NEAR-INFRARED OPTOELECTRONIC SYSTEM FOR SOIL ORGANIC CARBON PERCENTAGE ESTIMATION

Pablo Andrés Gómez Toloza

Ingeniero Electrónico

Universidad Industrial de Santander

Facultad de Ingenierías Fisicomecánicas

Escuela de Ingenierías Eléctrica, Electrónica y de Telecomunicaciones

Bucaramanga

2024

NEAR-INFRARED OPTOELECTRONIC SYSTEM FOR SOIL ORGANIC CARBON PERCENTAGE ESTIMATION

Pablo Andrés Gómez Toloza

Ingeniero Electrónico

Master thesis to qualify for the title of *Magister en Ingeniería Electrónica*

Advisor

Ph.D. Henry Arguello Fuentes

Co-Advisor

Ph.D. Hans Yecid García Arenas

Universidad Industrial de Santander

Facultad de Ingenierías Fisicomecánicas

Escuela de Ingenierías Eléctrica, Electrónica y de Telecomunicaciones

Bucaramanga

2024

DEDICATORY

To my family, for their love, support, and guidance.

To my love, Daniela, for being by my side through every step of this research.

ACKNOWLEDGMENTS

To my love, Daniela, for her unwavering support and encouragement through the most challenging moments.

My sincere gratitude to Professor Henry Arguello for his invaluable guidance and professional and academic insights.

To my co-advisor, Hans Garcia, for his dedication, mentorship, and invaluable role in shaping my personal and professional growth.

To the members of the HDSP group for their continued support and patience in our academic and professional development.

To the HDSP Optical Laboratory team, whose contributions were essential to the completion of this research.

Table of contents

List of figures

List of tables

Resumen

Titulo: Sistema optoelectrónico de infrarrojo cercano para la estimación del porcentaje de carbono orgánico en el suelo.

Director: Ph.D. Henry Arguello Fuentes, henarfu@uis.edu.co.

Co-Director: Ph.D Hans Yecid García Arenas, hans.garcia@correo.uis.edu.co.

Autor: Pablo Andrés Gómez Toloza, pablo2228330@correo.uis.edu.co.

Entidad: Universidad Industrial de Santander.

Interesado: High Dimensional Signal Processing Group (HDSP).

En el sector agrícola, controlar y monitorear las propiedades físico-químicas del suelo es de vital importancia ya que permite optimizar los cultivos. Si bien existen una amplia gama de propiedades físico-químicas del suelo, el porcentaje de carbono orgánico (COS), destaca ya que una correcta caracterización de los niveles de COS permite mejorar la productividad de los cultivos. Tradicionalmente, se han utilizado los métodos químicos para calcular el porcentaje de COS con alta precisión, pero con costos de análisis elevados y una respuesta lenta. En consecuencia, la espectroscopía de infrarrojo cercano (NIR) ha demostrado ser una herramienta útil para estimar características intrínsecas de muestras de suelo como el COS, a partir de información espectral adquirida utilizando sistemas optoelectrónicos. Sin embargo, para procesar estas firmas espectrales, se han implementado algoritmos computacionales que, a través de una arquitectura computacionalmente compleja y un gran conjunto de datos de firmas espectrales en el NIR estiman el porcentaje de COS. Por lo tanto, en este trabajo propone un sistema NIR optoelectrónico para la estimación del porcentaje de COS, basado en técnicas de procesamiento aplicadas en escenarios con pocas firmas espectrales de muestras de suelos colombianas.

Abstract

Title: Near-infrared optoelectronic system for soil organic carbon percentage estimation.

Advisor: Ph.D. Henry Arguello Fuentes, henarfu@uis.edu.co.

Co-Advisor: Ph.D Hans Yecid García Arenas, hans.garcia@correo.uis.edu.co.

Author: Pablo Andrés Gómez Toloza, pablo2228330@correo.uis.edu.co.

Entity: Universidad Industrial de Santander.

Stakeholder: High Dimensional Signal Processing Group (HDSP).

In agriculture, controlling and measuring soil physicochemical properties are vital since they allow crop optimization. Therefore, chemical methods have traditionally been used to calculate the percentage of organic carbon in soil samples (SOC) with high accuracy, but with high analysis costs and slow response. Consequently, near-infrared spectroscopy (NIR) has proven to be a useful tool, since it allows estimating intrinsic characteristics of soil samples such as SOC, from spectral information acquired using optoelectronic systems. However, to process these spectral signatures, computational algorithms have been implemented through a computationally complex architecture, high training times, and a large dataset of spectral signatures in the NIR labeled with their respective percentage of organic carbon, to estimate the SOC percentage. This work proposes an optoelectronic NIR system for SOC percentage estimation, which allows the SOC estimation,based on processing techniques applied when there are few labeled spectral signatures.

Introduction

Characterizing cultivated lands is fundamental in agriculture, as it enhances crop productivity, and efficiency Niu et al. (2024). Among the most crucial features for agricultural lands is the soil organic carbon (SOC) percentage, which provides information about soil fertility and moisture, among other soil characteristics, enabling the application of techniques to improve the land Pavlovic et al. (2024). Traditionally, chemical methods such as Calcination and Dichromate are used to estimate the SOC percentage, calculating it from the incineration or dissolution of the soil sample, thereby destroying the sample in the process. While these chemical methods provide high sensitivity in SOC percentage calculation, they present difficulties regarding transportation, costs, and analysis response times Nocita et al. (2015); Smith et al. (2019); Zhong et al. (2021). An alternative for soil quality analysis is soil characterization through spectral information. Spectral information refers to the amount of light reflected or emitted by an object across different electromagnetic spectrum wavelengths. The spectrum is crucial for soil characterization because it provides insights into its physical, chemical, and biological properties, such as mineral composition, presence of organic matter, SOC content, and texture. In particular, the estimation of SOC has been extensively studied in the state-of-the-art by jointly using spectral images in the near-infrared (NIR) range and computational algorithms. Spectral signatures in the NIR provide information that chemical methods usually ignore; specifically, soil's spectral response can significantly contribute to estimating soil properties and environmental conditions. Therefore, several spectral signature acquisition architectures and computational algorithms for SOC estimation have been developed

Diaz et al. (2024); Cao et al. (2024); Gomez et al. (2022); Wang et al. (2024); Sharma et al. (2024). Spectral acquisition methodologies in soil samples, such as Whiskbroom and Pushbroom, have been widely used because of their ease of implementation and high spectral resolution. However, these scanning methodologies have notable shortcomings in repeatability and reduced operator interaction with the sample. On the other hand, based on the rise of artificial intelligence, various computational algorithms have been developed to extract intrinsic features from a few spectral signatures and estimate the SOC percentage. However, the accuracy of these computational algorithms is directly related to the number of spectral signatures. Therefore, it has become necessary to implement a computational algorithm that achieves precise SOC estimation results in environments with few spectral signatures. Hence, this work presents an optoelectronic system that allows for SOC percentage estimation from a few spectral signatures in the NIR region. The proposed optoelectronic system is based on the assumption that the physicochemical composition of a soil sample is concentrated in its spectral signature in the NIR region. Therefore, this study introduces an optoelectronic system that allows the acquisition of spectral information in the NIR region from a soil sample in situ in a controlled and repeatable scenario. To achieve this, an automated acquisition protocol is implemented, calibrated, and designed for a Whiskbroom-type system, enabling the acquisition of repeatable spectral signatures with a low standard deviation. Additionally, a computationally low-complexity estimation algorithm is implemented, allowing the estimation of the SOC percentage from the limited number of acquired spectral signatures. Finally, a comparative analysis is conducted between the estimation algorithm results and the SOC values obtained from a chemistry laboratory employing the coefficient of determination (r^2) metric.

1. Objectives

General

- To implement an optoelectronic system to estimate the percentage of soil organic carbon SOC from near-infrared (NIR) spectral information employing a computational algorithm.

Specifics

1. To implement and calibrate an optoelectronic system that allows the acquisition of spectral signatures in the NIR range using a point-to-point spectral scanning protocol.

2. To acquire a database of spectral signatures in the NIR, of soil samples with previously labeled SOC.

3. To implement a computational algorithm to estimate of organic carbon based on a limited number of spectral signatures in the NIR range.

4. To verify the organic carbon percentage estimations of soil samples by comparing the results obtained on R^2 using the implemented algorithm with the physicochemical SOC results from chemical laboratories.

Impact of the Research and Contributions

Chapter 3 Optoelectronic System for NIR Spectral Signature Acquisition: This chapter introduces the design, implementation, and acquisition protocol of a whiskbroom-type optical system that enables controlled acquisition of spectral signatures from soil samples to estimate the percentage of SOC.

Chapter 4 Characterization and acquisition of spectral signature datasets: This chapter involves the optoelectronic system and protocol acquisition implemented in Chapter 3; a description and characterization of the terrain and the soil sample extraction protocol, as well as the acquisition and analysis of the spectral signatures database in the NIR range.

Chapter 5 SOC estimation algorithm from a limited number of NIR spectral signatures: This chapter presents the implementation of an algorithm enabling SOC percentage estimation using NIR spectral signatures, conducting comparisons with state-of-the-art algorithms for SOC estimation based on computational complexity.

Chapter 6Validation of SOC estimation compared to state-of-the-art methods: This chapter describes a quantitative comparison of SOC estimation using the proposed computational algorithm against state-of-the-art algorithms relative to the number of spectral signatures employed, as well as precision compared to traditional methods.

List of Publications

Published journal papers

1. Roman Jacome, **Pablo Gomez**, Henry Arguello, (2023), Middle output regularized end-to-end optimization for computational imaging, pp. 1421-1431, vol. 10, Issue 11, **Optica**.

Ongoing papers

1. Ariolfo Camacho, Hans Garcia, **Pablo Gomez**, Carlos Velasquez, Andres Jerez, Flavio Prieto, and Henry Arguello, Soil Organic Carbon Estimation through a Designed Optical Sensing System, **Submitted**

2. Kevin Arias, **Pablo Gomez**, Carlos Hinojosa, Juan Carlos Niebles, and Henry Arguello. Protecting Imaging from Manipulations with Deep Learning Optical Signature, **Submitted**

3. Manuel Herrera, David Morales Norato, **Pablo Gomez**, Hans Garcia, Hoover Rueda, and Henry Arguello, Implicit Fusion spectral Representation **Ongoing**.

4. **Pablo Gomez**, Roman Jacome, Hans Garcia and Henry Arguello, Spectral Authenticity for computational imaging, **Ongoing**.

Conference papers

1. **Pablo Gomez**, Hans Garcia, and Henry Arguello, (2023), Computational algorithm for soil organic carbon percentage estimation through NIR spectroscopy, **2023 Computational Optical Sensing and Imaging (COSI)**, paper JW2A.7.

2. **Pablo Gomez**, Ariolfo Camacho, and Henry Arguello, (2022), Design and Implementation of an Automated Protocol for Spectral Signatures Acquisition on Colombian Agricultural Soil Samples Into the Visible and Infrared Range, **2022 IEEE ANDESCON**, Barranquilla, Colombia, 2022, pp. 1-6, doi: 10.1109/ANDESCON56260.2022.9989753.

3. Sebastian Ardila, **Pablo Gomez**, Lineth Orduz, Robert Gomez, Jorge Bacca, Henry Arguello, and Hans Garcia, Low-cost optoelectronic system for IR spectral acquisition based on band selection, **XVIII National Meeting on Optics and the IX Andean and Caribbean Conference on Optics and its Applications (ENO-CANCOA)**, **presented**.

4. **Pablo Gomez**, Sebastian Ardila, Lineth Orduz, Robert Gomez, Henry Arguello, and Hans Garcia, Design of a pushbroom NIR optimized system for citrus spectral data acquisition, **XVIII National Meeting on Optics and the IX Andean and Caribbean Conference on Optics and its Applications (ENO-CANCOA)**, **presented**.

5. Sergio Urrea, **Pablo Gomez**, Karen Fonseca, Hans Garcia, and Henry Arguello, Mismatch Correction for End-to-End Designed Phase-Encoded-Based Spectral Imaging System, **2024 32th European Signal Processing Conference (EUSIPCO)**, **presented**.

6. **Pablo Gomez**, Roman Jacome, Emmanuel Martinez, Hans Garcia, Henry Arguello, Optical authenticity in pushbroom system for spectral information protection,2024 IEEE International Conference on Acoustics, Speech and Signal Processing, **submitted**.

Directed/co-directed thesis

1. **Undergrad thesis**, Diseño e implementación de un prototipo de banda transportadora para la identificación automática de productos cítricos mediante inteligencia artificial, Juliana Lucia Pineda Cardozo y Sergio Andrés Jimenez Buitrago.

2. Theoretical background

Agriculture has become one of the most influential sectors in the world Abbass et al. (2022), because of its growth in recent years. Specifically, in Colombia, agriculture is one of the most influential sectors, contributing 6.8 % of GDP, Figure 1. Therefore, the control and characterization of the soil quality are some of the most important challenges in agriculture. A correct characterization of the land increases the efficiency of crops and increases the yield and quality of agricultural products. It directly affects the farmers since it allows them to obtain greater retribution at the moment of its commercialization. Several factors influence agriculture, both external such as climate, irrigation, and internal such as fertility, humidity, and physicochemical components, among others. Among these characteristics of the soil, the most important are the physicochemical components, such as soil organic carbon (SOC), nitrogen (N), potassium (K), phosphorus (P), and pH. Finally, among these chemical components, the most influential in the optimal growth and fertility of the soil is the percentage of SOC Berhe et al. (2022). The main reason is that the rate of SOC is directly proportional to characteristics such as moisture, fertility, and porosity Wang et al. (2022), which determines the quality of agricultural products.

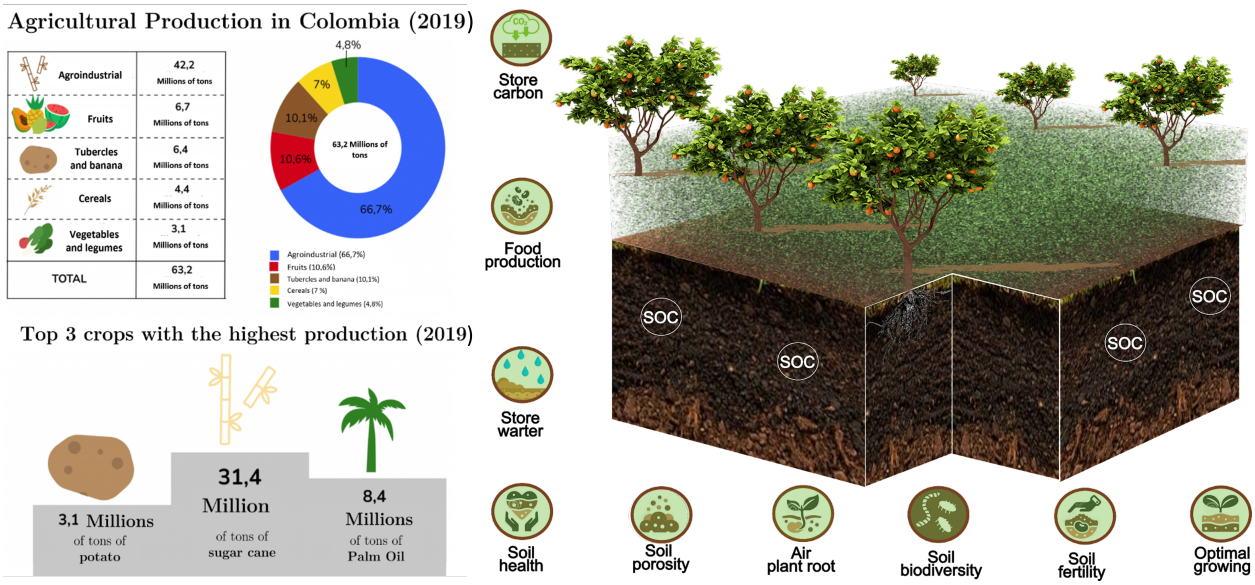


Figure 1. Influence of the agricultural sector in Colombia and the importance of SOC characterization and control. Image taken and modified from Gaget (2021).

Although the opportunity to characterize soils is valuable, it is important to highlight that this terrain analysis entails additional challenges and considerations. The task involves a significant technological investment in terms of image acquisition and processing, given that the need to analyze land extensions spanning several kilometers is being addressed. This process not only requires advanced hardware and software but also entails high computational costs due to the complexity, the amount of data involved in large-scale image processing, and the specialized personnel needed for the maintenance and inspection of these technologies. It is essential to recognize these aspects to fully understand the magnitude of the technology and resource investment required to successfully carry out soil characterization.

2.1. Traditional methods for SOC percentage estimation

Nowadays, several chemical methods have been developed to measure essential physico-chemical characteristics such as SOC. These traditional methods, calculate the physicochemical content of various molecules, such as SOC Nocita et al. (2015); Smith et al. (2019); Zhong et al. (2021), nitrogen (N) Jiang et al. (2017), calcium (CA), and even electrical characteristics of the soil such as cation exchange (CIC), among others.

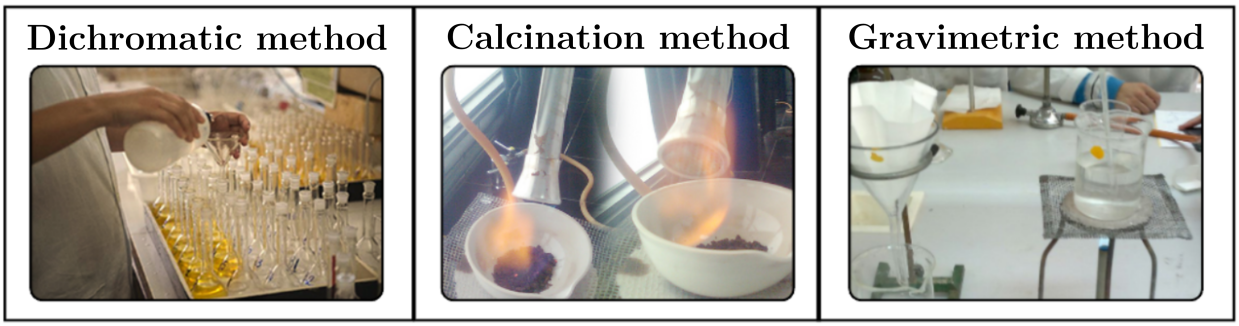


Figure 2. Traditional methods for the estimation of physicochemical components Izquierdo Bautista and Arevalo Hernandez (2021); Garcia Galvis and Ballesteros Gonzalez (2005).

2.1.1. Calcination Method. Calcination methods performs the calculation of various physicochemical properties through the ignition of the soil sample, to subsequently make a comparison of the weight lost in this process, Figure 2 Izquierdo Bautista and Arevalo Hernandez (2021), noting that the temperatures at which the combustion of each molecule occurs are already characterized and properly labeled.

2.1.2. Walkley & Black Method (Dichromate). Some methods used for the estimation of physicochemical properties rely on chemical reactions aimed at the liberation of CO_2 after combustion at high temperatures, thus allowing the determination of the total carbon present

in the soil. Among the most commonly used conventional methods in laboratories is wet oxidation, also known as the Walkley and Black method Garcia Galvis and Ballesteros Gonzalez (2005), or titrimetric determination Amacher et al. (1986), Figure 2. This technique is widely used due to its minimal requirements, as it does not demand sophisticated equipment and is less costly and more precise compared to other conventional methods. In particular, the Walkley and Black method involves the use of chemicals such as potassium dichromate and sulfuric acid. Although this technique offers multiple advantages, it is important to note that it is not environmentally friendly, as the reagents used have the potential to be pollutants, and soil samples are usually completely destroyed and discarded after the analysis is completed.

2.1.3. Gravimetric Method. Another widely used method is the loss-on-ignition (LOI) or gravimetric method, where organic matter is oxidized by heating to 375 °C or higher, depending on the soil type Yanosky and Macintosh (2001). The estimation is done by measuring the gravimetric mass loss of the organic matter relative to its initial weight, as shown in Figure 2 . However, its precision will depend on the type of furnace, the duration of the process, the ignition temperature, and the type of soil being analyzed, as this technique is not effective with all types of soils.

To perform these physicochemical analyses, soil samples must be directly collected in the field and then transported to specialized soil chemistry laboratories for processing Bojago et al. (2023); Özbolat et al. (2023). These laboratories have highly qualified personnel and specific devices, which reveal some main characteristics; for instance, they allow obtaining highly reliable results regarding the physicochemical characterization of the sample since they offer a sensitivity

of 0.01 %, at the cost of around $200.000 (COP), but the response times of these physicochemical analyses are high, around 1 month, due to the need of transferring the soil samples from the field to the laboratories, and to the low existence of these soil chemistry laboratories, which represents a high demand for these specialized centers. Therefore, traditional estimation methods are available to obtain a physicochemical analysis result with high sensitivity and precision, although with high response times. This is inefficient in practice, as farmers need this vital information on soil composition as soon as possible to apply methodologies to optimize crop growth and quality.

2.2. Spectral information

Spectral information, specifically spectral signatures, offers numerous advantages and applications where it can achieve performance comparable or even better than traditional chemical methods for estimating soil physicochemical properties. Spectral methods are non-destructive, allowing for rapid, in-situ analysis without extensive sample preparation. They can provide continuous data over large areas, making them highly suitable for precision agriculture and environmental monitoring. Additionally, spectral methods can simultaneously measure multiple soil properties, whereas chemical methods are often time-consuming and limited to specific analyses, as shown in Figure 3.

Spectral information provides crucial insights into soil's physical, chemical, and biological properties, such as mineral composition, organic matter content, SOC content, and texture Vairavan et al. (2024). Scientifically, this works by analyzing the light reflected or emitted by soil across various wavelengths in the electromagnetic spectrum. These methods typically require a large number of spectral signatures to achieve accurate estimations. In such contexts, the abundan-

ce of data points allows for better calibration and more reliable results. However, in conditions with limited spectral data, it becomes more challenging to obtain SOC estimation results comparable to traditional chemical methods due to the reduced ability to capture soil variability and complexity. Thus, implementing a computational algorithm that can achieve precise SOC estimation with few spectral signatures is necessary for improving the performance of these methods in less data-rich environments.

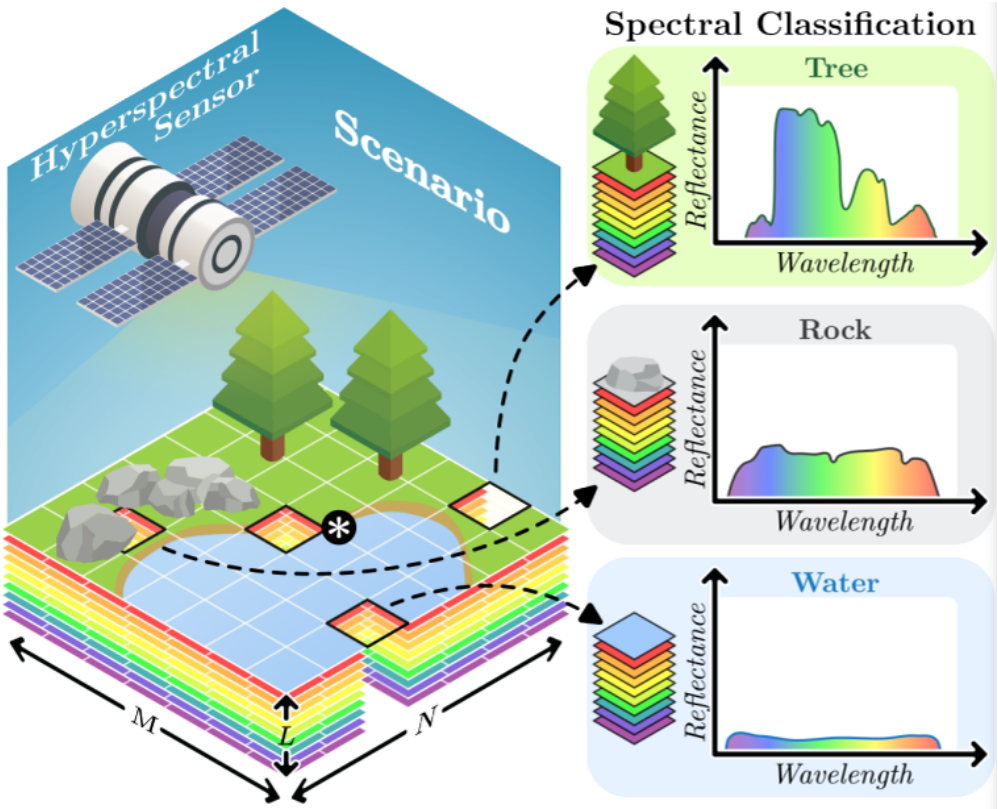


Figure 3. Spectral information and its difference depending on the type of material, taken from Bacca et al. (2023)

2.3. Spectroscopy Analysis

Spectroscopy serves as a powerful analytical tool for soil characterization, leveraging distinct electromagnetic spectra—UV (ultraviolet), VIS (visible), and NIR (near-infrared) each offering unique insights into soil composition, as shown in Figure 4. UV spectroscopy is instrumental in identifying specific organic and inorganic compounds in soils by exploiting high-energy photons that induce electronic transitions in molecules Baumann et al. (2021); Zhao et al. (2021). Its capability to discern chemical bonds is advantageous, but UV spectroscopy's effectiveness can be hindered by the presence of interfering organic matter, requiring careful sample preparation, and analysis conditions.

