IUSS Working Group WRB, 2015. World Reference Base for Soil Resources 2014, Update 2015. International Soil Classification System for Naming Soils and Creating Legends for Soil Maps. World Soil Resources Reports No. 106. FAO, Rome.

Khaledian, Y., & Miller, B. A. (2020). Selecting appropriate machine learning methods for digital soil mapping. Applied Mathematical Modelling, 81, 401-418. doi.org/10.1016/j.apm.2019.12.016

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., & Kenkel, B. (2020). caret: Classification and Regression Training. R package version 6.0-86. Avaliable at: https://cran. r-project. org/web/packages/caret/caret. pdf.

Lal, R., Bouma, J., Brevik, E., Dawson, L., Field, D. J., Glaser, B., Hatano, R., Hartemink, A. E., Kosaki, T., Lascelles, B., Monger, C., Muggler, C., Ndzana, G. M., Norra, S., Pan, X. Paradelo, R., Reyes-Sánchez, L. B., Sandén, T., Singh, B. R., Spiegel, H., Yanai, H. & Zhang, J. (2021). Soils and sustainable development goals of the United Nations (New York, USA): An IUSS perspective. Geoderma Regional, e00398. doi:10.1016/j.geodrs.2021.e00398.

Lorna Jane Mitchell. PHP Web Services: APIs for the Modern Web. 2016

Manohar, H. M., & Appaiah, S. (2017). Stabilization of FIFO system and Inventory Management. International Research Journal of Engineering and Technology, 4(6), 5631-5638.

Nocita, M., Kooistra, L., Bachmann, M., Müller, A., Powell, M., & Weel, S. (2011). Predictions of soil surface and topsoil organic carbon content through the use of laboratory and field spectroscopy in the Albany Thicket Biome of Eastern Cape Province of South Africa. Geoderma, 167, 295-302. doi:10.1016/j.geoderma.2011.09.018.

Nocita, M., Stevens A., van Wesemael, B., Aitkenhead, M., Bachmann M., Barthès, B., Ben-Dor, E., Brown, D.J., Clairotte, M., Csorba, A., Dardenne, P., Demattê, J.A.M, Genot, V., Guerrero, C., Knadel,, M., Montanarella, L., Noon, C., Ramirez-Lopez, L., Robertson, J., Sakai, H., Soriano-Disla, J.M., Shepherd, K.D., Stenberg, B., Towett, E.K., Vargas, R., Wetterlind, J., 2015. Soil spectroscopy: An alternative to wet chemistry for soil monitoring. In Advances in agronomy. 132, 139-159. https://doi.org/10.1016/bs.agron.2015.02.002

Orgiazzi, A., Ballabio, C., Panagos, P., Jones, A., & Fernández-Ugalde, O. (2018). LUCAS Soil, the largest expandable soil dataset for Europe: a review. European Journal of Soil Science, 69(1), 140-153. doi: 10.1111/ejss.12499

Padarian, J., Minasny, B., & McBratney, A. B. (2020). Machine learning and soil sciences: A review aided by machine learning tools. Soil, 6(1), 35-52. doi:10.5194/soil-2019-57

Quinlan, J.R. (John R., 1992. Learning with continuous classes, in: 5th Australian Joint Conference on Artificial Intelligence. pp. 343–348. https://doi.org/10.1142/9789814536271

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Ramirez-Lopez, L., Behrens, T., Schmidt, K., Stevens, A., Demattê, J. A. M., & Scholten, T. (2013). The spectrum-based learner: A new local approach for modeling soil vis–NIR spectra of complex datasets. Geoderma, 195, 268-279. doi:10.1016/j.geoderma.2012.12.014.

Ramirez-Lopez, L., Wadoux, A. C., Franceschini, M. H. D., Terra, F. S., Marques, K. P. P., Sayão, V. M., & Demattê, J. A. M. (2019). Robust soil mapping at the farm scale with vis–NIR spectroscopy. European Journal of Soil Science, 70(2), 378-393.

