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

Online near-infrared spectroscopy for soil attributes prediction

Ricardo Canal Filho

Dissertation presented to obtain the degree of Master in Science. Area: Agricultural Systems Engineering

Piracicaba
2023

Ricardo Canal Filho
Agronomist

Online near-infrared spectroscopy for soil attributes prediction

versão revisada de acordo com a Resolução CoPGr 6018 de 2011

Advisor:
Prof. Dr. **JOSÉ PAULO MOLIN**

Dissertation presented to obtain the degree of Master in Science. Area: Agricultural Systems Engineering

Piracicaba
2023

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

Canal Filho, Ricardo

Online near-infrared spectroscopy for soil attributes prediction / Ricardo Canal Filho - - versão revisada de acordo com a Resolução CoPGr 6018 de 2011. - - Piracicaba, 2023.

63 p.

Dissertação (Mestrado) - - USP / Escola Superior de Agricultura “Luiz de Queiroz”.

1. Agricultura de precisão 2. Espectroscopia de reflectância difusa 3. Sensoriamento proximal do solo 4. Aprendizado de máquina 5. Variabilidade do solo I. Título

ACKNOWLEDGMENTS

I would like to thank the National Council for Scientific and Technological Development (CNPq) for the scholarship conceded. Furthermore, it is necessary to greet the company Spectral Solutions Ltda. and the IBRA laboratory for the partnership that allowed this project to be accomplished.

I would also like to thank the University of São Paulo (USP), represented by the Luiz de Queiroz College of Agriculture (ESALQ) and the graduate program in Agricultural Systems Engineering for the opportunity on being part of this. Even with all the problems, internal and external, that we know exist and affect these institutions, the front for science and principally education they represent are things I have faith is worth living for.

To professor Molin: thank you for believing not only in what we do, but especially on who we are. For the support when things did not come off as I expected in these years.

To the employees of Biossystems Engineering department who helped me along the course, and took part on the development of this study.

To my LAP colleagues who were open to share knowledge and opportunities. I will always appreciate how you helped me out in this last year.

To my family and Fernanda who support me on working with what I believe. With special gratitude to my sister, of blood, dreams and, now, also professionally! You were there in the toughest moment.

And a reminder for when I came back here. There was neither happiness nor misery at all; only the comparison of one state with another.

“Attendre et espérer”

SUMMARY

RESUMO

Espectroscopia no infravermelho próximo para a predição de atributos do solo em tempo real

A agricultura de precisão (AP) baseia-se na identificação da variabilidade espacial e temporal dos atributos que influenciam a produção agrícola. Nesse sentido, técnicas que permitam monitorar o solo e as culturas em alta densidade espacial vêm sendo estudadas pela comunidade de AP. A espectroscopia de reflectância difusa (DRS) é um técnica que permite, sobretudo na região do infravermelho próximo (NIR), coletar espectros de solo direto no campo, utilizando sensores embarcados em máquinas agrícolas. O uso dessa técnica permite coletar pontos em alta densidade espacial que, em conjunto com o aprendizado de máquina (ML), se transformam em dados quali-quantitativos dos atributos do solo. Entretanto, em solos tropicais, principalmente no Brasil, essa área ainda é pouca desenvolvida em comparação a estudos, por exemplo, da Austrália, Estados Unidos e Europa. O projeto de pesquisa dessa dissertação foi proposto no âmbito de ampliar o desenvolvimento da técnica nos solos tropicais brasileiros. Uma área experimental da Universidade de São Paulo, em Piracicaba-SP, foi utilizada para a coleta de espectros de solo em tempo real no infravermelho próximo. Foram testados diferentes modelos estatísticos para predição de atributos químicos e físicos do solo. Protocolos de calibração e de uso da DRS em campo foram avaliados. Os principais resultados desta dissertação foram organizados em três capítulos. O primeiro aborda protocolos de calibração quanto ao uso de técnicas de pré-procesamento do espectro e diferentes modelos estatísticos. Os resultados sugerem que o uso de dados brutos em conjunto com modelos de redução de dimensionalidade do espectro multivariado do solo oferecem a estratégia mais eficiente para calibração dos modelos preditivos. O segundo capítulo abordou a inserção de amostras de diferentes áreas na calibração dos modelos de ML. Os resultados mostraram predições mais robustas quando modelos foram calibrados apenas com amostras da própria área experimental, denotando a importância da calibração local para uso da DRS NIR. No terceiro e último capítulo, a área foi revisitada em um segundo dia de coleta espectral, três semanas após a primeira, seguindo os mesmos critérios experimentais e instrumentais. Os modelos de ML calibrados no primeiro dia foram testados para predição dos atributos do solo com espectros da segunda coleta. Reportou-se baixa capacidade preditiva dos modelos neste caso, indicando a necessidade de calibrações locais não só no espaço, mas também no tempo, para que a técnica desempenhe corretamente. Os resultados reportados provam o potencial da técnica para a agricultura, pois mostram que é possível a predição de atributos do solo com espectros NIR coletados diretamente no campo. Ainda, este trabalho pode auxiliar no desenvolvimento das práticas de AP, e oferecer diretrizes para futuras pesquisas que busquem o desenvolvimento da DRS para predição de atributos do solo em tempo real, a fim de estabelecer seu uso em larga escala na agricultura.

Palavras-chave: Agricultura de precisão, Espectroscopia de reflectância difusa, Sensoriamento proximal do solo, Aprendizado de máquina, Variabilidade do solo

ABSTRACT

Online near-infrared spectroscopy for soil attributes prediction

Precision agriculture (PA) is based on the identification of spatial and temporal variability of the attributes that influence agricultural production. In this sense, techniques that allow monitoring soil and crops in high spatial density have been studied by the PA community. Diffuse reflectance spectroscopy (DRS) is a technique that allows, especially in the near-infrared (NIR) region, to acquire online soil spectra, embedding sensors in agricultural machines. The use of this technique allows data acquisition in high spatial density, which, together with machine learning (ML), are transformed into quali-quantitative data of soil attributes. However, in tropical soils, especially in Brazil, this research area is still poorly developed compared to studies from Australia, the United States of America and Europe. The research project of this dissertation was proposed to expand the development of the technique in Brazilian tropical soils. An experimental area of the University of São Paulo, in Piracicaba-SP, was used to acquire online soil NIR spectra. Different statistical models were tested to predict soil chemical and physical attributes. Calibration and use protocols of DRS in the field were evaluated. The main findings of this dissertation were organized into three chapters. The first one addresses calibration protocols regarding the use of spectrum preprocessing techniques and different statistical models. The results suggest that the use of raw data combined with dimensionality reduction statistical models offer the most efficient strategy for calibration of predictive models. The second chapter addressed the insertion of samples from different areas in the calibration of ML models. The results showed more robust predictions when models were calibrated only with samples from the experimental area itself, denoting the importance of local calibration for the use of DRS NIR in online acquisition. In the third and last chapter, the area was revisited on a second day of spectral acquisition, three weeks after the first one, following the same experimental and instrumental criteria. The ML models calibrated on the first day were tested for prediction of soil attributes with spectra from the second day of acquisition. Low predictive performance of the models was reported in this scenario, indicating the need for local calibrations not only in space, but also in time, for the technique to perform properly. The results reported in this dissertation prove the potential of the technique for agriculture, as they show that it is possible to predict soil attributes with online NIR spectra. Furthermore, this work can help in the development of PA practices, and offer guidelines for future research that seek the development of DRS for prediction of soil attributes in the field, to establish its large-scale use in agriculture.

Keywords: Precision agriculture, Diffuse reflectance spectroscopy, Proximal soil sensing, Machine learning, Soil variability

LIST OF ACRONYMS

A	range
AI	artificial intelligence
C0	nugget
C1	sill
Ca	calcium
CEC	cation exchange capacity
DRS	diffuse reflectance spectroscopy
ET	extra trees
g kg^{-1}	grams per kilogram
GNSS	Global Navigation Satellite System
ISO	International Organization for Standardization
K	potassium
km h^{-1}	kilometers per hour
Lasso	least absolute shrinkage and selection operator
LV	latent variable
m s^{-1}	meters per second
MAE	mean absolute error
Mg	magnesium
ML	machine learning
NIR	near-infrared
OM	organic matter
P	phosphorus
PA	precision agriculture
PC	principal components
PCA	principal components analysis
PCR	principal components regression
pH	potential of hydrogen
PLSR	partial least squares regression
PP1	preprocessing sequence 1
PP2	preprocessing sequence 2
PSS	proximal soil sensing
R^2	coefficient of determination
RD	raw data
RF	random forest
RMSE	root mean squared error
RPIQ	ratio of performance to interquartile distance
SG	Savitzky-Golay
V	basis saturation
VNIR	visible and near-infrared
X^2	chi-squared

1. GENERAL INTRODUCTION

The International Society of Precision Agriculture defines precision agriculture (PA) as "a management strategy that takes account of temporal and spatial variability to improve sustainability of agricultural production" (ISPA, 2022). For this strategy to be adopted, techniques capable of identifying and mapping agricultural fields are essential, especially aspects related to soil and crops.

Soil is a natural asset on which agriculture depends because it provides mechanical support, water, oxygen and nutrients that plants absorb. Soil characteristics can limit or leverage agricultural production and soils of different geological formations and different formation times have different pedogenetic characteristics (Topp et al., 1997). In addition, the topographic position and the use and management by which this soil has passed over the years influence its physical and chemical characteristics (Zörb et al., 2014; Fontoura et al., 2019; Lu et al., 2021), which highlights the importance of identifying its attributes and the variability that exists among them.

That is why soil sampling was established as a necessity for agriculture to base the planning and decisions of the productive steps. Once collected, soil attributes data are used to map the field (AbdelRahman et al., 2020). The characterization of spatial dependence is done using geostatistics tools, which consider the values of attributes associated with the geographic position of each collection (De Iaco et al., 2022).

In geostatistical analysis, the range of the variogram is the main indicator to guide the sample density that must be used to correctly encompass the spatial variability of an attribute. The range varies for each analyzed attribute, and also with soil characteristics and area management (Vieira, 2000; Sória et al., 2018). With this knowledge, the practice of sampling, especially regarding the sample density, is repeatedly questioned by researchers, as it is far from ideal (Wollenhaupt et al., 1994; Montanari et al., 2012; Cherubin et al., 2014; Cherubin et al., 2015), which does not adequately represent the spatial variability of soil attributes.

Both collecting a sample and analyzing it represent a cost. Due to this, those responsible for the decision-making in the agricultural production chain resist increasing the sampling density in the traditional way due to the increase in cost and time that it would represent. Alternative techniques have been explored to increase the density of data about soil attributes, to find a faster method, capable of diluting the cost per acquisition and reducing the amount of inputs required for analysis (Molin & Tavares, 2019). Within this context, the diffuse reflectance spectroscopy (DRS) has been tested in an attempt to predict the attributes of a soil sample through the spectral signature that it emits when in contact with a certain type of energy (Kuang et al., 2012).

The near-infrared region (NIR) shows potential for use, as it is the range of the spectrum that expresses primary interactions, the so-called fundamental vibrations, and secondary interactions, the overtones, with soil attributes (Pasquini, 2018). These interactions occur in the form of energy absorption, reflection or transmission, and can be related to soil attributes in terms of quantity and quality (Stenberg et al., 2010).

Using the soil spectra, researchers have been applying multivariate statistical models that predict the values of interest from features called predictors. These techniques have been time-consuming and costly in the past. However, the recent development of technologies, leveraging machine learning (ML) methods and the beginning of intensive use of artificial intelligence (AI) in multiple knowledge areas, has been boosting the use of multivariate statistics for prediction models (Sharma et al., 2020).

The DRS NIR is already widespread for research. However, its use is mainly documented by collecting samples from the field and taking them to a laboratory, where these samples are, in most cases, treated, and only then the spectrum of this soil sample is acquired (Lacerda et al., 2016; Demattê et al., 2017; Cezar et al., 2019).

However, to actually represent a change in agricultural production, this technique needs to act directly in the field, with all the challenges that a field operation represents, such as variations in moisture, temperature, non-uniformity of particles and others. All this while maintaining the maximum quality of the collected spectrum to enable the calibration of robust ML prediction models.

Both soil scientists and PA researchers were dedicated to identify and overcome the main challenges of using DRS NIR in the field (Shonk et al., 1991; Sudduth & Hummel, 1993; Shibusawa et al., 2001; Stenberg et al., 2007; Ben Dor et al., 2008; Mouazen et al., 2009), adapting the technology for proximal soil sensing (PSS) (Viscarra Rossel et al., 2011). This aspect was named the DRS for online spectra acquisition. The identification and quantification of soil attributes using online NIR spectra became possible. Some researchers have already reported success in using the DRS NIR for inference in the field, as is the case of Mouazen & Kuang (2016), who monitored the variability of phosphorus (P) in the soil of an UK agricultural field during three consecutive seasons. They were able to identify the sources of variability, and basing decision-making on the information obtained from the data, guiding a successful site-specific application of P, which was able to increase the homogeneity of the attribute in the area.

In Brazilian tropical soils, however, few studies have been reported applying DRS to acquire online spectra. Franceschini et al. (2018) studied the effects of external factors and potential spectral correction for online visible and near-infrared (VNIR) soil attributes prediction, aiming the quantification of lime requirement. The study indicated that online soil spectra had potential for soil properties characterization, although advances in sensing solutions and chemometric methods applied in this context would be required.

Eitelwein et al. (2022) used the strategy of calibrating ML models with VNIR spectra in a 15% portion of a commercial field in mid-west Brazil, trying to extrapolate the calibration to the entire area. For this, they sought to encompass the maximum variability of the attributes desired when choosing the calibration area. Despite reporting robust models in the calibration area for clay, organic matter (OM), cation exchange capacity (CEC), potential of hydrogen (pH), basis saturation (V), calcium (Ca), magnesium (Mg) and potassium (K), the extrapolation of these models to the entire area only reported good results for clay and OM, according to the evaluation parameters used by the authors.

The research project of this dissertation was developed in the sense of continuing the advances of the DRS for PSS in tropical soils. This study has the hypothesis that the use of DRS NIR in the field, supported by ML techniques, can generate diagnoses of spatialization of soil attributes, if the correct guidelines for calibrating the prediction models are considered. The work was planned in four distinct stages: acquisition of online NIR spectra in an experimental area, and soil sampling associated with spectral acquisition; laboratory spectra acquisition of field samples and physico-chemical analysis; testing of different statistical techniques for ML models calibration; mapping the spatial variability of the predicted attributes using geostatistics tools.

The objectives that coincide with the stages planned for this study were: to use the DRS NIR directly in the field to predict soil attributes; to determine the prediction potential of different statistical models using soil online NIR spectra; to identify the best procedures for ML models calibration using online NIR spectra, considering from where the samples that compose the initial dataset come from; to characterize the spatio-temporal stability of the use of DRS NIR for attribute prediction outside controlled laboratory conditions. The main findings of this study were organized into chapters, that follow.

References

AbdelRahman, M. A., Zakarya, Y. M., Metwaly, M. M., & Koubouris, G. (2020). Deciphering soil spatial variability through geostatistics and interpolation techniques. Sustainability, 13(1), 194.

Ben-Dor, E., Heller, D., & Chudnovsky, A. (2008). A novel method of classifying soil profiles in the field using optical means. Soil Science Society of America Journal, 72(4), 1113-1123.

Cezar, E., Nanni, M. R., Guerrero, C., da Silva Junior, C. A., Cruciol, L. G. T., Chicati, M. L., & Silva, G. F. C. (2019). Organic matter and sand estimates by spectroradiometry: Strategies for the development of models with applicability at a local scale. **Geoderma**, 340, 224-233.

Cherubin, M. R., Santi, A. L., Eitelwein, M. T., Menegol, D. R., Ros, C. O. D., Pias, O. H. D. C., & Berghetti, J. (2014). Eficiência de malhas amostrais utilizadas na caracterização da variabilidade espacial de fósforo e potássio. **Ciência Rural**, 44, 425-432.

Cherubin, M. R., Santi, A. L., Eitelwein, M. T., Amado, T. J. C., Simon, D. H., & Damian, J. M. (2015). Dimensão da malha amostral para caracterização da variabilidade espacial de fósforo e potássio em Latossolo Vermelho. **Pesquisa Agropecuária Brasileira**, 50(2), 168-177.

De Iaco, S., Hristopulos, D. T., & Lin, G. (2022). Geostatistics and Machine Learning. **Mathematical Geosciences**, 1-7.

Demattê, J. A. M., Ramirez-Lopez, L., Marques, K. P. P., & Rodella, A. A. (2017). Chemometric soil analysis on the determination of specific bands for the detection of magnesium and potassium by spectroscopy. **Geoderma**, 288, 8-22.

Eitelwein, M. T., Tavares, T. R., Molin, J. P., Trevisan, R. G., de Sousa, R. V., & Demattê, J. A. M. (2022). Predictive Performance of Mobile Vis–NIR Spectroscopy for Mapping Key Fertility Attributes in Tropical Soils through Local Models Using PLS and ANN. **Automation**, 3(1), 116-131.

Franceschini, M. H. D., Demattê, J. A. M., Kooistra, L., Bartholomeus, H., Rizzo, R., Fongaro, C. T., & Molin, J. P. (2018). Effects of external factors on soil reflectance measured on-the-go and assessment of potential spectral correction through orthogonalisation and standardisation procedures. **Soil and Tillage Research**, 177, 19-36.

Fontoura, S. M. V., de Castro Pias, O. H., Tiecher, T., Cherubin, M. R., de Moraes, R. P., & Bayer, C. (2019). Effect of gypsum rates and lime with different reactivity on soil acidity and crop grain yields in a subtropical Oxisol under no-tillage. **Soil and Tillage Research**, 193, 27-41.

International Society of Precision Agriculture (ISPA). Precision Agriculture Definition. https://www.ispag.org/about/definition. Acessado em 05 de novembro de 2022.

Kuang, B., Mahmood, H. S., Quraishi, M. Z., Hoogmoed, W. B., Mouazen, A. M., & van Henten, E. J. (2012). Sensing soil properties in the laboratory, in situ, and on-line: a review. **Advances in Agronomy,** 114, 155.

Lacerda, M. P., Demattê, J. A., Sato, M. V., Fongaro, C. T., Gallo, B. C., & Souza, A. B. (2016). Tropical texture determination by proximal sensing using a regional spectral library and its relationship with soil classification. **Remote Sensing**, 8(9), 701.

Lu, Y., Gao, Y., Nie, J., Liao, Y., & Zhu, Q. (2021). Substituting chemical P fertilizer with organic manure: effects on double-rice yield, phosphorus use efficiency and balance in subtropical China. **Scientific Reports**, 11(1), 1-13.

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(SPE), 126-147.

Montanari, R., Souza, G. S. A., Pereira, G. T., Marques, J. U. N. I. O. R., Siqueira, D. S., & Siqueira, G. M. (2012). The use of scaled semivariograms to plan soil sampling in sugarcane fields. **Precision Agriculture**, 13(5), 542-552.

Mouazen, A. M., Maleki, M. R., Cockx, L., Van Meirvenne, M., Van Holm, L. H. J., Merckx, R., ... & Ramon, H. (2009). Optimum three-point linkage set up for improving the quality of soil spectra and the accuracy of soil phosphorus measured using an on-line visible and near infrared sensor. **Soil and Tillage Research**, 103(1), 144-152.

Mouazen, A. M., & Kuang, B. (2016). On-line visible and near infrared spectroscopy for in-field phosphorous management. **Soil and Tillage Research**, 155, 471-477.

Pasquini, C. (2018). Near infrared spectroscopy: A mature analytical technique with new perspectives–A review. **Analytica Chimica Acta**, 1026, 8-36.

Sharma, R., Kamble, S. S., Gunasekaran, A., Kumar, V., & Kumar, A. (2020). A systematic literature review on machine learning applications for sustainable agriculture supply chain performance. **Computers & Operations Research**, 119, 104926.

Shibusawa, S. (2001). Soil mapping using the real-time soil spectrophotometer. In Precision Agriculture'01, Proc., **3rd Euro. Conf. Precision Agriculture**, Agro Montpellier (pp. 485-490).

Shonk, J. L., Gaultney, L. D., Schulze, D. G., & Van Scoyoc, G. E. (1991). Spectroscopic sensing of soil organic matter content. **Transactions of the ASAE**, 34(5), 1978-1984.

Sória, J. E., Tavanti, R. F. R., Alves, M., Andreotti, M., & Montanari, R. (2018). Scaled semivariogram in the sample planning of soils cultivated with sugarcane. **Journal of Agricultural Science**, 10(9), 315-325.

Stenberg, B., Rogstrand, G., Bolenius, E., & Arvidsson, J. (2007). On-line soil NIR spectroscopy: identification and treatment of spectra influenced by variable probe distance and residue contamination. **Precision agriculture**, 7, 125-131.

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.

Sudduth, K. A., & Hummel, J. W. (1993a). Portable, near-infrared spectrophotometer for rapid soil analysis. **Transactions of the ASAE**, 36(1), 185-193.

Topp, G. C., Reynolds, W. D., Cook, F. J., Kirby, J. M., & Carter, M. R. (1997). Physical attributes of soil quality. In **Developments in Soil Science** (Vol. 25, pp. 21-58). Elsevier.

Vieira, S. R. (2000). Geoestatística em estudos de variabilidade espacial do solo. **Tópicos em ciência do solo**. Viçosa: Sociedade Brasileira de Ciência do Solo, 1, 1-53.

Viscarra Rossel, R. A., Adamchuk, V. I., Sudduth, K. A., McKenzie, N. J., & Lobsey, C. (2011). Proximal soil sensing: an effective approach for soil measurements in space and time. In **Advances in agronomy** (Vol. 113, pp. 243-291). Academic Press.

Wollenhaupt, N. C., Wolkowski, R. P., & Clayton, M. K. (1994). Mapping soil test phosphorus and potassium for variable-rate fertilizer application. **Journal of Production Agriculture**, 7(4), 441-448.