VIS spectroscopy operates within the visible range, allowing for precise identification of soil pigments and minerals based on their absorption and reflection properties. This method provides excellent spectral resolution, making it ideal for distinguishing among different chemical species in the visible spectrum. However, VIS spectroscopy may struggle to detect more complex organic compounds and is sensitive to variations in sample granularity and preparation techniques Caten et al. (2016).

2.3.1. Spectroscopy on NIR. Particularly, NIR spectroscopy has emerged and developed rapidly in recent decades as an essential method for the analysis of samples in agriculture. These NIR spectral signatures permit the analysis of the composition of the soil samples, such as SOC in Figure 5. Spectroscopy on the NIR is based on the principle that organic carbon percentage information is mainly concentrated in the NIR spectral range Angelopoulou et al. (2019).

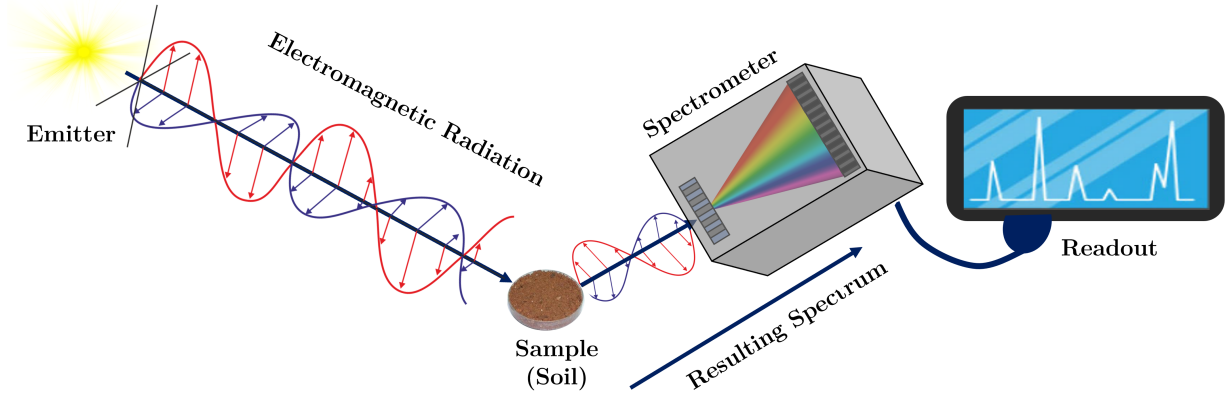


Figure 4. Acquisition of spectral signatures by electromagnetic spectral radiation in a soil sample.

From these spectral signatures in the NIR, optical systems have been developed for acquisition, such as airborne platforms Yang et al. (2021), on satellites Angelopoulou et al. (2019), or locally portable systems Hutengs et al. (2019); Seema et al. (2020). These systems use the concepts of reflectance and absorbance of light in soil samples to obtain a wide range of spectral signatures in a fast, simple, and field-applicable manner. Additionally, a spectral signature is acquired through sensors that capture the continuous signal and discretize it into a certain number of bands in a specific spectral range. Various sensors have been developed that are constructed from molecules that vary depending on the spectral range being used. Specifically in the NIR, specialized sensors have been developed due to the nature of the spectrum. Among the most commonly used molecules are InGaAs, and HgCdTe. The principle of operation of these sensors is based on the conversion of photons to electrons, as it is done in an analogous way in the visible sensors (CMOS). Consequently, these systems are isolated from undesirable external elements such as stray light and noise, among others, to increase their repeatability and the reproducibility of the capture process in controlled, external locations such as in the field Maiwald et al. (2022); Chabrillat et al. (2019);

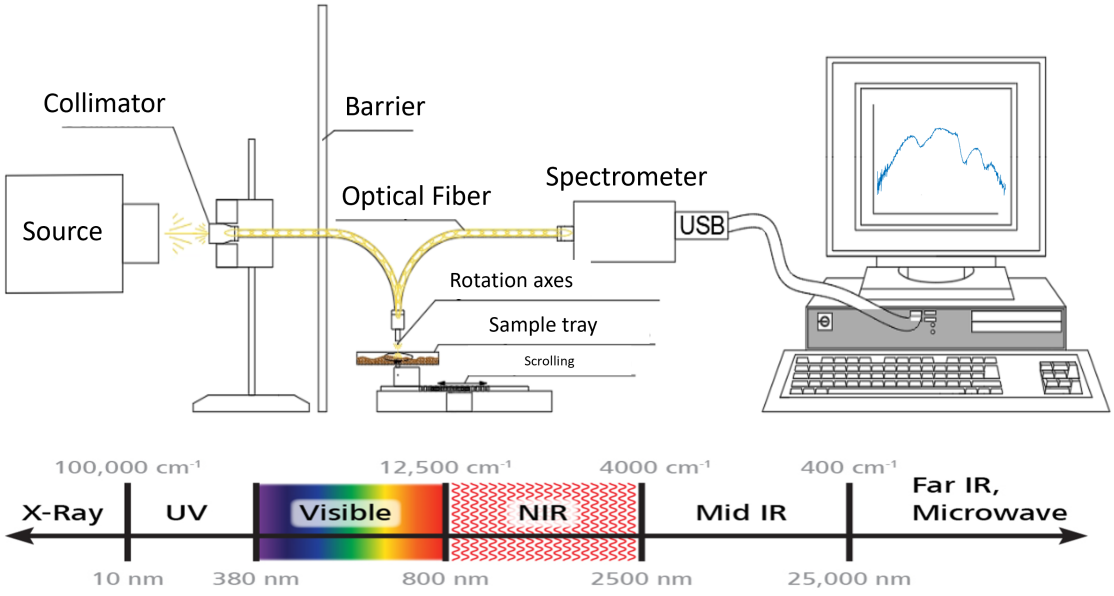


Figure 5. Optoelectronic system of acquisition of spectral signature setup, in the near infrared (NIR) spectral range.

Ben Dor et al. (2015).

2.4. Spectral information acquisition methodology

There are various methods of data acquisition in soil spectral analysis, each offering different advantages depending on the application. These methods may provide better spatial or spectral resolution, enabling more detailed insights into soil properties. Figure 6 shows the different optical systems to acquire these SI through the combination of dispersive elements such as prisms, 2D detectors, and color filters. These methods vary in their ability to capture spectral information, with some focusing on high-resolution spatial data while others emphasize spectral detail.

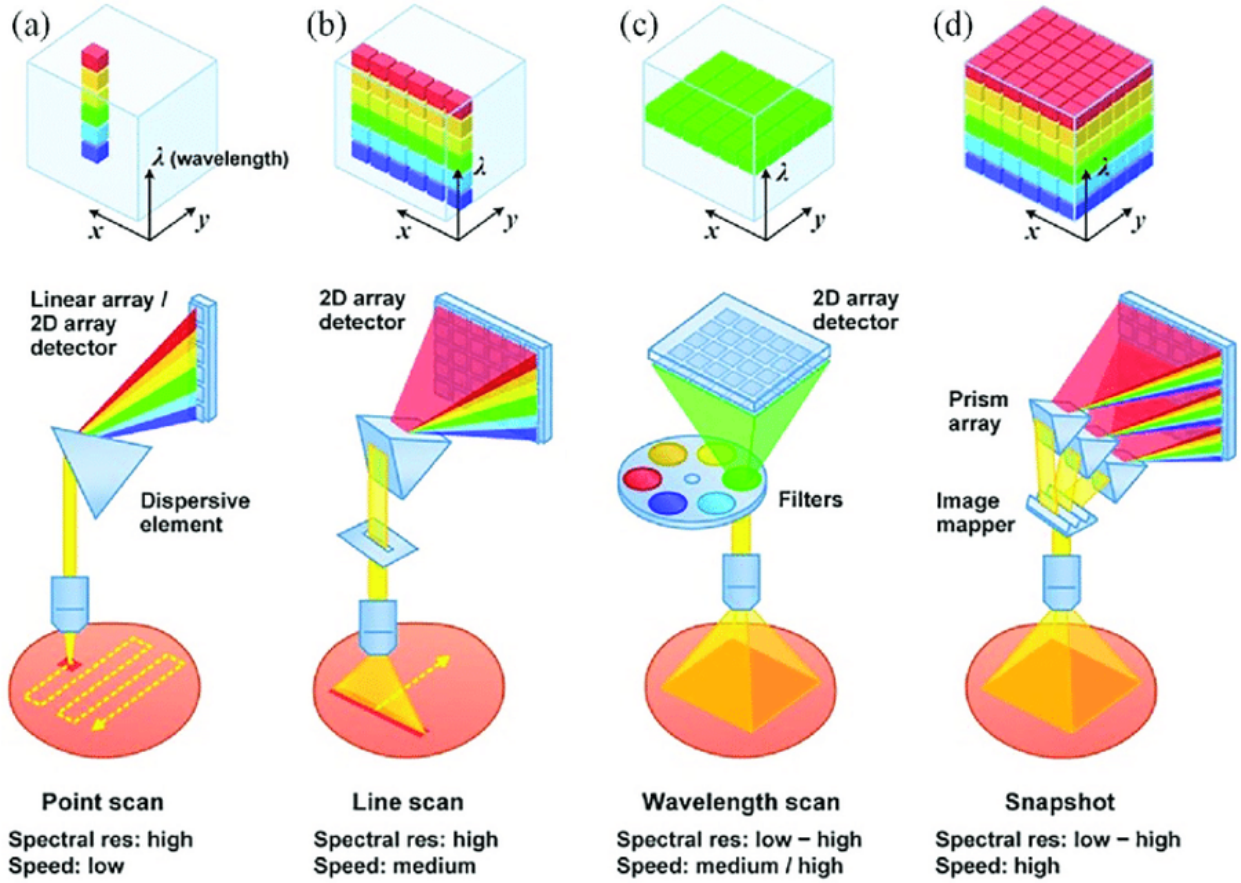


Figure 6. Point scanning (A), line scanning (B), wavelength scanning (C), and single capture (D) methods for SI. Adapted from Wang et al. (2017)

Specifically, systems include those based on scanning and wide-field imaging. The first one, line scanning, shown in Figure 6B, sweeps the scene using a line-shaped monochromatic laser to move along an *x* or *y* dimension and capture the wavelengths of multiple points simultaneously. The aforementioned method achieve high spectral resolution per line, allowing accurate demultiplexing and quantification of the image samples, but exhibit low acquisition speed.

From the difficulties previously exposed for scanning methods, another type of system was developed for SI capture known as wide-field, which uses an illumination of the region and detects

the light emitted from a group of pixels within an area using a 2D matrix in the detector, as seen in Figure 6C and 6D. In the first case of wavelength scanning Gat (2000), a two-dimensional image acquisition is performed by sequentially capturing a full λ wavelength band of the scene using color filters. On the other hand the single capture method Arce et al. (2013) uses multiple prisms to capture both spatial and spectral information in a single shot. On the other hand, point scanning method Vane et al. (1993) is performed through a focused laser beam in the form of a point. For the point scan, shown in Figure 6A, a point-shaped monochromatic laser is used to capture the wavelengths along the x and y spatial dimensions of the scene. Compared to the previously mentioned methods, Whiskbroom spectral acquisition offers higher spectral resolution and greater repeatability. However, this acquisition type suffers significant shortcomings in automation, stability, and repeatability of the acquisition process. Therefore, it is essential to implement a system designed to obtain stable spectral signatures of soil samples in the NIR.

2.5. Computational algorithms for soil organic carbon percentage estimation

Various computational algorithms have been developed for estimating SOC using NIR spectroscopy. Based on mathematical modeling and computational algorithms, it has been demonstrated that it is possible to estimate the percentage of SOC employing a large amount of spectral signatures in the NIR. These computational algorithms have verified that it is possible to perform the extraction of intrinsic characteristics of the spectral signatures through changes of representation bases, denoising, and smoothing Yeh et al. (2023); Caten et al. (2016). Additionally, to verify the estimation quality of these computational algorithms, several metrics have been developed in the state-of-the-art. Among the most widely used is the r-squared (r^2) metric, since it allows us to

analyze quantitatively the behavior of the regression model, as well as to efficiently conclude how well this regression fits the real data. This metric is defined as follows

$$r^2 = \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}. \tag{1}$$

where N is the number of estimated values, y_i is the percentage of real carbon, $\hat{y}_i$ is the percentage of carbon estimated by the regression model and $\bar{y}$ is the mean of all values. Furthermore, another metric widely used in the state of the art to evaluate and analyze regression models is known as the mean square error (MSE), which allows measuring the mean square error of the predictions, as shown in the equation (2)

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2. \tag{2}$$

Based on these metrics, various algorithms have been used to estimate SOC, which can be divided into two main groups: traditional algorithms and data-driven algorithms.

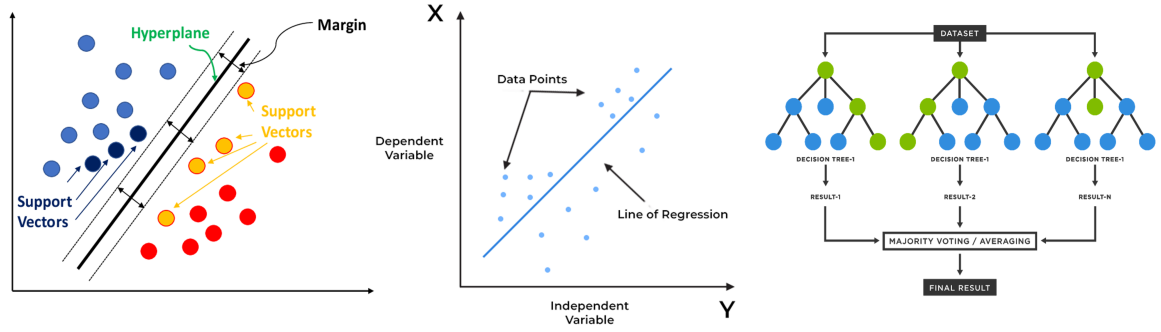


Figure 7. Traditional state-of-the-art SOC percentage estimation algorithms, cited from (a) SVM Xu et al. (2021), (b) RF Santana et al. (2017) and (c) Linear Regression Ramirez et al. (2021)

2.5.1. Traditional Algorithms. In the state-of-the-art, several computational algorithms have been developed and implemented for the estimation of SOC percentage from NIR spectral signatures, which have shown a good performance of 0.7 on r^2 metric, comparable to the values offered by chemical methods Shi et al. (2014). These estimation algorithms are based on a group of processing, which allows highlighting intrinsic characteristics of the signature and making predictions based on specific features and phenomena of spectral signatures. The most commonly used processing methods are spectral signature transformations, changes in representation bases, normalization, standardization, smoothing, and denoising Kinoshita et al. (2016). Several algorithms have been implemented from this type of processing applied to spectral signatures through supervised learning to allow the extraction of spectral signature characteristics. Through this process, spectral information estimates the percentage of SOC based on previously labeled spectral values and signatures. Among the widely used supervised algorithms there is linear regression (Figure 7 (c)), which, from the input data, searches for a linear equation that best describes the correlation between the independent and dependent variables. On the other hand, the Random Forest (RF), Figure 7 (b), an algorithm is also a supervised learning technique, which generates several decision trees on the training data set, finally obtaining a robust model based on a series of previously trained decisions. Additionally, the Support Vector Machine (SVM), Figure 7 (a), the algorithm performs a correlation analysis of the data in a high-dimensional feature space so that the data points can be characterized, even when they cannot be linearly separated. Also, the partial least squares regression (PLSR) method Shi et al. (2014), has been merged with principal com-

ponent analysis (PCA) Sato et al. (2014), which is based on rotating the coordinate system of the initial variables to new orthogonal axes, ensuring that this new representation plane coincides with the direction of maximum variance of the input data. Finally, the most recent automated learning methods such as the single-layer perceptron (SLP) and multiple-layer perceptron (MLP), work as a type of interconnected layers of neurons, highlighting that the output of the neurons of one layer become inputs to the next layer, allowing the estimation of SOC percentage since it is possible that the relationship between the spectral signature and the SOC percentage is not exactly linear.

2.5.2. Deep computational algorithms for SOC estimation. Recently, based on the rise of artificial intelligence, several neural network architectures have been developed that allow the estimation of SOC percentage from large amounts of NIR spectral signatures. These neural networks are based on the principle of unsupervised learning, which resembles the functioning of the human brain. Neural networks start from a large amount of data, where intrinsic feature extraction is performed in different domains or representation bases. The main feature of these architectures is that they act as a black box, which is fed with a large set of data, in our case, NIR spectral signatures and the neural network is responsible for learning the best way to estimate the percentage of SOC, either by modifying the spectral signature through range selection, transformations, denoising, smoothing, among others Zhong et al. (2021); Zhang et al. (2022). Among these SOC estimation neural networks, the best known are the recurrent neural networks (RNN), which allow the extraction of features to estimate the SOC through a large dataset of spectral signatures. Additionally, convolutional neural networks (CNN), Figure 8 (a), have been implemented, which consist of an input layer, an output layer, and several hidden layers. These layers perform

operations that modify the spectral signatures to learn their particular characteristics. These kinds of architectures have been widely used in various computational tasks due to their robustness in extracting features regardless of the complexity of the information used. More recently, the Very Deep Convolutional Networks for Large Scale Image Recognition (VGG) were proposed, Figure 8 (b), which are based on a group of blocks composed of an incremental number of convolutional layers, together with the interleaving of max-pooling blocks between convolutions. Although this type of VGG Zhong et al. (2021) architecture has been designed mainly for image estimation and classification tasks, the state-of-the-art has modified this architecture to work with spectral signatures, obtaining promising results concerning SOC estimation.

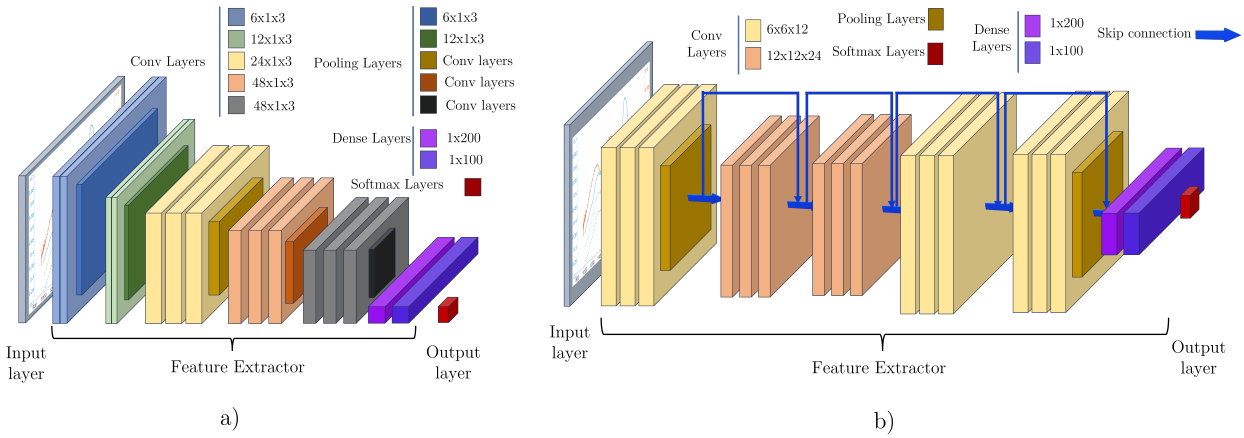


Figure 8. State-of-the-art deep computational algorithms, which allow the estimation of SOC. a) VGG neural network from the literature for feature extraction, b) ResNet neural network architecture.

2.5.3. Performance of computational algorithms based on mount of data. Finally, it is necessary to highlight that the accuracy of these methods is directly proportional to the amount of data that exists along with the computational complexity, Figure 9, therefore, it is one of

the most relevant characteristics to take into account, since the training times and the performance of the results obtained will have a great variability and a direct correlation with the amount of data, Figure 9, which exceeds approximately 1.000 spectral signatures Ng et al. (2020).

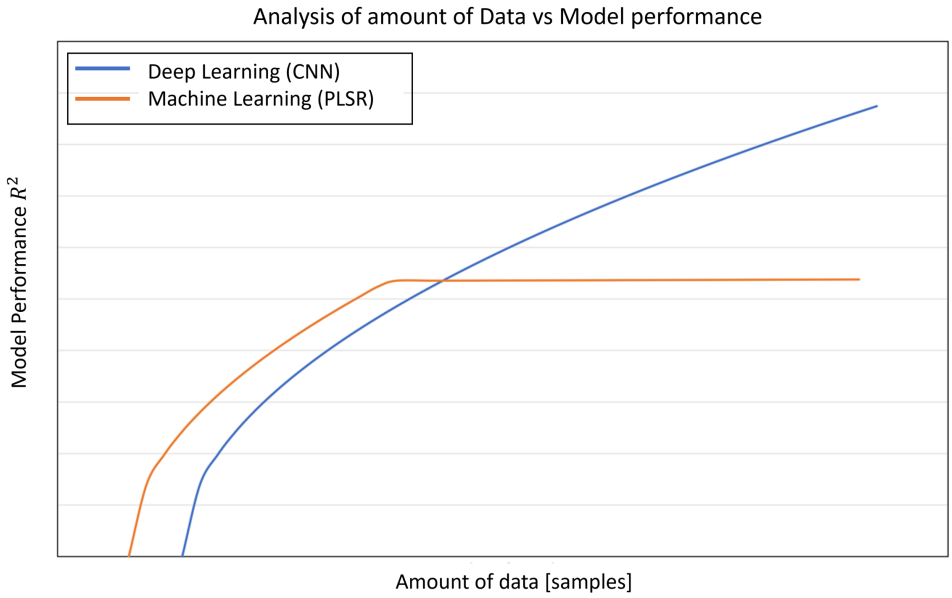


Figure 9. Correlation analysis of performance of computational algorithms vs. the amount of data consumed, cited from the Ng et al. (2020)

For this reason, it is necessary to adapt SOC estimation algorithms to achieve precision comparable to both traditional chemical methods and data-driven estimation algorithms. In various applications, particularly in Colombia, which has diverse thermal zones, spectral signatures can vary significantly depending on the altitude, and climate of the region where they are acquired. These variations are due to the differences in the temperature, humidity, and vegetation across different places. As a result, algorithms used in these applications must provide optimal results

even with limited data, as gathering large datasets from every thermal zone is often impractical.

3. Optoelectronic System for NIR Spectral Signature Acquisition

This section presents a design, implementation, and acquisition protocol of a whiskbroom-type optical system that enables controlled acquisition of spectral signatures from soil samples to estimate the percentage of SOC. Specifically, we initially present the theoretical modeling of a Whiskbroom system, along with each stage of the prototype construction for acquisition developed throughout the project, highlighting key characteristics of each stage.

3.1. Mathematical Modeling of Whiskbroom-Type Optical System

To achieve proper design and implementation of a whiskbroom-type optical system, it is essential to conduct mathematical modeling to simulate the system's operation. Therefore, a detailed mathematical modeling of the whiskbroom-type optical system is required. This modeling enables the simulation of the system's operation and comprehension of its behavior under various conditions. Some aspects to consider in the modeling include the system optics, interaction with soil samples, the lighting source, utilized detectors, and mechanical components for scanning. Specifically, this acquisition method is mathematically modeled as follows

$$y(x,\lambda) = \gamma_{\text{fiber}}(x,\lambda)\gamma_{\text{lens}}(x,\lambda)\mathbf{f}(x,\lambda). \tag{3}$$

In this expression, $y(x,\lambda)$ represents the intensity of light detected by the spectrometer at position x and wavelength λ. The term $\mathbf{f}(x,\lambda)$ denotes the intensity of light originating from the scene containing soil samples at position x and wavelength λ. Additionally, $\gamma_{\text{lens}}(x,\lambda)$ characteri-

zes the transfer function of the collimating lens, while $\gamma_{\text{fiber}}(x, \lambda)$ represents the transfer function describing the propagation of light through the optical fiber. These functions collectively describe the interactions between the optical components and the light as it passes through the system, providing a comprehensive understanding of the system's optical behavior.