Rizzo, R., Medeiros, L. G., de Mello, D. C., Marques, K. P., de Souza Mendes, W., Silvero, N. E. Q., Dotto, A. C., Bonfatti, B. R., & Demattê, J. A. (2020). Multi-temporal bare surface image associated with transfer functions to support soil classification and mapping in southeastern Brazil. Geoderma, 361, 114018. doi.org/10.1016/j.geoderma.2019.114018

Shi, Z., Wang, Q., Peng, J., Ji, W., Liu, H., Li, X., & Viscarra-Rossel, R. (2014). Development of a national VNIR soil-spectral library for soil classification and prediction of organic matter concentrations. Science China Earth Sciences, 57(7), 1671-1680. doi.org/10.1007/s11430-013-4808-x

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

Stenberg, B., Viscarra-Rossel R.A., Mouazen, A.M., Wetterlind, J. (2010). Visible and near infrared spectroscopy in soil science. Advances in agronomy. 107, 163-215. https://doi.org/10.1016/S0065-2113(10)07005-7

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

Stevens, A., Ramirez-Lopez, L., 2020. prospectr: Functions for Processing and Sample Selection of Spectroscopic Data.

Stoner, E. R., & Baumgardner, M. F. (1981). Characteristic variations in reflectance of surface soils. Soil Science Society of America Journal, 45(6), 1161-1165. doi:10.2136/sssaj1981.03615995004500060031x.

Tasneem, S., & Ammar, R. (2012). Performance Study of a Distributed Web Server: An Analytical Approach. Journal of Software Engineering and Applications, Vol. 5 No. 11, 2012, pp. 855-863. doi: 10.4236/jsea.2012.511099.

Terra, F. S., Rossel, R. A. V., & Demattê, J. A. (2019). Spectral fusion by Outer Product Analysis (OPA) to improve predictions of soil organic C. Geoderma, 335, 35-46.

Tuğrul, K. M. (2019). Soil management in sustainable agriculture. In Soil Management and Plant Nutrition for Sustainable Crop Production. IntechOpen. doi.org/10.5772/intechopen.88319.

Viscarra-Rossel, R., & McBratney, A. B. (1998). Soil chemical analytical accuracy and costs: implications from precision agriculture. Australian Journal of Experimental Agriculture, 38(7), 765-775. doi:10.1071/ea97158

Viscarra-Rossel, R., Cattle, S. R., Ortega, A., & Fouad, Y. (2009). In situ measurements of soil colour, mineral composition and clay content by vis–NIR spectroscopy. Geoderma, 150(3-4), 253-266. doi:10.1016/j.geoderma.2009.01.025.

Viscarra-Rossel, R., Behrens, T., Ben-Dor, E., Brown, D. J., Demattê, J. A. M., Shepherd, K. D., Shi, Z., Stenberg, B., Stevens, A., Adamchuk, V., Aichi, H., Barthes, B. G., Barthomeus, H. M., Bayer, A. D., Bernoux, M., Bottcher, K., Brodský, L., Du, C. W., Chappell, A., Fouad, Y., Genot, V., Gomez, C., Grunwald, S., Gubler, A., Guerrero, C., Hedley, C. B., Knadel, M., Morrás, H. J. M., Nocita, L., Ramirez-Lopez, L., Roudier, P., Rufasto Campos, E. M., Sanborn, P., Sellito, V. M., Sudduth, K. A.,Rawlins, B.G., Walter, C., Winowiecki, L. A., Hong, S. Y., & Ji, W. (2016). A global spectral library to characterize the world's soil. Earth-Science Reviews, 155, 198-230. doi:10.1016/j.earscirev.2016.01.012.

Wetterlind, J., & Stenberg, B. (2010). Near-infrared spectroscopy for within-field soil characterization: small local calibrations compared with national libraries spiked with local samples. European Journal of Soil Science, 61(6), 823-843. doi:10.1111/j.1365-2389.2010.01283.x

Zhang, X. G., Huang, B., Ji, J. F., Hu, W. Y., Sun, W. X., & Zhao, Y. C. (2012). Quantitative prediction of soil salinity content with Visible-Near infrared Hyper-spectra in northeast china. Spectroscopy and Spectral Analysis, 32(8), 2075-2079.

Zhang, X., & Huang, B. (2019). Prediction of soil salinity with soil-reflected spectra: A comparison of two regression methods. Scientific reports, 9(1), 1-8.