Zörb, C., Senbayram, M., & Peiter, E. (2014). Potassium in agriculture–status and perspectives. **Journal of Plant Physiology**, 171(9), 656-669

2. EFFICIENCY OF PREPROCESSING TECHNIQUES APPLIED TO ONLINE NEAR-INFRARED SOIL SPECTRA IN DIMENSIONALITY REDUCTION STATISTICAL REGRESSION

Abstract

In soil science, near-infrared region (NIR) is the most studied electromagnetic spectrum for predicting attributes of agronomic interest using diffuse reflectance spectroscopy (DRS). The methods used for prediction vary both in the statistical model used and in the treatment applied to field spectra before prediction. There is no consensus in the literature about which model to use nor if preprocessed spectra prediction statistically differs from the use of raw spectra, and how both affect the data processing steps. As the amount of data acquired in the agricultural production grows, efficient data processing protocols are necessary to leverage PA techniques in large scale. This study was proposed to evaluate five statistical models for soil attributes prediction using online NIR spectra, and then compare the use of raw and preprocessed spectra to fit predictive models, evaluating how data processing was affected. Online NIR spectra were acquired in a Brazilian tropical soil and used in two calibration strategies: only local samples (Local calibration), and gathering samples from other fields (Global calibration). Each calibration strategy tested the use of raw data, and two different preprocessing sequences. For each soil fertility attribute considered, that were clay, sand, organic matter (OM), cation exchange capacity (CEC), potential of hydrogen (pH) and potassium (K), the study tested three dimensionality reduction statistical models: partial least squares regression (PLSR), principal components regression (PCR), and least absolute shrinkage and selection operator (Lasso), and two non-linear regressors: random forest (RF) and extra trees (ET). All dimensionality reduction statistical models outperformed the non-linear, except the K prediction using ET. The PCR is highlighted as presented the best parameters in 42% of models calibrated, and also close parameters to the best observed in other 50%. The prediction using raw and preprocessed spectra presented no statistical difference. However, an average increase of 200% was observed in the processing time demanded for prediction. The results reported in the scenario evaluated suggest the use of raw spectra as the most efficient strategy for online NIR spectra prediction of soil attributes.

Keywords: chemometrics; near-infrared spectroscopy; data analysis, machine learning, spectra preprocessing.

2.1 Introduction

DRS in the NIR region is a potential method for wide application in agriculture for soil attributes prediction. This electromagnetic spectral region expresses primary and secondary energy-soil interactions (Stenberg et al., 2010). These interactions occur as absorption, reflection, or transmission features, and can be related to soil attributes in quality and quantity (Nocita et al., 2015).

Established in the soil science, DRS NIR has its main use in laboratory. However, one of the challenges is the technique to work directly in the field, acquiring the called online spectra, to enhance the spatial density of soil data. The soil sampling density often adopted in agriculture is ineffective to identify spatial patterns in agricultural areas (Wollenhaupt et al., 1994; Montanari et al., 2012; Cherubin et al., 2014; Cherubin et al., 2015), leading to questionable decision-making and consequently inefficient management of agriculture inputs.

Researchers are dedicated to leverage NIR spectra for soil attributes prediction using multivariate statistical methods (Pasquini, 2018). However, two main points remain unsolved. The statistical models used for calibration of prediction models are not a consensus. Nor the type of method, such as linear or non-linear models, or how to approach the multidimensionality of soil spectra, reducing the features or working with the entire spectra. Although partial least squares regression (PLSR) is often the most cited method applied to ML in soil spectroscopy (Bellon-Maurel & McBratney, 2011; Rossel & Behrens, 2010), there is other methods available and not a single method recognized as the best. The performance seems to vary due to diverse agricultural conditions and also in different regions/soil types.

The other point is data pretreatment. The use of preprocessing techniques is commonly observed before fitting the predictive models (Franceschini et al., 2018; Munnaf et al., 2021a; Zhang et al., 2021). These techniques aim to reduce noise, highlight features, and extract useful information from the raw data (Dotto et al., 2018). Nevertheless, their use implies in a greater computational processing cost, fact that need to be accounted for the development of AI systems aiming the real time prediction and intervention. Some of the methods applied, as PLSR, are dimensionality reduction techniques, which can cope with multivariate data, and aid the noisy, redundant, and irrelevant data removal (Velliangiri et al., 2019).

These data modeling steps are often neglected and few studies discussing it are available. Either which statistical model to use, evaluating the differences presented by the methods, and also the assessment of spectra preprocessing performance needs to be addressed. These considerations are important since to increase the density of information about the soil, especially considering PSS (Viscarra Rossel et al., 2011) and online DRS, the database collected in commercial areas will escalate. The studies on this research area already report from 50 to 300 sampling points per hectare. Therefore, an efficient data processing protocol is necessary to promote the technique in large scale.

In this scenario, this study evaluated five ML methods applied in a database of online NIR spectra acquired in a Brazilian tropical field. Also, verified the feasibility of preprocessing techniques applied on two of the best performing ML calibrations from the previous step, assessing the data processing cost as the time demanded by the machine to perform a prediction, and its trade-off in models' performance.

2.2 Materials and Methods

2.2.1. Soil spectral acquisition and laboratory analysis

The experiment was conducted in a sandy loam agricultural area of 6.0 ha in Piracicaba, São Paulo state, Brazil (22°43'03.51"S, 47°36'50.03"W) where online NIR spectra were acquired. In the last three years, soybean was cultivated during summer season following a fallow system during winter.

In November 2021, a structure mounted in the tractor's three-point hydraulic hitch carried a subsoiler shank in 0.15 m depth, opening a furrow. The shank tip smoothened the bottom of the furrow. A steel armored case coupled in the back of the shank carried the spectrophotometer MicroNIR OnSite-W (Viavi Solutions Inc., California, EUA), acquiring spectra from 908.1-1676.2 nm, with a resolution of 6.2 nm, totaling 125 wavelengths. The spectra were acquired through a sapphire window, exported via USB, and converted via an Ethernet cable connected to a notebook. A 99% reflectance disc was used as maximum reflectance reference, and the own system had a minimum reflectance measuring system. A global navigation satellite system (GNSS) antenna Ag-Star (Novatel, Calgary, Canada) with TerraStar-C correction signal (Hexagon, São Paulo, Brazil) was used to track spectra geographical coordinates.

The spectrophotometer software performs a principal component analysis (PCA), excluding samples that are outside the established confidence region, generating an average spectrum every 10 seconds. The acquisition lines were separated by 12 m. Following an operation speed of 0.583 m s^{-1} (2.1 km h^{-1}) and the spectrophotometer acquisition time, 383 online NIR spectra was acquired in the area. A descriptive analysis aiming to exclude errored acquisition points, like field borders, was carried out before the data modeling.

One day after spectral acquisition, 72 soil samples were collected in the bottom of the furrow left by the shank tip. Aiming to calibrate prediction models with online spectra, the sampling region needed to match the spectral transect acquired. All start and end momentum of spectral acquisition points, signaled in the software, was simultaneously demarcated during field operation. The product of operation speed and acquisition time resulted in the transect length to be sampled (Figure 1).

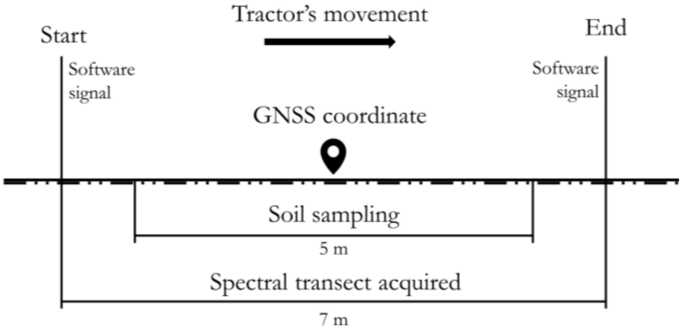


Figure 1. Schematic representation of soil spectral acquisition and associated sampling region.

Soil physical-chemical analysis were performed in a commercial laboratory. The attributes considered in this study and its respective analysis method were clay and sand – NaOH dispersant, OM - oxidation, CEC – sum of basis plus total acidity, pH – $CaCl_2$ and K - resin.

2.2.2. Data modeling

Data modeling was conducted in the software Jupyter Notebook (Kluyver et al., 2016; Python Software Foundation, 2022). A descriptive analysis, aiming to exclude sampling errors due to the field operation, as during maneuvers, resulted in the exclusion of 80 spectra. The dataset remained with 303 spectra.

The calibrations built tested two approaches. Only spectra from the study area were used (Local dataset) and online spectra from experimental field gathered with laboratory spectra from other two fields (Global dataset). The samples from the two other fields were at the laboratory database and their respective descriptions can be found in Eitelwein (2017). This strategy is suggested to enhance models' performance due to data augmentation in the calibration (Munnaf et al., 2019; Guerrero et al., 2021).

The modeling tested the dimensionality reduction statistical models PLSR, principal components regression (PCR) and least absolute shrinkage and selection operator (Lasso), due to the multivariate character of soil spectra (Stenberg et al., 2010). Random forest (RF) and extratrees (ET), two non-linear regression models, were also tested.

PLSR is a common technique applied in soil science (Bellon-Maurel & McBratney, 2011). It reduces data dimensionality producing latent variables (LV), named X scores (Wold et al., 2001). The combinations of latent variables generate the linear regression model (Kuang et al. 2015). The y-residuals, or deviations between measured and predicted values, are also obtained (Equation I).

$$y = bX + e \qquad \text{Eq. I,}$$

where, y = vector of response variables, b = vector of regression coefficients, X = matrix of independent variables and e = vector that indicates y-residuals.

PCR occurs in three steps: a) PCA in the data matrix; b) a linear regression is applied to obtain the vector of estimated regression coefficients; c) the loadings (eigenvectors) of PCA are used to obtain the estimators (β) of PCR, as shown in Equation II.

$$\widehat{\beta_k} = V_k \widehat{\delta_k} \qquad \text{Eq. II,}$$

where, $\hat{\beta}$ = PCR estimator, k belongs to {1, ..., p}, p = number of covariants, V = orthonormal set of eigenvectors, $\hat{\delta}$ = coefficients vector of estimated regressors.

Lasso is a penalty regression l1-norm that aims to find β = {βj}, values that minimizes Equation III (Tibshirani, 2011).

$$\sum_{i=1}^{N}(y_i - \sum_j x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p}|\beta_j| \qquad \text{Eq. III,}$$

where, xij = normalized predictors, yi = central response values, i = 1, 2, ..., N and j = 1, 2, ..., p.

RF regression is a combination of decision trees, where each tree depends on a randomly sampled vector (Breiman, 2001). It reduces the bias basing the result on various decision trees (Blanco et al. 2018).

Statistical model ET, also known as Extremely Randomized Trees, differs from RF because has a robuster randomization of decision trees (Geurts et al., 2006). The ET model calculates an optimal value to be used as data matrix splitting rules. This difference leads to more diverse trees and fewer splits when training the model. The randomness obtained additionally allows lower processing cost to work with ET and reduction of the variance of the model, but with slightly higher bias (Eslami et al., 2019).

Data was split in the proportion of 75% for calibration and 25% for validation, using k-fold cross-validation, with k = 10 (Jung et al., 2018). In this technique, each sample is used one time in the validation set, and k-1 times in calibration, ensuring a less biased estimation of the model performance.

Statistical metrics used was the coefficient of determination (R^2), the root mean squared error (RMSE) and the mean absolute error (MAE). The higher the R^2 values, and lower the RMSE and MAE, it is implied that the model had a better performance.

2.2.3. Preprocessing techniques

After the data modeling step, the method with the best results reported and the PLSR, as the most observed regression model in the literature of soil spectroscopy, were used to test the feasibility of preprocessing techniques. Raw data (RD), preprocessing sequence 1 (PP1) and preprocessing sequence 2 (PP2) datasets were created to compare the predictive performance applying preprocessing techniques (Table 1). The techniques were tested in sequences as observed in the literature (Guerrero et al., 2021; Munnaf et al. 2019; Tavares et al., 2020).

Table 1. Datasets created to test the application of preprocessing techniques.

Acronym	Preprocessing techniques order						
RD			--------------------------------				
PP1	MA	+	1st SG derivative	+	2nd SG derivative	+	smoothing SG
PP2	MA	+	MN (0,1)	+	1st SG derivative	+	smoothing SG

RD: raw data; PP1: preprocessing sequence 1; PP2: preprocessing sequence 2; MA: moving average; MN: maximum normalization; SG: Savitzky-Golay.

Preprocessing methods applied were the moving average, to reduce the effect of noisy wavebands; maximum normalization, a method to scale and uniform the distribution of variation (Rinnan et al., 2009); first and second Savitzky-Golay (SG) derivatives, used to reduce noise and highlight spectral features and possible hidden information (Ben Dor et al., 1995); and SG smoothing algorithm.

A 95% Kruskal-Wallis test was applied in the prediction values from RD, PP1 and PP2. This statistic is used to assess the similarity of k groups, presenting the degrees of freedom, the chi-squared (X^2), and the calculated p-value. The test is significant if p-value < 0.05.

In this study, the processing cost was considered as the time demanded by the machine to perform a prediction. The *datetime* function within *datetime* library (Python Software Foundation, 2022) was used to measure the time needed for ten consecutive predictions (repetitions) of each attribute considered. The time of each repetition was calculated by the average of the prediction of the six attributes evaluated. Data were processed in a notebook with the specifications: SSD Kingston NV1 2280 NVMe, IntelCore i5 octa-core 1.60 GHz processor, 8 Gb of DDR4 2,666 MHz memory.

2.3 Results and Discussion

2.3.1. Statistical model selection

Global dataset presented the best R^2 in validation using raw spectral data, while Local dataset presented the lowest errors of prediction (Table 2). In general, R^2 gives the idea of the adjustment of the model comparing the predicted and observed values. The lower the errors of prediction, the higher R^2 is expected to be. However, especially for Local models, some cases were out of this rule. OM prediction with PLSR had RMSE of 3.48 g kg^{-1} and MAE of 2.75 g kg^{-1}, with R^2 0.22. Lasso prediction for the same attribute presented RMSE of 4.28 g kg^{-1} and MAE of 3.09 g kg^{-1}, with R^2 0.54. For CEC, PCR and Lasso models, respectively, presented 3.54 mmol$_c$ kg^{-1} and 3.04 mmol$_c$ kg^{-1} of RMSE, 3.82 mmol$_c$ kg^{-1} and 3.13 mmol$_c$ kg^{-1} of MAE, but the R^2 values were 0.70 for PCR to 0.39 for Lasso.

For both datasets, PCR and Lasso performed better in nine out of 12 prediction models calibrated. PLSR performed better for clay of Local and sand of Global datasets. Overall, the dimensionality reduction models outperformed the non-linear models tested, RF and ET. The K prediction with ET of Local dataset were the only non-linear model with the best R^2, RMSE and MAE.

Soil electromagnetic spectra is multivariate (Stenberg et al., 2010; Nocita et al., 2015), sensors reported in literature will acquire from 100 up to 2,200 features (Nawar & Mouazen, 2019; Coblinski et al., 2021; Eitelwein et al., 2022). Few algorithms are able to train powerful models if the number of observations (n) in the dataset is lower than the number of features (p); this is known as the “Curse of Dimensionality”. Thus, p needs to be reduced for efficient data modeling (Velliangiri et al., 2019). PCR uses the feature extraction method for dimensionality reduction. Since no feature is excluded, dimension can be decreased without losing much information of the initial dataset (Jolliffe, 2011; Velliangiri et al., 2019). This method is demonstrably robust to noisy, sparse, and possibly mixed valued covariates (Agarwal et al., 2019). These factors can explain the performance this model achieved in this study using online NIR soil spectra, and corroborate other studies that applied PCR to laboratory NIR spectra (Chang et al., 2001; Pudełko & Chodak, 2020; Wei et al., 2022). PCR obtained the best numerical parameters in five of the 12 prediction models calibrated, and values close to the best calibration method in clay prediction of Local dataset, and clay, OM, CEC and K of Global dataset.

Lasso regression uses all predictors variables inserted in calibration, but weights each one of them, penalizing those with lower importance in the variance of target variable, being so considered as a dimensionality reduction model (Tibshirani, 2011). However, it differs from PCR and PLSR, as these last will not use all the original variables and will perform the linear regression using as predictors the features extracted from original dataset (principal components or LVs), selected in calibration. This reduces considerably the processing cost. A PCR calibration and prediction took an average of 2% of the time demanded to perform a Lasso prediction. Therefore, in this study, the PCR was selected as the most efficient model tested for soil attributes prediction using DRS NIR.

2.3.2. Prediction performance using raw and preprocessed data

The validation parameters of prediction models testing raw and preprocessed data are presented in Table 3. For PCR modeling, PP2 did not presented any models with the best validation parameters, in both Local and Global datasets. Raw data had the best parameters for sand, CEC and pH predictions using Global dataset, and clay, sand and CEC using Local dataset, totalizing six models. The modeling applying PP1 performed better for clay, OM and K prediction with Global dataset, and for OM, pH and K with Local dataset.

The best parameters for Global dataset prediction using PLSR were divided into the three strategies: raw data had the best numerical performance for physical attributes clay and sand, PP1 for CEC and K, and PP2 for OM and pH. For Local dataset prediction applying PLSR, PP1 was the one without any best reported predictions. Clay, OM and CEC had the best parameters using raw data, and sand, pH and K using PP2.

No pattern on the best strategy (raw data, or preprocessing sequences) was observed using either PCR or PLSR. The results varied for dataset and attribute predicted. This can imply that it is difficult to assess the direct effects of preprocessing techniques applied to soil spectroscopy prediction modeling. Other studies that compared preprocessing techniques for soil attributes prediction using visible and NIR spectra also struggle to define a best method (Benedet et al., 2020; Wei et al., 2022).

Preprocessing techniques are commonly applied in soil spectroscopy to remove noisy and redundant data, and highlight important features. However, the best numerical parameters observed, as previously highlighted in this study, are usually the method authors adopt to define which strategy to follow (Benedet et al., 2020; Munnaf et al., 2021a; Munnaf et al., 2021b; Wang et al., 2020). Although best values of the given metrics used in this research area can be defined as the highest R^2 and ratio of performance to interquartile distance (RPIQ), or the lowest errors of prediction (RMSE or MAE), when it comes to close values, as observed in the test of RD, PP1 and PP2, the existence or not of a proven statistical difference needs to be addressed. The main parameters used for soil spectroscopy ML evaluation are metrics of mean values. Therefore, the 95% Kruskal-Wallis test was carried out, and detected no statistical difference among the predicted values using the three strategies for any soil attribute considered (Table 4).

For an efficient ML modeling, any step added to the process must be justified, as it inevitably represents a higher processing cost, requiring either more time for prediction or the use of more sophisticated machines, making the application of the technique more expensive and hindering its leverage in agriculture. Thus, the efficiency of the acquisition, processing and analysis must be considered. The processing time for the machine to perform ten predictions of the six attributes considered in this study was measured, using the strategies RD, PP1, and PP2. The time required for prediction with preprocessed spectra was, on average, 200% higher than from using raw spectral data (Table 5).

Table 2. Validation parameters of principal components regression (PCR), least absolute shrinkage and selection operator (Lasso), partial least squares regression (PLSR), random forest (RF) and extratrees (ET) in Local and Global datasets.

	PCR				Lasso			PLSR				RF			ET		
	NC	R^2	RMSE	MAE	R^2	RMSE	MAE	LV	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE
								Local dataset									
clay	4	0.15	21.61	15.88	**0.16**	**20.97**	**15.26**	1	-0.08	29.76	22.96	0.04	24.93	19.43	0.04	24.93	19.41
sand	1	0.02	24.01	17.67	0.00	24.35	18.49	**6**	**0.06**	**20.72**	**17.14**	0.07	25.27	20.02	0.02	28.86	23.02
OM	**9**	**0.71**	**3.34**	**2.48**	0.54	4.28	3.09	7	0.22	3.48	2.75	0.00	5.85	4.33	0.15	5.95	4.46
CEC	**9**	**0.70**	**3.54**	**3.04**	0.39	3.82	3.13	4	0.03	5.39	3.71	0.00	4.92	3.62	0.21	5.12	3.61
pH	**4**	**0.03**	**0.29**	**0.24**	0.01	0.30	0.27	4	-0.26	0.50	0.38	0.15	0.40	0.34	0.02	0.38	0.31
K	2	0.03	1.00	0.78	0.00	2.99	2.95	3	-0.22	0.94	0.76	0.21	1.09	0.87	**0.38**	**0.86**	**0.76**
								Global dataset									
clay	8	0.94	31.45	23.77	0.94	32.16	24.20	**7**	**0.94**	**29.22**	**23.82**	0.68	69.58	49.22	0.65	80.26	60.20
sand	**10**	**0.97**	**36.37**	**28.47**	0.97	40.19	29.92	10	0.97	39.12	30.19	0.75	106.58	60.52	0.69	117.29	58.19
OM	10	0.80	3.15	2.39	**0.80**	**3.08**	**2.31**	10	0.77	3.15	2.45	0.39	4.91	3.57	0.33	5.16	3.62
CEC	5	0.77	11.18	8.44	**0.78**	**10.97**	**8.28**	5	0.65	15.58	10.75	0.71	13.67	9.89	0.69	14.07	9.74
pH	**3**	**0.51**	**0.40**	**0.26**	0.50	0.41	0.27	1	0.29	0.45	0.33	0.50	0.40	0.31	0.46	0.41	0.31
K	5	0.70	1.57	1.11	**0.70**	**1.57**	**1.06**	3	0.63	1.72	1.16	0.57	1.85	1.23	0.52	1.96	1.22

clay: g kg^{-1}; sand: g kg^{-1}; OM: organic matter, g kg^{-1}; CEC: cation exchange capacity, mmol$_c$ kg^{-1}; pH: potential of hydrogen, dimensionless; K: potassium, mmol$_c$ kg^{-1}; NC: number of principal components applied in regression; R^2: coefficient of determination; RMSE: root mean squared error; MAE: mean squared error; LV: number of latent variables applied in regression. Highlighted in bold are the models that presented the best statistical parameters on validation.

Table 3. Results presented in validation of partial least squares regression (PLSR) and principal components regression (PCR) prediction models testing the use of raw data (RD) and two different spectra preprocessing sequences (PP1 and PP2) in Local and Global datasets.