3.2. Selection of commercial optoelectronic components

3.2.1. Optical commercial components. In the design process of a spectral information acquisition system, meticulous consideration of several pivotal characteristics is imperative. These include, but are not limited to, the determination of the spectral range to be employed, the requisite sensors for spectral signature acquisition, the spectral resolution of the sensor, and the spectral regions necessary for SOC estimation.

Of paramount importance in the design of a Whiskbroom system is the selection of appropriate illumination sources. Within the market, various illumination types are available, each characterized by distinct spectral efficiencies and stability profiles. Among the prevalent illumination sources are fluorescent, Light light-emitting diode (LED), halogen, and tungsten variants. It is worth noting that while LED and fluorescent lighting solutions exhibit suboptimal spectral responses within the NIR range and often demonstrate limited stability across this spectrum, halogen % tungsten illuminations are favored choices for implementing optical systems targeting the Visible and Near-Infrared (VNIR) range.

Furthermore, meticulous consideration of the aforementioned factors, particularly the selection of illumination sources and the method of scene displacement, is pivotal in the design and optimization of Whiskbroom-type spectral acquisition systems. Such considerations ensure the

attainment of reliable and high-quality spectral data critical for a diverse array of applications, including SOC estimation in soil science and environmental monitoring endeavors.

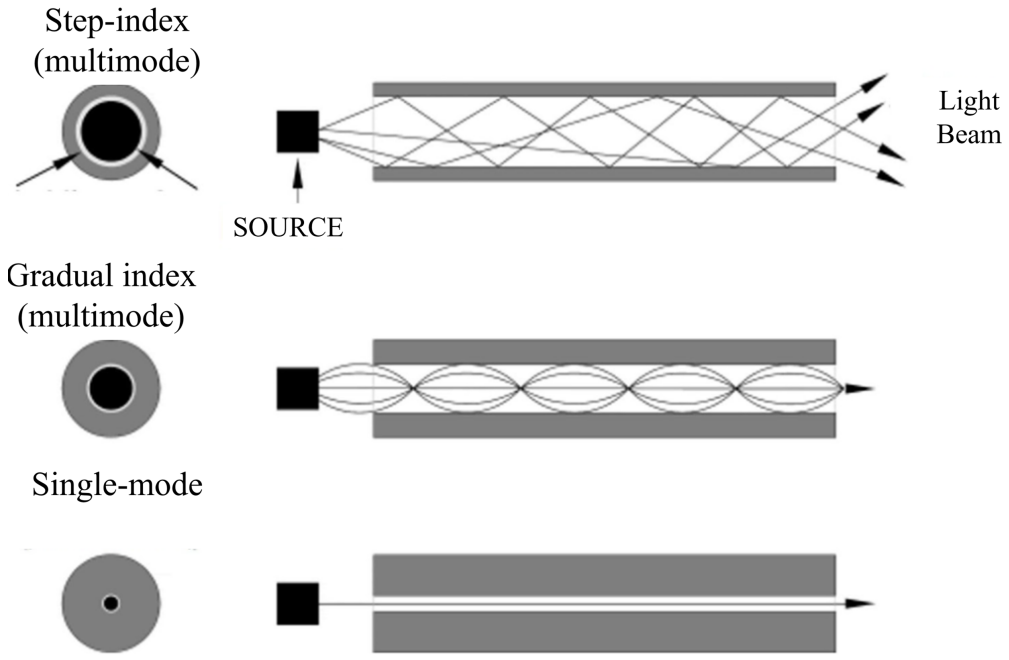


Figure 10. Types of optical fibers: single-mode and multimode fibers, taken from Tian et al. (2024).

Additionally, the Whiskbroom spectral acquisition architecture can be constructed in a basic form using a light source illuminating the scene, a collimating lens responsible for reducing the scene size to a smaller area, an optical fiber receiving the collimated light rays, and a device dispersing the light for subsequent intensity measurement via a detector.

Within the acquisition system, the optical fiber serves as a mechanism for transmitting data in the form of light pulses. In optical fibers, the core consists of extremely thin fibers of pure glass, surrounded by a coating with a refractive index comparable to that of glass. Moreover, it serves as an optimal transmission mechanism due to its low susceptibility to external electromag-

netic interference, as well as low losses due to dispersion. As depicted in Figure 10, depending on the mode of light propagation, optical fibers are classified as either single-mode or multimode fibers. In single-mode fibers, light propagates through total internal reflection with lower attenuation compared to multimode fibers. However, the main disadvantage of this type of fiber is the core size, which in many cases has a diameter close to 8 [μm], requiring greater precision in the angle of incidence of light rays along the optical path. On the other hand, multimode fibers have core diameters ranging from 50 to 60 [μm].

Next, it is necessary to define the optical components of the instrument that disperses light. Typically, this device consists of an array of collimating lenses and mirrors, a diffraction grating or prism, and a photodetector. This optical configuration is known as a spectrometer. A spectrometer enables the measurement of the interaction between electromagnetic radiation and matter as a function of wavelength. Depending on the optical elements comprising it, the spectrometer can acquire spectral information in different regions of the electromagnetic spectrum such as ultraviolet, visible, and/or infrared.

It is important to highlight that the selection of the diffraction grating in the spectrometer determines the spectral resolution and allows for the acquisition of the wavelengths of interest. This is because it separates the incident polychromatic light into its constituent wavelengths. Therefore, the diffraction grating is one of the main optical elements in the point acquisition methodology to measure the spectral response associated with SOC in a soil sample.

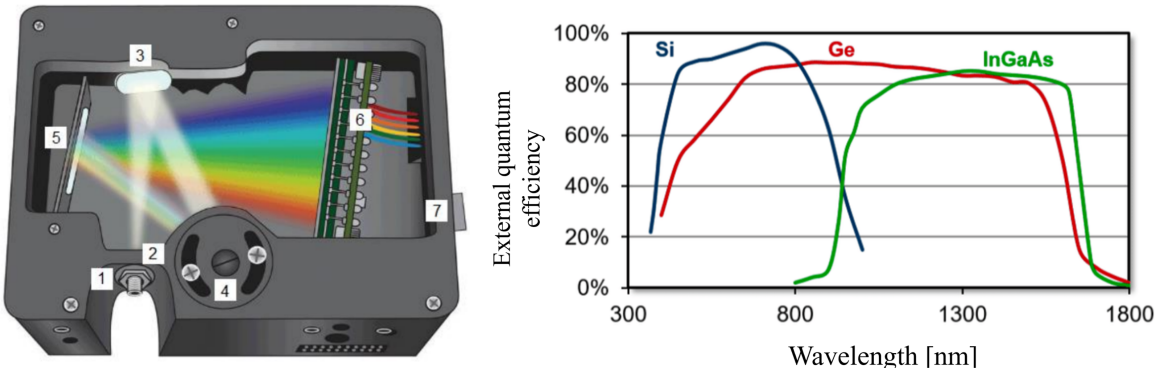
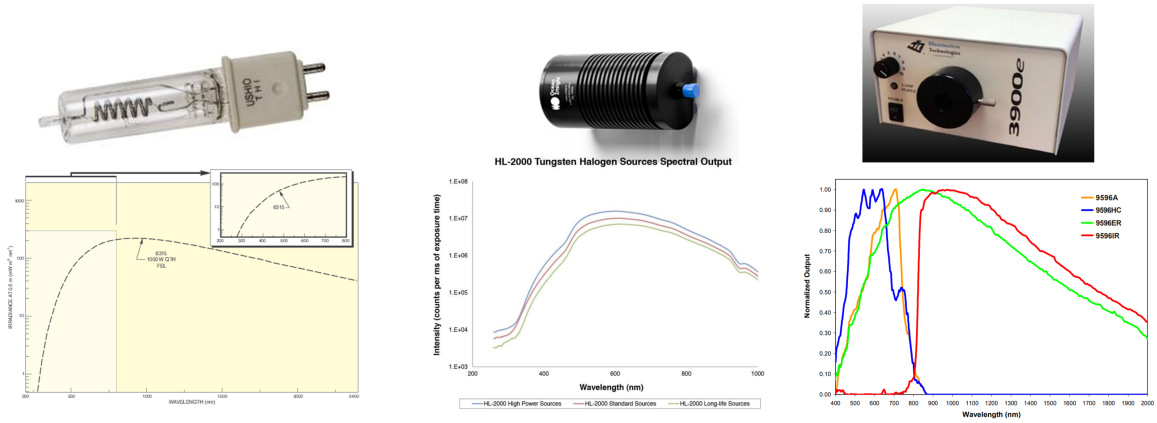


Figure 11. Components of a spectrometer typically include. 1) Light input. 2) Fixed entrance slit. 3) Collimating mirror to capture all the light. 4) Diffraction grating. 5) Focusing mirror. 6) Spectral information detector. 7) USB port for data transfer. And external quantum efficiency of three photodetectors fabricated with Silicon (Si), Germanium (Ge), and an alloy of Gallium Arsenide and Indium Arsenide (InGaAs).

Regarding the sensor used for spectral information acquisition, various photoelectric detection devices are employed to measure the spectral response resulting in the interaction between the light source and the soil sample. One of the most critical parameters of the sensors to be used is the material from which they are constructed, as this parameter determines the spectral region in which they can operate. While sensors constructed from Silicon and Germanium typically exhibit high quantum efficiency in the visible spectral region (400 [nm] - 1000 [nm]), detectors fabricated from Indium-Gallium-Arsenide (InGaAs) alloys allow spectral information acquisition in the NIR region (900 [nm] to 2.500 [nm]), Figure 11. In summary, to design an optimal Whiskbroom acquisition system, it is necessary to consider specifications such as spatial resolution, which is determined by the resolution of the optomechanical displacement elements, and spectral resolution, which is directly correlated with optical components such as the collimator, fibers, spectrometer,

and the spectral range of the illumination used.

The first element to be selected is the lighting source. The differentiating factors between each lighting source are cost, power, and spectral efficiency, as shown in Figure 12. From this market lighting sources analysis, it is necessary to highlight that there are various types of lighting, including LED and fluorescent, which were discarded from the outset because they have a non-constant spectral response across the spectral range of interest, 900-2500 [nm]. On the other hand, the type of lighting that meets the requirement is those designed with halogen or tungsten lamps. Specifically, various illuminations with the indicated spectral range were found in the market, as seen in Figure 12. However, as mentioned earlier, the power of the lighting sources is an important criterion to consider. Therefore, the lighting source in Figure 12, section b), is discarded since its power is much lower than the other two. On the other hand, while lighting source a) meets the spectral response and power entirely, there is a risk for system operators as its efficiency range extends to the UV, which is dangerous as it can cause serious skin diseases due to exposure. That said, the 3900E source from Illumination Technologies is selected because it has spectral efficiency in the range of interest, which is 900-2500 [nm], and the electrical power of this source is sufficient to illuminate the soil sample and acquire spectral signatures.



a) Tungsteno + Halogen (200 – 2400 [nm])

b) Tungsteno + Halogen (360 – 2400 [nm])

c) Halogen (400 – 2200 [nm])

Figure 12. Types of lighting sources and luminous efficiency according to wavelength, a) Tungsten-halogen lamp with spectral response from UV to NIR, b) Tungsten-halogen lamp with spectral response from UV to NIR, c) Tungsten-halogen lamp with spectral response from VIS to SWIR. Taken from Piccini et al. (2024); Illumination Technologies (2024)

Additionally, since it is necessary to condense the illumination reflected by the soil sample when illuminated by the aforementioned lamp, a condenser lens must be selected. In the market, there are various collimators classified into spectral ranges in which they are optimized, as shown in Figure 13. Specifically, for the implementation of the Whiskbroom system, it is essential to highlight the main characteristic of having high spectral efficiency in the NIR range, in our case from 900-2500 [nm]. Figure 13 shows a group of collimators that meet this main characteristic.

Figure 13. Types of collimators based on spectral efficiency, taken from Umeda et al. (2017)

After the collimator, the light reflected by the soil sample propagates along the optical fibers, which are separated based on their spectral efficiency. For the implemented system, a set of optical fibers efficient in the NIR range is required. Additionally, due to the Whiskbroom design, bifurcated fibers are necessary. Therefore, the selected optical fibers are the QBIF1000-NIR-BX, as they have a spectral efficiency of 900-2.500 [nm], Figure 14 c).

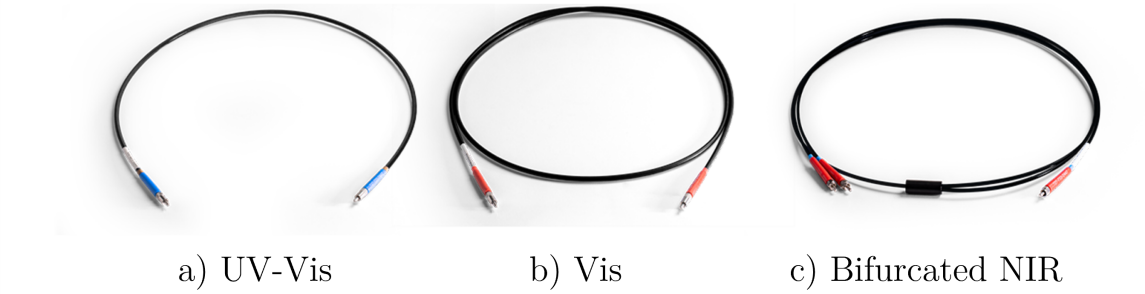


Figure 14. Optical fiber types according to spectral efficiency: a) UV-Vis optimized fiber, b) Vis optimized fiber, c) Bifurcated NIR fiber, taken from Mickelson (2018)

On the other hand, concerning the spectrometers for acquiring spectral information, they are again divided depending on the spectral range. Specifically, in the market, the differentiating

parameters of spectrometers are resolution and the spectral range for which the device was built, as shown in Figure 15. Taking this into account, the spectrometer chosen for the implementation of the Whiskbroom system is the NirQUEST reference Hamamatsu G9208-512W InGaAs linear array from Ocean Optics. The main reason is its acquisition range, which is 900-2500 [nm]. It is noteworthy that this spectrometer is the right choice because its spectral resolution is 3.5 [nm], meeting the spectral resolution requirements, and the range, where SOC information is concentrated in the state of the art. Highlighting that, in this case, a bifurcated optical fiber was used. This fiber features a core that guides light from the illumination source, which propagates internally towards the sample and exits at a predetermined distance. The same exit path is used to acquire the reflection of the spectral information, which is then conveyed to the sensor. This specific optical fiber was selected because its structure ensures that the tip illuminating the soil sample provides homogeneous illumination and optimal collection. This is achieved by having a single-mode core for illumination at the center, surrounded radially by a set number of modes that capture the spectral information reflected from the soil sample.

Figure 15. Types of spectrometers available commercially, depending on the application and spectral range, taken from Oliveira et al. (2024). a) Spectrally efficient spectrometer for the UV-VIS range, b) Spectrally efficient spectrometer for the VIS-NIR range, c) Spectrally efficient spectrometer for the SWIR range

3.2.2. Electromechanical commercial elements. Moreover, successful implementation of the Whiskbroom acquisition methodology necessitates precise and repeatable displacement of the soil sample. Contemporary approaches typically involve two primary methods: angular and linear displacement. Angular displacement entails rotating the soil sample in a circular trajectory, enabling spectral information acquisition at various angular positions with respect to a fixed radius. This mechanism further allows for radial sweeping of the soil sample, thereby facilitating comprehensive scene coverage. Conversely, linear scene displacement methodologies involve the orchestrated movement of the scene to predefined (x, y) positions through a series of actuators. This approach offers precise control over the spatial positioning of spectral acquisitions, directly influencing the spatial resolution of the acquired data. The precision of actuator movement directly correlates with the spatial resolution achievable during scene scanning.

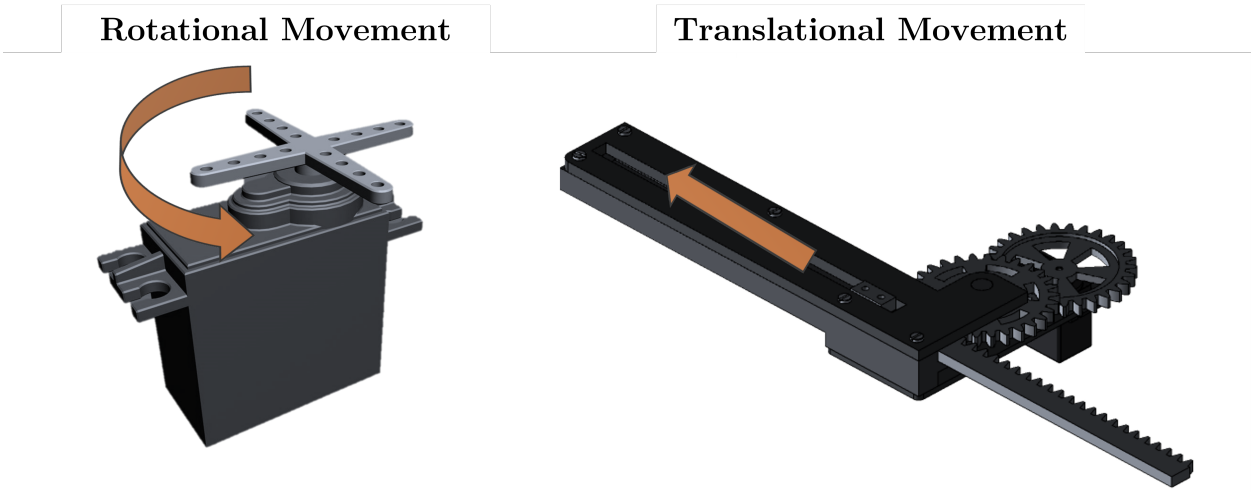


Figure 16. Spatial scene scanning methodologies using various actuators. Left: Rotational or angular scanning. Right: Translational scanning.

For this purpose, there are various types of motors available commercially, classified primarily by power consumption, torque, rotation range, and step resolution, as shown in Figure 17.

a) Stepper motor b) Servomotor

Figure 17. Types of low-power consumption motors: a) Stepper motor, b) Servomotor.

After selecting the electromechanical elements available in the market to implement the Whiskbroom acquisition system, construction begins. For optimal implementation of the acquisition system, it is necessary to design a structure in which all elements are positioned precisely to ensure the repeatability of soil spectral signature acquisition over time. Based on the point acquisition Whiskbroom type, this structure needs to be designed in a way that it is fully isolated from external light sources where spectral information acquisition takes place. Additionally, another requirement for the acquisition system structure is that it allows for the automation of the acquisition protocol, minimizing user intervention and ensuring the reliability of the acquired spectral signatures.

3.3. Implementation of cartesian scanning Whiskbroom prototype

To achieve controlled acquisition of spectral signatures, a two-stage optomechanical system was constructed preliminary. This system facilitates the controlled acquisition of spectral signatures from a soil sample contained within a Petri dish, employing a point-by-point scanning method across Cartesian and rotational displacements. This setup is depicted in Figure 18.

Figure 18. Prototype 1: Construction and implementation of an automated spectral acquisition system.

The first acquisition prototype was built based on a set of basic requirements. The first of these is the necessity, due to the Whiskbroom acquisition methodology, to keep the acquisition stage completely dark and to automate the process of acquiring spectral signatures by reducing technician and user interaction through Cartesian displacement. For this purpose, the acquisition mechanical system was initially designed, consisting of 2 arms and an acquisition base. The basic operation is that each of the arms has a servomotor, which is positioned at 90° where they join, to perform soil sample movement through a set of 3D-printed rails and acquire the largest possible number of spectral signatures, specifically 1.520 spectral signatures. It is important to highlight that the decision was made to move or rotate the soil sample while keeping the optical fibers fixed, as the calibration of the distances between the optical fiber and the soil sample significantly influences the spectral response obtained. Additionally, although the mechanical system was built to sweep Cartesian-ally, acquisition times increased to 2-3 hours per soil sample, which is unfea-

sible. Therefore, it was decided to implement an additional improvement with a servomotor in the acquisition base to increase the acquisition speed by rotating the soil sample, reducing acquisition times to 1 hour for 1.520 spectral signatures.

On the other hand, to address the requirement of eliminating external light, an enclosure was proposed in which the distribution of the optomechanical acquisition system is located. It is worth noting that the dimensions of the prototype were affected by the size of the displacement arms for acquisition, resulting in an enclosure with dimensions of $45 \times 50 \times 60$ [cm]. The optical acquisition elements, such as the NirQuest spectrometer, the Arduino control board, and the lighting system with the Illumination Technology 3900E source, are located in the upper part. This illuminates a group of bifurcated fibers with a collimator, fixing the distances and focusing direction to maintain experiment repeatability. On the other hand, in the lower compartment, the mechanical acquisition system is located, consisting of the 2 displacement arms, 3 servomotors, the soil sample acquisition base, and the positioning base for the spectralon.

Figure 19. The variability in the acquisition of spectral signatures obtained from Prototype 1.

Although the designed and implemented optoelectronic system allows for the semi-automated acquisition of spectral signatures from soil samples, the variability and repeatability of the acquisition process are affected, as observed in Figure 19. Additionally, one of the most notable shortcomings is the acquisition times, which amount to 1 hour. Therefore, it is possible to improve automation and reduce acquisition times through a better optoelectronic system design.

3.4. Design and implementation of polar scanning Whiskbroom prototype

Given the shortcomings of the Cartesian scanning acquisition system, an optomechanical polar system was developed, which allows for the acquisition of fewer spectral signatures in less time while maintaining the stability and repeatability of the acquisition process. It is important to note that user interaction occurs only at one stage. To reduce acquisition times, the spectral

correlation of the 1.520 spectral signatures acquired by the previously implemented system from a soil sample was analyzed using the Spectral Angle Mapper (SAM) metric. This metric allows for the calculation of the degree of difference or spectral correlation between the signatures. The results, as observed in Figure 20, highlight that the SAM metric value does not exceed 0.06°. Therefore, it is possible to conclude that a reliable group of spectral signatures can be obtained with a smaller number of spectral signatures.

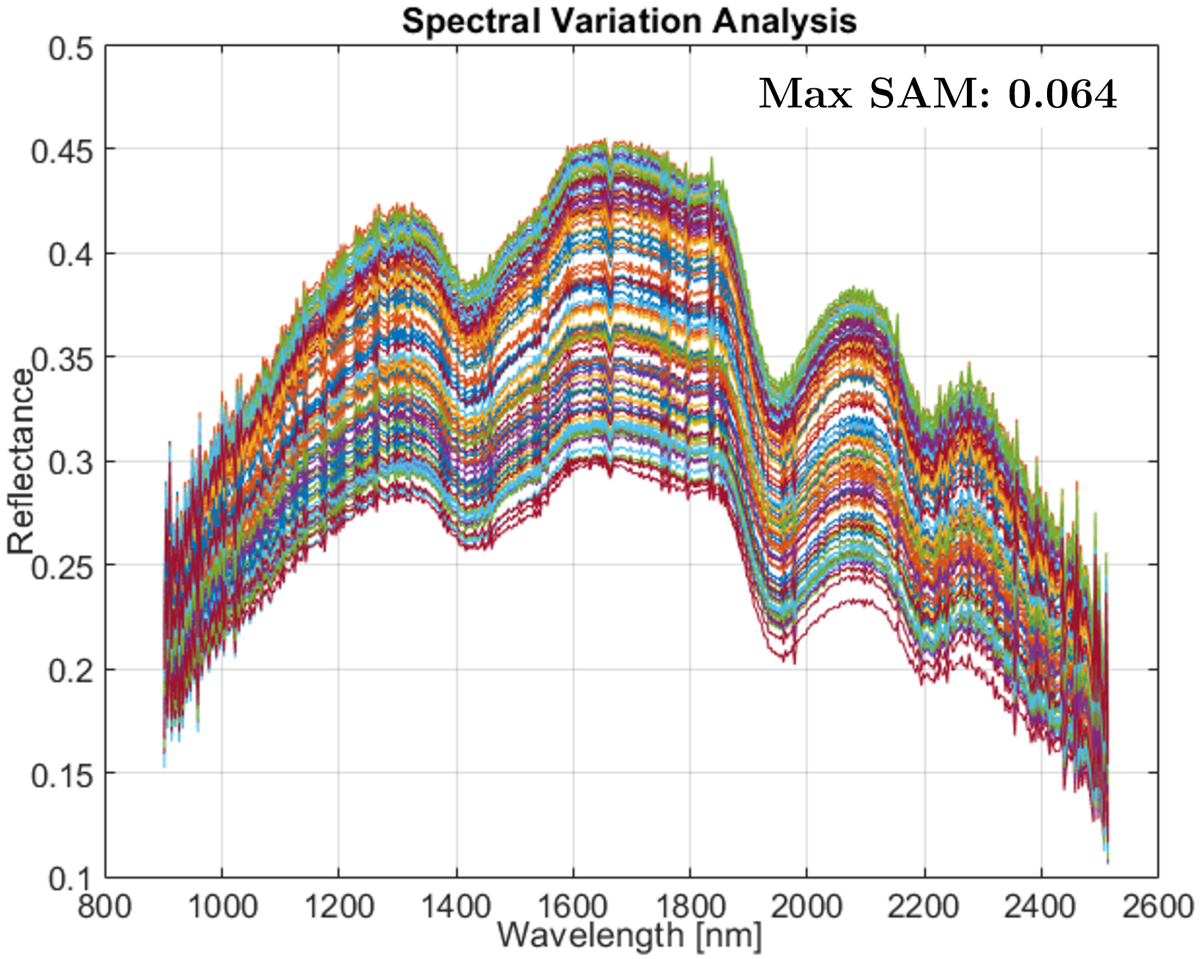


Figure 20. Spectral variability analysis of 1520 acquired spectral signatures.