		Global dataset												Local dataset											
		RD				PP1				PP2				RD				PP1				PP2			
		NC	R^2	RMSE	MAE	NC	R^2	RMSE	MAE	NC	R^2	RMSE	MAE	NC	R^2	RMSE	MAE	NC	R^2	RMSE	MAE	NC	R^2	RMSE	MAE
PCR	clay	8	0.94	31.45	23.77	**8**	**0.95**	**28.97**	**23.21**	10	0.94	32.33	25.46	**4**	**0.15**	**21.61**	**15.88**	3	0.12	21.77	16.74	1	0.04	20.44	15.97
	sand	**10**	**0.97**	**36.37**	**28.47**	8	0.97	36.90	30.46	10	0.97	42.34	33.12	**1**	**0.02**	**24.01**	**17.67**	4	0.02	25.48	19.47	2	0.02	25.62	19.35
	OM	10	0.80	3.15	2.39	**9**	**0.80**	**3.04**	**2.20**	4	0.80	3.12	2.46	9	0.71	3.34	2.48	**9**	**0.71**	**3.24**	**2.41**	8	0.24	4.82	3.62
	CEC	**5**	**0.77**	**11.18**	**8.44**	3	0.76	11.54	8.28	3	0.76	11.59	8.93	**9**	**0.70**	**3.54**	**3.04**	9	0.58	4.29	3.43	8	0.10	5.27	4.16
	pH	**3**	**0.51**	**0.40**	**0.26**	2	0.49	0.41	0.29	5	0.51	0.40	0.30	4	0.03	0.29	0.24	**3**	**0.02**	**0.28**	**0.23**	3	0.03	0.28	0.23
	K	5	0.70	1.57	1.11	**3**	**0.72**	**1.52**	**1.03**	3	0.69	1.59	1.14	2	0.03	1.00	0.78	**5**	**0.21**	**0.92**	**0.77**	2	0.06	0.96	0.75
		LV	R^2	RMSE	MAE	LV	R^2	RMSE	MAE	LV	R^2	RMSE	MAE	LV	R^2	RMSE	MAE	LV	R^2	RMSE	MAE	LV	R^2	RMSE	MAE
PLSR	clay	**7**	**0.94**	**29.22**	**23.82**	4	0.93	31.68	23.78	3	0.80	57.04	40.54	**1**	**-0.08**	**29.76**	**22.96**	3	-0.21	31.49	24.26	1	-0.19	31.18	22.10
	sand	**10**	**0.97**	**39.12**	**30.19**	16	0.96	44.76	36.98	7	0.96	48.38	36.42	6	0.06	20.72	17.14	2	0.12	19.99	16.51	**2**	**0.15**	**19.65**	**15.64**
	OM	10	0.77	3.15	2.45	8	0.78	2.94	2.37	**9**	**0.80**	**2.87**	**2.17**	**7**	**0.22**	**3.48**	**2.75**	4	-0.08	4.10	3.18	5	-0.08	4.11	3.27
	CEC	5	0.65	15.58	10.75	**2**	**0.69**	**10.88**	**7.80**	3	0.56	16.58	11.09	**4**	**0.03**	**5.39**	**3.71**	5	-0.16	5.90	4.46	7	-0.28	6.20	4.95
	pH	1	0.29	0.45	0.33	1	0.29	0.47	0.35	**3**	**0.48**	**0.37**	**0.28**	4	-0.26	0.50	0.38	1	-0.12	0.47	0.32	**1**	**-0.09**	**0.47**	**0.32**
	K	3	0.63	1.72	1.16	**3**	**0.81**	**1.44**	**1.04**	3	0.68	1.72	1.28	3	-0.22	0.94	0.76	15	-0.50	1.04	0.76	**2**	**-0.06**	**0.87**	**0.71**

clay: g kg^{-1}; sand: g kg^{-1}; OM: organic matter, g kg^{-1}; CEC: cation exchange capacity, $mmol_c$ kg^{-1}; pH: potential of hydrogen; K: potassium, $mmol_c$ kg^{-1}; NC: number of principal components applied in regression; R^2: coefficient of determination; RMSE: root mean squared error; MAE: mean squared error; LV: number of latent variables applied in regression. Highlighted in bold are the models that presented the best statistical parameters on validation.

Table 4. Kruskal-Wallis test results for comparison of physicochemical soil attributes predicted attributes using online NIR spectra.

	p-value	
	Local	Global
clay	0.734	0.958
sand	0.998	0.472
OM	0.849	0.157
CEC	0.735	0.522
pH	0.970	0.079
K	0.988	0.980

OM: organic matter; CEC: cation exchange capacity; pH: potential of hydrogen; K: potassium.

Once a ML model is calibrated, when an online spectrum is acquired, it can be directly applied to prediction. Preprocessing sequences applied before prediction will consume extra time; that can be justified if it represents a gain in the accuracy. However, this was not observed in this study. The addition of the steps of preprocessing resulted in twice the processing cost, here assessed as the time required for prediction, with no gain in predictive accuracy.

Even the existence of automatic methods to test preprocessing techniques and fit predictive models using soil spectra, such as the "all-possibilities approach", defined as "an extremely computer power-consuming method" (Kopacková et al., 2017), its applicability is still questionable. This relapses in the efficiency of the processing protocol, especially for online DRS soil attributes prediction, which the use of ML calibrations can develop AI systems to predict soil attributes and intervein in the field in real time. For this achievement, standard, rapid, simple and efficient processing protocol, preferably reducing human interference, needs to be developed (Rossi et al., 2022; Wei et al., 2022). Therefore, it is suggested that the use of spectrum preprocessing techniques in the scenario evaluated in this study was not feasible.

Table 5. Average time processing cost demanded by the machine to predict ten times the six attributes evaluated in this study.

	RD	PP1	PP2
		seconds	
1	0.1346	0.2692	0.2692
2	0.0470	0.0997	0.1012
3	0.1037	0.2074	0.2074
4	0.1166	0.2356	0.2333
5	0.1306	0.2586	0.2639
6	0.1126	0.2422	0.2422
7	0.1346	0.2558	0.2558
8	0.1152	0.2315	0.2315
9	0.0985	0.1757	0.2010
10	0.1198	0.2397	0.2397
Mean	0.1113	0.2215	0.2245
%	-	199%	202%

RD: raw data; PP1: preprocessing sequence 1; PP2: preprocessing sequence 2; %: % of time compared with the model of lower processing time.

Future works should further investigate whether non-linear models can perform better with augmented number of samples, possibly outperforming the dimensionality reduction statistical models. Also, investigate the differences observed in the prediction using raw and preprocessed spectra, to confirm if the reported in this study is repeated for other areas.

2.4 Conclusions

Five statistical methods were applied to online NIR spectra prediction of soil attributes in a Brazilian tropical field. The dimensionality reduction statistical models performed better than the non-linear models. PCR models was mostly more accurate than PLSR models, and more efficient than Lasso models since it consumed less computer processing time.

No group of predicted values, for any of the six attributes evaluated, presented statistical difference from the others. The use of preprocessing techniques did not reach the expected objective of aiding the model to be more accurate. However, the application of these techniques increased the time required by the machine to perform a prediction using field spectra. It is then suggested that raw data was the most efficient for the scenario evaluated in this study.

References

Agarwal, A., Shah, D., Shen, D., & Song, D. (2019). On robustness of principal component regression. **Advances in Neural Information Processing Systems**, 32.

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

Ben-Dor, E., & Banin, A. (1995). Near-infrared analysis as a rapid method to simultaneously evaluate several soil properties. **Soil Science Society of America Journal**, 59(2), 364-372.

Benedet, L., Faria, W. M., Silva, S. H. G., Mancini, M., Demattê, J. A. M., Guilherme, L. R. G., & Curi, N. (2020). Soil texture prediction using portable X-ray fluorescence spectrometry and visible near-infrared diffuse reflectance spectroscopy. **Geoderma**, 376, 114553.

Blanco, C.M.G.; Gomez, V.M.B.; Crespo, P. (2018) Spatial prediction of soil water retention in a Páramo landscape: Methodological insight into machine learning using random forest. **Geoderma**. 316, 100–114.

Breiman, L. (2001). Random forests. **Machine Learning**, 45(1), 5-32.

Chang, C. W., Laird, D. A., Mausbach, M. J., & Hurburgh, C. R. (2001). Near-infrared reflectance spectroscopy–principal components regression analyses of soil properties**. Soil Science Society of America Journal**, 65(2), 480-490.

Cherubin, M. R., Santi, A. L., Eitelwein, M. T., Menegol, D. R., Ros, C. O. D., Pias, O. H. D. C., & Berghetti, J. (2014). Eficiência de malhas amostrais utilizadas na caracterização da variabilidade espacial de fósforo e potássio. **Ciência Rural**, 44, 425-432.

Cherubin, M. R., Santi, A. L., Eitelwein, M. T., Amado, T. J. C., Simon, D. H., & Damian, J. M. (2015). Dimensão da malha amostral para caracterização da variabilidade espacial de fósforo e potássio em Latossolo Vermelho. **Pesquisa Agropecuária Brasileira**, 50(2), 168-177.

Coblinski, J. A., Inda, A. V., Dematte, J. A., Dotto, A. C., Gholizadeh, A., & Giasson, E. (2021). Identification of minerals in subtropical soils with different textural classes by VIS–NIR–SWIR reflectance spectroscopy. **Catena**, 203, 105334.

Dotto, A. C., Dalmolin, R. S. D., ten Caten, A., & Grunwald, S. (2018). A systematic study on the application of scatter-corrective and spectral-derivative preprocessing for multivariate prediction of soil organic carbon by Vis-NIR spectra. **Geoderma**, 314, 262-274.

Eitelwein, M. T. (2017). Sensoriamento proximal de solo para a quantificação de atributos químicos e físicos (Doctoral dissertation, Universidade de São Paulo).

Eitelwein, M. T., Tavares, T. R., Molin, J. P., Trevisan, R. G., de Sousa, R. V., & Demattê, J. A. M. (2022). Predictive Performance of Mobile Vis–NIR Spectroscopy for Mapping Key Fertility Attributes in Tropical Soils through Local Models Using PLS and ANN. **Automation**, 3(1), 116-131.

Eslami, E.; Salman, A. K.; Choi, Y.; Sayeed, A.; & Lops, Y. (2019). A data ensemble approach for real-time air quality forecasting using extremely randomized trees and deep neural networks. **Neural Comput. Appl.**, vol. 6, Mar. 2019, doi: 10.1007/s00521-019-04287-6.

Franceschini, M. H. D., Demattê, J. A. M., Kooistra, L., Bartholomeus, H., Rizzo, R., Fongaro, C. T., & Molin, J. P. (2018). Effects of external factors on soil reflectance measured on-the-go and assessment of potential spectral correction through orthogonalisation and standardisation procedures. **Soil and Tillage Research**, 177, 19-36.

Geurts, P.; Ernst, D.; & Wehenkel, L. (2006) Extremely randomized trees. **Machine Learning**, vol. 63, no. 1, pp. 3–42. doi: 10.1007/s10994- 006-6226-1.

Guerrero, A., De Neve, S., & Mouazen, A. M. (2021). Data fusion approach for map-based variable-rate nitrogen fertilization in barley and wheat. **Soil and Tillage Research**, 205, 104789.

Jolliffe, I. (2011). Principal component analysis (pp. 1094-1096). Springer Berlin Heidelberg. RESUME SELİN DEĞİRMECİ Marmara University, Goztepe Campus ProQuest Number: ProQuest). Copyright of the Dissertation is held by the Author. All Rights Reserved, 28243034, 28243034.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B. E., Bussonnier, M., Frederic, J., ... & Willing, C. (2016). Jupyter Notebooks-a publishing format for reproducible computational workflows (Vol. 2016, pp. 87-90).

Kuang, B., Tekin, Y., & Mouazen, A. M. (2015). Comparison between artificial neural network and partial least squares for on-line visible and near infrared spectroscopy measurement of soil organic carbon, pH and clay content. **Soil and Tillage Research**, 146, 243-252.

Montanari, R., Souza, G. S. A., Pereira, G. T., Marques, J. U. N. I. O. R., Siqueira, D. S., & Siqueira, G. M. (2012). The use of scaled semivariograms to plan soil sampling in sugarcane fields. **Precision Agriculture**, 13(5), 542-552.

Munnaf, A. M., Nawar, S., & Mouazen, A. M. (2019). Estimation of Secondary Soil Properties by Fusion of Laboratory and On-Line Measured Vis–NIR Spectra. **Remote Sensing**, 11(23), 2819.

Munnaf, M. A., Guerrero, A., Nawar, S., Haesaert, G., Van Meirvenne, M., & Mouazen, A. M. (2021a). A combined data mining approach for on-line prediction of key soil quality indicators by Vis-NIR spectroscopy. **Soil and Tillage Research**, 205, 104808.

Munnaf, M. A., Haesaert, G., Van Meirvenne, M., & Mouazen, A. M. (2021b). Multi-sensors data fusion approach for site-specific seeding of consumption and seed potato production. **Precision Agriculture**, 22(6), 1890-1917.

Nawar, S., & Mouazen, A. M. (2019). On-line vis-NIR spectroscopy prediction of soil organic carbon using machine learning**. Soil and Tillage Research**, 190, 120-127.

Nocita, M., Stevens, A., van Wesemael, B., Aitkenhead, M., Bachmann, M., Barthès, B., ... & Wetterlind, J. (2015). Soil spectroscopy: An alternative to wet chemistry for soil monitoring. In **Advances in Agronomy** (Vol. 132, pp. 139-159). Academic Press.

Pasquini, C. (2018). Near infrared spectroscopy: A mature analytical technique with new perspectives–A review. **Analytica chimica acta**, 1026, 8-36.

Python Software Foundation (2022). https://www.python.org/psf/

Pudełko, A., & Chodak, M. (2020). Estimation of total nitrogen and organic carbon contents in mine soils with NIR reflectance spectroscopy and various chemometric methods. **Geoderma**, 368, 114306.

Rinnan, Å., Van Den Berg, F., & Engelsen, S. B. (2009). Review of the most common preprocessing techniques for near-infrared spectra. **TrAC Trends in Analytical Chemistry**, 28(10), 1201-1222.

Rossel, R. V., & Behrens, T. (2010). Using data mining to model and interpret soil diffuse reflectance spectra. **Geoderma**, 158(1-2), 46-54.

Rossi, P.; Mangiavacchi, P.L.; Monarca, D.; Cecchini, M. Smart Machinery and Devices for Reducing Risks from Human-Machine Interference in Agriculture: A Review. In **Safety, Health and Welfare in Agriculture and Agro-food Systems**; Biocca, M., Cavallo, E., Cecchine, M., Failla, S., Romano, E., Eds.; Springer: Cham, Switzerland, 2022; pp. 195–204.

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.

Tavares, T. R., Molin, J. P., Javadi, S. H., Carvalho, H. W. P. D., & Mouazen, A. M. (2020). Combined use of vis-NIR and XRF sensors for tropical soil fertility analysis: Assessing different data fusion approaches. Sensors, 21(1), 148.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. **Journal of the Royal Statistical Society**, 73, 273-282.

Velliangiri, S., Alagumuthukrishnan, S., & Thankumar J. S. I. (2019). A review of dimensionality reduction techniques for efficient computation. **Procedia Computer Science**, 165, 104-111.

Viscarra Rossel, R. A., Adamchuk, V. I., Sudduth, K. A., McKenzie, N. J., & Lobsey, C. (2011). Proximal soil sensing: an effective approach for soil measurements in space and time. In **Advances in agronomy** (Vol. 113, pp. 243-291). Academic Press.

Wang, Y. P., Lee, C. K., Dai, Y. H., & Shen, Y. (2020). Effect of wetting on the determination of soil organic matter content using visible and near-infrared spectrometer. **Geoderma**, 376, 114528.

Wei, M. C. F., Canal Filho, R., Tavares, T. R., Molin, J. P., & Vieira, A. M. C. (2022). Dimensionality Reduction Statistical Models for Soil Attribute Prediction Based on Raw Spectral Data. **Artificial Intelligence**, 3(4), 809-819.

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. **Chemometrics and Intelligent Laboratory Systems**, 58(2), 109-130.

Wollenhaupt, N. C., Wolkowski, R. P., & Clayton, M. K. (1994). Mapping soil test phosphorus and potassium for variable-rate fertilizer application. **Journal of production agriculture**, 7(4), 441-448.

Zhang, J., Guerrero, A., & Mouazen, A. M. (2021). Map-based variable-rate manure application in wheat using a data fusion approach. **Soil and Tillage Research**, 207, 104846.

3. SPATIAL DISTRIBUTION AS A KEY FACTOR FOR EVALUATION OF SOIL ATTRIBUTES PREDICTION AT FIELD LEVEL USING ONLINE NEAR-INFRARED SPECTROSCOPY

This section has been published in Frontiers in Soil Science journal (Canal Filho R and Molin JP. 2022. Spatial distribution as a key factor for evaluation of soil attributes prediction at field level using online near-infrared spectroscopy. *Front. Soil Sci.* 2:984963. doi: 10.3389/fsoil.2022.984963)

Abstract

In soil science, near-infrared (NIR) spectra are being largely tested to acquire data directly in the field. Machine learning (ML) models using these spectra can be calibrated, adding only samples from one field or gathering different areas to augment the data inserted and enhance the models' accuracy. Robustness assessment of prediction models usually rely on statistical metrics. However, how the spatial distribution of predicted soil attributes can be affected is still little explored, despite the fact that agriculture productive decisions depend on the spatial variability of these attributes. The objective of this study was to use online NIR spectra to predict soil attributes at field level, evaluating the statistical metrics and also the spatial distribution observed in prediction to compare a local prediction model with models that gathered samples from other areas. A total of 383 online NIR spectra were acquired in an experimental field to predict clay, sand, organic matter (OM), cation exchange capacity (CEC), potassium (K), calcium (Ca), and magnesium (Mg). To build ML calibrations, 72 soil spectra from the experimental field (local dataset) were gathered, with 59 samples from another area nearby, in the same geological region (geological dataset) and with this area nearby and more 60 samples from another area in a different region (global dataset). Principal components regression was performed using k-fold (k=10) cross-validation. Clay models reported similar errors of prediction, and although the local model presented a lower R^2 (0.17), the spatial distribution of prediction proved that the models had similar performance. Although OM patterns were comparable between the three datasets, local prediction, with the lower R^2 (0.75), was the best fitted. However, for secondary NIR response attributes, only CEC could be successfully predicted and only using local dataset, since the statistical metrics were compatible, but the geological and global models misrepresented the spatial patterns in the field. Agronomic plausibility of spatial distribution proved to be a key factor for the evaluation of soil attributes prediction at field level. Results suggest that local calibrations are the best recommendation for diffuse reflectance spectroscopy NIR prediction of soil attributes and that statistical metrics alone can mispresent the accuracy of prediction.

Keywords: soil variability, geostatistics, diffuse reflectance spectroscopy, machine learning, agriculture management.

3.1 Introduction

PSS is a relevant technique to make soil data acquisition faster and more cost effective (Viscarra Rossel et al., 2011; Wang et al., 2015). In this sense, many authors have studied techniques to be adapted for PSS. DRS in the VNIR has been largely tested to predict soil physical and chemical attributes (Pasquini, 2018; Molin & Tavares, 2019). The prediction can perform on primary NIR response attributes, which means attributes like clay and OM, that have direct spectral absorption patterns in this region or even on secondary response attributes that do not have direct patterns in NIR but can be predicted due to the construction of indirect calibrations.

The idea of using ML models of DRS NIR spectra for soil attributes prediction lies into the choice of the statistical model and then in the accurate prediction of these attributes. Dimensionality reduction models are often chosen due to the multidimensionality of soil spectra (Williams & Norris, 1987). Besides coping with multivariate data analysis (Velliangiri et al., 2019), dimensionality reduction models can sometimes smooth the values predicted, loosing extreme values that the model considers as outliers (Bellon-Maurel et al., 2010) and therefore needs careful implementation. In this sense, PCR is a multivariate method of simple implementation, which had its potential demonstrated since the beginning of studies for soil properties prediction using DRS. Authors reported successful prediction of this technique for diverse soil attributes, such as soil organic carbon, OM, pH, and macronutrients, such as total nitrogen and total and extractable P and K (Chang et al., 2001; Barthès et al., 2008, Christy et al., 2008; Wang et al., 2015; Morellos et al., 2016). Then, statistical metrics are being used for the assessment of ML model robustness (Vishwakarma et al. 2021), such as the R², which gives the idea of the variance portion of the data that the model is explaining; the RMSE and MAE, which represent the error of prediction the model offered; and the RPIQ, which is calculated using the RMSE and the range between first and third quantiles of the data.

However, PA has in its very definition the consideration of temporal and spatial variability of agricultural production (ISPA, 2022). This fact comes from the necessity of understanding the patterns of the variability in the field, since agriculture needs to adapt or act in the variability of production. Soil physical and chemical attributes have well-known relations and patterns defined by soil science in the study of agricultural soil fertility, and these relations are studied by means of the spatial dependence in geostatistics (Abdel Rahman et al., 2021). The range of a fitted variogram means the distance in which a point is still related, or spatial dependent, to another.

With this knowledge, investigations show that the relation between soil attributes will affect the construction of ML models. Early when DRS were tested for PSS, Stenberg et al. (2010) stated that prediction models using VNIR spectrum should consider only samples from the same morphopedological formation, since the variations in soil mineralogy will affect the spectral signature, and the model will not be able to accurately predict attributes with this variation. Nevertheless, studies have been reaching satisfactory prediction metrics in constructing models not only with the fusion of samples from the same geological region (Ulusoy et al., 2016; Franceschini et al., 2018) but also using samples from fields with different soil formations (Guerrero et al., 2021).

The statistical metrics are important to define the accuracy of a prediction model. However, the way the ML calibration affects the distribution of attributes should be considered with the same importance, since this distribution will directly affect the decision making in agriculture productive process. Hence, this study aimed to understand if the insertion of outside samples in the calibration of NIR soil attributes prediction models affect the spatial dependence of predicted values. The objective was to define whether the spatial distribution should be always taken into account when evaluating the quality of prediction from an ML model for both primary and secondary NIR response soil attributes.

3.2 Materials and Methods

The steps followed for this study development are summarized by the flowchart shown in Figure 1. These steps will be further explained in detail.

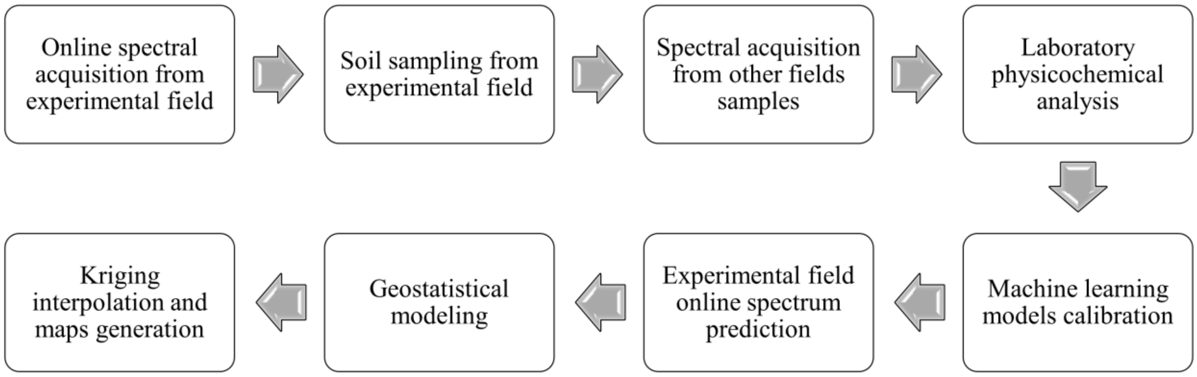

Online spectral acquisition from experimental field
Soil sampling from experimental field
Spectral acquisition from other fields samples
Laboratory physicochemical analysis
Machine learning models calibration
Experimental field online spectrum prediction
Geostatistical modeling
Kriging interpolation and maps generation

Figure 1. Flowchart of steps developed in this study.

3.2.1. Study area

The study area is located in Piracicaba, São Paulo state, Brazil (22°43'03.51"S, 47°36'50.03"W), where online NIR spectra were acquired for high spatial resolution prediction of soil attributes. Following the criteria of using another area from the same geological formation region, samples from another area of 3,300 m distance from the experimental field, described in Eitelwein (2017), were used (22°41'57.64"S, 47°38'33.13"W). For the composition of a dataset with samples from multiple geological formations, samples were added from an area located in Mato Grosso state, Brazil (14°06'05.02"S, 57° 46'01.66"W), also described in Eitelwein (2017) (Figure 2).

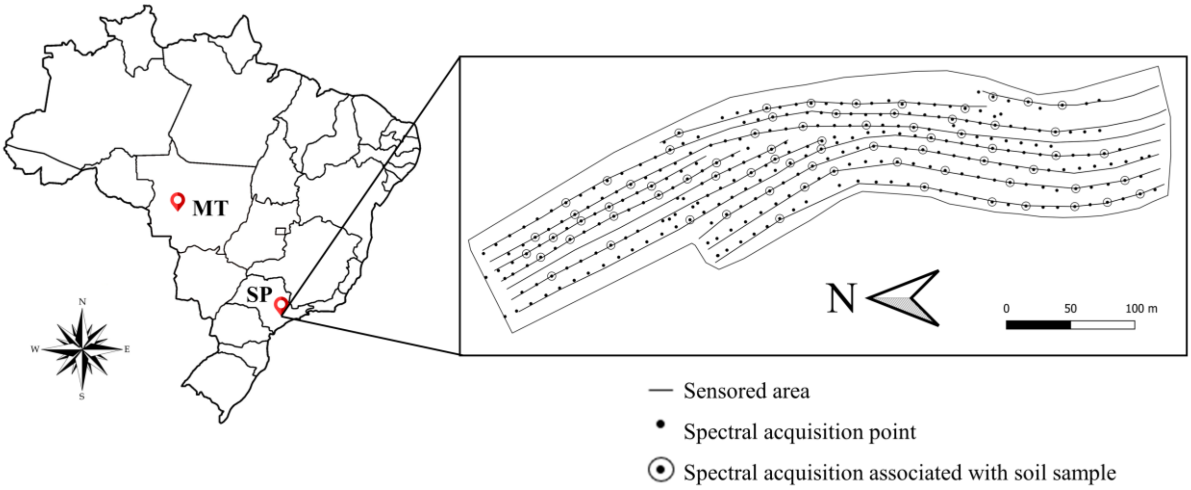

MT
SP
N
W
E
S
N
0
50
100 m
— Sensored area
● Spectral acquisition point
⊙ Spectral acquisition associated with soil sample

Figure 2. Location of areas from where samples were acquired for models' calibrations. Highlighted, located in Piracicaba, São Paulo State (SP), the experimental field shape, sensored transects, spectral points acquisition and associated soil samples. Samples from another field nearby were used to compose Geological dataset. Global dataset was built adding samples from another area, located in Mato Grosso (MT) state.

3.2.2. Online spectral acquisition and soil sampling

In November 2021, online soil spectral data were acquired using a structure mounted on the three-point hydraulic hitch of a tractor. A subsoiler shank was attached to this structure carrying a steel armored case that protects the NIR spectrophotometer (MicroNIR from VIAVI Solutions Inc., USA). The tip of the shank makes the 0.15-m-depth furrow, and the soil is smoothed by the bottom of the case, where the NIR spectrophotometer collects online soil spectra through a sapphire window at a spectral resolution of 908.1–1676.2 nm, every 6.2 nm, resulting in 125 different wavelengths. Spectra are collected at the base of the case, which were transported by a USB cable, converted for transmission via an ethernet cable, and recorded on a laptop computer.

A 99% reflectance disk was used as reference for white (maximum reflectance), and the equipment itself has an internal reference measurement for black (minimum reflectance). Each spectrum collected in the field was associated with its geographic coordinates using a GNSS Ag-Star (Novatel, Calgary, Canada) receiver with TerraStar-C differential correction (Hexagon, Alabama, USA).

The tractor traveled the area in the normal direction of the machine traffic, limited by the presence of terraces and with 12 m between each transect sensored. The spectrometer carries an internal data acquisition that groups spectra samples using principal components (PCs), excluding samples that are outside the confidence interval established in the software, and thus generates a spectrum by the mean. The acquisition time was 10 s each at a speed of 0.583 m s^{-1} (2.1 km h^{-1}), resulting in 383 online NIR spectra acquired. During the field operation, 72 random starting sensing points (12 samples ha^{-1}), indicated by the acquisition software, were demarcated and further sampled at the bottom of the furrow, excluding 1.0 m at the beginning and at the end of the transect, which aimed to overlap the area that corresponded to an online spectrum acquired (Figure 3). Those samples were submitted for laboratory analysis and used for model calibration. In addition, the density of 12 samples ha^{-1} allowed to generate maps from laboratory analysis to be used as counter proof of the models' prediction.

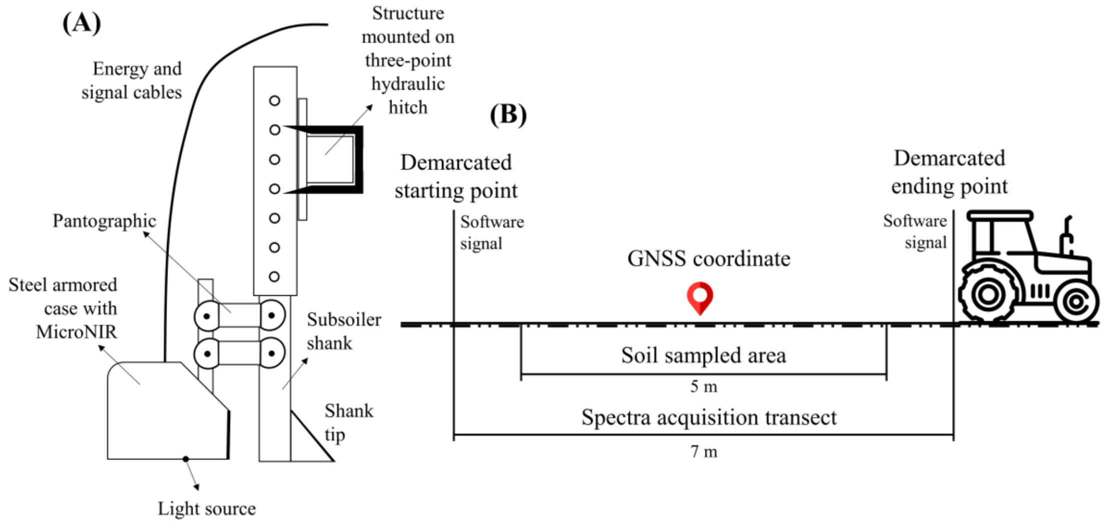


Figure 3. (A) Scheme of subsoiler shank carrying the spectrophotometer; (B) Scheme of spectral acquisition, associated soil sample and coordinate.

3.2.3. Soil physicochemical analysis

Soil physicochemical analysis were carried out on a commercial laboratory. The soil attributes that were considered and the respective analysis method were as follows: clay and sand, HMFS+NaOH; OM, oxidation; CEC, sum of basis (resin) plus soil total acidity (KCl); and Mg, K and Ca, resin. P models were discarded, as the preliminary analysis presented its independence distribution with primary NIR response attributes in the experimental area (Stenberg et al., 2010).

3.2.4. Prediction models calibration

The software Jupyter Notebook (Kluyver et al., 2016; Python Software Foundation, 2022) was used for data processing. Calibration models were built using three datasets: local—only the 72 samples from the experimental field; geological—adding 59 samples from a field of the same morphopedological region, nearby; and global—adding 60 samples from a field in Mato Grosso on the geological dataset.

Adding samples from other areas is a strategy adopted by researchers to augment the number of observations in the calibration, thus improving the accuracy of model (Munnaf et al., 2019; Guerrero et al., 2021; Zhang et al., 2021).

The statistical model used was the PCR. PCR is a dimensionality reduction model, indicated to build calibrations with soil spectra due to its multidimensionality characteristic and the possible collinearity among variables (Williams & Norris, 1987). Velliangiri et al. (2019) described that dimensionality reduction models, such as

PCR, can aid ML models in the removal of noisy and redundant data. Therefore, raw spectral data were used for models calibration in this study. Each dataset was randomly divided in the proportion of 70% for calibration and 30% for validation, using k-fold (k = 10) cross-validation (Jung, 2018), which is recommended for the evaluation of ML models to reduce bias. A random state in the software function was always set to ensure repeatability and that after the split, the same 21 samples from the experimental field would be used for the validation of all three calibration strategies. The assessment of the models' accuracy was performed using common metrics from the literature of soil attributes prediction using VNIR spectra: R^2, RMSE, MAE, and RPIQ. The parameters were evaluated as the higher the R^2 and RPIQ values and the lower the RMSE and the MAE values, the better was the model performance.

3.2.5. High spatial resolution prediction and data interpolation

The models calibrated were then used to predict the soil attributes considered using the online spectra acquired in the experimental field. A descriptive analysis aiming to exclude acquisition points, like field borders, was carried out before the prediction, which resulted in the use of 303 online spectra for prediction that were then used for data interpolation. Data of each attribute were individually interpolated by ordinary kriging, using the software VESPER (Minasny et al., 2006). The method used was block kriging, in 3.0 × 3.0 m pixels, and the minimum and maximum neighboring points for interpolation was determined as 4 and 300, respectively. Additional kriging parameters are available in Table A1 of Appendix. After kriging interpolation, the maps generated for each predicted attribute were exported to QGIS software (QGIS Development Team, 2022) for analysis and comparison.

3.3 Results and Discussion

3.3.1. Soil attributes correlation

The correlation observed among soil attributes can indicate that a secondary calibration that can be explored (Stenberg et al., 2010). The Pearson correlations of datasets used in this study are presented in Figure 4. For the local dataset, which only contains samples from the experimental field, the only primary–secondary NIR response attributes strong correlation observed is OM-CEC of 0.76. On the other hand, the geological and global datasets presented all common physicochemical correlations: clay and OM strongly and positively correlated to CEC and, consequently, to plant nutrients (Syers et al., 1970).

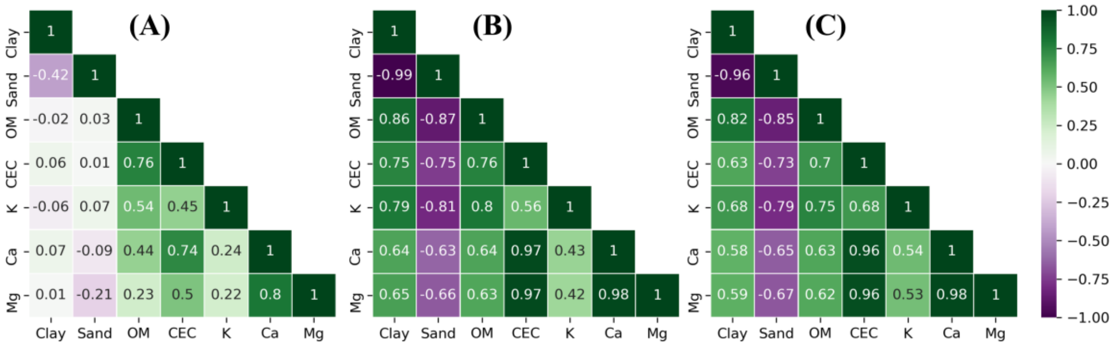


Figure 4. Pearson's correlation matrix of laboratory analysis from soil samples that composed the three strategies of datasets used in this study. (A) Local dataset; (B) Geological dataset; (C) Global dataset.

3.3.2. Prediction models performance

The results for k-fold cross-validation of local, geological, and global prediction models are presented in Table 1. The local model usually performed its best prediction using fewer principal components than geological and global calibrations. Lower values for prediction errors (RMSE and MAE) were observed for the local model for all soil attributes predicted, except OM. On the other hand, R^2 and RPIQ values for geological and global models overcame the local strategy, which presented R^2 >0.60 for only OM and CEC and its best RPIQ of 1.35 for Ca prediction, while both geological and global models surpassed RPIQ = 2.00 for all attributes predicted.

Table 1. Results of online prediction of soil clay, sand, organic matter (OM), cation exchange capacity (CEC) and calcium (Ca) using principal components regression (PCR) models developed for the different calibration strategies of only in-field samples (Local), adding samples from the same geological region (Geological) and from different geological regions (Global).

	Local					Geological					Global				
	NC	R^2	RMSE	MAE	RPIQ	NC	R^2	RMSE	MAE	RPIQ	NC	R^2	RMSE	MAE	RPIQ
Clay	4	0.17	19.88	15.08	0.67	7	0.97	25.83	20.71	11.34	8	0.95	28.96	23.25	9.58
Sand	1	0.05	23.41	17.53	0.12	9	0.97	30.45	25.33	15.76	10	0.97	36.74	29.1	12.64
OM	9	0.75	3.11	2.28	1.25	9	0.87	2.64	2.26	4.36	10	0.8	3.03	2.22	3.73
CEC	6	0.6	3.51	2.78	1.1	8	0.73	11.92	8.74	3.06	5	0.76	11.6	8.66	3.69
K	6	0.14	0.93	0.77	0.51	4	0.56	2.11	1.57	2.2	5	0.72	1.51	1.03	3.62
Ca	10	0.39	2.54	2.08	1.35	8	0.68	6.91	5.01	2.13	3	0.55	7.01	5.17	2.43
Mg	1	0.01	1.83	1.41	0.52	8	0.65	7.1	5.1	2.27	3	0.57	6.99	4.83	2.51

NC = number of principal components used in calibration; R^2 = coefficient of determination; RMSE = root mean square error; MAE = mean absolute error; RPIQ = ratio of performance to interquartile distance.

Note that RPIQ values variation follows R^2 values, departing from the prediction error presented by the model, since the smallest errors of the local model were not accompanied by better RPIQ values. This may imply that another parameter is needed to fully comprehend if the prediction model is sufficiently assertive to be used as a field technique for soil data acquisition. Agriculture is an activity that depends on the soil and its characteristics in deciding on productive steps. Not only the statistical distribution but also knowing the soil attributes content in the determined location is crucial for decision making (Abdel Rahman et al., 2021). The spatial dependence of an attribute is known to be described by geostatistics, fitting variograms with the samples of the area (de Iaco et al., 2022). In this sense, it is suggested that the comprehension of the predicted values variation can contribute to a precise decision-making process of DRS NIR as a technique applied in the context of PA, both in quantitative terms, by the error of prediction, and in qualitative terms, by evaluating the spatial distribution of predicted values.

However, before evaluating the models in terms of variation in values observed, defining what is implied in the construction of soil attributes ML models is needed. A set of 72 soil samples from the nearby area located in Piracicaba (SP, Brazil) added to build geological and global datasets was divided and submitted for analysis to four different commercial laboratories, aiming to verify the difference in values that a standard laboratory analysis of a soil sample can present. A mean variation of 21.4 g kg^{-1} for clay content and 24.4 g kg^{-1} for sand content was observed between the analysis of the four laboratories. For chemical attributes, the results were even more discrepant. The analysis of OM and CEC exhibited a maximum Pearson correlation coefficient of 0.51 between laboratories. It is noteworthy that the mean error of prediction of the models calibrated in this study presented lower values than the variation observed among the different laboratories. The complete analysis of the 72 soil samples from the four laboratories is available in Eitelwein (2017).

The certification of soil analytical laboratories in international level is a competence of the International Organization for Standardization (ISO) (ISO, 2022a; ISO, 2022b). The standards of procedures and certification include acceptable errors and calibration limits for soil testing. This means that every analysis, even from certified laboratories, is susceptible to errors in some scale, and stakeholders of agriculture production always dealt with these possible variations.

Finding the correct values instead of generalizing attributes and variability is an obvious goal of PA (ISPA, 2022), but the calibration of ML models depends on the reference values inserted in the calibration. DRS is directly related to the intrinsic content of an attribute of response in the determined electromagnetic spectrum region (Fang et al., 2018; Pasquini, 2018). Thus, if there is no consensus in the value inserted for calibration, a misbalance of predicted versus observed values occurs, and the models automatically incorporate errors of prediction in some magnitude. This could imply that while we use this basis for ML models using DRS in soil science (Barra et al., 2021), we will hardly reach an accuracy level that allows to find the exact same values due to the model input, one of the three main sources that can lead to output uncertainty (Huang et al., 2015). Instead, we should aim to minimize the errors of prediction as much as possible and look forward to the repeatability of distribution and the agronomic plausibility of predicted attributes distribution, assuming that variations of some kind, already present in current analytical methods used, will not overcome the benefits that the technique can offer. Therefore, we suggest that evaluating the predicted attributes in quantiles associated with the prediction errors (Malone et al., 2011; Ma et al., 2017; Vaysse et al., 2017) is an effective approach rather than equalizing categories (Somarathna et al., 2016; Franceschini et al., 2018; Pouladi et al., 2019).

PCR models presented a described characteristic of this statistical method of smoothing predicted values when compared to those inserted in calibration (Bellon-Maurel et al., 2010) (Table 2). Besides the loss of extreme values of all datasets, the major portion of the population followed the distribution (Velliangiri et al., 2019) (Figure 5). For clay and Mg prediction, the local model caused the major concentration of values when compared to geological and global predictions.

Table 2. Range of values observed for clay, organic matter (OM), cation exchange capacity (CEC), potassium (K), calcium (Ca) and magnesium (Mg) in laboratory analysis (Lab), and predicted values using online spectrum of experimental field from three different calibrations strategies of only in-field samples (Local), adding samples from the same geological region (Geo) and from different geological regions (Global).

	Clay			OM			CEC		
			g kg^{-1}					mmol$_c$ kg^{-1}	
	Min	Max	R	Min	Max	R	Min	Max	R
Laboratory	51	183	132	12	35	23	44	68	24
Local	89	149	60	12	26	14	45	70	25
Geological	79	175	96	11	29	18	37	74	37
Global	50	182	132	11	30	19	43	65	22
	K			Ca			Mg		
					mmol$_c$ kg^{-1}				
	Min	Max	R	Min	Max	R	Min	Max	R
Laboratory	0	5	5	10	34	24	5	22	17
Local	0	4	4	12	29	17	8	12	4
Geological	1	4	3	9	33	24	-1	19	20
Global	1	4	3	15	26	11	5	19	14

Min: minimum value observed; Max: maximum value observed; R: range (Max – Min)

The global model followed the exact range of values observed in the laboratory for its clay prediction, and presented the major concentration of values for Ca prediction. For OM and K, the three strategies presented similar population distribution, even though all three flattened the distribution curve observed in the values of laboratory analysis. CEC prediction is highlighted as the most similar distribution for all populations. Nevertheless, the range of predicted values places the local calibration as the closest to laboratory population.

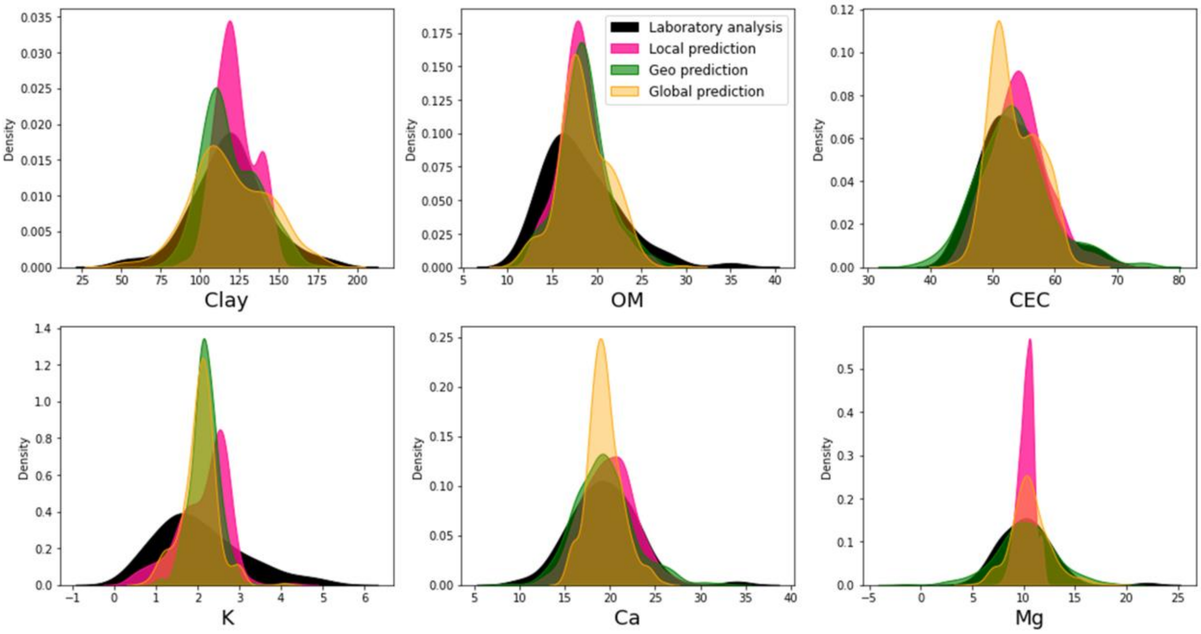


Figure 5. Kernel density estimate plots of clay, organic matter (OM), cation exchange capacity (CEC), potassium (K), calcium (Ca) and magnesium (Mg) for the attributes observed in the laboratory analysis of experimental area and predicted using online spectra on three strategies of calibration: only in-field samples (Local), adding samples from the same geological region (Geological) and from different geological regions (Global).

3.3.3. High spatial resolution prediction and data interpolation

The parameters of fitted variograms for clay, sand, OM, CEC, K, Ca, and Mg, using the three strategies of calibration, hardly presented similar values (Table 3). However, the nugget to total sill ratio (Cambardella et al., 1994) presented moderate spatial dependence for almost all predicted attributes. Only Ca prediction from the local dataset and sand prediction from the global dataset exhibited pure nugget effect, indicating the inexistence of spatial dependence on the distribution of these attributes contents on the experimental field. Regardless, for Ca, geological and global calibrations were able to find spatial dependence. The same was observed for sand, in which local and geological calibrations pointed spatial dependence. This indicates that the spatial distribution of predicted values can be affected depending on the calibration model, despite the prediction error presented, corroborating with the results found in Pouladi et al. (2019).