Additionally, to further reduce acquisition times, it is proposed to change the type of Cartesian spatial scanning performed on the soil sample to a polar scan. To carry out the acquisition through polar scanning, the pairs of servomotors responsible for spatial movement (x, y) were replaced with a stepper motor to position the soil sample at a specific radius. Furthermore, by using the servomotor located beneath the soil sample, it is possible to precisely position the acquisition point (R,θ). The resulting optical system can be seen in Figure 21.

Figure 21. Design, Distribution, and Implementation of a Polar Spectral Acquisition System

Additionally, to reduce the dimensions of the optoelectronic system, a new structure for the enclosure was designed and constructed, as shown in Figure 21, which consists of three specialized compartments, as depicted in Figure 21. In the first stage, the optoelectronic devices, such as the NIR spectrometer and the Arduino control board, are located. In the upper compartment, the bifurcated optical fiber and the light source are positioned, highlighting that the materials used can withstand the high temperatures generated by the light source. Lastly, there is the stage where the automated polar acquisition process takes place. This compartment has the stepper motor, the white reference, the servomotor, the fixed position of the bifurcated optical fiber, the soil sample, and the displacement rail. From the polar spectral acquisition system, a comparative analysis of deviation and repeatability in the acquisition process was conducted, comparing it with the previously implemented system, as shown in Figure 22. This analysis concluded that the standard deviation of the polar system, compared to the Cartesian system, is lower, as shown in Figure 22.

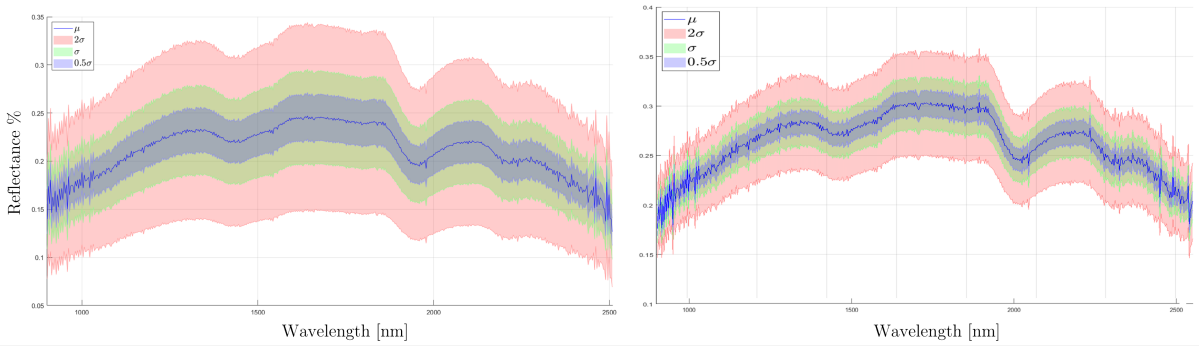


Figure 22. Comparison of results of spectral signature acquisition by polar and cartesian scanning. Analysis of standard deviation in the process of spectral signature acquisition by different scanning methods.

Figure 23. Improved Polar Scanning Spectral Acquisition System with Enhanced Enclosure Materials and Mechanical Component Construction Materials.

Finally, to enhance the quality of the implemented optical system, the construction material was changed from wood to acrylic, which withstands the high temperatures generated by the light source. Additionally, the construction of the mechanical elements was improved by replacing the used metal displacement rail with a custom-designed displacement rail and base, as shown in Figure 23. To confirm the stability, portability, and resistance to external environments outside the laboratory, the prototype for acquiring polar spectral signatures was tested in the field. Specifically, the prototype was transported and deployed on various sidewalks in the municipality of Simacota, Santander, where samples from citrus fruit crops were collected, and in-field acquisition was performed. It is noteworthy that the prototype's performance was satisfactory; despite being

in a high-temperature environment, it was possible to conduct the polar scanning of the sample and complete the acquisition. This resulted in the development of a portable acquisition prototype designed to address a key research objective: overcoming the challenges of transportation and the extended response times associated with traditional chemical methods for SOC estimation. The optical system developed enables in-situ spectral data acquisition with high repeatability.

3.5. Development of an Automated spectral acquisition protocol

To maintain the rigor of spectral acquisition experiments over time and ensure the reliability of the acquired spectral signatures, it is essential to develop an acquisition protocol that not only incorporates the optical system but also ensures temporal efficiency, meaning that it facilitates rapid data acquisition. For this reason, this work presents the protocols found in the literature regarding the acquisition of spectral signatures in soil samples and proposes an acquisition protocol comparing its reduction in variability over time and repetitive acquisition processes.

3.5.1. State of the Art Acquisition Protocol. In the document by Ben Dor et al. (2015) a standard protocol for spectroscopic measurements of soil in the laboratory is proposed, highlighting the importance of controlling both systemic and non-systemic factors that affect the quality of spectral signatures. This protocol suggests methods for stabilizing instrument components and adequately preparing samples, as well as meticulously recording conditions such as measurement geometry and the environmental setting. It recommends using direct contact probes on the sample or optical fiber in a dark box environment to avoid the heterogeneity introduced by surfaces such as the glass in Petri dishes. On the other hand, in the literature, a process for acquiring spectral signatures using a spectroradiometer has been described, from illuminating the samples

with artificial light to processing the data to reconstruct the characteristic spectral signature, including the preparation and arrangement of samples in the lab and sample collection in the field. Both approaches emphasize the need for careful handling and rigorous control to obtain reliable results in soil spectroscopy.

Various protocols have been developed in the literature for spectral data acquisition. For example, spectral data were acquired using the Fieldspec 3 spectroradiometer. All spectral signatures were collected following the same protocol: the spectral sensor was positioned 8 [*cm*] above the sample surface, scanning an area of approximately 2 [*cm*] and illuminated by two external 50W halogen lamps that provided the light source for the scene. These lamps were positioned 35 [cm] from the sample (non-collimated rays and a zenith angle of 30°C) with a 90°C angle between them. A standard Spectralon white plate was scanned every 20 minutes during the scans. For each trial, two replicates were obtained for each sample by rotating the Petri dish 180°. Each spectrum was averaged from 100 readings over 10 seconds. The spectral reflectance was transformed through continuum removal, a preprocessing step that eliminates continuous features from the spectra and is often used to isolate specific absorption features. Once the captured spectral data were corrected, spectral analysis of the samples was performed, starting with procedures such as principal component analysis (PCA) to reduce the dimensionality of the spectra, which also allows for the inclusion of stages for selecting relevant variables for the samples and the design and development of models for the spectral estimation of soil properties.

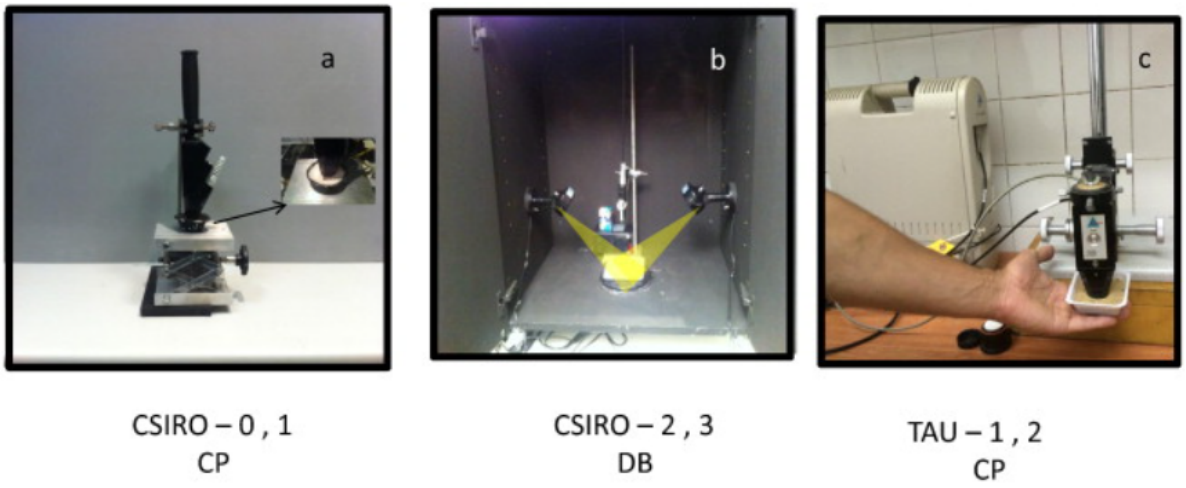


Figure 24. State-of-the-Art Protocols, Taken from Ben Dor et al. (2015).a) Protocol for spectral signature acquisition under normal lighting conditions, b) Protocol for spectral signature acquisition under controlled lighting conditions, c) Manual protocol for spectral signature acquisition under normal lighting conditions.

Additionally, as part of the efforts to standardize the acquisition of spectral signatures, protocols from various institutions, including the Czech University of Life Sciences Prague, used a contact spectral probe to perform soil measurements in five different agricultural areas in the Czech Republic Ben Dor et al. (2015). The measurement setup, an ASD spectrometer, and the light source were preheated for 30 minutes to mitigate noise caused by temperature changes in these electronic components. Soil samples were placed on 9 [cm] diameter Petri dishes, forming a 2 [cm] thick soil layer, prepared to minimize reflectance from the dish bottom and other external light sources (solar radiation) that could alter the characteristic spectral signature of the soil sample. The samples were leveled with a stainless steel blade to ensure a flat surface flush with the top of the Petri dish, as a smooth surface in soil spectral analysis ensures a high signal-to-noise ratio (SNR). All spectral readings were taken in the center of the samples (three replicas each) in a dark room to prevent

interference from contaminating light. The authors made a radiometric correction was performed using a white reference before the first scan and after every six measurements to ensure the accuracy of the collected data, this decision was taken after a statistical analysis of the variability of the optical system during the acquisition process.

3.5.2. Proposed automation protocol for the acquisition of spectral information.

To overcome these limitations and reduce operator interaction during spectral acquisition experiments, an automated protocol using the previously built Whiskbroom-type optical system was developed. The automation of the proposed protocol ensures that the optical elements interact automatically by moving parts with the support of a developed computational tool, allowing the acquisition of spectral signatures from each sample with minimal variation.

Specifically, Figure 25 shows the schematic of the spectral acquisition, grouped into its main stages (blue box). On the right is the corresponding automation for the development of each stage (green box), detailing the interaction of the mechanical part (servomotors) and the programmable control board with the parameterization and spectral acquisition interface responsible for managing the proposed automated spectral acquisition protocol. Specifically, the calibration process involves selecting the spectral acquisition parameters such as acquisition time, the number of spectral signatures per spatial point, and the type of spectrometer to be used to obtain reference spectral signatures. Additionally, the operator must inspect and confirm that the reference spectral signatures of the spectralon and black are not saturated and are within the normal range, [50.000-60.000] counts for white and [0-2.000] counts for black, respectively. If any anomalous values arise, the acquisition protocol allows for parameter adjustments to ensure optimal spectral

signature acquisition.

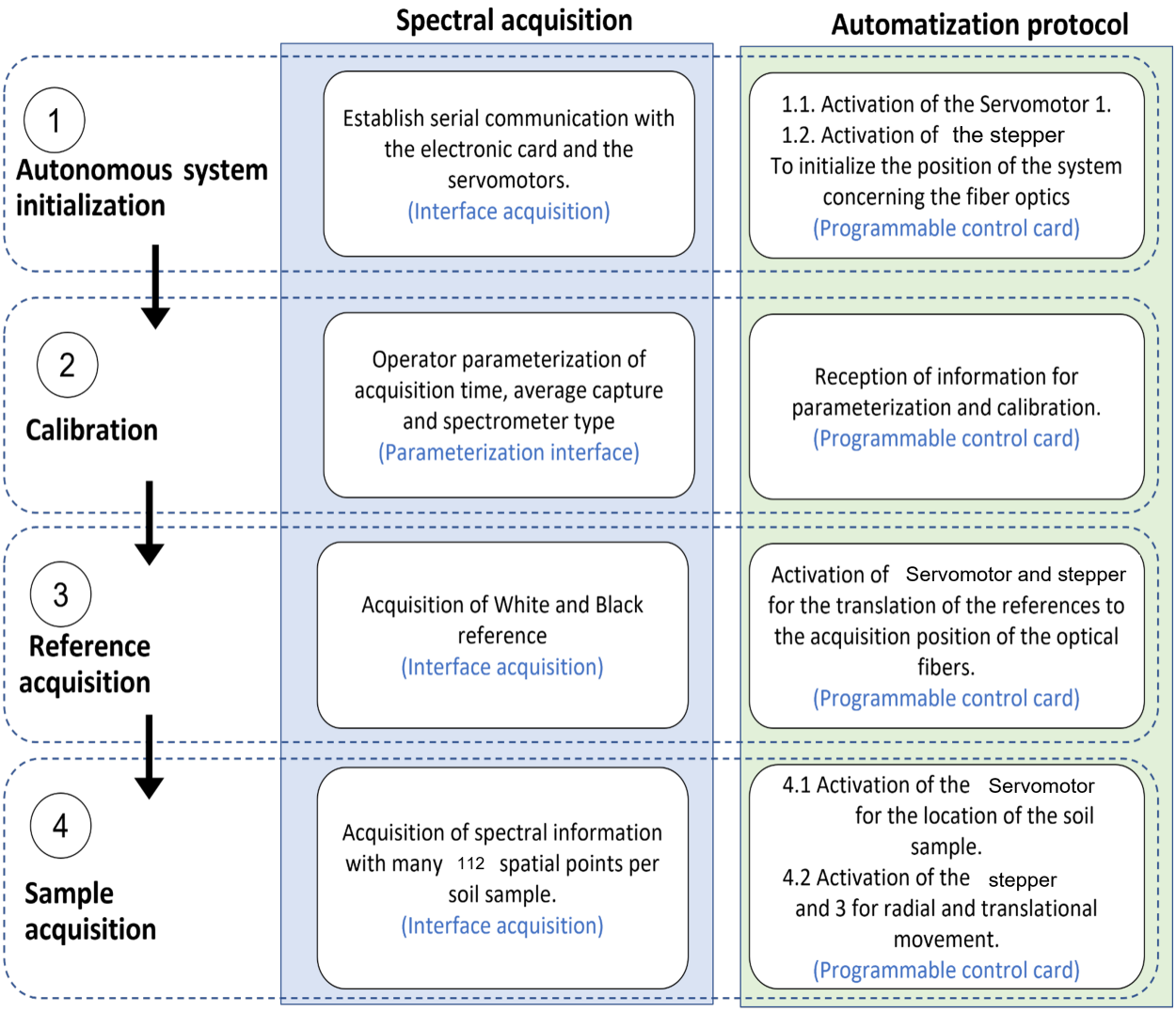


Figure 25. Proposed spectral acquisition protocol, including control of mechanical elements and information acquisition.

The implementation of the automated protocol for capturing spectral signatures of soil samples is divided into two main groups: the computational interface or tool (parameterization and acquisition) and the hardware automation that enables the spectral information acquisition process. The synchronization of these two areas generates an automated system for capturing spectral

signatures.

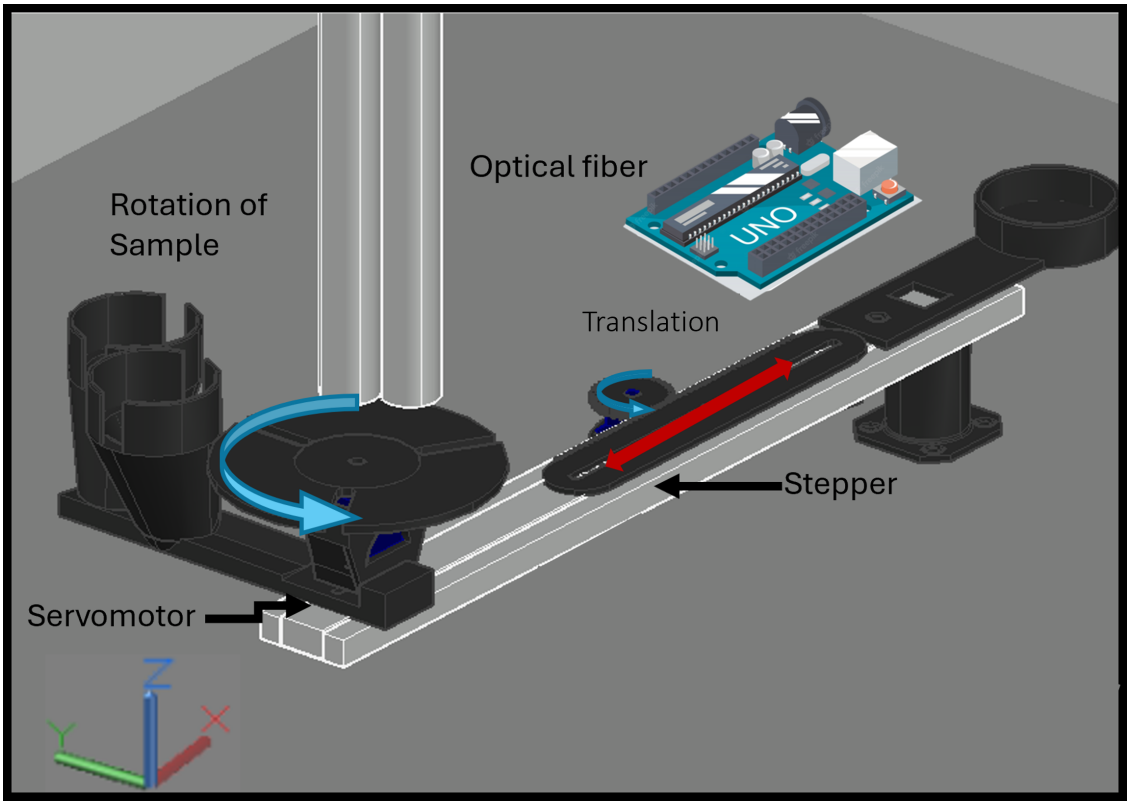


Figure 26. Optomechanical distribution of system acquisition.

This work designed and built an automated protocol for capturing spectral signatures, reducing operator interference in the acquisition process since the operator only intervenes in placing the sample and the Spectralon in the sample holders before starting the spectral acquisitions. The interactions of the components designed and implemented by 3D printing reduce construction costs. Automation allows 32 degrees of freedom (z,r,θ) with a range of 5 movements that enable the spectral measurement acquisition of the white reference (Spectralon) and the soil sample spectral signatures. All mechanical movements are controlled by an Arduino board, which has a serial connection with the developed interface. The designed interface allows parameterizing the inte-

gration time and specifying the number of spectral signatures to be averaged, enabling automatic initiation of the first spectral acquisition on the target reference. This is done using the translational movement achieved from the stepper motor, allowing the acquisition base to be moved to a perpendicular position relative to the optical fibers. After this, through the bifurcated optical fibers, which allow the light from the illumination source to be emitted and receive the reflection of the signal from the reference or the soil sample returned to the detectors (NIR spectrometer) using the same fiber.

Note that the fibers are always in a fixed position, and the automated system moves the reference and the soil sample to acquire the spectral measurement. To capture the black reference, the optical system is completely sealed, and the illumination from the light source is removed, allowing the acquisition of the black reference. The information obtained from the black-and-white references is saved for the post-processing stage, enabling the calculation of reflectance and/or absorbance. Subsequently, the process of acquiring spectral signatures on the soil samples begins with the operation of the servomotor and stepper motor, which move and position the soil sample perpendicularly to the optical fibers at a predetermined distance of 2 [cm]; after this, the spectral acquisition process on the soil sample is automatically initiated.

With the proposed scheme, it is possible to cover the distribution of acquisition points in a polar manner, allowing the acquisition of 112 spectral signatures, distributed over 28 angles and 4 different radii. Specifically, after each integration time, 5 spectral signatures are acquired in one position (r, θ); then, the sample is rotated at an angle of 12.85° using the servomotor, placing a new spatial point under the optical fibers for another 5 captures. Then, using the stepper motor, the

sample is moved approximately 0.5 [cm], and the process is repeated, achieving the acquisition of 112 spectral signatures per experiment for each soil sample and two spectral signatures of the references (white and black). As the proposed design of the entire spectral acquisition protocol is automated, interferences or errors generated by the laboratory technician's interaction at the time of acquisition are reduced. The capture time is optimized by acquiring a larger number of spectral signatures in less time and with less variation, improving the repeatability and reproducibility of each experiment; in Figure 26, a descriptive diagram of the mechanical system is presented.

3.5.3. Development of a computational tool for the management of the spectral information acquisition automation protocol. To offer a simple and friendly interaction with the user, a computational tool has been developed in Matlab consisting of an interface *Acquisition and Parameterization*. Fig. 27 illustrates an interface image for parameterization and acquisition of spectral signatures of samples. This interface has several areas of work; among these highlights is the possibility of using different spectrometers for the development of the capture experiment. It is also possible to parameterize the integration time since if you use a longer integration time, more light enters the sensor and can saturate the sensors. This interface also allows the acquisition of the black-and-white reference, which enables the calculation of reflectance, according to the equation (4).

The interface also allows the management of the Arduino control board to drive the rotors and rails that perform the sample movements in an automated way. The developed visualization interface allows the observation in real-time of the spectral signatures that are being captured (see Fig. 27. All the information of each experiment and sample, such as the references of the spec-

trometers used, the distance between the fibers and the sample, and sample characteristics such as color and texture, are stored as metadata (see Fig. 27). The metadata is used in the experiment's subsequent pre-processing and processing stages. The equation (4) is used to check the reflectance of each soil sample.

$$R = \frac{I - B}{W - B}, \tag{4}$$

where R is the reflectance of the spectral signature obtained, I is the intensity of the raw spectral signature of the sample to be studied, W is the white reference, and B is the black reference.

To confirm the stability, portability, and resistance to external environments outside the laboratory, the prototype for acquiring polar spectral signatures was tested in the field. Specifically, the prototype was transported and deployed on various sidewalks in the municipality of Simacota, Santander, where samples from citrus fruit crops were collected, and in-field acquisition was performed. It is noteworthy that the prototype's performance was satisfactory; despite being in a high-temperature environment, it was possible to conduct the polar scanning of the sample and complete the acquisition.