Table 3. Parameters of fitted variograms for clay, sand, organic matter (OM), cation exchange capacity (CEC), potassium (K), calcium (Ca) and magnesium (Mg) predicted values using online spectra of experimental field from three different calibrations strategies of only in-field samples (Local), adding samples from the same geological region (Geological) and from different geological regions (Global).

	Local			Geological			Global		
	C0	C1	A	C0	C1	A	C0	C1	A
Clay	115.8	67.2	230.3	187.5	167.5	52.2	213.1	500.1	139.9
Sand	7.0	3.4	196.9	154.1	126.0	33.5	-	-	-
OM	5.3	2.8	189.3	3.4	3.3	28.5	5.9	1.9	37.0
CEC	32.7	35.0	199.2	13.6	20.3	21.8	0.0	17.6	142.4
K	0.04	0.02	197.6	0.08	0.04	85.7	0.02	0.1	21.3
Ca	-	-	-	2.7	7.9	26.4	2.7	1.3	89.4
Mg	2.7	1.3	89.5	4.6	3.9	27.7	3.2	1.28	108.1

C0 = nugget; C1 = sill; A = range.

Clay prediction was not considerably affected by the addition of samples from outside areas, which could have happened due to the relation of clay and the fundamentals of NIR with soil mineralogy (Fang et al., 2018). Due to the direct response of this attribute in VNIR, other authors even reported satisfactory prediction in independent tests, extrapolating predictive models in scanned but previous unsampled agricultural areas (Eitelwein et al., 2022). The range presented by the three variograms fitted for clay prediction was discrepant: 230.3 m for local dataset, 52.2 m for geological dataset, and 139.9 m for the global dataset. Despite that, the ordinary kriging reached similar patterns and also similar values for the attribute (Figure 6), highlighting the variation amplitude observed in quantiles division, which is small. Class discrepancy of values was also lower than the MAE of prediction models (15.08 g kg^{-1} for local, 20.71 g kg^{-1} for geological, and 23.25 g kg^{-1} for global). The evaluation of R^2 and RPIQ would lead to the discarding of clay local model. However, the spatial distribution of predicted values alongside the error of prediction proved the ability of the local calibration to predict this attribute of primary response in NIR.

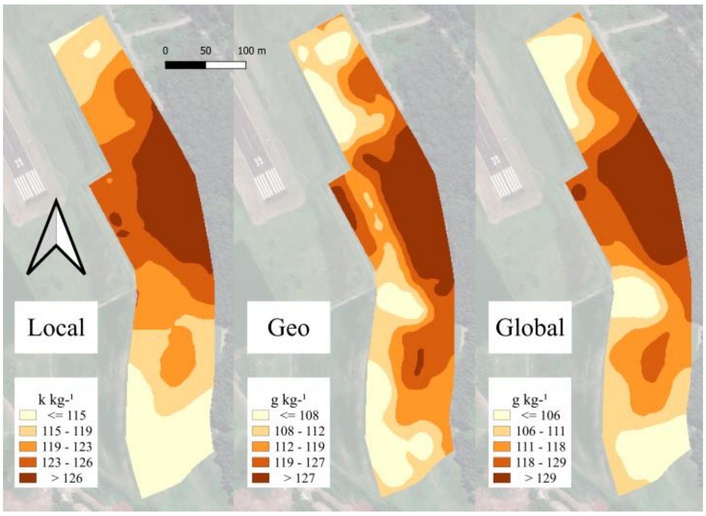


Figure 6. Maps of the five quantiles obtained by ordinary kriging for clay prediction using three different strategies of calibration only in-field samples (Local), adding samples from the same geological region (Geological) and adding samples from different geological regions (Global).

The prediction of OM is widely explored using DRS NIR due to the fact that OM is a primary response attribute in this region of electromagnetic spectrum, with its typical wavelength absorption being reported to comprise (nm) 1,660, 1,728, 1,754, 2,056, 2,264, 2,306, and 2,347 (Nocita et al., 2015). Its prediction can also arise in moist soil (not in field capacity) (Wang et al., 2020), a condition often observed in field soils. This is the most likely explanation for the satisfactory prediction of OM using the local, geological, and global dataset calibrations (Figure 7).

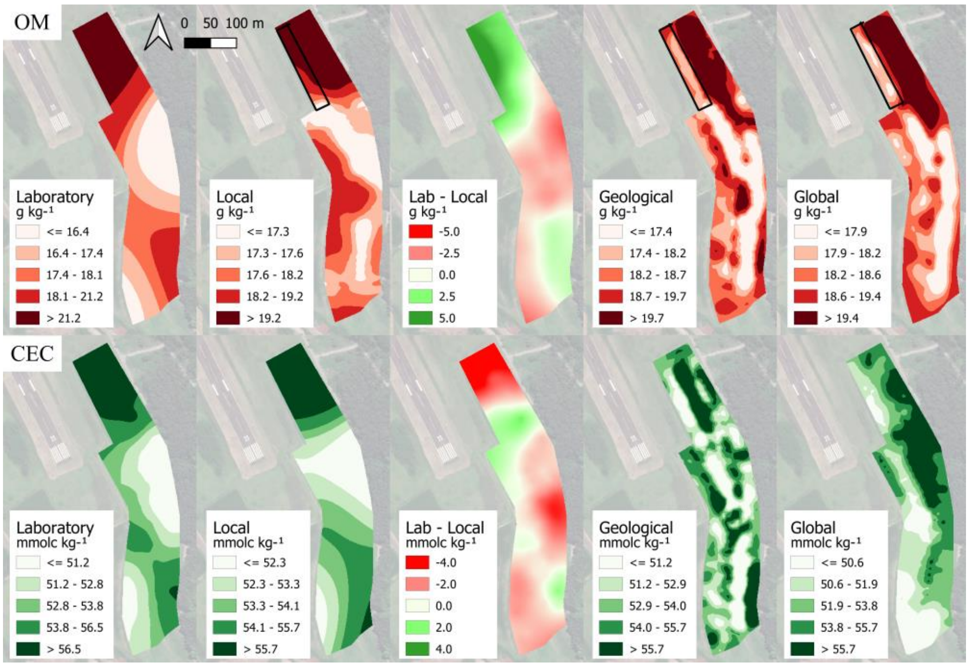


Figure 7. Maps of the five quantiles obtained by ordinary kriging of organic matter (OM), in red, and cation exchange capacity (CEC), in green. Maps are presented in order of: laboratory analysis; prediction model calibrated with only in-field samples (Local); difference of Laboratory and Local values (Lab – Local); prediction model calibrated adding samples from the same geological region (Geological); and adding samples from different geological regions (Global).

Although with different calibrations, the models reached similar patterns of distribution, which was also observed by Pouladi et al. (2019). As was observed for clay, the variation amplitude in quantiles distribution is small for OM prediction. The most divergent area was observed in the northwest portion of the field, where the local calibration pointed a zone of high OM content and geological and global calibrations pointed the opposite. In addition, the addition of outside samples in the calibration set clearly affected the spatial dependence of prediction of OM, since the range of the variograms expressively decreased. For local calibration, a spatial dependence until 189.3 m of distance was observed from one sampling point to another. For geological and global calibration, however, the spatial dependence was found until 28.5 and 37.0 m.

Although there were similarities among the maps of local, geological, and global calibrations, when comparing the map generated from the high density of 72 soil samples analyzed in the laboratory, it is noted that the local model presented the best fitted prediction for OM in the experimental field, similar to that reported by Stevens et al. (2013). An explanation for local better prediction can be that changes in iron oxide content can cancel variations in OM absorption features (Adar et al., 2014). The major portion of the area presented a difference of <2.5 g kg^{-1}. The greatest difference was observed in the same region that the local model disagreed with geological and global calibrations. Exactly in this region, laboratory analysis presented a single sample with 35 g kg^{-1} of OM content. The second highest OM content observed in the laboratory was 28 g kg^{-1}. The upper limit loss in the range presented for local calibration was clearly affected for this sample only (Table 2). The errors of prediction of this model (MAE = 2.28 g kg^{-1} and RMSE = 3.11 g kg^{-1}) were also increased due to what was quoted. Thus, it is assumed that a resampling of that area is needed to verify if the sample of 35 g kg^{-1} was accurate or it was an outlier due to the error in the sampling procedure/laboratory analysis (Hemingway, 1955). Nevertheless, if laboratory analysis showed greatest values than the local model, which classified the area as a high content one, geological and global calibrations are wrong in the assumption of a low OM content area.

CEC prediction was discrepant between the three models (Figure 7). The geological model reached an irregular distribution of patterns in the field. While local calibration presented a variogram range of 199.2 m and global calibration of 142.4 m, the geological model reduced the range to 21.8 m. The difference between local and global prediction stands for the inversion of patterns observed, changing high CEC values zones into low ones. Although the range of 24 mmol$_c$ kg^{-1} in CEC values was observed in the laboratory, followed by three datasets predictions (Table 2), the quantiles limits presented a small variation of 1.5–2.0 mmol$_c$ kg^{-1}.

Attributes that do not have direct spectral response in the region studied can be predicted if the attribute presents covariation with another of primary response (Stenberg et al., 2010). Thus, various authors have dedicated their attention to construct indirect VNIR calibrations to predict these soil attributes (Munnaf et al., 2019; Pätzold et al., 2020; Bönecke et al., 2021). The use of calibrations that compile soil samples from different areas to predict these attributes is a common practice, usually gathering data from the same morphopedological region (Stenber et al., 2010). Nevertheless, the strategy of putting together the areas from different regions is also observed and stated as an effective approach depending on the results demonstrated (Munnaf et al., 2021). In this study, although smaller prediction errors were obtained from local model prediction of CEC, geological and global models presented better R^2 and RPIQ and metrics similar to others (Ulusoy et al., 2016; Rehman et al., 2019; Chen et al., 2021). However, it is noted that the values obtained from different strategies led to different patterns of attributes spatialization in the field, affecting the spatial dependence as for the geological model or the inversion of patterns as for the global model.

The comparison between the kriging maps obtained for CEC analysis in the laboratory and that obtained using local model calibration leads to the conclusion that the local model was the only strategy among the three that successfully predicted the attribute. At the north of the area, the region of greatest discrepancy was observed, where even though the model accurately defined the region of higher CEC at the field, it downsized the value observed by the laboratory analysis, characteristics smoothing reported for PCR prediction models (Bellon-Maurel et al., 2010). It is highlighted that, although there was a small range of CEC values from both laboratory analysis (24 mmol$_c$ kg^{-1}) and local prediction (25 mmol$_c$ kg^{-1}) and a small variation amplitude in quantiles division, the local calibration was able to accurately identify the spatial patterns in the field.

The failure of the prediction of CEC for geological and global models, despite the considered good metrics presented by these two strategies, can be explained by the correlation observed between soil attributes (Kuang et al., 2012) (Figure 4). For only the experimental area, CEC had a strong correlation with OM of 0.76, which is a primary response attribute in NIR. Note that in the experimental area, CEC is almost independent from clay, with a correlation coefficient of −0.06. By the addition of samples from the area of the same geological region than the experimental field, the correlation of CEC with OM is maintained at 0.76. Yet, the model identifies a strong correlation with other primary NIR response attributes, where CEC and clay had a correlation coefficient of 0.75.

The similar effect happened for the global dataset. Although the kernel density estimation plots pointed the same statistical distribution of predicted values for all datasets (laboratory analysis, and local, geological, and global predictions) (Figure 5) and satisfactory metrics were presented (R^2, RMSE, MAE, and RPIQ) (Table 1), the prediction of CEC with neither geological nor global models was accurate, which places spatial distribution and agronomic plausibility of this distribution as a fundamental factor for classifying the model as robust or not. Even though other authors found a positive influence of creating calibrations from multiple fields (Carmon & Ben Dor, 2017; Munnaf et al., 2019), this was not the case for the one tested in this study when the field spatialization parameter was taken into account. This could also be possible due to the use of other techniques more related to fundamental vibrations of soil attributes in the spectra, like mid- infrared (Greenberg et al., 2022) or X-ray fluorescence (Qu et al., 2022), or other factors that were not investigated in this study.

The prediction of plant nutrients was not consistent for any of the datasets used for model calibration, and Ca maps represented the same patterns observed for K and Mg (Figure 8). Local and geological datasets resulted in a prediction without coherent spatial patterns, and for the global dataset, although the north portion of the area presented the same pattern and similar values to those observed in the laboratory, it may be assigned by chance, once the other patterns were not steady.

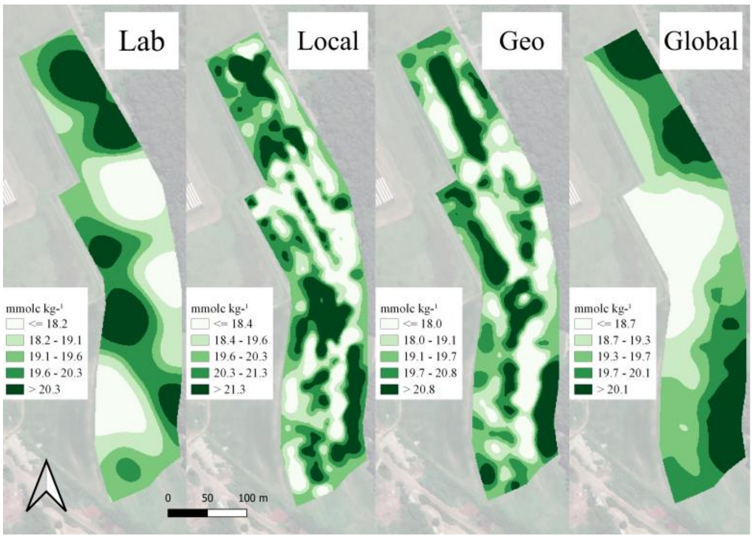


Figure 8. Maps of the five quantiles obtained by ordinary kriging of calcium (Ca) values of laboratory samples analysis (Lab) and the prediction models using three different strategies of calibration: only in-field samples (Local), adding samples from the same geological region (Geological) and adding samples from different geological regions (Global).

The unsuccessful prediction of Ca can be related to the correlations presented for this attribute (Figure 4), as it was for the successful prediction of CEC (Stenberg et al., 2010; Kuang et al., 2012). In the experimental field, Ca had an average correlation with OM of 0.44. This fact could explain the slightly better R^2 and RPIQ values presented in the validation of local Ca model, although this was not true for K (Table 1), even with a 0.55 positive correlation with OM presented in the local dataset. Nevertheless, this correlation magnitude proved to be insufficient to allow an accurate prediction using DRS NIR. For geological and global models, once outside samples were entered in the dataset, nutrient correlations were modified, presenting, in both cases, significant positive correlation with clay and OM and negative correlation with sand. Therefore, the ML models used in this study, which are helpful tools to deal with spectral data and correlations between soil attributes (Barra et al., 2021), were not able to perform a consistent prediction.

This study suggests that the correlation coefficient itself, even when corroborated with satisfactory statistical metrics on prediction models validation, cannot identify if a secondary response attribute can be predicted with DRS ML models (Marín-González et al., 2013). The correlation observed in the target area alone must be taken into account, and it is of high importance that this correlation is not twisted after the union of outside samples in the model calibration, which can cause the distortion of the attributes spatial distribution in the field.

3.4 Conclusions

Spatial distribution in terms of zones and agronomic plausibility of predicted values obtained from DRS NIR prediction models proved to be a key factor of robustness evaluation. Using R^2 and RPIQ without field spatialization is suggested to be a vulnerable strategy due to misleading decisions that these metrics would lead into in the present study. This study suggests to further investigate the spatialization of soil attributes predicted using NIR spectra in areas with greater variability. It is necessary to further check the weaknesses that ML models of NIR spectra calibrated with samples from more than one area presented in the spatialization of the predicted attributes. If the observed results in the present study are repeated for other agricultural fields, it may indicate that local models are the best recommendation for DRS used for field-scale PSS.

References

AbdelRah M. A. E., Zakarya, Y. M., Metwaly, M. M., & Koubouris, G. (2021). Deciphering soil spatial variability through geostatistics and interpolation techniques. **Sustainability** (Switzerland), 13(1). https://doi.org/10.3390/su13010194

Adar, S., Shkolnisky, Y., & Ben-Dor, E. (2014). Change detection of soils under small-scale laboratory conditions using imaging spectroscopy sensors. **Geoderma**, 216. https://doi.org/10.1016/j.geoderma.2013.10.017

Barra, I., Haefele, S. M., Sakrabani, R., & Kebede, F. (2021). Soil spectroscopy with the use of chemometrics, machine learning and preprocessing techniques in soil diagnosis: Recent advances–A review. In **TrAC - Trends in Analytical Chemistry** (Vol. 135). https://doi.org/10.1016/j.trac.2020.116166

Barthès, B. G., Brunet, D., Hien, E., Enjalric, F., Conche, S., Freschet, G. T., d'Annunzio, R., & Toucet-Louri, J. (2008). Determining the distributions of soil carbon and nitrogen in particle size fractions using near-infrared reflectance spectrum of bulk soil samples. **Soil Biology and Biochemistry**, 40(6). https://doi.org/10.1016/j.soilbio.2007.12.023

Bellon-Maurel, V., Fernandez-Ahumada, E., Palagos, B., Roger, J. M., & McBratney, A. (2010). Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy. In **TrAC - Trends in Analytical Chemistry** (Vol. 29, Issue 9). https://doi.org/10.1016/j.trac.2010.05.006

Bönecke, E., Meyer, S., Vogel, S., Schröter, I., Gebbers, R., Kling, C., Kramer, E., Lück, K., Nagel, A., Philipp, G., Gerlach, F., Palme, S., Scheibe, D., Zieger, K., & Rühlmann, J. (2021). Guidelines for precise lime management based on high-resolution soil pH, texture and SOM maps generated from proximal soil sensing data. **Precision Agriculture**, 22(2). https://doi.org/10.1007/s11119-020-09766-8

Cambardella, C. A., Moorman, T. B., Novak, J. M., Parkin, T. B., Karlen, D. L., Turco, R. F., & Konopka, A. E. (1994). Field-Scale Variability of Soil Properties in Central Iowa Soils**. Soil Science Society of America Journal**, 58(5). https://doi.org/10.2136/sssaj1994.03615995005800050033x

Carmon, N., & Ben-Dor, E. (2017). An Advanced Analytical Approach for Spectral - Based Modelling of Soil Properties. **International Journal of Emerging Technology and Advanced Engineering**, 7(3).

Chang, C.-W., Laird, D. A., Mausbach, M. J., & Hurburgh, C. R. (2001). Near-Infrared Reflectance Spectroscopy-Principal Components Regression Analyses of Soil Properties. **Soil Science Society of America Journal**, 65(2). https://doi.org/10.2136/sssaj2001.652480x

Chen, Y., Gao, S., Jones, E. J., & Singh, B. (2021). Prediction of Soil Clay Content and Cation Exchange Capacity Using Visible Near-Infrared Spectroscopy, Portable X-ray Fluorescence, and X-ray Diffraction Techniques. **Environmental Science and Technology**, 55(8). https://doi.org/10.1021/acs.est.0c04130

Christy, C. D. (2008). Real-time measurement of soil attributes using on-the-go near infrared reflectance spectroscopy. **Computers and Electronics in Agriculture**, 61(1). https://doi.org/10.1016/j.compag.2007.02.010

de Iaco, S., Hristopulos, D. T., & Lin, G. (2022). Special Issue: **Geostatistics and Machine Learning. Mathematical Geosciences**, 54(3). https://doi.org/10.1007/s11004-022-09998-6

Eitelwein, M. T. (2017). Proximal soil sensing: quantification of physical and chemical soil attributes. Luiz de Queiroz College of Agriculture - University of São Paulo.

Eitelwein, M. T., Tavares, T. R., Molin, J. P., Trevisan, R. G., de Sousa, R. V., & Demattê, J. A. M. (2022). Predictive Performance of Mobile Vis–NIR Spectroscopy for Mapping Key Fertility Attributes in Tropical Soils through Local Models Using PLS and ANN. **Automation**, 3(1). https://doi.org/10.3390/automation3010006

Fang, Q., Hong, H., Zhao, L., Kukolich, S., Yin, K., & Wang, C. (2018). Visible and Near-Infrared Reflectance Spectroscopy for Investigating Soil Mineralogy: A Review. In **Journal of Spectroscopy** (Vol. 2018). https://doi.org/10.1155/2018/3168974

Franceschini, M. H. D., Demattê, J. A. M., Kooistra, L., Bartholomeus, H., Rizzo, R., Fongaro, C. T., & Molin, J. P. (2018). Effects of external factors on soil reflectance measured on-the-go and assessment of potential spectral correction through orthogonalisation and standardisation procedures. **Soil and Tillage Research**, 177. https://doi.org/10.1016/j.still.2017.10.004

Greenberg, I., Seidel, M., Vohland, M., Koch, H. J., & Ludwig, B. (2022). Performance of in situ vs laboratory mid-infrared soil spectroscopy using local and regional calibration strategies. **Geoderma**, 409. https://doi.org/10.1016/j.geoderma.2021.115614

Guerrero, A., de Neve, S., & Mouazen, A. M. (2021). Data fusion approach for map-based variable-rate nitrogen fertilization in barley and wheat. **Soil and Tillage Research**, 205. https://doi.org/10.1016/j.still.2020.104789

Hemingway, R. G. (1955). Soil-sampling errors and advisory analyses. **The Journal of Agricultural Science**, 46(1). https://doi.org/10.1017/S0021859600039563

Huang, J., Zare, E., Malik, R. S., & Triantafilis, J. (2015). An error budget for soil salinity mapping using different ancillary data. **Soil Research**, 53(5). https://doi.org/10.1071/SR15043

International Organization for Standardization. (n.d.-a). General requirements for the competence of testing and calibration laboratories. ISO Standard No. 17025. Retrieved June 23, 2022, from https://www.iso.org/standards.html

International Organization for Standardization. (n.d.-b). Laboratory testing of soil. ISO Standard No. 17892. Retrieved June 23, 2022, from https://www.iso.org/standards.html

International Society of Precision Agriculture. (2022). Precision Agriculture definition. Retreved June 02, 2022, from https://www.ispag.org/about/definition

Jung, Y. (2018). Multiple predicting K-fold cross-validation for model selection. **Journal of Nonparametric Statistics**, 30(1). https://doi.org/10.1080/10485252.2017.1404598

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., & Willing, C. (2016). Jupyter Notebooks—a publishing format for reproducible computational workflows. Positioning and Power in Academic Publishing: Players, Agents and Agendas - Proceedings of the 20th **International Conference on Electronic Publishing**, ELPUB 2016. https://doi.org/10.3233/978-1-61499-649-1-87

Kuang, B., Mahmood, H. S., Quraishi, M. Z., Hoogmoed, W. B., Mouazen, A. M., & van Henten, E. J. (2012). Sensing soil properties in the laboratory, in situ, and on-line. A review. In **Advances in Agronomy** (Vol. 114). https://doi.org/10.1016/B978-0-12-394275-3.00003-1

Ma, Y., Minasny, B., & Wu, C. (2017). Mapping key soil properties to support agricultural production in Eastern China. **Geoderma Regional**, 10. https://doi.org/10.1016/j.geodrs.2017.06.002

Malone, B. P., McBratney, A. B., & Minasny, B. (2011). Empirical estimates of uncertainty for mapping continuous depth functions of soil attributes. **Geoderma**, 160(3–4). https://doi.org/10.1016/j.geoderma.2010.11.013

Marín-González, O., Kuang, B., Quraishi, M. Z., Munóz-García, M. Á., & Mouazen, A. M. (2013). On-line measurement of soil properties without direct spectral response in near infrared spectral range. **Soil and Tillage Research**, 132. https://doi.org/10.1016/j.still.2013.04.004

Minasny, B., McBratney, A. B., & Whelan, B. M. (2006). VESPER version 1.62. Australian Centre for Precision Agriculture.