3.5.4. Acquisition Protocol Stability Results. The established protocol for spectral signature acquisition ensures a consistent, reliable, and reproducible process for obtaining spectral data from soil samples. By minimizing user interaction, the protocol enhances stability and control over the acquisition process, thereby ensuring the fidelity of the acquired spectral signatures. This approach not only promotes efficiency but also mitigates potential sources of va-

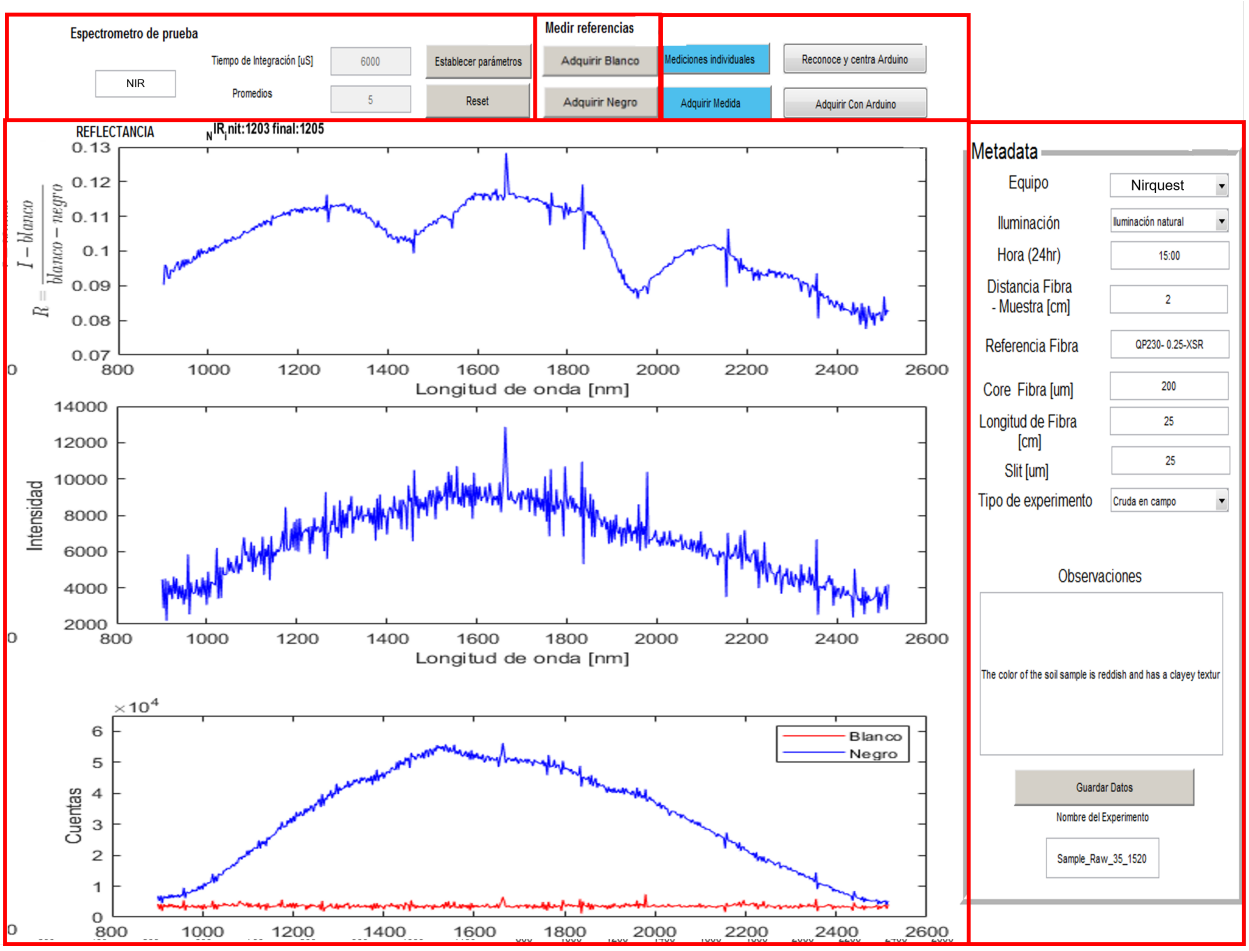


Figure 27. Interface developed for the acquisition and parameterization of spectral signatures in the NIR range.

riability, leading to a higher level of confidence in obtaining the spectral data. Consequently, the protocol represents a robust framework for acquiring spectral signatures with minimal interference and maximal reliability.

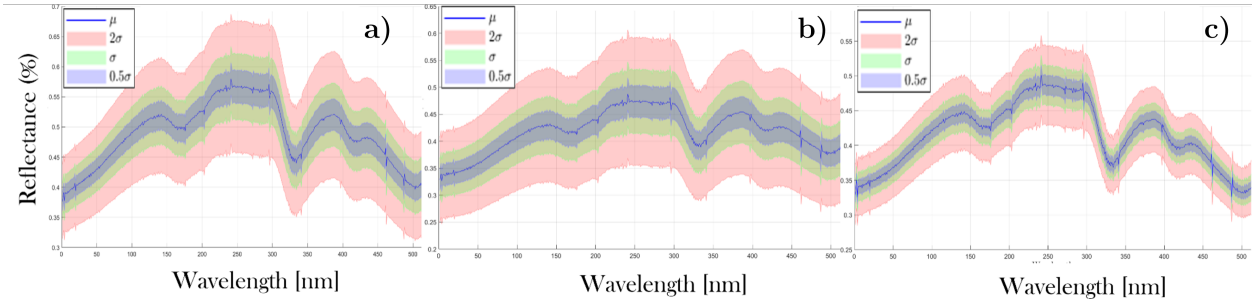


Figure 28. Variance obtained from the experiments performed on the different prototypes and the final protocol. In (a), the spectral variability of the Cartesian optical system is shown, which captures 1,520 spectral signatures. In (b), the spectral variability of the Polar V1 optical system is observed, also capturing 1,520 spectral signatures. In (c), the spectral variability of the Polar V2 optical system is depicted, which captures 112 spectral signatures..

Through the experiments carried out, both in the initial prototypes and in the proposed protocol, the following results were obtained, Fig. 28, where the variance of the reflectance of the spectral signatures obtained from the NIR spectrometer is observed, exposed from a box diagram, showing in this way the variability of the samples in the same experiment, the atypical data and the average of the same. In the results acquired by the NIR spectrometer, initially from the first prototype, significant variability is obtained between signatures acquired in the same experiment, with a standard deviation of 0.06 of reflectance, highlighting that the value of the mean in comparison with the variance is visibly unbalanced, which is attributed to the low value of repeatability that this prototype has. Likewise, the values obtained by the second prototype have values of standard deviation of approximately 0.05 of reflectance, but compared with the previous prototype, the mean

Metrics	**Protocol**			
	Cartesian		**Polar**	**Proposed**
Acquisition Time [min]	120	12	20	**10**
Spectral Signatures	1520	112	112	**112**
σ	0.06	0.034	0.0557	**0.0286**

Table 1. Results of ablation analysis of different implemented protocols

improves its stability and tendency. Finally, the variability of the proposed prototype is presented in the table below, and it is observed that its variance decreases significantly, with a maximum value of the standard deviation of 0.028, even when the number of acquired signatures is reduced in the cartesian scanning system. The uncertainty throughout the experiment remained constant, thus concluding that the proposed protocol has a better response in terms of variability during the process of capturing spectral signatures due to automation.

Additionally, a quantitative analysis is performed on the capture time and the number of spectral signatures captured in each experiment implemented and the proposed system. It can be seen that the proposed system dramatically exceeds prototype #1 due to a large number of samples and the reduction of the capture time. Likewise, for experiment #2, the acquisition time is longer than the proposed method due to the complete automation of the capture process. Therefore, it is observed that the proposed system has excellent stability throughout the soil spectral signature acquisition experiment, together with the low cost due to the fact that it was designed and built from 3D printing parts.

This chapter concludes with the design, and implementation of an automated optoelectronic

system that allows the acquisition of spectral signatures from a soil sample in the NIR using a Whiskbroom-type scanning methodology with a specific protocol of acquisition.

4. Characterization and acquisition of spectral signature dataset

In this chapter, employing the optoelectronic system and protocol acquisition implemented in Chapter 3, a description and characterization of the terrain and the soil sample extraction protocol were conducted, as well as the protocol for extracting soil samples from crops and the type of crop, among other characteristics, were analyzed. Emphasizing the distribution of the SOC percentages obtained at the end, which are of vital importance for the training, validation, and testing of the estimation algorithms.

4.1. Sample soil collection treatment

This research was carried out in Simacota, Santander, Colombia, an area noted for its mountainous terrain and extensive citrus farms. The geographic location of the study area, specifically at the WGS84 coordinates (6° 26 N; 73° 21 W), as shown in Figure 29. In this region, soil samples were systematically collected based on specific criteria like topography, land usage, and existing vegetation. Eight evenly spaced soil samples were obtained after pinpointing the suitable extraction zones.

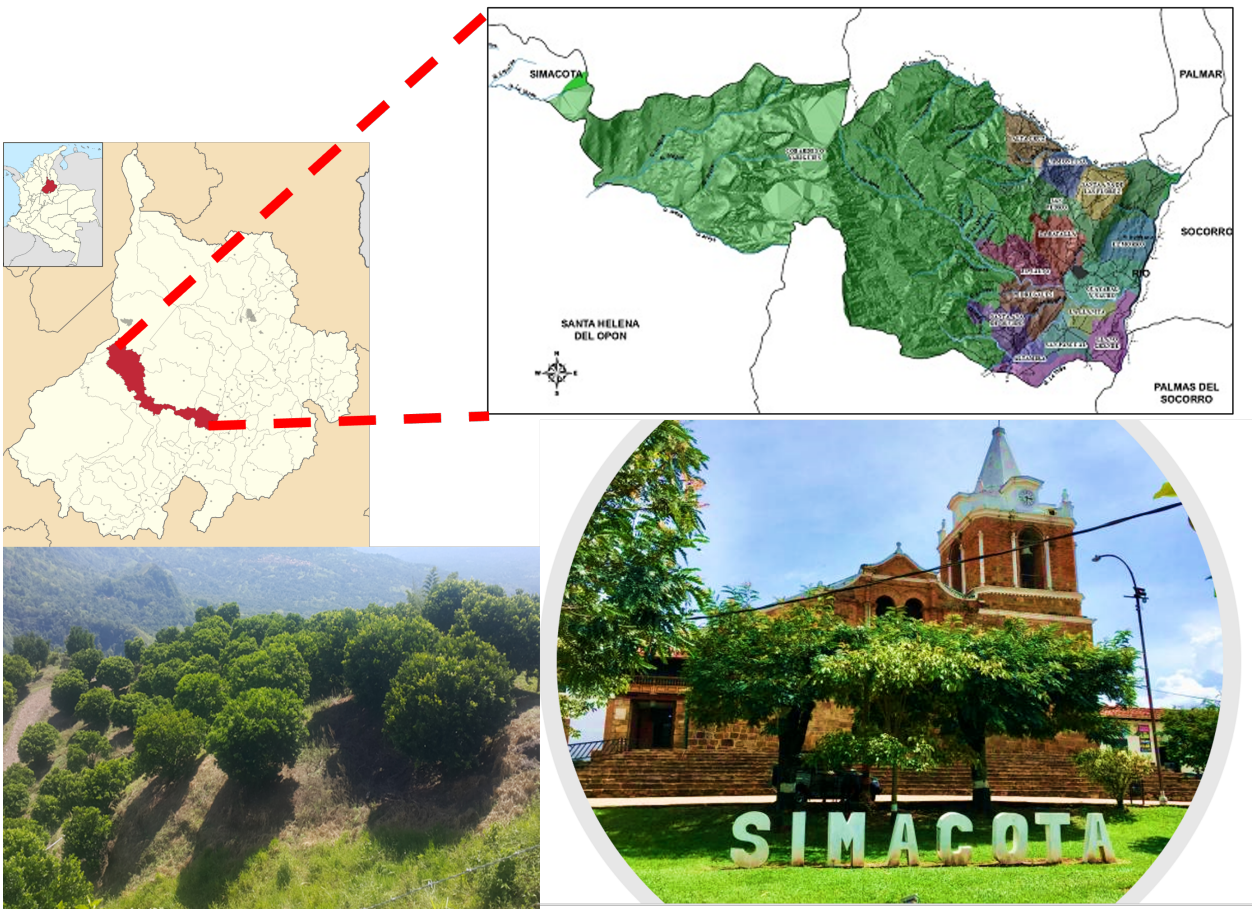


Figure 29. Geographical location in which the SOC estimation study is performed

Each sample was collected from a depth of 30 [cm] to minimize contamination from surface residues, across each one-hectare designated area. Following collection, the soil underwent a standardized preparation process involving drying, grinding, and sieving, resulting in a uniform sample devoid of moisture and uneven particles, thus ensuring consistent spectral properties. Additionally, the collected soil samples were separated into three major groups to acquire a diverse database depending on the conditions of the soil samples. The group labeled as $Raw(C)$ corresponds to the unprocessed soil samples. The group labeled as Soil (S) corresponds to the soil samples proces-

sed by the chemistry laboratory, which is necessary to obtain the SOC percentage by traditional methods. Finally, the group labeled as HDSP (H) corresponds to the soil samples treated within the HDSP group, which involved a process of crushing, sieving, and drying, emulating a traditional process that farmers could perform on their farms. This approach aims to acquire a database as realistic as possible. In addition, each of the soil samples was characterized physicochemically by a laboratory specialized in soil analysis, providing diverse properties, among which the percentage of SOC stands out 30.

	LABORATORIO QUÍMICO DE SUELOS CONVENIO UIS-GOBERNACIÓN DE SANTANDER RESULTADO ANÁLISIS DE SUELOS		Código: F-AA-03 Versión: 01 Página 1 de 1

Cliente	Joel Sánchez		Fecha de recepción de la muestra		Diciembre 02 de 2021		Departamento	Santander		Finca	El recuerdo						
Entidad	----------		Fecha de Análisis		Diciembre de 2021		Municipio	Simacota		Cultivo	Cítricos						
Dirección	N.S.		Fecha de Emisión de Resultado		Enero 14 de 2021		Vereda	El Salto									
Análisis solicitado	Caracterización	X	Elementos menores	X	Azufre	N.S.C.	C.I.C	X		C.E	X						

Código muestra	pH unid	%C	P (ppm)	Ca	Mg	Na	K	Al	% Arena	% Limo	% Arcilla	Textura	B	Fe	Mn	Cu	Zn	S	CIC meq/100g	CE mhos/cm
				meq/100g suelo									(ppm)							
21-01	6,8	3,11	1,06	26,0	0,60	0,03	0,12	N.A	48	22	30	Fco-Arcillo-Arenoso	N.S.C.	14,9	2,79	0,30	0,32	N.S.C.	27,0	0,15

Figure 30. Response of physicochemical analysis to soil samples, with characteristics such as pH, calcium, texture, and SOC.

4.2. Data SOC percentage distribution

After processing the soil samples, they are labeled with the percentage of organic carbon, obtained by the chemical laboratory, by methods such as calcination or dichromate, with a resolution of 0.01 %. Figure 31 displays the distribution of SOC results for the 100 soil samples, revealing that the average SOC varies between 1 % and 2 %, which equates to 10 to 20 [g/kg]. It is necessary to highlight that in the literature study by Wang et al. (2022), soil samples containing less than 120 [g/Kg] ($< 12\,\%$) SOC are typically rich in minerals. This high mineral content complicates

the spectral and non-destructive estimation of SOC, as minerals can affect the reflection of light from the organic components within the soil Wang et al. (2022). Specifically, in this case, taken into account minerals such as potassium, calcium, and magnesium the average of analyzed soil samples provides values above the standard for high mineral content Georgiou et al. (2022).

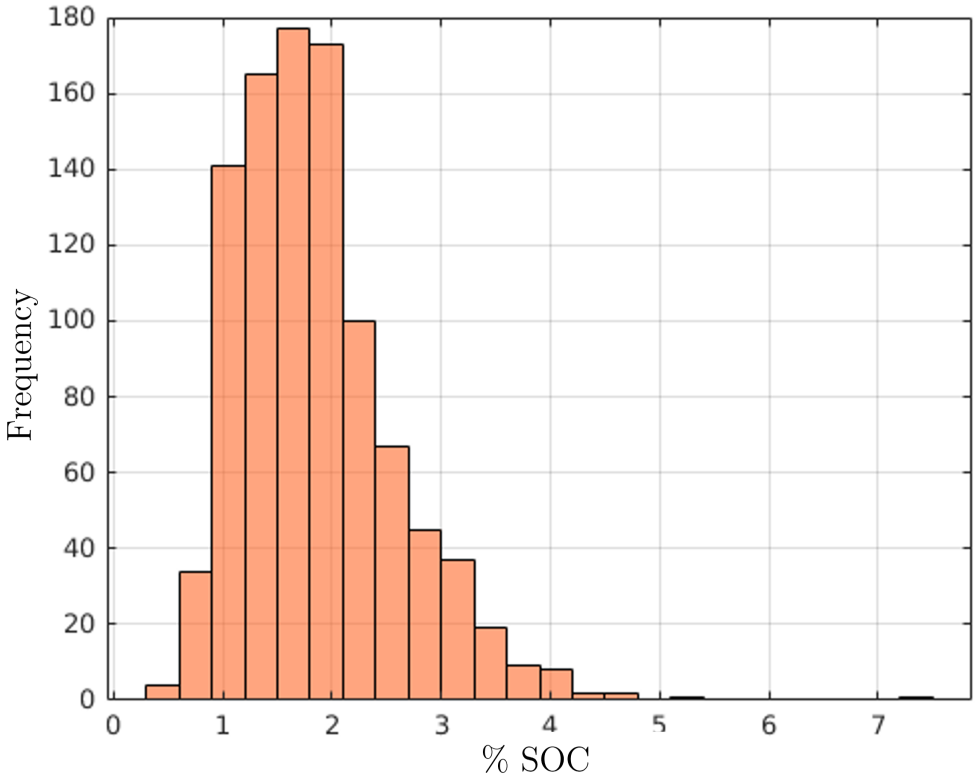


Figure 31. SOC percentage distribution of soil samples.

4.3. Spectral Dataset Acquisition

The acquisition of the NIR spectral signatures database was conducted over a two-year period. The primary reason for this extended acquisition period is that the soil samples were collected sequentially within the same timeframe to characterize the terrains of the Simacota municipality and provide farmers with the SOC percentage at least three times a year. This process resulted in

a total of 3.000 NIR spectral signatures, with 1.000 from each treatment of the soil samples: C, S, and H, in both intensity (Counts) and reflectance.

It is important to highlight that throughout the acquisition process of spectral signatures, geometric, electronic, optical, and mechanical calibration of the Whiskbroom-type system was ensured. This calibration allowed the acquisition of temporally partially invariant spectral signatures. Additionally, as shown in Figure 32 a), the SOC percentage is not directly related to the reflectance percentage, which underscores the challenge of estimating SOC percentage using traditional regression spectroscopy methods. Highlighting that in Figure 32 a) the spectral signature is observed with processing. At the spectral extremes of 900 [nm] and 2.500 [nm], the dynamic range in which the spectral signature can be represented decreases, an intrinsic factor of the sensor. Consequently, noise increases in these ranges, and it was decided to exclude these ranges since they do not allow for correct characterization of the soil sample, as shown in Figure 32 b).

Additionally, a spectral variation analysis was conducted between the dataset acquired from raw soil samples (C) and treated soil samples (H), as shown in Figure 33. Specifically, in this case, a pair of spectral signatures were selected, both corresponding to the same soil sample; however, one was acquired without any treatment, and the other was obtained after the sample was crushed, sieved, and dried as previously described. This comparison reveals that spectral information varies, both in reflectance level Figure 33 a-b and in specific peaks when normalized to the same level Figure 33 c-d. This variation arises because the composition of the acquired sample changes due to the treatment process, affecting factors such as moisture content and homogeneity. Therefore, it was concluded that the treated (H) spectral dataset will be used in this case, as it ensures greater

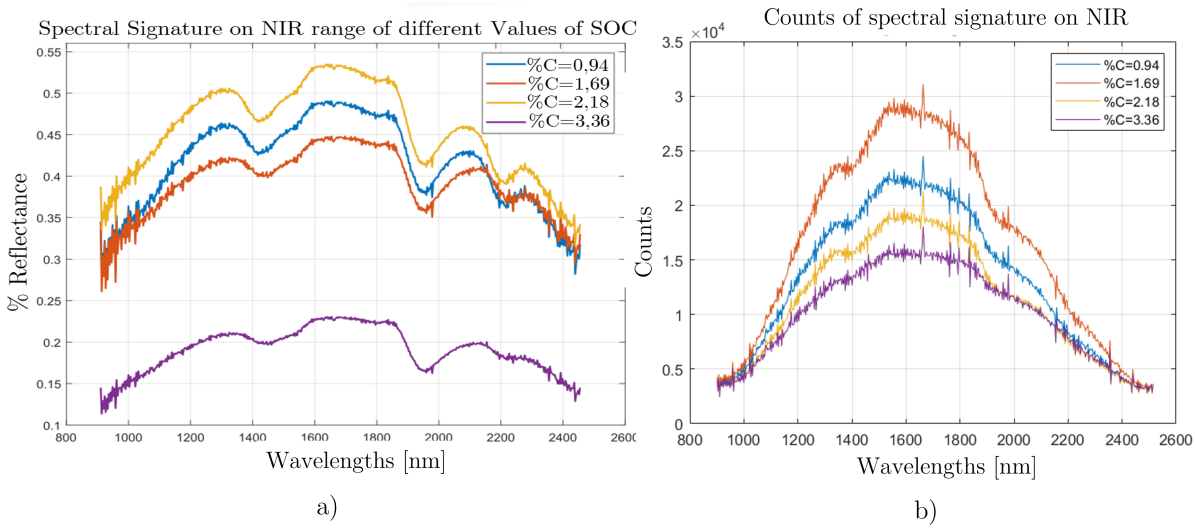


Figure 32. SOC percentage analysis based on percentage reflectance. a) Reflectance of spectral signature on NIR with different SOC percentages, and b) Counts of spectral signatures on NIR range with different SOC percentages.

consistency in the soil sample's homogeneity and, consequently, in the acquired signature.

This chapter concludes with the characterization of the soil sample collection process over two years in the municipality of Simacota, Santander, as well as the creation of a database of NIR spectral signatures totaling 3.000 signatures, divided equally among the three data processing groups. Additionally, an analysis of temporal variability in the acquisition process was performed. This confirms the stability of the acquisition protocol and the optoelectronic system based on the acquired database.

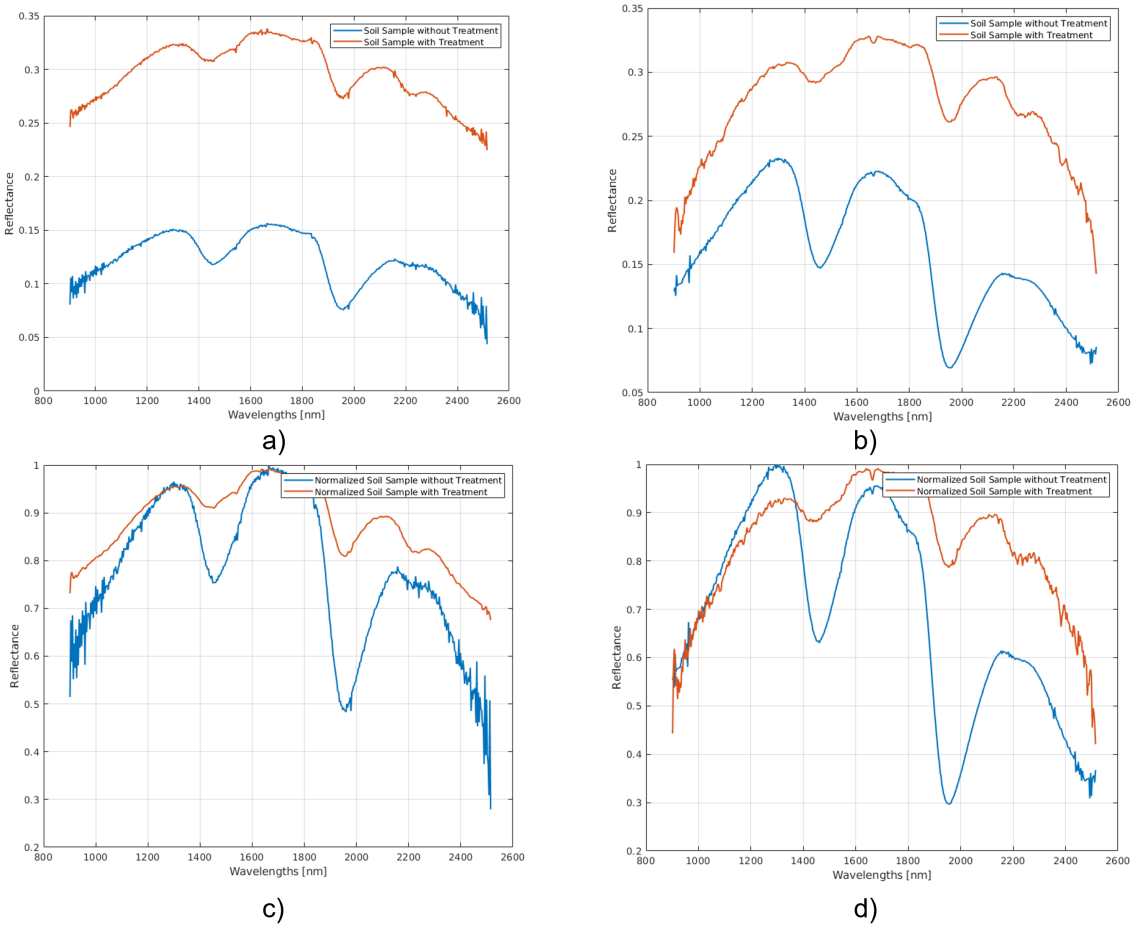


Figure 33. Spectral signature comparison between Raw soil sample and HDSP soil sample. a) Spectral signatures with same percentage carbon as 2.5 %, b) Spectral signatures with same percentage carbon as 3.5 %, c) Spectral signatures normalized with percentage carbon as 2.5 %, and d) Spectral signatures normalized with percentage carbon as 3.5 %

5. Computational algorithm of SOC estimation by spectral signature

Considering that the next phase of this research involves estimating SOC from spectral signatures, a literature review was conducted on methods previously used for this task. Specifically, the literature introduces multiple architectures designed to extract intrinsic features from spectral signatures and estimate the SOC percentage within the NIR spectrum. It is important to note that many of the developed algorithms can estimate the SOC percentage with high precision because they are trained on large datasets. However, they are not easily replicable for other types of spectral signatures, as these signatures exhibit characteristic shapes depending on the soil type and terrain from which they were collected.

Therefore, developing a computational algorithm that can estimate the SOC percentage in an environment with a limited number of spectral signatures is essential to achieve estimation results comparable to traditional chemical methods. The computational algorithms used for this purpose are classified into two major groups: machine learning and deep learning, depending on the type of learning methodology employed. However, it has been widely demonstrated that preprocessing the data used to train the estimation models significantly improves their performance Shi et al. (2023).

5.1. Preprocessing of spectral signature to enhance the performance of neural networks

Following the methodology found in the literature on SOC estimation, various preprocessing techniques were applied to the spectral signature, as they enhance the extraction of important features Zhang et al. (2020); Rozenstein et al. (2014). Specifically in this case, since the spectral

signature comes from a soil sample containing various physicochemical components, the task involves highlighting subtle changes in the spectral signature that correspond to SOC content. In this research work, various treatments and processing methods were analyzed. Below, they are listed along with a brief explanation of each method. In Figure 34, an example of the processes performed is shown, with the figure in the top left of the first row representing reflectance without any processing.

- **Normalization.** The normalization of spectral signatures in the NIR of soil samples is a crucial step to ensure comparability and accuracy in spectral data interpretation. In this study, normalization was performed for each experiment to adjust for variations between samples and minimize the effects of external conditions or instrumentation during data acquisition. This process ensures that observed differences in spectral signatures truly reflect the chemical and physical variations in the soils, thereby enabling a more accurate assessment of SOC and other physicochemical parameters. Normalization is essential for stabilizing the data and facilitating further analysis and comparison in soil characterization studies, as shown in Figure 34 in the top right of the first row.

- **Standarization.** The standardization of spectral signatures in the NIR of soil samples is an essential procedure aimed at enhancing the consistency and reliability of the data across different experiments. In this research, standardization was applied individually for each experimental setup to mitigate the effects of varying conditions and equipment sensitivities. This practice helps in normalizing the range of spectral data, ensuring that the variation

among spectral signatures accurately reflects differences in soil composition, not discrepancies in measurement conditions. Such standardization is crucial for effective comparison and analysis of soil properties, particularly for precise SOC estimation and other related soil characteristics, as shown in Figure 34 in the bottom left of the second row.

- **Savitzky-Golay Filter.** The Savitzky-Golay filter operates by fitting successive sub-sets of adjacent data points with a low-degree polynomial using linear least squares. This method effectively preserves higher momentums, such as the height and sharpness of peaks, which are critical in SOC estimation. It smooths the spectral data from soil samples by minimizing noise while maintaining the structural integrity of the spectral features Shi et al. (2023). This capability is particularly useful for enhancing the precision of SOC predictions by ensuring that the relevant spectral characteristics are accurately represented and not distorted by random noise, as shown in Figure 34 in the bottom right of the second row.

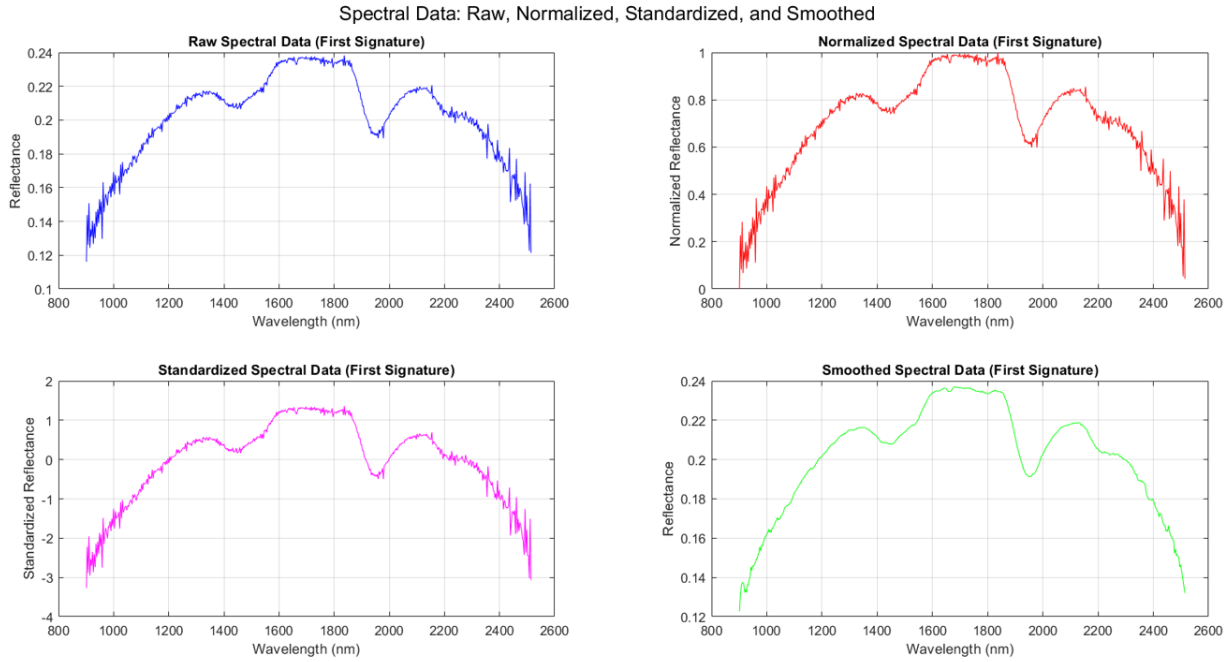


Figure 34. All preprocessing steps applied to the acquired spectral signatures.

- **Data augmentation.** Additionally, data augmentation techniques were implemented, taking advantage of the fact that each soil sample provided 112 spectral signatures. This approach ensures that data augmentation leverages the intrinsic variability within the spectral data from each sample, rather than generating synthetic variations. By doing so, the data augmentation process maintains the inherent characteristics of the spectral signatures, enhancing the robustness and generalizability of the machine learning models employed for SOC estimation. This method effectively increases the dataset size while preserving the natural differences and patterns present in the original spectral data from the soil samples.

    - **Mean.** The mean technique involves smoothing the spectral data by averaging each point with its neighbors. This method is crucial for SOC estimation as it helps to reduce

random noise in the spectral signatures from soil samples. By averaging the data points over a defined window, the moving average filter clarifies the underlying trends in the SOC content by dampening the effects of outlier values and noise. This smoothing process is vital for enhancing the accuracy of SOC predictions, ensuring that the true spectral characteristics indicative of organic carbon content are more pronounced.

- **Median.** The median filter is applied to spectral data to reduce noise while preserving edges by replacing each point with the median of neighboring points in a defined window. This method is particularly effective for SOC estimation as it helps to maintain the essential characteristics of the spectral signatures from soil samples. By using the median rather than the mean, the filter minimizes the impact of extreme outliers that could distort the data, ensuring a more accurate representation of the soil's organic carbon content. The preservation of sharp features within the spectral data is crucial for accurately identifying and quantifying SOC.
- **By spatial distribution.** Lastly, taking advantage of the acquisition process developed in this work, a type of data augmentation based on the scanning method and spectral data acquisition was implemented. Specifically, the acquisition protocol was designed to easily identify the position of each acquisition point within a soil sample. Thus, the acquisition space for the 112 signatures per sample was divided into a Cartesian plane, segmented into four quadrants. From these four quadrants, the data augmentation process was conducted by extracting the average of the signatures from each quadrant

(28 signatures per quadrant). This method enhances the dataset by incorporating spatially diverse spectral information, allowing for a more robust estimation of SOC by capturing variations across different sections of the sample.

- **Spectral Range of Interest.** Additionally, considering that literature has demonstrated the existence of spectral regions where chemical information regarding organic carbon content in the NIR is condensed, spectral range selection was conducted for training the aforementioned algorithms. Highlighting that the literature includes analyses of spectral ranges of interest, where information on certain physicochemical properties is most concentrated, it is noted that in the infrared range from 1,400 to 2,500 nm, the highest values are obtained in the estimation process of SOC SORIANO DISLA et al. (2014). This included the full spectrum (900-2.500 [nm]) and two specific spectral ranges: RE1, ranging from 1.500 to 1.800 [nm], and RE2, covering the spectral range from 2.000 to 2.300 [nm]. By focusing on these specific spectral regions, the algorithms can better capture the relevant chemical information related to soil organic carbon, potentially improving their accuracy in SOC estimation, an example is shown in Figure 35. Highlighting that in this case, the preprocessed spectral signature underwent normalization, and instead of using the mean, the median was utilized, resulting in a change in the spectral profile.

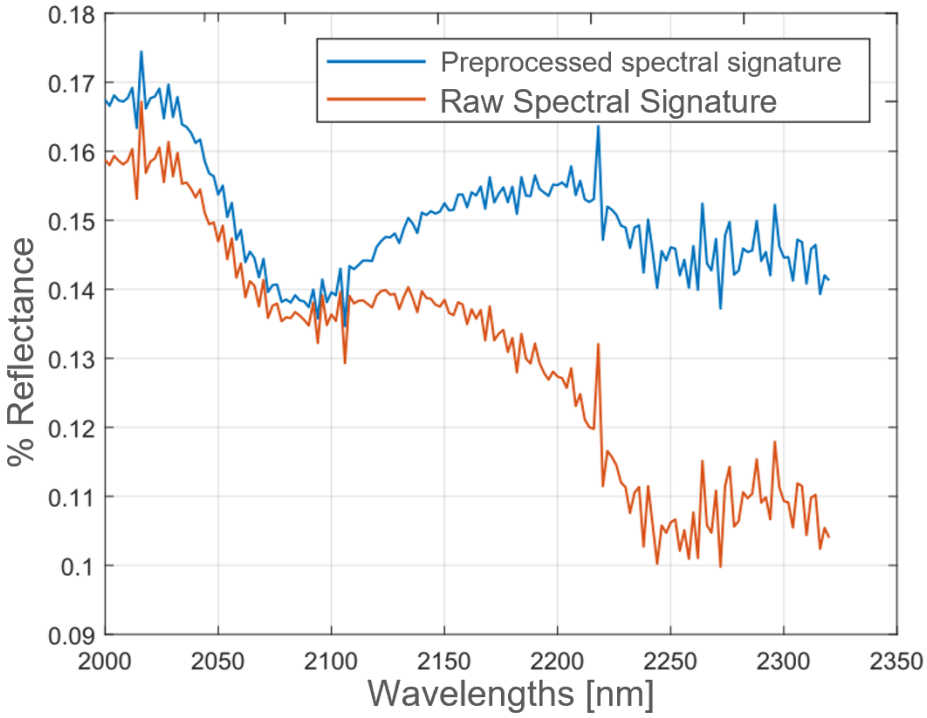


Figure 35. Example of selection of spectral range of interest (RE2) in specific spectral signature.

5.2. Machine learning models for SOC estimation by spectral signature in NIR range

The machine learning regression models implemented for this study corresponded to artificial intelligence techniques widely used in the scientific community: linear regression, random forest (RF), support vector machine (SVM) for regression, partial least squares regression (PLSR), and the k-nearest neighbors (KNN) method.

Linear regression models the relationship between SOC content and NIR spectral data using a straightforward linear approach, making it an excellent starting point for exploratory analysis due to its simplicity and ease of interpretation. Support Vector Machine (SVM), on the other hand, pro-

vides a more robust estimation by using SVM to model both linear and non-linear relationships, effectively managing the high-dimensional space of NIR spectral data to predict SOC levels accurately. This is complemented by Random Forest (RF), an ensemble method that enhances prediction accuracy and stability by averaging the results of multiple decision trees, thereby handling non-linear data effectively and providing resistance against overfitting.

For more complex data interactions, Partial Least Squares Regression (PLSR) is employed to handle multicollinearity among NIR spectral predictors, optimizing the extraction of relevant features by maximizing the covariance between responses and predictors. Meanwhile, the K-Nearest Neighbors (KNN) method offers a straightforward, intuitive approach by predicting SOC based on the closest spectral signatures in the training dataset, which assumes that similar NIR spectra correspond to similar SOC contents. Both PLSR and KNN are crucial for refining SOC estimations where the relationships within the data are not just about presence or absence but about the degree of similarity and interaction among spectral features. While in the literature machine learning algorithms achieve comparable results to traditional chemical methods, they have significant shortcomings in that they require specific tuning depending on the quality and, more specifically, the quantity of data. In other words, if there is a smaller amount of data available for estimating SOC, the performance of the models decreases considerably.

5.3. Deep learning models for SOC estimation by spectral signature in NIR range

Although various deep learning architectures have been developed in the literature, in this case study, the implementation and adaptation of two widely used architectures for SOC estimation from a large number of spectral signatures were performed Carvalho et al. (2024); Egeonu and

Jia (2024). Two advanced deep neural network architectures were employed to extract intrinsic features from spectral signatures for estimating physicochemical properties. Specifically, the VGG and Recurrent Neural Network (RNN) models were used. These architectures utilize convolutional neural networks for robust feature extraction and pattern recognition. They were selected due to their proven effectiveness in deriving complex and significant features from spectral data Miao et al. (2024). The RNN is particularly adept at capturing sequential information in spectral data, which is essential for modeling the temporal or spatial structures inherent in spectral signatures. On the other hand, the VGG architecture is celebrated for its depth and capacity to learn hierarchical data representations, allowing it to extract crucial features at various levels of abstraction Huan et al. (2024). Together, these two complementary architectures in the literature provide a thorough toolkit for spectral data analysis and feature extraction, making them highly effective for accurate SOC estimation from spectral signatures in this study.

VGG network architecture

This architecture is primarily used for image classification and object recognition tasks, this is a type of convolutional neural network that has been influential in the field of computer vision. It is distinguished by its deep structure, featuring numerous convolutional layers arranged sequentially, each utilizing small 3×3 convolutional filters and interspersed with max-pooling layers. This consistent architecture with compact filters allows VGG to develop detailed hierarchical representations of input images, capturing elemental features like edges and textures, as well as more complex attributes such as shapes and objects.

The VGG architecture is notably suited for estimating SOC from spectral signatures due

to its proficiency in learning layered data representations. From the succession of convolutional and max-pooling layers, VGG effectively extracts features at various abstract levels. This capacity for hierarchical feature extraction helps the model identify intricate patterns and nuances in spectral data, crucial for detecting variations linked to SOC content. The depth of the architecture and its comprehensive feature learning capabilities make it ideal for managing the complex, high-dimensional nature of spectral data. The VGG model shines in isolating key features from spectral signatures, enhancing the precision of SOC estimation in soil samples. Fine-tuning of hyperparameters such as batch size, dropout, learning rate, epochs, beta, gamma, and seed further optimizes the model's performance. The final model configuration includes a batch size of 32, a learning rate of 5×10^{-4}, 250 training epochs, a fixed dropout rate of 0,01, and specific beta and gamma values of 1,05 and 0 respectively, with seeds $0, 707, 976, 10, 42$, ensuring optimal results in each iteration. Specifically, this architecture is shown in Figure 8 a).

Resnet network architecture

This architecture in literature is commonly used for image recognition and classification tasks in computer vision. However, this architecture has been proved that can be made a SOC estimation from spectral signatures because it can effectively extract discriminative features from complex data. By leveraging its deep architecture and residual connections, the ResNet can learn hierarchical representations of spectral data, capturing low-level and high-level features relevant to SOC content. This enables the ResNet to accurately estimate SOC by modeling the complex relationship between spectral signatures and SOC content in soil samples. For the Resnet network's training aimed at estimating organic carbon, a fine-tuning of hyperparameters was conducted. This

included adjustments to batch size, dropout, learning rate, epochs, beta, gamma, and seed. The fine-tuning concluded after each iteration by establishing the optimal combination of parameters: a batch size of 32, a learning rate of 5×10^{-4}, 250 training epochs, a constant dropout rate of 0,01, and beta and gamma values set at 1,05 and 0, respectively. Additionally, the process utilized five specific seeds $\{0, 707, 976, 10, 42\}$, the same as VGG architecture. Specifically, this architecture is shown in Figure 8 b).

5.4. Optimized computational algorithm for small spectral soil datasets

Considering that state-of-the-art algorithms in both machine learning and deep learning are optimized to estimate SOC percentage from a large amount of data, and emphasizing the significant limitation in this work due to the small number of acquired spectral signatures available, a low-complexity neural network architecture is proposed. This architecture allows for feature extraction from spectral signatures, exploiting preprocessing properties such as smoothing and spectral range selection to estimate SOC percentage.

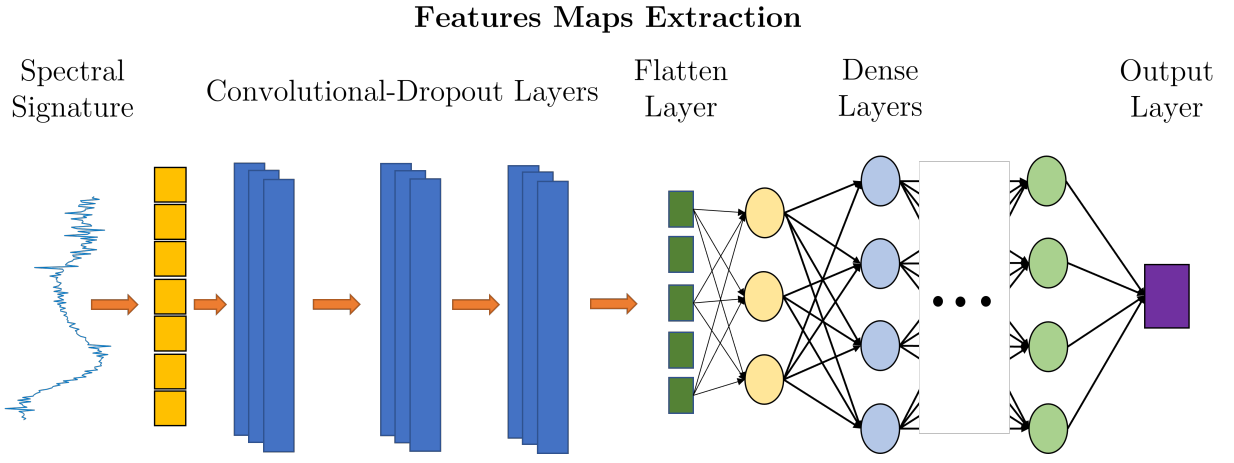


Figure 36. Proposed neural network architecture for SOC estimation using a reduced group of NIR spectral signatures

Specifically, the proposed neural network implemented is depicted in Figure 36. The main feature of the proposed architecture is that it allows for the extraction of features from an NIR soil spectral signature using a reduced group of convolutional and dense layers, achieving a total of 1 million parameters. This is significantly lower compared to state-of-the-art algorithms such as VGG and ResNet, which typically have around 100 million parameters Miao et al. (2024); Egeonu and Jia (2024) . It is important to highlight that for the neural network to achieve optimal results, preprocessing of the acquired data is necessary, which will be presented next.

In conclusion, this chapter culminates with an ensemble of computational algorithms, encompassing both machine learning and deep learning methodologies, tailored towards estimating the percentage of SOC from NIR spectral signatures. Additionally, it encapsulates the design and implementation of a computational algorithm meticulously optimized for SOC estimation from a reduced dataset. Furthermore, it encapsulates the various data processing techniques applied to the spectral database acquired in the laboratory via the implemented optical system, leveraging the devised protocol.

6. Validation

6.1. Literature Review on Metrics for SOC Estimation

In the literature, several metrics are commonly used to evaluate the accuracy and effectiveness of models for SOC estimation. Below are the most frequently reported metrics, each explained with its respective equations, advantages, and disadvantages.

6.1.1. Mean Absolute Error (MAE). MAE calculates the average absolute difference between predicted ($\hat{y}_i$) and actual (y_i) values

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|. \tag{5}$$

MAE is straightforward to understand and interpret, providing a direct measure of average model error. Unlike MSE and RMSEP, MAE does not penalize larger errors more significantly, which may not highlight models that perform poorly on outliers.

6.1.2. Pearson's Correlation Coefficient (r). This coefficient measures the strength and direction of the linear relationship between predicted ($\hat{y}_i$) and actual (y_i) values

$$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}}. \tag{6}$$

Pearson's r provides insight into the linear relationship between variables, which can be useful for model diagnostics. However, it only measures linear relationships, so it may not capture the full

performance of models with non-linear dependencies.

6.1.3. Mean Relative Error (MRE). MRE calculates the average of the absolute relative errors between predicted and actual values, expressed as a percentage

$$\text{MRE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|. \tag{7}$$

MRE provides a relative measure of model error, which can be useful for comparing performance across different scales. However, MRE can be problematic when actual values are close to zero, leading to disproportionately large error values.

6.1.4. Mean Squared Error pre (MSE) and Root Mean Squared Error prediction(RMSEP). MSE measures the average of the squares of the errors between the predicted ($\hat{y}_i$) and actual (y_i) values. n denotes the number of data points

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{8}$$

MSE penalizes larger errors more significantly, which can be useful for identifying models that perform poorly on outliers. However, it can be heavily influenced by outliers, leading to a skewed assessment of model performance.

RMSEP is the square root of MSE and provides an error measure in the same units as the test data

$$\text{RMSEP} = \sqrt{\text{MSE}}. \tag{9}$$

RMSEP is easier to interpret than MSE since it is in the same units as the target variable. Like MSE, RMSEP is sensitive to outliers and can overestimate model error due to large deviations.

6.1.5. Coefficient of Determination (r^2). r^2 measures the proportion of the variance in the dependent variable explained by the independent variables. $\bar{y}$ denotes the mean of the observed data

$$r^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{10}$$

r^2 provides a clear indication of how well the independent variables explain the variability of the dependent variable. However, it can be misleading if the model is overfitted, as it may show a high value even if the model generalizes poorly to new data.

In this study, the r^2 was selected as the fundamental metric for evaluating the performance of SOC estimation models due to its ability to measure the proportion of variance in the dependent variable that is predictable from the independent variables. r^2 provides a clear and intuitive indication of how well the model explains the variability of the data, making it an essential metric for assessing the overall goodness of fit. Additionally, r^2 is widely recognized and used in regression analysis, facilitating comparisons with other studies in the literature.

As an auxiliary metric, the RMSEP was chosen because it provides an error measure in the same units as the original data, offering a straightforward interpretation of the model's prediction accuracy. RMSEP complements r^2 by quantifying the average magnitude of the prediction errors, thereby providing additional insight into the model's performance, particularly in terms of how closely the predictions match the actual values. This dual-metric approach ensures a comprehensive

evaluation of the model, balancing the proportion of explained variance with the magnitude of prediction errors, and thus enhancing the robustness of the performance assessment.

6.2. Analysis of literature datasets for SOC estimation

Several datasets containing spectral signatures in the NIR range are available for training algorithms to estimate SOC.

6.2.1. LUCAS (Land Use/Cover Area Frame Statistical Survey) dataset. The LUCAS dataset stands out due to its extensive coverage and detailed spectral information. The LUCAS dataset includes a comprehensive collection of soil samples from across Europe, characterized by a wide range of soil properties and environmental conditions. It provides high-resolution NIR spectral data, which is crucial for developing robust SOC estimation models. The LUCAS dataset, specifically the most recent version available, provides comprehensive spectral and soil property data across Europe. It includes spectral signatures measured in the Near-Infrared (NIR) range and possibly other bands, capturing detailed information about soil composition and characteristics. The LUCAS dataset, specifically the most recent version available, provides comprehensive spectral and soil property data across Europe. It includes spectral signatures measured in the Near-Infrared (NIR) range and possibly other bands, capturing detailed information about soil composition and characteristics. This dataset encompasses a wide array of physical and chemical soil properties such as pH, organic carbon content, texture, and nutrient levels. The spectral signatures are acquired annually from numerous geographically distributed sampling points, ensuring a robust representation of soil variability across the continent. It is important to note that the data provided by the European dataset already includes smoothing, which must be considered

when training computational algorithms. This dataset is invaluable for environmental studies, offering both spectral data for remote sensing applications and detailed soil properties essential for agricultural and ecological research.

The main advantages of the LUCAS dataset include its large sample size, geographic diversity, and standardized measurement protocols, which enhance the reliability and comparability of the data. Additionally, the dataset is publicly accessible, promoting transparency and reproducibility in research. However, there are some limitations to consider. The dataset's primary focus on European soils may limit its applicability to other regions with different soil characteristics. Furthermore, the quality of the spectral data can vary due to differences in soil preparation and measurement techniques.