Molin, J. P., & Tavares, T. R. (2019). Sensor systems for mapping soil fertility attributes: Challenges, advances, and perspectives in brazilian tropical soils. **Engenharia Agricola**, 39(specialissue). https://doi.org/10.1590/1809-4430-ENG.AGRIC.V39NEP126-147/20190126

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

Munnaf, M. A., Guerrero, A., Nawar, S., Haesaert, G., van Meirvenne, M., & Mouazen, A. M. (2021). A combined data mining approach for on-line prediction of key soil quality indicators by Vis-NIR spectroscopy. **Soil and Tillage Research**, 205. https://doi.org/10.1016/j.still.2020.104808

Munnaf, M. A., Nawar, S., & Mouazen, A. M. (2019). Estimation of secondary soil properties by fusion of laboratory and on-line measured Vis-NIR spectra. **Remote Sensing**, 11(23). https://doi.org/10.3390/rs11232819

Nocita, M., Stevens, A., van Wesemael, B., Aitkenhead, M., Bachmann, M., Barthès, B., Dor, E. ben, Brown, D. J., Clairotte, M., Csorba, A., Dardenne, P., Demattê, J. A. M., Genot, V., Guerrero, C., Knadel, M., Montanarella, L., Noon, C., Ramirez-Lopez, L., Robertson, J., … Wetterlind, J. (2015). Soil Spectroscopy: An Alternative to Wet Chemistry for Soil Monitoring. **Advances in Agronomy**, 132. https://doi.org/10.1016/bs.agron.2015.02.002

Pasquini, C. (2018). Near infrared spectroscopy: A mature analytical technique with new perspectives – A review. In **Analytica Chimica Acta** (Vol. 1026). https://doi.org/10.1016/j.aca.2018.04.004

Pätzold, S., Leenen, M., Frizen, P., Heggemann, T., Wagner, P., & Rodionov, A. (2020). Predicting plant available phosphorus using infrared spectroscopy with consideration for future mobile sensing applications in precision farming. **Precision Agriculture**, 21(4). https://doi.org/10.1007/s11119-019-09693-3

Pouladi, N., Møller, A. B., Tabatabai, S., & Greve, M. H. (2019). Mapping soil organic matter contents at field level with Cubist, Random Forest and kriging. **Geoderma**, 342. https://doi.org/10.1016/j.geoderma.2019.02.019

Python Software Foundation. (2022). Python.

QGIS Development Team. (2022). QGIS Geographic Information System. Open Source Geospatial Foundation.

Qu, M., Liu, H., Guang, X., Chen, J., Zhao, Y., & Huang, B. (2022). Improving correction quality for in-situ portable X-ray fluorescence (PXRF) using robust geographically weighted regression with categorical land-use types at a regional scale. **Geoderma**, 409. https://doi.org/10.1016/j.geoderma.2021.115615

Rehman, H. U., Knadel, M., Jonge, L. W., Moldrup, P., Greve, M. H., & Arthur, E. (2019). Comparison of Cation Exchange Capacity Estimated from Vis–NIR Spectral Reflectance Data and a Pedotransfer Function. **Vadose Zone Journal**, 18(1). https://doi.org/10.2136/vzj2018.10.0192

Somarathna, P. D. S. N., Malone, B. P., & Minasny, B. (2016). Mapping soil organic carbon content over New South Wales, Australia using local regression kriging. **Geoderma Regional**, 7(1). https://doi.org/10.1016/j.geodrs.2015.12.002

Stenberg, B., Viscarra Rossel, R. A., Mouazen, A. M., & Wetterlind, J. (2010). Visible and Near Infrared Spectroscopy in Soil Science. In **Advances in Agronomy** (Vol. 107, Issue C). 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). https://doi.org/10.1371/journal.pone.0066409

Syers, J. K., Campbell, A. S., & Walker, T. W. (1970). Contribution of organic carbon and clay to cation exchange capacity in a chronosequence of sandy soils. **Plant and Soil**, 33(1). https://doi.org/10.1007/BF01378202

Ulusoy, Y., Tekin, Y., Tümsavaş, Z., & Mouazen, A. M. (2016). Prediction of soil cation exchange capacity using visible and near infrared spectroscopy. **Biosystems Engineering**, 152. https://doi.org/10.1016/j.biosystemseng.2016.03.005

Vaysse, K., & Lagacherie, P. (2017). Using quantile regression forest to estimate uncertainty of digital soil mapping products. **Geoderma**, 291. https://doi.org/10.1016/j.geoderma.2016.12.017

Velliangiri, S., Alagumuthukrishnan, S., & Thankumar Joseph, S. I. (2019). A Review of Dimensionality Reduction Techniques for Efficient Computation. **Procedia Computer Science**, 165. https://doi.org/10.1016/j.procs.2020.01.079

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

Vishwakarma, G., Sonpal, A., & Hachmann, J. (2021). Metrics for Benchmarking and Uncertainty Quantification: Quality, Applicability, and Best Practices for Machine Learning in Chemistry. In **Trends in Chemistry** (Vol. 3, Issue 2). https://doi.org/10.1016/j.trechm.2020.12.004

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

Wang, Y., Huang, T., Liu, J., Lin, Z., Li, S., Wang, R., & Ge, Y. (2015). Soil pH value, organic matter and macronutrients contents prediction using optical diffuse reflectance spectroscopy. **Computers and Electronics in Agriculture**, 111. https://doi.org/10.1016/j.compag.2014.11.019

Wang, Y. P., Lee, C. K., Dai, Y. H., & Shen, Y. (2020). Effect of wetting on the determination of soil organic matter content using visible and near-infrared spectrometer. **Geoderma**, 376. https://doi.org/10.1016/j.geoderma.2020.114528

Williams, P., & Norris, K. (1987). Near-infrared technology in the agricultural and food industries. **American Association of Cereal Chemists**, Inc. https://www.cabdirect.org/cabdirect/abstract/19892442443

Zhang, J., Guerrero, A., & Mouazen, A. M. (2021). Map-based variable-rate manure application in wheat using a data fusion approach. **Soil and Tillage Research**, 207. https://doi.org/10.1016/j.still.2020.104846

Appendix

Table A1. Variograms and kriging additional parameters reported for clay, sand, organic matter (OM), cation exchange capacity (CEC) and calcium (Ca) for the different calibration strategies of only in-field samples (Local), adding samples from the same geological region (Geological) and from different geological regions (Global).

	Local				Geological				Global			
	Model	RMSE	SE Pred		Model	RMSE	SE Pred		Model	RMSE	SE Pred	
			min	max			min	max			min	max
Clay	Lin	3.65	2.46	4.54	Exp	6.14	5.57	10.57	Sph	16.73	5.82	11.11
Sand	Lin	0.14	4.53	9.75	Exp	5.80	5.51	9.96	-	-	-	-
OM	Lin	0.23	0.54	1.00	Lin	0.61	0.89	1.79	Sph	1.20	0.95	1.34
CEC	Gaus	5.04	0.63	2.06	Lin	3.03	2.12	4.36	Mat	0.61	1.34	2.85
K	Lin	0.01	0.04	0.09	Sph	0.00	0.09	0.17	Exp	0.01	0.11	0.36
Ca	-	-	-	-	Sph	0.24	0.47	0.98	Exp	0.08	0.48	0.85
Mg	Sph	0.01	0.25	0.62	Sph	0.18	1.12	1.90	Exp	0.09	0.47	0.83

Model: variogram model; RMSE: root mean square error of fitting variogram; SE Pred: minimum (min) and maximum (max) kriging prediction associated error; Lin: linear with sill; Gaus: gaussian; Sph: spherical; Exp: exponential; Mat: matern.

4. ONLINE NEAR-INFRARED SOIL ATTRIBUTES MAPPING REQUIRE LOCAL CALIBRATION IN SPACE AND TIME

Abstract

Near-infrared (NIR) region can be used for diffuse reflectance spectroscopy (DRS) direct in agricultural areas as an alternative for acquiring soil data faster and more cost-effective. Building machine learning (ML) calibrations, soil attributes can be predicted in high spatial density. However, for these calibrations to work it is commonly used a higher density of soil samples analyzed in laboratory than the adopted in agricultural production. There are no reports on if a ML model calibrated on this basis can consistently perform over the time, being useful more than once, or require calibration on every online spectra acquisition. Therefore, this study's objective was to acquire online NIR spectra in an agricultural field where a previous spectral acquisition and ML models were built and validated, assessing the performance of predictive models over the time, using the same sensor, operational and instrumentational conditions. Two spectral acquisitions were made separated by 21 days over a fallow Brazilian tropical soil. A total of 140 spectra were acquired on each day, and used for comparisons. Physical and chemical soil properties were predicted using principal components regression models calibrated in day 1. Spectra characteristics, such as morphology, features and intensity, were compared, and also the attributes predicted values and maps interpolated by ordinary kriging. Besides the core of DRS NIR seemed to be maintained analyzing the spectra acquired in both days, neither high correlated or overlapped spectra presented similar values of predictions. The only Pearson's correlation coefficient (r) significative at 99% was for calcium prediction, of 0.22 for the comparison of the entire 140 spectra prediction of each day. For clay, organic matter and cation exchange capacity, that presented a robust prediction in day 1, the r values ranged from -0.14 to 0.32, but were not significative. The maps generated showed no similar attributes spatial distribution, hindering the use for agricultural management decisions. Soil moisture is suggested to take a role as a source of variation, but the analysis of residual maps and the likely water dynamic based on the altimetry map of the area may indicate that were not the only factor actuating. Other environmental variables should be considered to identify the variations observed in online NIR spectra acquired in same experimental conditions. If overcome of these variations do not succeed, the reported in this study suggest that online NIR spectra ML models require local calibrations in space and time.

Keywords: agricultural field operations; machine learning; environmental factors; diffuse reflectance spectroscopy.

4.1. Introduction

Aiming the sustainability of production systems, the use of inputs of a supply chain needs to be optimized. In this sense, efforts are being made in PA towards the identification of spatial and temporal variability of agricultural systems to support management decisions (International Society of Precision Agriculture, 2022). The soil is an essential part of agriculture, providing water, nutrients, air and mechanical sustentation for plants. Its intrinsic and extrinsic variability identification and management practices regulate the variability of agriculture production (Molin and Tavares, 2019; Yin et al., 2021) and ensure continued soil fertility (Johnston and Poulton, 2018). Therefore, techniques to turn soil data acquisition faster, more efficient and cost-effective have been an interest of soil scientists and precision agriculture researchers (Molin and Tavares, 2019; Viscarra Rossel et al., 2011).

DRS is a technique of energy-matter interaction that allow to capture inherent data about an object. NIR region is an alternative for DRS application that has been proven its potential in soil science (Nocita et al., 2015). Tested for the first time in laboratory, researchers have identified specific wavelengths of interaction between diverse soil properties with NIR spectra, such as mineralogy (Fang et al., 2018), texture (Bönecke et al., 2021), soil OM and organic carbon (Bönecke et al., 2021; Munnaf et al., 2021; Wang et al., 2015). These were called the primary NIR response attributes (Stenberg et al., 2010). It was established that other attributes could also be related to NIR spectra, if a covariation with one of primary response occurs. Then, studies have related NIR spectra with CEC, pH and plant nutrients, like soil nitrogen, P, K, Ca and Mg (Munnaf et al., 2019; Munnaf and Mouazen, 2021; Yang et al., 2020), calling them as secondary or indirect NIR response attributes.

As the development of this research area advanced, the introduction of ML techniques allowed to quantify soil attributes using DRS (Barra et al., 2021; Morellos et al., 2016). The idea is to use the electromagnetic spectra with laboratory analysis as reference values, training models that will latter need only the soil spectra to predict the sample attributes.

Once the technique was validated, an adjacent research area started trying to adapt the DRS NIR for PSS (Viscarra Rossel et al., 2011), acquiring soil spectra direct from agricultural areas in high spatial density, the so-called online spectra (Ben-Dor et al., 2008; Mouazen et al., 2009; Stenberg et al., 2007). The possible advantages are the primary goals of acquiring soil data faster, reducing the cost of acquisition per sample, the laboratory reagents waste and the laborious work that laboratory spectral analysis demand, as drying, grinding and sieving soil. Researchers have been reporting results showing that is possible to map soil attributes using DRS NIR online spectra, with most studies being reported in temperate regions (Kuang et al., 2012) and few studies exploring the tropical soils (Eitelwein et al., 2022; Franceschini et al., 2018). The online NIR spectra was used to build ML calibrations to quantify all the same attributes that were previously tested in laboratory, both the primary and the secondary NIR response attributes (Munnaf and Mouazen, 2021; Yang et al., 2020).

However, the ML calibrations reported in these studies invariably use a higher density of soil samples acquired in commercial scale. To achieve the benefits of fine-scale soil mapping of optimization of input distribution, reaching resource use efficiency, profitability and sustainability of agricultural production systems, understanding the calibration protocols the technique require is crucial. This includes comprehending if the online NIR spectra is stable over the time, allowing repeatability of products generated. Researchers identified the importance of local calibrations for predicting soil attributes using visible and NIR spectra (Brown, 2007; Canal Filho and Molin, 2022; Stenberg et al., 2010), and although efforts are being made trying to overcome this limitation, no definitive strategy is established (Gogé et al., 2014; Stevens et al., 2013; Wetterlind et al., 2010). In this sense, it is also needed to assess the usefulness of ML calibrations based on DRS NIR spectra in terms of spatial and time specificity. No studies were found reporting the prediction of a ML calibration of online NIR spectra over the time, assessing if the models require local calibrations in time to perform properly.

We hypothesized that if online NIR spectra present spatiotemporal stability, even the models with poor prediction performance would present similar patterns in the field, denoting that a model calibrated using online NIR spectra can repeat the products generated for soil attributes prediction along the time. Therefore, the objective of this study was to acquire online NIR spectra in an agricultural field where a previous spectral acquisition and ML models were built and validated, assessing the performance of predictive models over the time, using the same sensor, operational and instrumentational conditions.

4.2. Materials and Methods

This study was carried out following the steps presented in Figure 1, that will be further explained in the sections 4.2.1 to 4.2.4.

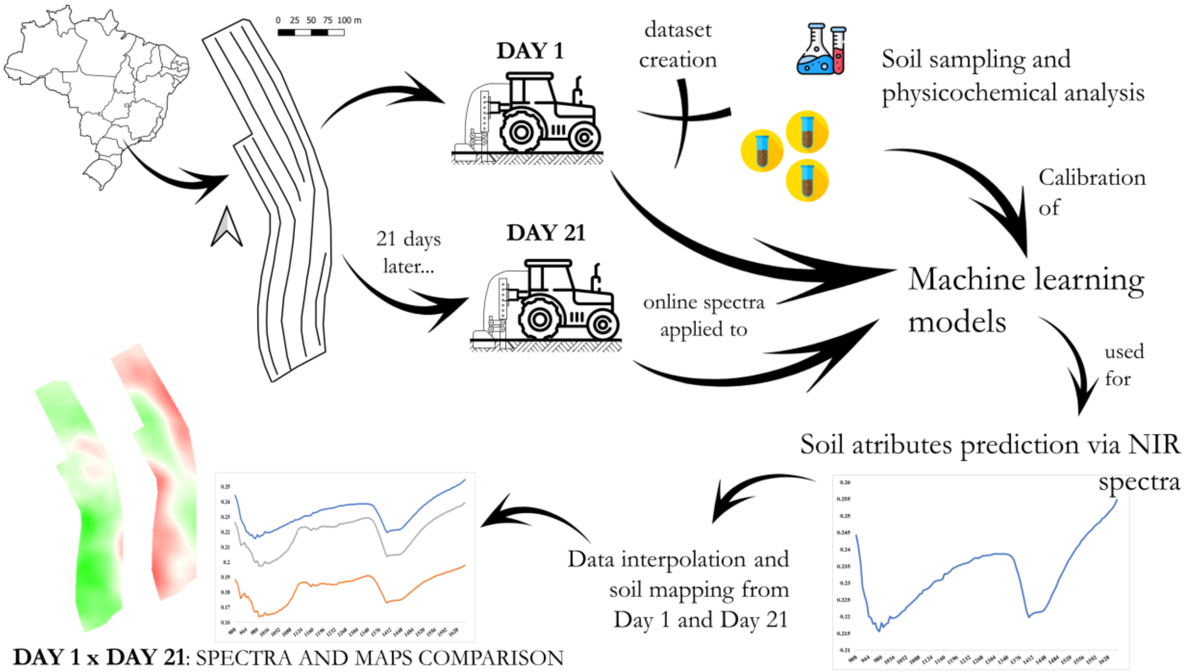

0 25 50 75 100 m
DAY 1
dataset creation
Soil sampling and physicochemical analysis
Calibration of
21 days later...
DAY 21
online spectra applied to
Machine learning models
used for
Soil atributes prediction via NIR spectra
Data interpolation and soil mapping from Day 1 and Day 21
DAY 1 x DAY 21: SPECTRA AND MAPS COMPARISON

Figure 1. Graphical abstract of the methodology developed in this study.

4.2.1. Study area

The study area is a sandy loam experimental field of 6.0 ha, from University of São Paulo (USP), located in Piracicaba, São Paulo state, Brazil (22°43'03.51"S, 47°36'50.03"W). In the last three years, a soybean-fallow system was conducted. The study happened in November, 2021, at the end of the fallow phase, before soybean seeding. During the 21 days that separated the two spectral acquisitions, no operations took place in the area.

Water topsoil dynamic is quite dependent of altimetry (Lee et al., 2011; Orth et al., 2013; Uebbing et al., 2017), and soil moisture is known to affect NIR spectra (Pasquini, 2018; Wang et al., 2020). The area presents a mid-west highest point of 586 m, with declivity direction to the south, with lowest point in 576 m. Two north-south terraces divide the area and delimit the lowest portions at the east. The north portion of the area is a plateau with little altitude variation (Figure 2).

4.2.2. Online spectra acquisition and soil sampling

Online soil spectral data were acquired using a subsoiler shank making a 0.15-m-depth furrow, attached to a structure mounted on the three-point hydraulic hitch of a tractor. This shank carried a steel armored case protecting the NIR spectrophotometer (MicroNIR from VIAVI Solutions Inc., USA), that acquired in 125 wavelengths from 908.1 to 1676.2 nm with a spectral resolution of 6.14 nm. The complete description of how the spectrophotometer calibrates and acquire online spectra can be found in Canal Filho & Molin (2022). Acquisition points were georeferenced using a GNSS Ag-Star (Novatel, Calgary, Canada) receiver with TerraStar-C differential correction (Hexagon, Alabama, USA).

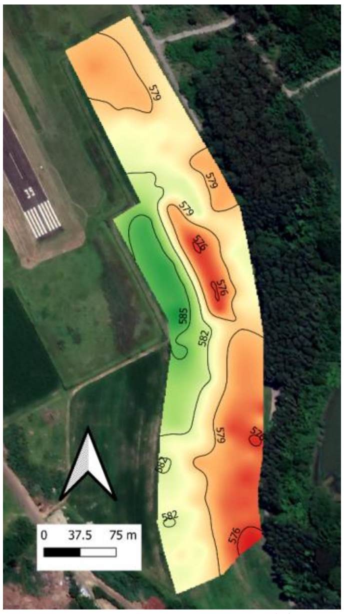


Figure 2. Altimetry map of the experimental field, with respective contour lines. The higher portions are colored in green tones, and the lower portions, in red.

Two subsequent spectral acquisitions occurred in the area, separated by 21 days. Day 1 was used to validate the methods of spectra acquisition, associated soil sampling, calibrate and evaluate the robustness of ML models for prediction. The tractor traveled the area limited by the presence of terraces, making 12 acquisition lines in day 1. The speed was set to 0.583 m s^{-1} (2.1 km h^{-1}). A total of 383 spectral points were acquired in day 1, followed by a filtering process of measurement errors that excluded 80 spectra. The complete report for this study can be found in Canal Filho and Molin (2022). Data acquisition in day 21 followed the same experimental and instrumental conditions from day 1. Once the models were calibrated and validated, in day 21 only the odd lines of day 1 were acquired, resulting in six acquisition lines and 140 spectra. Then, for the comparison between the spectra and product generated in both days, day 1 of acquisition was reduced to the same lines of day 21, resulting in the same number (140 points) of soil spectra to be compared.

Soil sampling occurred in day 1, and was used as reference values to compose the dataset for ML models calibration and evaluation. The soil was sampled in the bottom of the furrow left by the subsoiler shank, in the spectral acquisition transect, to ensure that soil analysis was correspondent to the area previously sensed. For this, the spectrophotometer acquisition time of one spectrum (10 s) was multiplied by the operation speed, resulting in the transect length sensed (7 m). As the software indicated in real-time the starting point of each spectrum, the demarcation of 72 random starting points allowed the afterward soil sampling. To reduce uncertainty, one meter was discarded at the beginning and end of each transect sampled, resulting in a five-meter-long soil sample along the furrow. The attributes considered in laboratory analysis for posterior ML models' prediction were clay, OM, CEC, pH, P, K, Ca and Mg.

In each day of acquisition, ten soil samples were randomly collected at the bottom of the furrow left by the subsoiler shank to monitor the soil moisture. These samples were collected during the field operation, sealed and weighted at field condition. Then, taken to a forced ventilation oven at 105ºC during 72 h to obtain their dry weight (Teixeira et al., 2017). The soil moisture at each day was considered as the mean value for the ten samples collected.