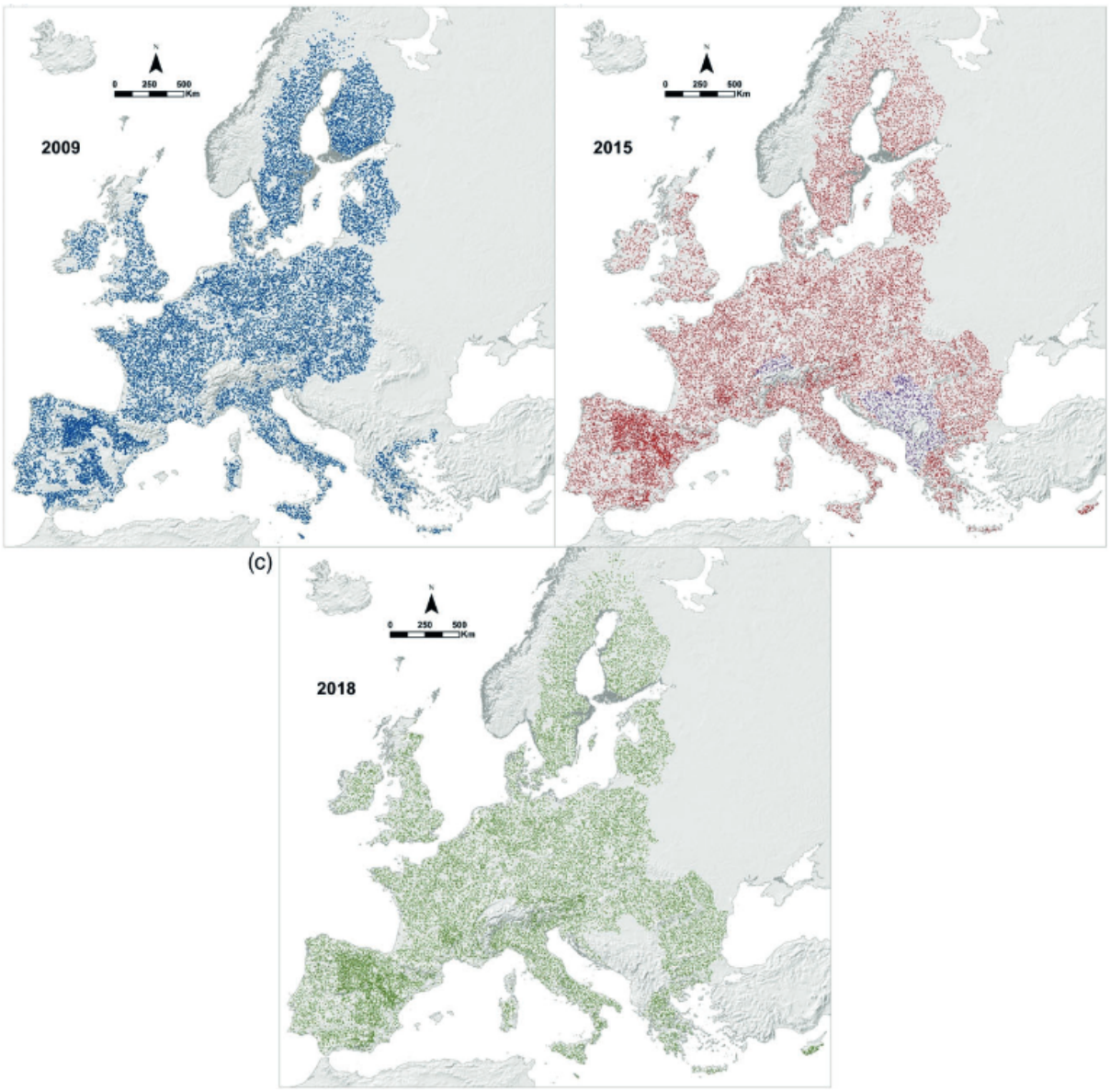


Figure 37. Temporary sampling throughout Europe for the acquisition of spectral signatures of soil samples with various labeled characteristics, including SOC percentage, taken from Orgiazzi et al. (2017)

Finally, similar to the HDSP dataset, an analysis was conducted on the distribution of carbon percentages present in LUCAS, as shown in Figure 38, highlighting that the data mean lies between 10 and 20% SOC. This is attributed to the acquisition method and the type of agricultural lands

that were sampled. This aspect is crucial when training SOC estimation algorithms, as a wide variability range in SOC percentages suggests that the spectral variability will also be greater.

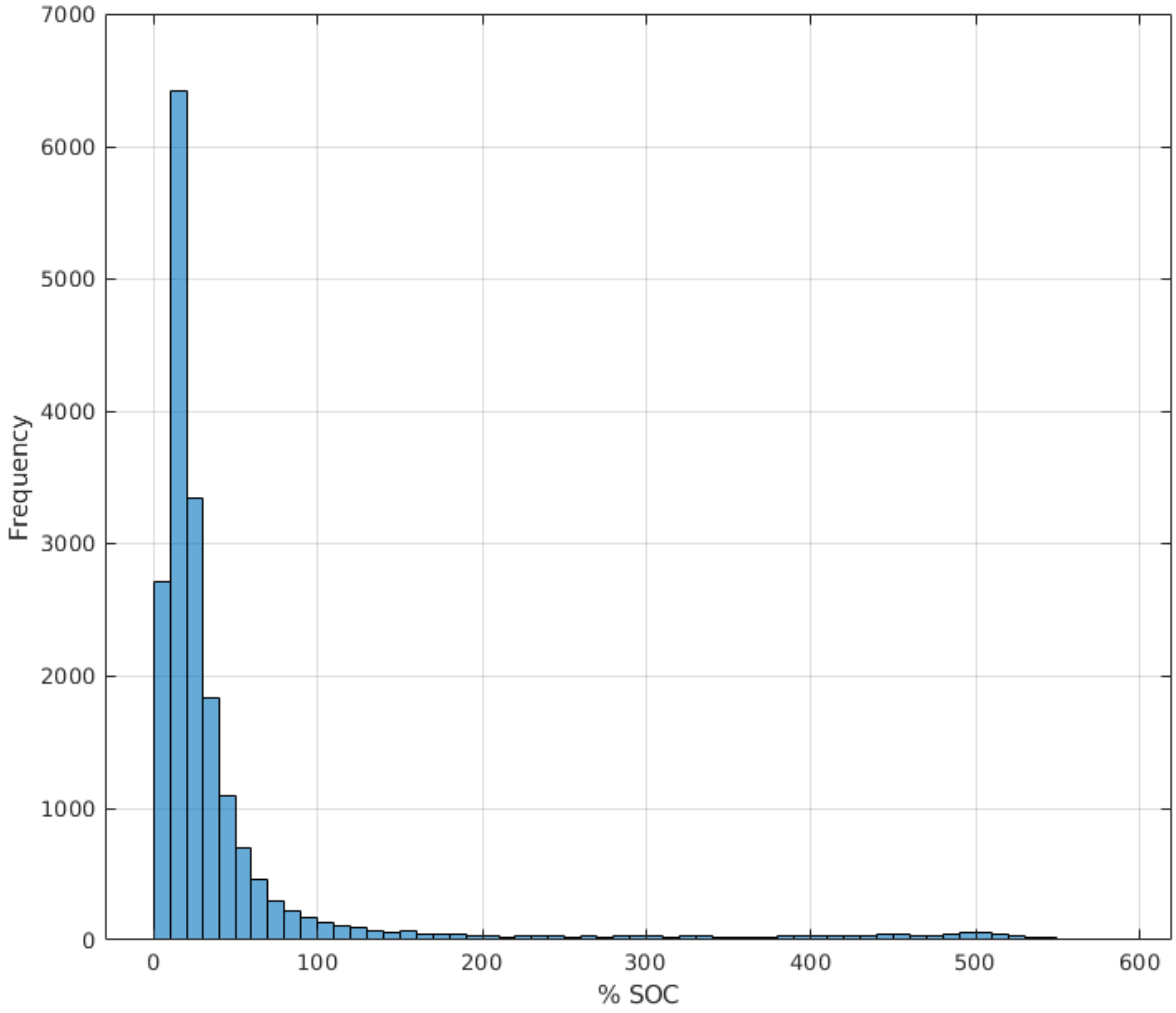


Figure 38. Distribution analysis for SOC, taken from Orgiazzi et al. (2017)

6.2.2. Acquired Colombian spectral soil dataset. On the other hand, the dataset acquired in Colombia by the developed optoelectronic system, as described in Chapter 4, presents a contrast with the LUCAS dataset. Specifically, it contains a smaller quantity of spectral signature

data, which complicates the task of SOC estimation. However, it is noteworthy that the spectral variability among each dataset entry is lower, as will be analyzed further.

6.2.3. Comparison of spectral and temporal variability of the datasets. To better understand the dataset used for comparing results obtained by computational algorithms, two analyses were conducted. The first analysis focused on temporal variation, where a series of spectral signatures were captured at or near the same point for each dataset. For the dataset acquired in Colombia by the implemented optoelectronic system, three captures were conducted. In contrast, for the LUCAS dataset, annual measurements were taken at random points, selecting three geographically close locations, as shown in Figure 39. This approach revealed that the temporal variability of spectral signatures in the Colombian dataset is lower compared to the European dataset.

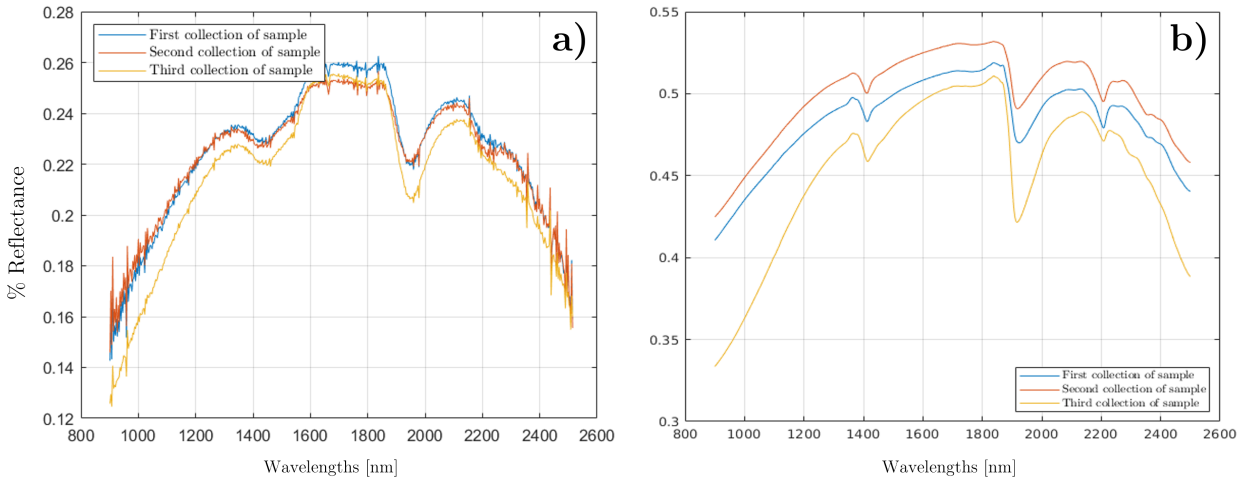


Figure 39. Temporal variability analysis based on percentage reflectance. a) Three spectral signatures on the Colombian dataset, b) Three spectral signatures on the LUCAS Dataset.

Furthermore, the next conducted analysis was a spectral comparison of signatures with the same percentage of carbon within the same dataset and across datasets, as depicted in Figure 40. This comparison highlighted that, as expected, the morphology of signatures acquired on another continent differs significantly from those obtained in Colombia. Consequently, there is a need to adapt computational algorithms to achieve better performance in estimating SOC percentage from a reduced number of spectral signatures.Additionally, the spectral difference is attributed to the acquisition process, instruments used, crops in the area where the soil sample was taken, and intrinsic soil properties in Europe.

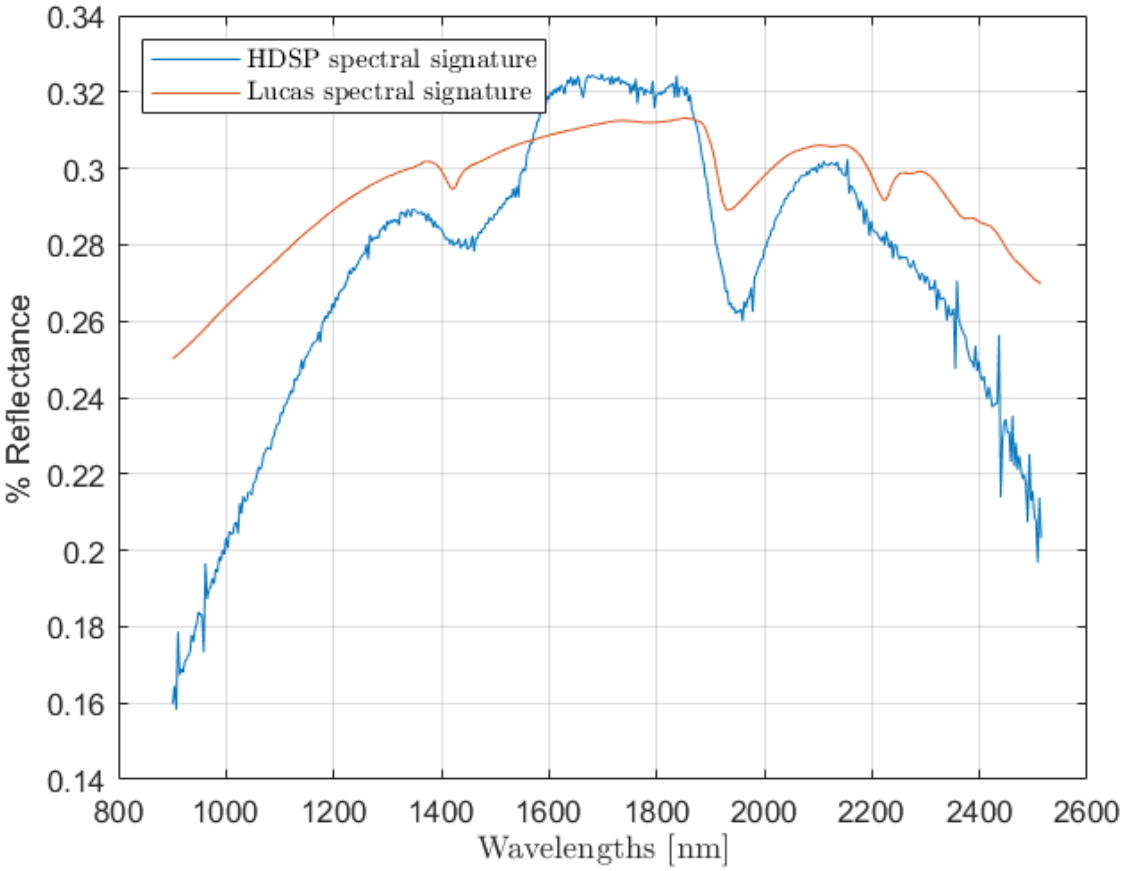


Figure 40. Analysis of spectral variability between the selected datasets, ensuring a SOC percentage of 4.5 without preprocessing tecnique.

6.3. Data preprocessing work pipeline

Given the wide range of preprocessing techniques that can be applied to the data used for SOC estimation, which includes both the data acquired by the implemented optoelectronic system and the European database, a series of combinations are proposed to validate the performance of the estimation algorithms. Various preprocessing methods, such as smoothing, spectral range selection, and normalization, will be systematically combined to assess their impact on the accuracy and

reliability of the SOC predictions. This approach aims to identify the most effective preprocessing strategies for enhancing the performance of computational models in estimating SOC by trying all possible combinations of preprocessing techniques, as shown in Figure 41.

Specifically, based on the preprocessing techniques mentioned in Chapter 5, and considering the analysis of using both reflectance and absorbance, a total of 72 possible combinations of treatments for the spectral signatures were obtained, as shown in Figure 41. These combinations include various methods such as smoothing, spectral range selection, normalization, and other preprocessing steps. By evaluating these different preprocessing strategies, we aim to determine the most effective approach for enhancing the performance of the SOC estimation algorithms.

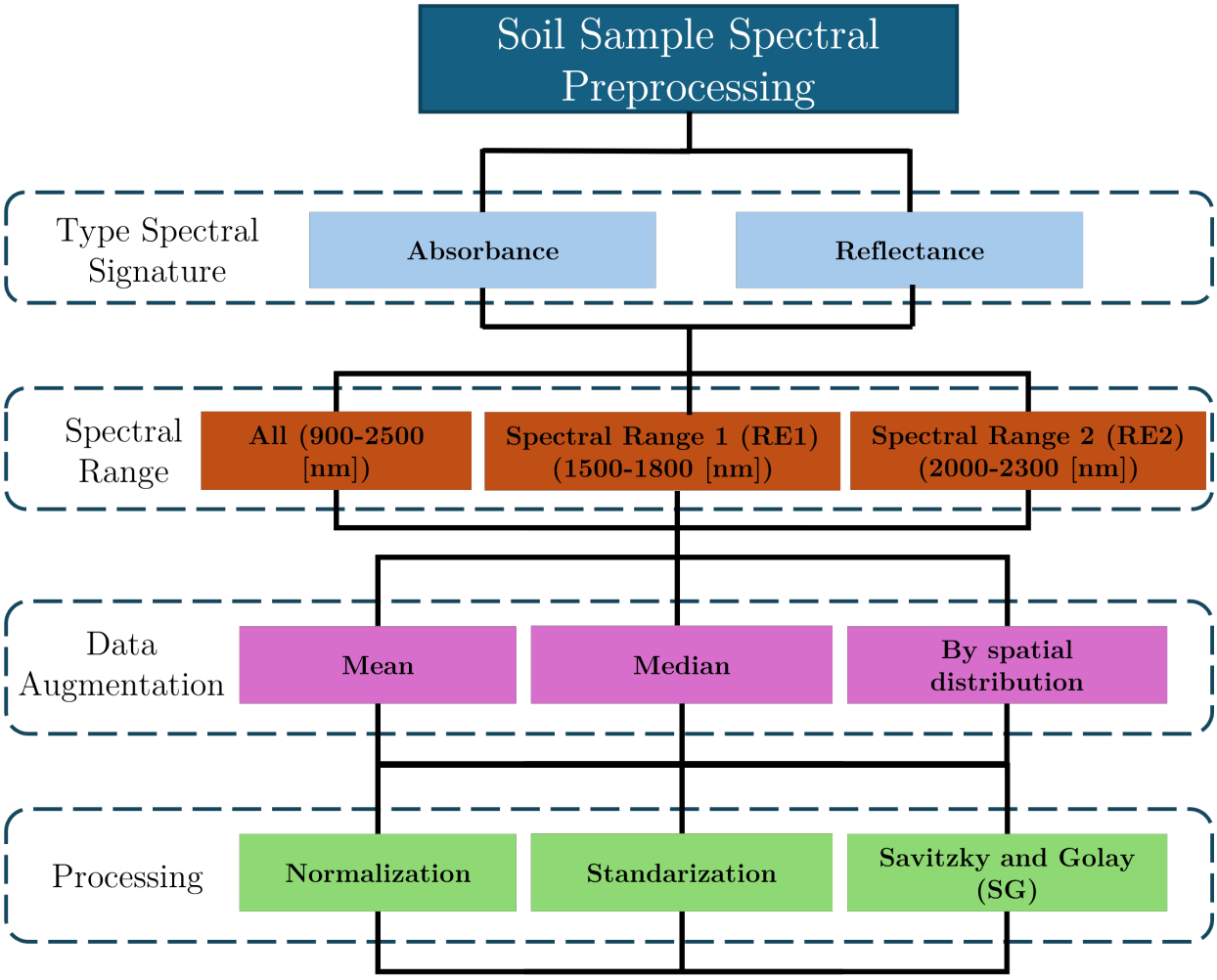


Figure 41. All possible combinations of spectral signature preprocessing are presented before training the SOC estimation models.

Among the configurations that must be highlighted in the training algorithms are the decay learning rate and ensuring that the best model is saved. Additionally, to ensure the reproducibility of the estimation models' performance, the k-fold cross-validation technique was applied. These configurations are crucial for optimizing the training process, as the decay learning rate helps to prevent overfiting during training, while saving the best model ensures that the most accurate and

robust version is retained. The k-fold cross-validation further enhances the reliability of the model evaluation by dividing the dataset into k subsets, allowing each subset to be used as a validation set while the remaining subsets are used for training, thereby providing a comprehensive assessment of the model's performance. The k-fold cross-validation technique was applied to ensure the reproducibility and robustness of the estimation models' performance. Specifically, this method was implemented for two datasets: one consisting of 1.000 spectral signatures in the NIR range, acquired in the optics laboratory, and another comprising 20.000 spectral signatures from the European LUCAS dataset. A k value of 5 was chosen for both datasets. This means that each dataset was divided into 5 subsets, or folds. For each iteration, one fold was used as the validation set while the remaining 4 folds were used for training the model. This process was repeated 5 times, with each fold serving as the validation set once. By using k-fold cross-validation, ensured that every data point was used for both training and validation, providing a comprehensive evaluation of the model's performance and mitigating the risk of overfitting, particularly in datasets of varying sizes and characteristics.

Based on this, Table 2 presents the RMSEP and r^2 results of the best preprocessing combinations for machine learning algorithms such as Linear Regression, PLSR, MLP, SLP, SVM, RF, deep learning algorithms VGG, ResNet, and the proposed algorithm, for both the LUCAS dataset and the dataset acquired in the optics laboratory, taking into account all results of estimation with all possible combinations of preprocessing. Highlighting that the spectral signature datasets were used in their entirety for this analysis.

The SOC estimation results vary according to the dataset used. Specifically, for the available

Preprocessing and architecture	*LUCAS dataset*				*Acquired dataset*			
	r^2		*RMSEP*		r^2		*RMSEP*	
	Val	*Test*	*Val*	*Test*	*Val*	*Test*	*Val*	*Test*
ML(SLP_ Abs_All_ mean_Raw)	0,66 ±0.07	0.631 ±0.03	0,473 ±0.09	0.504 ±0.07	0,674 ±0.05	0.651 ±0.01	0,464 ±0.06	0.515 ±0.03
ML(PLSR_ Ref_RE1_ median_SG)	0,641 ±0.07	0.617 ±0.02	0,512 ±0.03	0.55 ±0.01	0,651 ±0.01	0.634 ±0.05	0,479 ±0.07	0.534 ±0.05
ML(SVM_ Ref_RE1_ mean_SG)	0,573 ±0.07	0.553 ±0.09	0,601 ±0.03	0.623 ±0.01	0,601 ±0.04	0.587 ±0.09	0,51 ±0.07	0.602 ±0.06
DL(VGG_ Ref_RE2_ mean_norm)	**0,787 ±0.02**	**0,754 ±0.03**	**0,213 ±0.04**	**0,313 ±0.07**	0,761 ±0.07	0,732 ±0.05	0,216 ±0.08	0.297 ±0.02
DL(VGG_ Ref_RE1_ mean_standar)	0,744 ±0.07	0,732 ±0.03	0,279 ±0.04	0.356 ±0.07	0,739 ±0.01	0,711 ±0.02	0,256 ±0.07	0.324 ±0.08
DL(Resnet_ Abs_RE1_ median_SG)	0,756 ±0.09	0,686 ±0.01	0,37 ±0.03	0.419 ±0.03	0,709 ±0.03	0,684 ±0.06	0,298 ±0.06	0.367 ±0.07
Proposed (Ref_RE2_ mean_SG)	0.679 ±0.07	0.675 ±0.04	0,401 ±0.01	0.456 ±0.06	**0,791 ±0.07**	**0.743 ±0.06**	**0,201 ±0.09**	**0,246 ±0.02**
Proposed (Ref_RE1_ mean_SG)	0,659 ±0.07	0,641 ±0.03	0,456 ±0.01	0.479 ±0.01	0,756 ±0.04	0,731 ±0.03	0,256 ±0.05	0.297 ±0.01
Proposed (Ref_RE1 _mean_norm)	0,648 ±0.05	0,627 ±0.07	0,497 ±0.01	0.501 ±0.07	0,703 ±0.02	0,681 ±0.07	0,307 ±0.08	0.341 ±0.02

Table 2. SOC percentage estimation results for each type of computational algorithms. Highlighting that the order of writing is as follows: Algorithm type_Architecture type_Data type (Ref: Reflectance and Abs: Absorbance)_Spectral range type (All: 900-2500 [nm], RE1: 1500-1800 [nm] and RE2: 2000-2300 [nm])_Data augmentation type (Mean, Median, spatial distribution)_Processing type(Normalization, Standardization, Sgolay-Filter).