4.2.3. Machine learning models and data interpolation

ML models for soil attributes prediction were calibrated using PCR, a well described technique applied to cope with multivariate data (Agarwal et al., 2021; Chang et al., 2001). The number of PCs used in regression was set following the criteria of reducing the RMSE, and using at most 10 PCs to avoid model's overfitting (Seasholtz and Kowalski, 1993; Tracy et al., 2016). Data modeling was developed using Jupyter Notebook software (Kluyver et al., 2016; Python Software Foundation, 2022). The dataset composed of online near spectra from day 1 acquisition and analysis of corresponding soil samples was divided in the proportion of 70% for calibration and 30% for validation. To reduce the bias of the results reported, k-fold cross-validation (k = 10) was applied (Jung, 2018). The function random state was set to n = 456 to ensure repeatability of results after validation.

The strategy of using only local samples, from the experimental area, yielded the best predictions, and therefore were applied in this study. Further description of ML models applied in this study can be found in Canal Filho and Molin (2022). These same models were applied to the online spectra acquired in day 21.

After the prediction of the 140 points for both days of acquisition, data of each attribute were individually interpolated by ordinary kriging, using the software VESPER. Variograms were fitted within the software, that provides RMSE and AIC index for model adjustment, and gives the parameters nugget (C0), sill (C1) and range (A) (Minasny et al., 2006), that were further compared between both days. The method used was block kriging, in 3.0 × 3.0 m pixels, and the minimum and maximum neighboring points for interpolation was to software minimum and maximum values, of 4 and 300, respectively. The kriging results were converted to a raster format to be exported and analyzed in a geographic information system software.

4.2.4. Spectra stability and product analysis

4.2.4.1. Spectra and prediction values analysis

Firstly, a 99% significance Pearson's correlation analysis was used to compare the spectra and predicted attributes values from day 1 and day 21. The 125 wavelengths measured were the variables compared for spectra analysis. The 20 highest correlated spectra pairs (day 1 x day 21) were identified to be compared in: spectra characteristics (morphology, intensity, absorption features), location of spectra acquisition, and prediction values generated from those spectra. For prediction values comparison, the nearest neighbors were joined in pairs, yielding 140 pairs used for correlation analysis.

Secondly, at the field operation, two spectra would hardly be acquired at exactly the same point. Although tractor's operator, operation speed and acquisition lines were strictly the same, variations in orientation or border maneuvers could offset the spectra from day 21 from the location of day 1. This would hinder the direct comparison of day 1 x day 21 as: spectrum$_{\text{day1}}$ 1 x spectrum$_{\text{day21}}$ 1; spectrum$_{\text{day1}}$ n x spectrum$_{\text{day21}}$ n; …; spectrum$_{\text{day1}}$ 140 x spectrum$_{\text{day21}}$ 140. To undermine the distance between two acquisition points, an ellipse buffer of 2.5 m radius in the direction of tractor's movement was created to extract overlap points. As the soil sampling were carried in the length of five meters to match the transect of spectral acquisition, this buffer had the purpose of selecting the points that overlapped each other, being collected at the same location in relation to the length of the sensed transect.

4.2.4.2. Comparison of maps generated by ordinary kriging

The maps comparison is aimed to assess the similarity between the final products of DRS NIR soil attributes prediction from day 1 and day 21. For that, C0, C1 and A parameters from the variograms fitted for day 1 and day 21 values were compared. Also, kriged values converted to raster format were exported to QGIS software (QGIS Development Team, 2022). The raster calculator was used to subtract the values contained in both maps. Individually, each attribute difference of prediction was obtained, always subtracting the map of day 1 from the map of day 21. Positive values in the residual map mean day 21 overestimated the prediction from day 1. Negative values mean day 21 prediction underestimated day 1 prediction. That analysis allows to spatialize the differences of prediction in the experimental area.

4.3. Results and Discussion

The results for k-fold cross-validation of ML models used in this study are presented in Table 1. Primary response attributes, clay and OM, have well-known wavelengths of response (Fang et al., 2018; Nocita et al., 2015). As previously shown in Canal Filho and Molin (2022), both clay and OM ML models allowed the mapping of experimental area, although clay reported a considered low $R^2 = 0.17$. Of secondary response attributes, only CEC achieved similar field patterns of those observed in laboratory analysis, presenting an R^2 of 0.60 and RMSE of 3.51 $mmol_c\ kg^{-1}$. The poor parameters, especially of R^2, of 0.03 for pH, 0.02 for P, 0.14 for K, 0.39 for Ca and 0.01 for Mg, was confirmed as poor predictions results, although comparable errors of prediction and variance explained with clay, OM and CEC.

Table 1. Results of principal components regression models k-fold cross validation for clay, organic matter (OM), cation exchange capacity (CEC), potential of hydrogen (pH), phosphorus (P), potassium (K), calcium (Ca) and magnesium (Mg).

	unit	min		max	range	R^2	RMSE	MAE	NC	% var
Clay	$g\ kg^{-1}$	51	-	183	132	0.17	19.88	15.08	4	24.01
OM		12	-	35	23	0.75	3.11	2.28	9	40.69
CEC	$mmol_c\ kg^{-1}$	45	-	68	24	0.60	3.51	2.78	6	26.70
pH	-	4.1	-	6.8	2.7	0.03	0.32	0.27	4	12.09
P	$mg\ kg^{-1}$	5	-	68	63	0.02	9.39	8.59	3	18.42
K	$mmol_c\ kg^{-1}$	0.4	-	5	4.6	0.14	0.93	0.77	6	32.68
Ca		10	-	34	24	0.39	2.54	2.08	10	58.62
Mg		5	-	22	17	0.01	1.83	1.41	1	5.85

min: minimum value inserted in calibration; max: maximum value inserted in calibration; range: range of values inserted in calibration; R^2: coefficient of determination; RMSE: root mean squared error; NC: number of principal components used in regression; % var: percentage of total of outcome variance explained.

4.3.1. Spectra and prediction values analysis

The mean spectra from day 1 and day 21 proved that the intensity of reflectance was higher in day 1 (Figure 3). One possible factor contributing for that is the soil moisture. In day 1, the soil gravimetric moisture (θ) was 41.6 $g\ g^{-1}$, while, in day 21, θ was 69.5 $g\ kg^{-1}$. Water has an effect on DRS NIR spectra of augmenting absorption (Morellos et al., 2016; Nocita et al., 2015, 2013; Wang et al., 2020). The interaction of energy with matter can happen as transmission, reflexion or absorption, and one of it is a function of the other (Kortüm et al., 1963). The higher the absorption, the lower the reflectance. Even a tender rise in soil moisture could be perceived, reducing the mean reflectance values from approximately 0.135-0.145 to 0.110-0.120. Nevertheless, other factors can be contributing for the mean reduced intensity observed, such as environmental factors not considered in this study (sunlight radiation, temperature, etc.).

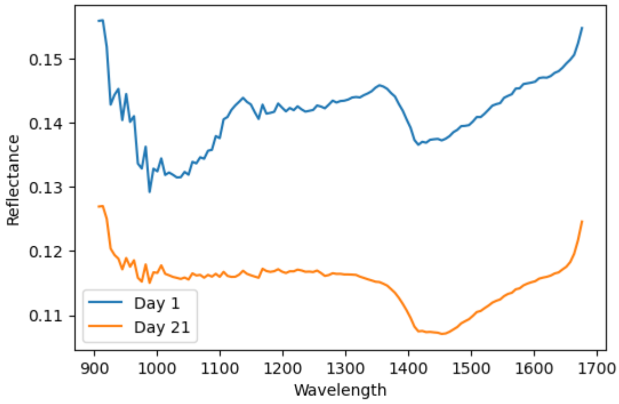


Figure 3. Mean spectra from day 1 and day 21 representing the intensity of reflectance on each acquisition day, obtained by averaging the reflectance values of the wavelengths read by the sensor.

The 20 most correlated spectra from day 1 and day 21 had their Pearson's correlation coefficient (r) ranging from 0.76-0.91, all were significative at 99%. Analyzing the spatialization of the 20 most correlated pairs, it is observed that 12 of those are at a distance of 20 up to 40 m from the other (Figure 4). None pair had a distance between each other lower than 20 m. The other pairs were far from each other. The pair represented by the blue square, for example, was separated by up to 500 m. This imply that the correlation of 8 out of the 20 most correlated pairs (or 40%), was not due to the proximity in the field, as it was expected and observed for the other 60%. However, this analysis indicates the need of further investigate if the variation of clay and OM in the experimental area can be the cause for the distant but highly correlated spectra. If two different portions had the same content (in quantity and quality) of primary NIR response attributes, it would be logic that the spectra from these two portions would be correlated since the sensor would perceive the same wavelengths of response (Kuang et al., 2012; Pasquini, 2018; Stenberg et al., 2010).

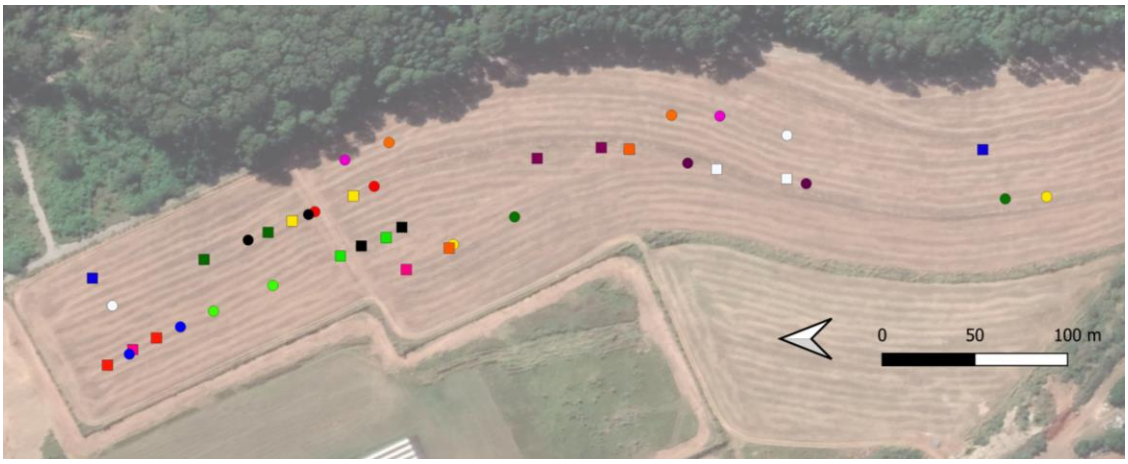


Figure 4. Distribution of the 20 highest correlated spectra pairs (day 1 x day 21) in the experimental field. Each pair is composed of a day 1 spectrum and its corresponding day 21 correlated spectrum., and are represented by the same geometric shape (square or circle) and color.

The spectra intensity was higher in day 1 on 16 of the 20 most correlated pairs, corroborating the mean spectra obtained on the entire population (140 from day 1 and 140 from day 21), and the implication of the higher soil moisture observed in acquisition of day 21, reducing the reflectance (Morellos et al., 2016; Nocita et al., 2015, 2013; Wang et al., 2020). For the spectra morphology, similar shapes and absorption features can be observed along the wavelengths of NIR spectra. Figure 5 exemplifies two pairs of correlated spectra separated by approximately 30 m. As it was observed from other authors, the primary attributes have characteristic wavelengths of interaction (Nocita et al., 2015; Pasquini, 2018). Even the acquisition happened separated by 21 days, the sensor perceived the same soil-spectra peculiarities, implying that the core of DRS technique was maintained in both acquisition days.

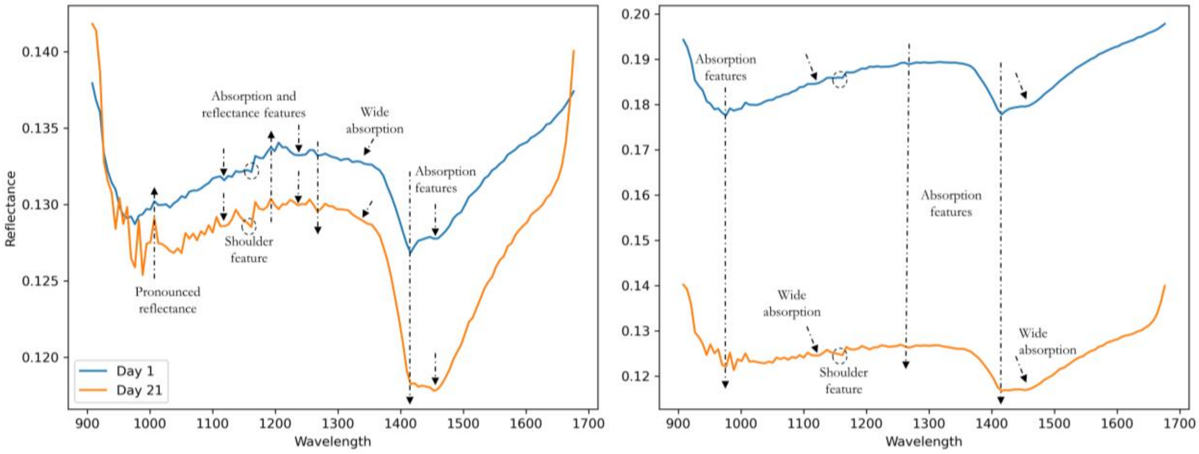


Figure 5. Spectra morphology analysis of two pairs of high-correlated spectra presenting comparable shape and absorption features.

Almost all the predicted values of day 1 and day 21 presented low correlation (Table 2). The total population had its strongest positive correlation for Ca prediction, with r = 0.22, which was the only positive r value significative at 99%. As Ca is a secondary NIR response attribute, and did not show adequate quantification or spatialization by the models used in this study, indicating no causal relation in the prediction, its exceptional 99% significative correlated prediction in both days might have happened by chance.

Table 2. Pearson's correlation coefficient of predicted values of entire day 1 and day 21 populations (Total – nearest neighbor), of the most correlated spectra pairs (20 most correlated) and the 32 pairs of day and day 21 overlap spectra extracted using an ellipse of 2.5 m radius (Overlap).

	Clay	OM	CEC	pH	P	K	Ca	Mg
Total	-0.02	-0.03	-0.14	-0.26*	-0.56*	-0.01	0.22*	0.20
20 most correlated	-0.06	0.19	0.32	0.04	-0.80*	-0.03	0.24	0.19
Overlap	-0.10	0.09	-0.11	-0.19	-0.73*	-0.22	0.33	0.13

OM: organic matter; CEC: cation exchange capacity; pH: potential of hydrogen; P: phosphorus; K: potassium; Ca: calcium; Mg: magnesium. *99% significative correlation.

Despite P and pH also presenting significative r, as it were negative r values, it is not of interest for the use of the ML models, denoting that zones of high values were inverted into zones of low P and pH values. P prediction had an alternance of distribution comparing the two days, represented by the negative r values of -0.56 for total population, and also observed in the most correlated and in the overlap spectra, with r = -0.80 and -0.73, respectively.

The attribute P has no direct response in NIR spectra, and its successful prediction is hardly obtained with this technique, particularly in tropical soils, because of the great adsorption specificity of this attribute with iron and aluminum oxides, and 1:1 clay minerals such as kaolinite (Pavinato et al., 2020). Usually, the best predictions of P are observed in temperate soils, especially those under organic fertilization, since the organic amendments compete for site-specific P adsorption, leaving more P labile (Mouazen and Kuang, 2016). In this situation, a positive correlation is frequently observed between P and OM. Being the last one a direct NIR response attribute, that allow to map P using the technique. In this context, the P prediction was not expected to work properly in the present study.

However, the inversion of patterns in P prediction can further indicate that, when the model identifies no pattern for an attribute, it will randomly assign values depending on the conditions of the spectra. Therefore, using DRS NIR spectra for indirect calibrations have to be strictly used after a covariation analysis with primary attributes in the desirable area (Chang et al., 2001; Stenberg et al., 2010). As a random attribution of values will follow, there is the risk of still presenting a reasonable prediction, but the absence of causality in this situation puts in risk the leverage of the technique as a reliable PA tool.

The correlation of predicted values for the 20 most correlated spectra is an indication that, however the spectra presenting correlation, similar shape and absorption features, as previously observed, the ML models consider other characteristics for attribute's quantification. Perchance, the spectra intensity is one of the major characteristics considered, as properties like moisture, mineralogy, clay content and OM directly affect the reflectance intensity (Stenberg et al., 2010; Terra et al., 2018).

The overlap predictions represent those spectra acquired in the same position in the area. The greatest positive correlation was observed for Ca prediction, with r = 0.33. Nevertheless, attributes like clay, OM, CEC and Mg were nearly independent between day 1 and day 21 predictions, while pH, P and K had negative correlation, whose are also not desirable for the use of a ML calibration of DRS NIR spectra over the time. Especially for primary NIR attributes, these contents are well-known for being stable in an agricultural area over a short period of time, either for clay that changes along soil weathering stages (Jackson and Sherman, 1953), or for OM even with long-term applications and conservationist management (Lu et al., 2021; Wang et al., 2019). The predictions for these attributes separated by 21 days, as it was made in this study, is expected to reach similar values.

The kernel density estimation plots show the distribution of predicted spectra from day 1 and day 21 (Figure 6). Clay prediction presented a similar distribution pattern. However, day 21 prediction tended to overestimate clay content in comparison with day 1. This may be in line with the observed for the mean spectra of both days. Both water and clay had a positive influence in absorption features of NIR spectra and negative influence in spectra intensity (Terra et al., 2018; Wang et al., 2020), and overtones of water and clay can also be observed in the same wavelengths (Nocita et al., 2015; Stenberg et al., 2010). Soil moisture was slightly higher in acquisition of day 21. As physical properties of soils dictate that the higher the clay content, the higher the water-holding capacity due to micropores augmentation (Rasa et al., 2018), it is suggested that the model identified the greater absorption features and lower intensity in day 2, and attributed that to clay.

Since a small change in soil moisture could have presented an influence in prediction, alternatives used to deal with the variation in soil moisture for NIR laboratory acquired spectra, such the external parameter orthogonalization (Wijewardane et al., 2016), the normalized soil moisture index (Nocita et al., 2013), direct standardization or orthogonal signal correction (Franceschini et al., 2018), should be also considered for online NIR spectra calibrations. Nevertheless, these strategies often use consecutive spectral acquisition of soil samples in different moisture contents. Authors resort to soil drying and rewetting and then build ML calibrations that consider the moisture level in prediction. This may be challenger to apply for online spectra, since the acquisition need to be made in field conditions.

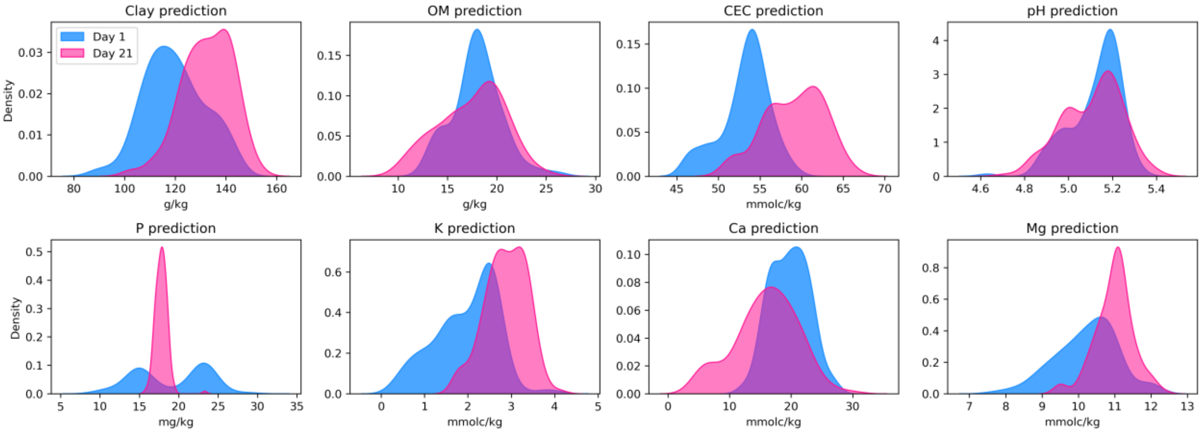
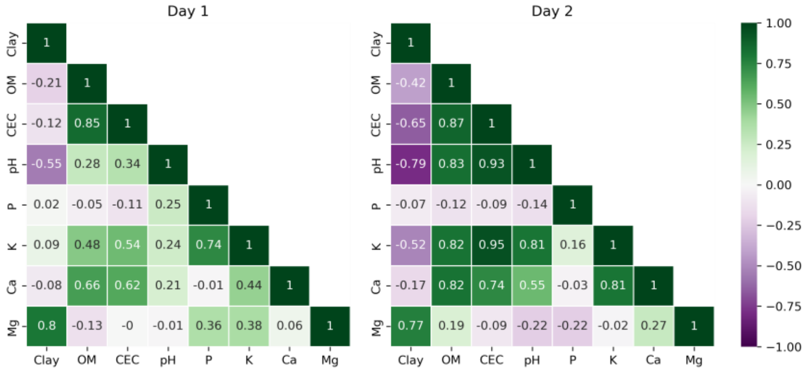


Figure 6. Kernel density estimation plots of the attributes predicted using the 140 online spectra from each acquisition day.

Other attributes that presented a similar pattern as clay, of overestimation on day 21, are CEC, K and Mg. This can also be related to the above described. However, the correlation among soil attributes predicted in both days suggest that it is now related to OM content (Figure 7). OM also has a property of water absorption and regulates negative charges in soil (Shepherd et al., 2002; Soane, 1990). CEC is directly related to the proportion of negative charges per mass unit, and also K and Mg contents, as their cation forms will be attracted by the negative charges of soil, absorbed by plants and extracted in laboratory. However, this was not true for Ca prediction, that are also a cation and can be directly related to the availability of negative charges, and therefore with OM and clay content.



Day 1

	Clay	OM	CEC	pH	P	K	Ca	Mg
Clay	1							
OM	-0.21	1						
CEC	-0.12	0.85	1					
pH	-0.55	0.28	0.34	1				
P	0.02	-0.05	-0.11	0.25	1			
K	0.09	0.48	0.54	0.24	0.74	1		
Ca	-0.08	0.66	0.62	0.21	-0.01	0.44	1	
Mg	0.8	-0.13	-0	-0.01	0.36	0.38	0.06	1

Day 2

	Clay	OM	CEC	pH	P	K	Ca	Mg
Clay	1							
OM	-0.42	1						
CEC	-0.65	0.87	1					
pH	-0.79	0.83	0.93	1				
P	-0.07	-0.12	-0.09	-0.14	1			
K	-0.52	0.82	0.95	0.81	0.16	1		
Ca	-0.17	0.82	0.74	0.55	-0.03	0.81	1	
Mg	0.77	0.19	-0.09	-0.22	-0.22	-0.02	0.27	1

Figure 7. Pearson's correlation among soil attributes predicted using online NIR spectra from day 1 and day 21 acquisitions.