European dataset, LUCAS, better results are obtained using state-of-the-art algorithms, particularly VGG. This is mainly due to the large amount of data available for model generalization, as well as the high variability in SOC percentages and the variety of crops from which the spectral signatures were extracted. However, when these deep learning algorithms, and even machine learning algorithms, are applied to an environment with fewer spectral signatures, the estimation accuracy decreases. In such cases, the computational algorithm specifically implemented for environments with fewer spectral signatures, such as the Colombian context, performs better. This is because the spectral variability of the signatures is much lower and the range of SOC percentages obtained in this research is narrower. Therefore, a computational algorithm, preprocessing, and learning model that can extract very fine features from the spectral signatures corresponding to SOC percentages is necessary.

6.4. Analysis of the influence of the number of spectral signatures on SOC estimation

Lastly, to validate the behavior of the adapted and proposed computational algorithms concerning the number of available spectral signatures for training, an analysis of the results was conducted by varying the number of samples used. This analysis was performed for both the dataset acquired in Colombia and the available European dataset. By systematically adjusting the sample size, we aimed to understand how the quantity of spectral signatures influences the performance and robustness of the SOC estimation algorithms.

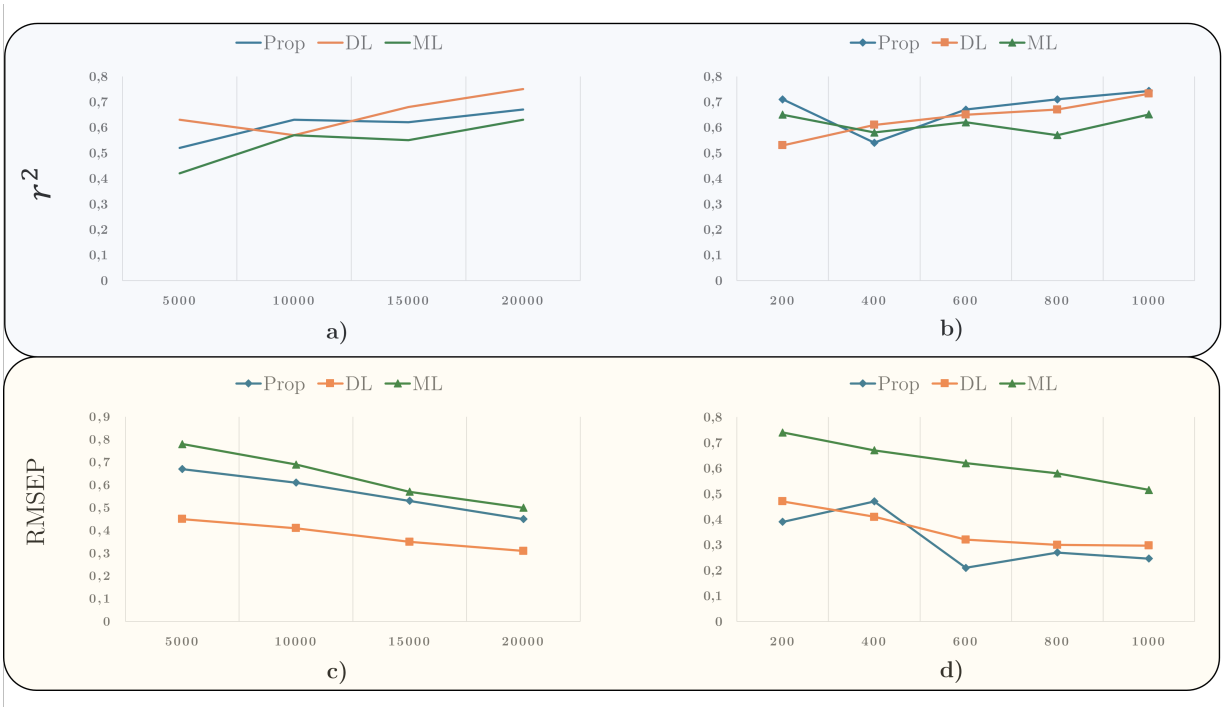


Figure 42. Behavioral analysis of the SOC estimation process depending on the number of spectral signatures used for model training, where **Prop** corresponds to the proposed network, **DL** corresponds to the best deep learning architecture with the best preprocessing combination, and **ML** corresponds of the best Machine learning architecture with best preprocessing combination. a) and c) Results of r^2 and RMSEP in the estimation of SOC from the European LUCAS dataset. b) and d) Results of r^2 and RMSEPP in the estimation of SOC from the Colombian dataset acquired in the HDSP laboratory. It is noteworthy that the best combination of dataset processing from the ablation analysis was used for each type of algorithm such as DL, ML, and the one proposed in the research work.

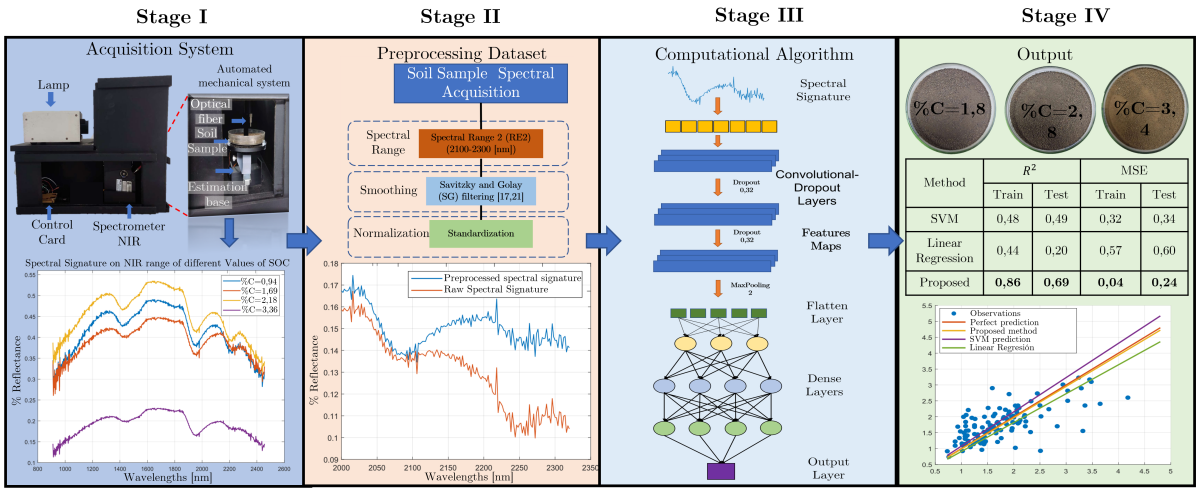
In Figure 42, it can be observed that the accuracy in SOC estimation improves when using the European dataset (LUCAS) to train the algorithms. The results indicate that deep learning algorithms (DL) significantly outperform machine learning algorithms (ML) and the proposed algorithm. However, when applied in a national context where there is a limitation in the number of spectral signatures available (specifically 5 % of the European dataset), the SOC estimation accuracy is higher with the proposed computational algorithm. This is primarily due to several factors,

including enhanced feature extraction capabilities of the proposed algorithm through database processing, which highlights subtle spectral features.

Another contributing factor to these results is the extensive variability present in the European dataset, encompassing spectral, spatial, and temporal aspects, crop types, extraction methods, precipitation diversity, and physical-chemical composition. Therefore, the proposed computational algorithm achieves higher precision in SOC estimation for national spectral signatures compared to algorithms found in the literature.

7. Conclusions and Future Work

This master's thesis research encompasses all the processes required to estimate SOC percentage using an optical computational system. Specifically, the modeling, implementation, calibration, and automation of a whiskbroom-type polar optoelectronic system were carried out, allowing the acquisition of 112 spectral signatures in the field while maintaining experiment repeatability due to its low variability of 0.02. Additionally, using this optical system, a dataset of 1,000 soil samples from the municipality of Simacota was acquired over time and labeled with their physicochemical characteristics, notably the SOC percentage.



Method	R^2		MSE	
	Train	Test	Train	Test
SVM	0,48	0,49	0,32	0,34
Linear Regression	0,44	0,20	0,57	0,60
Proposed	**0,86**	**0,69**	**0,04**	**0,24**

Figure 43. Summary of the computational optical system for SOC percentage estimation in each of the phases carried out in this master's thesis research.

Finally, state-of-the-art computational algorithms for estimating SOC percentage were analyzed, implemented, and adapted, alongside proposing a computational algorithm designed to perform well in environments with limited samples. Based on this, a comparative and ablation study

of SOC estimation performance was conducted for both the acquired dataset and a public dataset with a larger number of samples. The proposed algorithm demonstrated superior estimation in environments with a low number of spectral signatures and comparable accuracy to traditional computational algorithms in environments with a large number of spectral signatures.

Future work includes the systematic improvement of the acquisition system, and analyzing the feasibility of acquiring a smaller number of spectral signatures per soil sample. Additionally, considering that all spectral signatures have other types of physicochemical properties, a correlation analysis between each property could be performed. This would provide additional information to improve SOC estimation.

Bibliography

Abbass, K., Qasim, M. Z., Song, H., Murshed, M., Mahmood, H., and Younis, I. (2022). A review of the global climate change impacts, adaptation, and sustainable mitigation measures. *Environmental Science and Pollution Research*, 29(28):42539–42559.

Amacher, M. C., Henderson, R. E., Brupbacher, R. H., and Jr., J. E. S. (1986). Dichromate-oxidizable and total organic carbon contents of representative soils of the major soil areas of louisiana. *Communications in Soil Science and Plant Analysis*, 17(10):1019–1032.

Angelopoulou, T., Tziolas, N., Balafoutis, A., Zalidis, G., and Bochtis, D. (2019). Remote sensing techniques for soil organic carbon estimation: A review. *Remote Sensing*, 11(6).

Arce, G. R., Brady, D. J., Carin, L., Arguello, H., and Kittle, D. S. (2013). Compressive coded aperture spectral imaging: An introduction. *IEEE Signal Processing Magazine*, 31(1):105–115.

Bacca, J., Martinez, E., and Arguello, H. (2023). Computational spectral imaging: a contemporary overview. *Journal of the Optical Society of America A*, 40(4):C115.

Baumann, P., Lee, J., Frossard, E., Schönholzer, L., Diby, L., Hgaza, V., Kiba, D., Sila, A., Sheperd, K., and Six, J. (2021). Estimation of soil properties with mid-infrared soil spectroscopy across yam production landscapes in west africa. *Soil*.

Ben Dor, E., Ong, C., and Lau, I. C. (2015). Reflectance measurements of soils in the laboratory: Standards and protocols. *Geoderma*, 245-246:112–124.

Berhe, A., Carrillo, Y., Cavagnaro, T., Chen, D., Chen, Q., Román Dobarco, M., Dijkstra, F., Field, D., Grundy, M., He, J.-Z., Hoyle, F., Kögel-Knabner, I., Lam, S., Marschner, P., Martinez, C., Mcbratney, A., McDonald-Madden, E., Menzies, N., and Minasny, B. (2022). Ensuring planetary survival: the centrality of organic carbon in balancing the multifunctional nature of soils. *Critical Reviews in Environmental Science and Technology*, 52:1–17.

Bojago, E., Delango, M. W., and Milkias, D. (2023). Effects of soil and water conservation practices and landscape position on soil physicochemical properties in anuwa watershed, southern ethiopia. *Journal of Agriculture and Food Research*, 14:100705.

Cao, Y., Yang, W., Li, H., Zhang, H., and Li, M. (2024). Development of a vehicle-mounted soil organic matter detection system based on near-infrared spectroscopy and image information fusion. *Measurement Science and Technology*, 35(4):045501.

Carvalho, M., Cardoso-Fernandes, J., Lima, A., and Teodoro, A. C. (2024). Convolutional neural networks applied to antimony quantification via soil laboratory reflectance spectroscopy in northern portugal: Opportunities and challenges. *Remote Sensing*, 16(11).

Caten, A., Dalmolin, R., Dotto, A., Moura-Bueno, J., Boeing, E., Safanelli, J., Silva, W., and Boesing, B. (2016). Digital soil morphometrics via a low-cost radiometer for estimating soil organic carbon and texture. *Springer Environmental Science and Engineering*, pages 249–257.

Chabrillat, S., Gholizadeh, A., Neumann, C., Berger, et al. (2019). Standards and protocols for the

reflectance measurements of soils in the laboratory: Influence of different laboratory humidity conditions and set-ups. *Geophysical Research Abstracts*, 21.

Diaz, F. J., Ahmad, A., Parra, L., Sendra, S., and Lloret, J. (2024). Low-cost optical sensors for soil composition monitoring. *Sensors*, 24(4):1140.

Egeonu, D. and Jia, B. (2024). Performance evaluation and comparison of deep neural network models for african soil properties prediction. *Communications in Soil Science and Plant Analysis*, 55:1–22.

Gaget, H. (2021). 10 datos sobre la agricultura en colombia (y 2 napas) mas colombia.

Garcia Galvis, J. and Ballesteros Gonzalez, M. I. (2005). Evaluacion de parametros de calidad para la determinacion de carbono organico en suelos. *Revista Colombiana de quimica*, 34:201 – 209.

Gat, N. (2000). Imaging spectroscopy using tunable filters: a review. In *Wavelet Applications VII*, volume 4056, pages 50–64. International Society for Optics and Photonics.

Georgiou, K., Jackson, R., Vindušková, O., Abramoff, R., Ahlström, A., Feng, W., Harden, J., Pellegrini, A., Polley, H., Soong, J., Riley, W., and Torn, M. (2022). Global stocks and capacity of mineral-associated soil organic carbon. *Nature Communications*, 13:3797.

Gomez, P., Camacho, A., and Arguello, H. (2022). Design and implementation of an automated protocol for spectral signatures acquisition on colombian agricultural soil samples into the visible and infrared range. In *2022 IEEE ANDESCON*, pages 1–6.

Huan, J., Yuan, J., Zhang, H., Xu, X., Shi, B., Zheng, Y., Li, X., Zhang, C., Hu, Q., Fan, Y., et al. (2024). Identification of agricultural surface source pollution in plain river network areas based on 3d-eems and convolutional neural networks. *Water Science & Technology*, 89(8):1961–1980.

Hutengs, C., Seidel, M., Oertel, F., Ludwig, B., and Vohland, M. (2019). In situ and laboratory soil spectroscopy with portable visible-to-near-infrared and mid-infrared instruments for the assessment of organic carbon in soils. *Geoderma*, 355:113900.

Illumination Technologies, I. (2024). 3900e dc regulated lightsource with digital light feedback.

Izquierdo Bautista, J. and Arevalo Hernandez, J. J. (2021). Determinacion del carbono organico por el metodo quimico y por calcinacion. *Ingenieria y Region*, 26:20–28.

Jiang, Q., Li, Q., Wang, X., Wu, Y., Yang, X., and Liu, F. (2017). Estimation of soil organic carbon and total nitrogen in different soil layers using vnir spectroscopy: Effects of spiking on model applicability. *Geoderma*, 293:54–63.

Kinoshita, R., Roupsard, O., Chevallier, T., Albrecht, A., Taugourdeau, S., Ahmed, Z., and van Es, H. M. (2016). Large topsoil organic carbon variability is controlled by andisol properties and effectively assessed by vnir spectroscopy in a coffee agroforestry system of costa rica. *Geoderma*, 262:254–265.

Maiwald, M., Sowoidnich, K., and Sumpf, B. (2022). Portable shifted excitation raman difference spectroscopy for on-site soil analysis. *Journal of Raman Spectroscopy*, 53.

Miao, T., Ji, W., Li, B., Zhu, X., Yin, J., Yang, J., Huang, Y., Cao, Y., Yao, D., and Kong, X. (2024). Advanced soil organic matter prediction with a regional soil nir spectral library using long short-term memory–convolutional neural networks: A case study. *Remote Sensing*, 16(7):1256.

Mickelson, A. (2018). Guided wave optics. In Guenther, B. D. and Steel, D. G., editors, *Encyclopedia of Modern Optics (Second Edition)*, pages 221–228. Elsevier, Oxford, second edition edition.

Ng, W., Minasny, B., Mendes, W. D. S., and Demattê, J. A. M. (2020). The influence of training sample size on the accuracy of deep learning models for the prediction of soil properties with near-infrared spectroscopy data. *SOIL*, 6(2):565–578.

Niu, S., Lyu, X., Gu, G., Peng, W., Wang, Y., Xue, P., and Solodovnikov, S. Y. (2024). A framework for quantification and integration of green development of cultivated land in china: From the perspective of adaptability-vitality-resistance. *Land Degradation & Development*.

Nocita, M., Stevens, A., van Wesemael, B., Aitkenhead, M., Bachmann, M., et al. (2015). Chapter four - soil spectroscopy: An alternative to wet chemistry for soil monitoring. *Advances in Agronomy*, 132:139–159.

Oliveira, K. M. d., Gonçalves, J. V. F., Furlanetto, R. H., Oliveira, C. A. d., Mendonça, W. A., Haubert, D. d. F. d. S., Crusiol, L. G. T., Falcioni, R., Oliveira, R. B. d., Reis, A. S., et al. (2024). Predicting particle size and soil organic carbon of soil profiles using vis-nir-swir hyperspectral imaging and machine learning models. *Remote Sensing*, 16(16):2869.

Orgiazzi, A., Ballabio, C., Panagos, P., Jones, A., and Fernández-Ugalde, O. (2017). Lucas soil, the largest expandable soil dataset for europe: A review. *European Journal of Soil Science*, 69.

Özbolat, O., Sánchez-Navarro, V., Zornoza, R., Egea-Cortines, M., Cuartero, J., Ros, M., Pascual, J. A., Boix-Fayos, C., Almagro, M., de Vente, J., et al. (2023). Long-term adoption of reduced tillage and green manure improves soil physicochemical properties and increases the abundance of beneficial bacteria in a mediterranean rainfed almond orchard. *Geoderma*, 429:116218.

Pavlovic, M., Ilic, S., Ralevic, N., Antonic, N., Raffa, D. W., Bandecchi, M., and Culibrk, D. (2024). A deep learning approach to estimate soil organic carbon from remote sensing. *Remote Sensing*, 16(4):655.

Piccini, C., Metzger, K., Debaene, G., Stenberg, B., Götzinger, S., Borvka, L., Sandén, T., Bragazza, L., and Liebisch, F. (2024). In-field soil spectroscopy in vis–nir range for fast and reliable soil analysis: A review. *European Journal of Soil Science*, 75(2):e13481.

Ramirez, P., Calderon, F., Haddix, M., Lugato, E., and Cotrufo, M. F. (2021). Using diffuse reflectance spectroscopy as a high throughput method for quantifying soil c and n and their distribution in particulate and mineral-associated organic matter fractions. *Frontiers in Environmental Science*, 9.

Rozenstein, O., Paz Kagan, T., Salbach, C., and Karnieli, A. (2014). Comparing the effect of preprocessing transformations on methods of land-use classification derived from spectral soil

measurements. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, PP:1–12.

Santana, F. B., De Souza, A., and Poppi, R. (2017). Visible and near infrared spectroscopy coupled to random forest to quantify some soil quality parameters. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 191.

Sato, J., Figueiredo, and et al (2014). Methods of soil organic carbon determination in brazilian savannah soils. *Scientia Agricola*, 71:302–308.

Seema, Ghosh, A., Das, B., and Reddy, N. (2020). Application of vis-nir spectroscopy for estimation of soil organic carbon using different spectral preprocessing techniques and multivariate methods in the middle indo-gangetic plains of india. *Geoderma Regional*, 23:e00349.

Sharma, N. A., Kumar, K., Chand, R. R., and Kabir, A. (2024). Utilizing hyperspectral imaging with machine learning techniques for soil analysis. *Computational Intelligence Based Hyperspectral Image Analysis*.

Shi, X., Song, J., Wang, H., Lv, X., Zhu, Y., Zhang, W., Bu, W., and Zeng, L. (2023). Improving soil organic matter estimation accuracy by combining optimal spectral preprocessing and feature selection methods based on pxrf and vis-nir data fusion. *Geoderma*, 430:116301.

Shi, Z., Wang, Q., Peng, J., Ji, W., Liu, H., Li, X., and Viscarra Rossel, R. (2014). Development of national vnir soil-spectral library for soil classification and the predictions of organic matter. *Science China Earth Sciences*, 57:1671–1680.

Smith, P., Soussana, J.-F., Angers, D., Schipper, L., Chenu, C., Rasse, D., Batjes, N., van Egmond, F., McNeill, S., Kuhnert, M., Arias-Navarro, C., Olesen, J., Chirinda, N., Fornara, D., Wollenberg, E., Álvaro Fuentes, J., Sanz-Cobena, A., and Klumpp, K. (2019). How to measure, report and verify soil carbon change to realise the potential of soil carbon sequestration for atmospheric greenhouse gas removal. *Global Change Biology*, 26.

SORIANO DISLA, J., Janik, L., Viscarra Rossel, R., Macdonald, L., and McLaughlin, M. (2014). The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties. *APPLIED SPECTROSCOPY REVIEWS*, 49:139–186.

Tian, C., Chen, X., Ren, Y., Yang, Y., Wang, M., and Bai, X. (2024). Spectral characteristics and displacement sensing of u-shaped single-mode–multimode–single-mode fiber structure. *Sensors*, 24(10):3184.

Umeda, T., Miyaji, N., Nakazawa, S., Miwa, K., Wagatsuma, K., Motegi, K., Takiguchi, T., and Koizumi, M. (2017). A comparison of planar sensitivity and spatial resolution among different collimators and energy windows on 223ra imaging. *Japanese Journal of Radiological Technology*, 73:1132–1139.

Vairavan, C., Kamble, B., Durgude, A., Ingle, S. R., and Pugazenthi, K. (2024). Hyperspectral imaging of soil and crop: A review. *Journal of Experimental Agriculture International*, 46(1):48–61.

Vane, G., Green, R. O., Chrien, T. G., Enmark, H. T., Hansen, E. G., and Porter, W. M. (1993).

The airborne visible/infrared imaging spectrometer (aviris). *Remote sensing of environment*, 44(2-3):127–143.

Wang, S., Guan, K., Zhang, C., Lee, D., Margenot, A. J., Ge, Y., Peng, J., Zhou, W., Zhou, Q., and Huang, Y. (2022). Using soil library hyperspectral reflectance and machine learning to predict soil organic carbon: Assessing potential of airborne and spaceborne optical soil sensing. *Remote Sensing of Environment*, 271:112914.

Wang, Y. W., Reder, N. P., Kang, S., Glaser, A. K., and Liu, J. T. (2017). Multiplexed optical imaging of tumor-directed nanoparticles: a review of imaging systems and approaches. *Nanotheranostics*, 1(4):369.

Wang, Z., Chen, S., Lu, R., Zhang, X., Ma, Y., and Shi, Z. (2024). Non-linear memory-based learning for predicting soil properties using a regional vis-nir spectral library. *Geoderma*, 441:116752.

Xu, M., Chu, X., Fu, Y., Wang, C., and Wu, S. (2021). Improving the accuracy of soil organic carbon content prediction based on visible and near-infrared spectroscopy and machine learning. *Environmental Earth Sciences*, 80.

Yang, X., Bao, N., Li, W., Liu, S., Fu, Y., and Mao, Y. (2021). Soil nutrient estimation and mapping in farmland based on uav imaging spectrometry. *Sensors*, 21(11).

Yanosky, J. and Macintosh, D. (2001). A comparison of four gravimetric fine particle sampling methods. *Journal of the Air and Waste Management Association*, 51:878–84.

Yeh, Y.-F., Dhurumraj, T., and Ramnarain, U. (2023). Representations of the nature of science in south african physical sciences textbooks on electricity and magnetism. *Science & Education*, 32(5):1537–1559.

Zhang, B., Gao, S., Jia, F., Liu, X., and Li, X. (2020). Categorization and authentication of beijing-you chicken from four breeds of chickens using near-infrared hyperspectral imaging combined with chemometrics. *Journal of Food Process Engineering*, 43.

Zhang, W., Kasun, L. C., Wang, Q. J., Zheng, Y., and Lin, Z. (2022). A review of machine learning for near-infrared spectroscopy. *Sensors*, 22(24).

Zhao, D., Arshad, M., Wang, J., and Triantafilis, J. (2021). Soil exchangeable cations estimation using vis-nir spectroscopy in different depths: Effects of multiple calibration models and spiking. *Computers and Electronics in Agriculture*, 182:105990.

Zhong, L., Guo, X., Xu, Z., and Ding, M. (2021). Soil properties: Their prediction and feature extraction from the lucas spectral library using deep convolutional neural networks. *Geoderma*, 402:115366.