The correlation analysis of populations predicted by both days shows that, especially observing the correlations with primary response attributes, no relationship was inverted (positive turn to negative, or the contrary). But it shows a pattern of intensification in correlations on day 21 prediction, either for negative or positive values. Despite this study suggest that water content can be related to this aspect observed, two main points remain: 1) The water could intensify the values predicted. However, this may be not the only explanation for the failure of day 21 spectra prediction using ML models calibrated in day 1. The analysis of predicted values, as shown in Table 2, are almost independent from each other; 2) This may imply that other environmental factors can be interfering in spectra acquisition, which can explain the differences between the two days' prediction. Since is harder to control circumstances of a field operation than of a laboratory, any peculiarity can change the aspects of acquired spectra.

The ML models developed seem to be very dependent on these circumstances, and a further investigation on factors as sunlight, soil temperature, air temperature, etc., needs to be carried out. This means to investigate if there is a possibility of providing the models the necessary data to deal with it, guaranteeing more stability in online NIR spectra (therefore, in ML calibrations that depend on them) over the time. The answers for these questions may clarify the need for real-time calibrations of NIR soil spectra for ML prediction.

4.3.2. Comparison of maps generated

The parameters of fitted variograms for day 1 and day 21 predictions highlight the differences between the two days (Table 3). Only Mg kept its theoretical model over the predictions, using exponential model in both. However, the C0 = 0.42 and A = 46.4 for day 1 changed for C0 = 0.02 and A = 17.5 in day 21 prediction. For primary response attributes, OM had a similar A, of 23.0 m for day 1 and 21.1 m for day 21. But for clay it increased in 50%, 24.1 m in day 1 and 36.8 m in day 21. CEC had the closer values for day 1 and day 21, for either C0, C1 and A. For pH and plant nutrients considered, the changes in variogram parameters were expected not only because of the differences above described in soil spectra of two days, but also due to the inability of ML models to predict these attributes in first place, as reported in Canal Filho and Molin (2022). A possible explanation is described by Huang et al. (2015), that outputs generated from gathered data and various processes have sources of errors that can accumulate, especially when applying machine learning to chemistry data (Vishwakarma et al., 2021).

Table 3. Parameters of fitted variograms for predicted attributes using Day1 and Day2 spectral acquisitions for clay, organic matter (OM), cation exchange capacity (CEC), pH and soil available nutrients, phosphorus (P), potassium (K), calcium (Ca) and magnesium (Mg).

	Day 1				Day 21			
	Model	C0	C1	A	Model	C0	C1	A
Clay	Sph	13.59	102.6	24.1	Exp	42.71	59.70	36.8
OM	Exp	1.82	3.85	23.0	Gau	0.00	10.18	21.1
CEC	Exp	0.00	8.50	18.5	Gau	0.00	10.82	20.5
pH	Lin	0.00	0.01	88.9	Exp	0.00	0.02	16.7
P	Gau	0.24	21.40	32.9	Exp	0.03	0.80	37.6
K	Gau	0.10	0.46	27.3	Lin	0.01	0.19	35.9
Ca	Sph	5.62	4.50	95.6	Gau	0.00	26.79	25.2
Mg	Exp	0.42	0.32	46.4	Exp	0.02	0.26	17.5

C0: nugget; C1: sill; A: range; ratio; Sph: spherical; Exp: exponential; Gau: gaussian; Lin: linear with sill. Str: strong; Mod: moderate

The maps generated from ordinary kriging and subtracted using the formula day 21 – day 1 are presented in Figure 8. How it was previously suggested, the water took a role in prediction differences between the two days, but was probably not the only factor of influence. Clay residual map had a tendency of being positive, showing the pattern of day 21 in overestimation in comparison with day 1 prediction. However, if the water was the only factor actuating, the water dynamic in topsoil would appear in clay residual map, since the greater soil moisture would affect the NIR spectra, highlighting the regions where altimetry would conduct the water such as described in Lee et al. (2011), Orth et al. (2013), and Uebbing et al. (2017). The lower portions of the area, following the altimetry and agricultural terraces, were not where the higher differences of prediction appeared. OM had a more random residual distribution than the observed for clay, with more portions where day 21 underestimated the prediction in day 1, despite the greater soil moisture and related wavelengths of OM and water in NIR spectra (Nocita et al., 2015; Wang et al., 2020). This corroborates the analysis that water content in the soil was not the only factor actuating in the differences observed between both days' spectra, predicted values and maps interpolated.

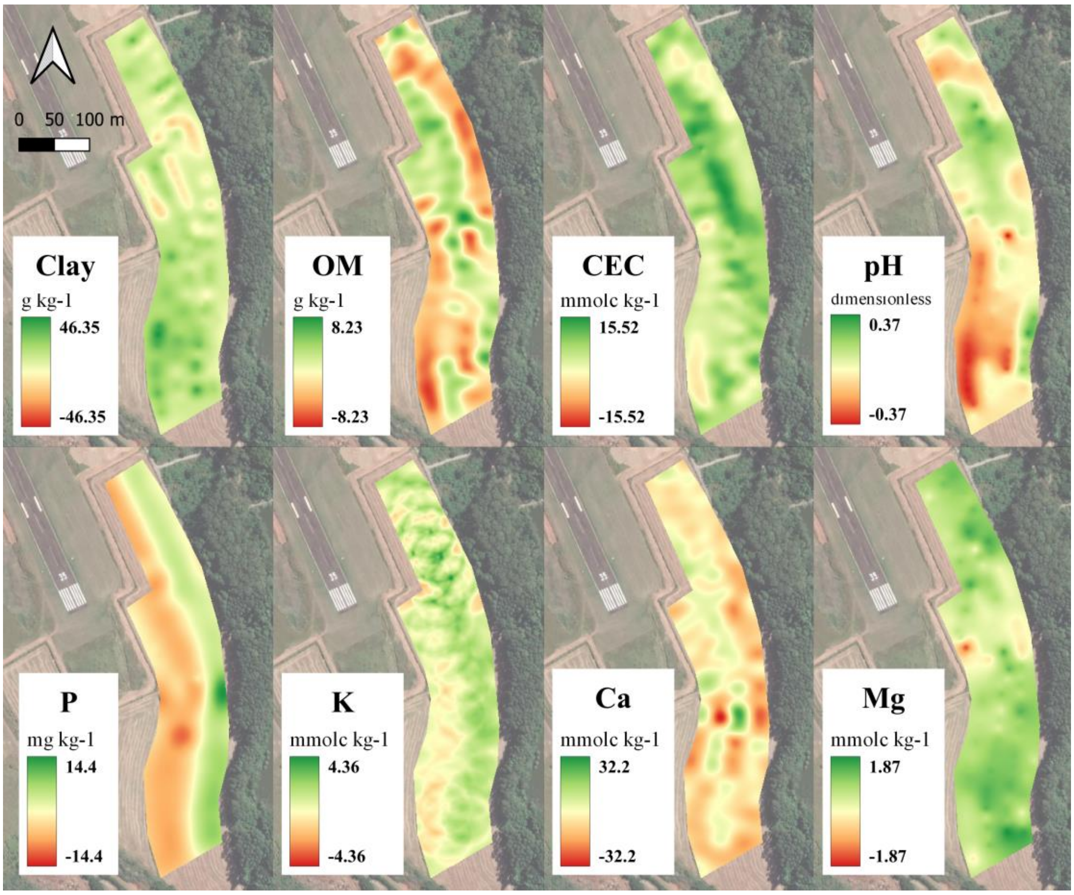


Figure 8. Maps of each evaluated attribute demonstrating the spatialization of the differences observed in the predictions of the two days, calculated by subtracting Day 21 - Day 1.

How it was suggested, this study may lead to different analysis of online NIR spectra. Environmental variables can be considered. ML models using different information that can be simultaneously acquired in field operations need to be tested. Other factors, not mentioned in this study, can also be proposed as sources of variation. The elucidation of these questions may clarify to the DRS community if there are strategies to consolidate the stability of online NIR spectra over the time. On the contrary, if the reported in this study prevail, the need for local calibrations in space and time to predict soil attributes using DRS in NIR region will be proven, and further strategies to deal with this will be necessary to leverage the technique into agricultural production.

4.4. Conclusions

An agricultural area where online NIR spectra and ML models for soil attributes prediction were validated was revisited, maintaining the operational and instrumentational parameters. NIR spectra morphology was preserved, but spectra intensity has changed, what can be related to soil moisture variation between both days. Highly correlated spectra were observed in close but also distant acquisition locations. However, high correlation did not mean similar values predicted, presenting almost independent values. Overlapped spectra between the two days also presented independent values in prediction. Therefore, ML models calibrated in the first day did not performed as consistent when used spectra acquired in the second day, suggesting the requirement of local calibrations for DRS NIR prediction both in space and time. The distribution of residuals between day 1 and day 21 suggested that the soil moisture was not the only issue of variation. Other factors, such as environmental, need to be investigated to address if the variation between different days of acquisition can be overcome.

References

Agarwal, A., Shah, D., Shen, D., Song, D., 2021. On Robustness of Principal Component Regression. **J Am Stat Assoc** 116. https://doi.org/10.1080/01621459.2021.1928513

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

Ben-Dor, E., Heller, D., Chudnovsky, A., 2008. A Novel Method of Classifying Soil Profiles in the Field using Optical Means. **Soil Science Society of America Journal** 72. https://doi.org/10.2136/sssaj2006.0059

Bönecke, E., Meyer, S., Vogel, S., Schröter, I., Gebbers, R., Kling, C., Kramer, E., Lück, K., Nagel, A., Philipp, G., Gerlach, F., Palme, S., Scheibe, D., Zieger, K., Rühlmann, J., 2021. Guidelines for precise lime management based on high-resolution soil pH, texture and SOM maps generated from proximal soil sensing data. **Precision Agriculture** 22. https://doi.org/10.1007/s11119-020-09766-8

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. https://doi.org/10.1016/j.geoderma.2007.04.021

Canal Filho, R., Molin, J.P., 2022. Spatial distribution as a key factor for evaluation of soil attributes prediction at field level using online near-infrared spectroscopy. **Frontiers in Soil Science** 2. https://doi.org/10.3389/fsoil.2022.984963

Chang, C.-W., Laird, D.A., Mausbach, M.J., Hurburgh, C.R., 2001. Near-Infrared Reflectance Spectroscopy-Principal Components Regression Analyses of Soil Properties. **Soil Science Society of America Journal** 65. https://doi.org/10.2136/sssaj2001.652480x

Eitelwein, M.T., Tavares, T.R., Molin, J.P., Trevisan, R.G., de Sousa, R.V., Demattê, J.A.M., 2022. Predictive Performance of Mobile Vis–NIR Spectroscopy for Mapping Key Fertility Attributes in Tropical Soils through Local Models Using PLS and ANN. **Automation** 3. https://doi.org/10.3390/automation3010006

Fang, Q., Hong, H., Zhao, L., Kukolich, S., Yin, K., Wang, C., 2018. Visible and Near-Infrared Reflectance Spectroscopy for Investigating Soil Mineralogy: A Review. **Journal of Spectroscopy**. https://doi.org/10.1155/2018/3168974

Franceschini, M.H.D., Demattê, J.A.M., Kooistra, L., Bartholomeus, H., Rizzo, R., Fongaro, C.T., Molin, J.P., 2018. Effects of external factors on soil reflectance measured on-the-go and assessment of potential spectral correction through orthogonalisation and standardisation procedures. **Soil and Tillage Research** 177. https://doi.org/10.1016/j.still.2017.10.004

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. https://doi.org/10.1016/j.geoderma.2013.07.016

Huang, J., Zare, E., Malik, R.S., Triantafilis, J., 2015. An error budget for soil salinity mapping using different ancillary data. **Soil Research** 53. https://doi.org/10.1071/SR15043

International Society of Precision Agriculture, 2022. Precision Agriculture definition [WWW Document].

Jackson, M.L., Sherman, G.D., 1953. Chemical Weathering of Minerals in Soils. **Advances in Agronomy** 5. https://doi.org/10.1016/S0065-2113(08)60231-X

Johnston, A.E., Poulton, P.R., 2018. The importance of long-term experiments in agriculture: their management to ensure continued crop production and soil fertility; the Rothamsted experience. **European Journal of Soil Science** 69. https://doi.org/10.1111/ejss.12521

Jung, Y., 2018. Multiple predicting K-fold cross-validation for model selection. **Journal of Nonparametric Statistics** 30. https://doi.org/10.1080/10485252.2017.1404598

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., Willing, C., 2016. Jupyter Notebooks—a publishing format for reproducible computational workflows, in: **Positioning and Power in Academic Publishing:** Players, Agents and Agendas - Proceedings of the 20th International Conference on Electronic Publishing, ELPUB 2016. https://doi.org/10.3233/978-1-61499-649-1-87

Kortüm, G., Braun, W., Herzog, G., 1963. Principles and Techniques of Diffuse-Reflectance Spectroscopy. Angewandte **Chemie International Edition** in English 2. https://doi.org/10.1002/anie.196303331

Kuang, B., Mahmood, H.S., Quraishi, M.Z., Hoogmoed, W.B., Mouazen, A.M., van Henten, E.J., 2012. Sensing soil properties in the laboratory, in situ, and on-line. A review, in: **Advances in Agronomy**. https://doi.org/10.1016/B978-0-12-394275-3.00003-1

Lee, H., Beighley, R.E., Alsdorf, D., Jung, H.C., Shum, C.K., Duan, J., Guo, J., Yamazaki, D., Andreadis, K., 2011. Characterization of terrestrial water dynamics in the Congo Basin using GRACE and satellite radar altimetry. **Remote Sensing Environment** 115. https://doi.org/10.1016/j.rse.2011.08.015

Lu, Y., Gao, Y., Nie, J., Liao, Y., Zhu, Q., 2021. Substituting chemical P fertilizer with organic manure: effects on double-rice yield, phosphorus use efficiency and balance in subtropical China. **Scientific Reports** 11. https://doi.org/10.1038/s41598-021-87851-2

Minasny, B., McBratney, A.B., Whelan, B.M., 2006. VESPER version 1.62.

Molin, J.P., Tavares, T.R., 2019. Sensor systems for mapping soil fertility attributes: Challenges, advances, and perspectives in brazilian tropical soils. **Engenharia Agricola** 39. https://doi.org/10.1590/1809-4430-ENG.AGRIC.V39NEP126-147/20190126

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

Mouazen, A.M., Kuang, B., 2016. On-line visible and near infrared spectroscopy for in-field phosphorous management. **Soil and Tillage Research** 155. https://doi.org/10.1016/j.still.2015.04.003

Mouazen, A.M., Maleki, M.R., Cockx, L., van Meirvenne, M., van Holm, L.H.J., Merckx, R., de Baerdemaeker, J., Ramon, H., 2009. Optimum three-point linkage set up for improving the quality of soil spectra and the accuracy of soil phosphorus measured using an on-line visible and near infrared sensor. **Soil and Tillage Research** 103. https://doi.org/10.1016/j.still.2008.10.006

Munnaf, M.A., Guerrero, A., Nawar, S., Haesaert, G., van Meirvenne, M., Mouazen, A.M., 2021. A combined data mining approach for on-line prediction of key soil quality indicators by Vis-NIR spectroscopy. **Soil and Tillage Research** 205. https://doi.org/10.1016/j.still.2020.104808

Munnaf, M.A., Mouazen, A.M., 2021. Development of a soil fertility index using on-line Vis-NIR spectroscopy. **Computers and Electronics in Agriculture** 188. https://doi.org/10.1016/j.compag.2021.106341

Munnaf, M.A., Nawar, S., Mouazen, A.M., 2019. Estimation of secondary soil properties by fusion of laboratory and on-line measured Vis-NIR spectra. **Remote Sensing** (Basel) 11. https://doi.org/10.3390/rs11232819

Nocita, M., Stevens, A., Noon, C., van Wesemael, B., 2013. Prediction of soil organic carbon for different levels of soil moisture using Vis-NIR spectroscopy. **Geoderma** 199. https://doi.org/10.1016/j.geoderma.2012.07.020

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

Orth, R., Koster, R.D., Seneviratne, S.I., 2013. Inferring soil moisture memory from streamflow observations using a simple water balance model. **Journal of Hydrometeorology** 14. https://doi.org/10.1175/JHM-D-12-099.1

Pasquini, C., 2018. Near infrared spectroscopy: A mature analytical technique with new perspectives – A review. **Analytica Chimica Acta**. https://doi.org/10.1016/j.aca.2018.04.004

Pavinato, P.S., Cherubin, M.R., Soltangheisi, A., Rocha, G.C., Chadwick, D.R., Jones, D.L., 2020. Revealing soil legacy phosphorus to promote sustainable agriculture in Brazil. **Scientific Reports** 10. https://doi.org/10.1038/s41598-020-72302-1

Python Software Foundation, 2022. Python [WWW Document].

QGIS Development Team, 2022. QGIS Geographic Information System.

Rasa, K., Heikkinen, J., Hannula, M., Arstila, K., Kulju, S., Hyväluoma, J., 2018. How and why does willow biochar increase a clay soil water retention capacity? Biomass **Bioenergy** 119. https://doi.org/10.1016/j.biombioe.2018.10.004

Seasholtz, M.B., Kowalski, B., 1993. The parsimony principle applied to multivariate calibration. **Analytica Chimica Acta** 277. https://doi.org/10.1016/0003-2670(93)80430-S

Shepherd M.A.*, Harrison, R., Webb, J., 2002. Managing soil organic matter – implications for soil structure on organic farms. **Soil Use Management** 18. https://doi.org/10.1079/sum2002134

Soane, B.D., 1990. The role of organic matter in soil compactibility: A review of some practical aspects. **Soil and Tillage Research** 16. https://doi.org/10.1016/0167-1987(90)90029-D

Stenberg, B., Rogstrand, G., Bölenius, E., Arvidsson, J., 2007. On-line soil NIR spectroscopy: Identification and treatment of spectra influenced by variable probe distance and residue contamination, in: **Precision Agriculture** 2007 - Papers Presented at the 6th European Conference on Precision Agriculture, ECPA 2007.

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

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. https://doi.org/10.1371/journal.pone.0066409

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

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

Tracy, T., Fu, Y., Roy, I., Jonas, E., Glendenning, P., 2016. Towards machine learning on the Automata processor, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). https://doi.org/10.1007/978-3-319-41321-1_11

Uebbing, B., Forootan, E., Braakmann-Folgmann, A., Kusche, J., 2017. Inverting surface soil moisture information from satellite altimetry over arid and semi-arid regions. **Remote Sensing Environment**. https://doi.org/10.1016/j.rse.2017.05.004

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

Vishwakarma, G., Sonpal, A., Hachmann, J., 2021. Metrics for Benchmarking and Uncertainty Quantification: Quality, Applicability, and Best Practices for Machine Learning in Chemistry. **Trends Chem**. https://doi.org/10.1016/j.trechm.2020.12.004

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

Wang, H., Xu, J., Liu, X., Zhang, D., Li, L., Li, W., Sheng, L., 2019. Effects of long-term application of organic fertilizer on improving organic matter content and retarding acidity in red soil from China. **Soil and Tillage Research** 195. https://doi.org/10.1016/j.still.2019.104382

Wang, Y.P., Lee, C.K., Dai, Y.H., Shen, Y., 2020. Effect of wetting on the determination of soil organic matter content using visible and near-infrared spectrometer. **Geoderma** 376. https://doi.org/10.1016/j.geoderma.2020.114528

Wetterlind, J., Stenberg, B., Söderström, M., 2010. Increased sample point density in farm soil mapping by local calibration of visible and near infrared prediction models. **Geoderma** 156. https://doi.org/10.1016/j.geoderma.2010.02.012

Wijewardane, N.K., Ge, Y., Morgan, C.L.S., 2016. Moisture insensitive prediction of soil properties from VNIR reflectance spectra based on external parameter orthogonalization. **Geoderma** 267. https://doi.org/10.1016/j.geoderma.2015.12.014

Yang, M., Mouazen, A., Zhao, X., Guo, X., 2020. Assessment of a soil fertility index using visible and near-infrared spectroscopy in the rice paddy region of southern China. **European Journal of Soil Science** 71. https://doi.org/10.1111/ejss.12907

Yin, H., Cao, Y., Marelli, B., Zeng, X., Mason, A.J., Cao, C., 2021. Soil Sensors and Plant Wearables for Smart and Precision Agriculture. **Advanced Materials**. https://doi.org/10.1002/adma.202007764

5. FINAL REMARKS

The use of diffuse reflectance spectroscopy in near-infrared region for online spectra acquisition and soil attributes prediction was tested in this study. The main findings were organized into three chapters, whose remarks can be synthetized as:

a) Procedures for data modeling were assessed. Dimensionality reduction statistical techniques outperformed the non-linear ones for machine learning models calibration, highlighting the efficiency and robustness of principal components regression models to deal with the multivariate character of soil spectra. The common applied spectra preprocessing techniques was tested and did not attend the expectation of creating more accurate models, suggesting it did not aid in the identification of noisy, redundant and irrelevant data.

b) The best strategy reported for the calibration of machine learning models was to use only soil samples from the area desired for attributes prediction. The spatial distribution of predicted attributes, particularly in the sense of agronomic plausibility of this distribution, proved to be a key factor for evaluation alongside the already established statistical parameters. These lead to the recommendation of local calibrations in space for the use of diffuse reflectance spectroscopy in proximal soil sensing.

c) The characterization of spatio-temporal stability of online near-infrared spectra proved that this factor is a challenger condition to the leverage of the technique into agricultural production. The prediction models calibrated in one day of soil online spectral acquisition did not consistently perform in a posterior soil acquisition that followed the same experimental and instrumentational parameters. Spectra characteristics, predicted values and soil mapping presented discrepancies between the two days. Besides other variables should be investigated to define if there are strategies to overcome the low stability of online near-infrared spectra, this can indicate for the recommendation of local calibrations not only in space, but also in time for the use of diffuse reflectance spectroscopy in proximal soil sensing.

The results obtained and presented in this work provide guidelines for the application of diffuse reflectance spectroscopy in agriculture, allowing to conclude that the use of online near-infrared spectra for soil attributes prediction is possible due to the predictive quality of both quantification and spatialization of the values obtained. The findings contribute to the advancement of these tools as techniques to support precision agriculture practices and suggest the orientation of future research that should be dedicated to the test of the technique for inference in the field, in the decision-making of the productive steps, in order to establish its use in large scale.