THESIS

Mohamed Yassine Zaghdoudi

MSc Environmental Engineering

Gödöllő

2023

Prediction of certain soil parameters using reflectance spectroscopy and legacy soil data

Primary Supervisor: Dr. Ádám Csorba

Author: Mohamed Yassine Zaghdoudi

Faculty of Agricultural and Environmental Science

Gödöllő

2023

TABLE OF CONTENTS

TABLE OF FIGURES

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to Dr. Ádám Csorba, my thesis supervisor, for his patient guidance, unwavering support, and helpful counsel. His vast knowledge and wealth of experience have inspired me throughout my academic career and everyday life. Additionally, I want to thank my parents and my friends for everything. Without their amazing support and understanding over the past few years, this accomplishment would not have been possible.

1. Introduction

Soil is an essential part of the Earth's critical zone. Since it offers a range of ecological services and is an essential component for numerous human endeavors, it is a crucial natural resource for supporting life on Earth and economic growth [2]. More specifically, soil acts as a water filter, distributes nutrients to plants, and serves as a source of food, fiber, and energy for humans. It also contributes to human well-being and living conditions while storing carbon and influencing greenhouse gas emissions. Moreover, soil significantly impacts our climate. Therefore, preserving and sustainably managing soil are vital actions required to address pressing global concerns like ensuring food availability, mitigating climate change effects, combating environmental degradation, overcoming water scarcity issues, and safeguarding biodiversity [3].

Soil is likewise seen as a heterogeneous system, with complicated processes and systems that are difficult to completely explain. Furthermore, such functions of soil are being jeopardized by massive pressures from urbanization and deterioration, as well as agroecological balances and food security [4].

Soil assessment necessitates complicated analytical procedures with many features at numerous sites. unfortunately, soil surveyors occasionally lack consistency in their analytical and methodology practices [3], hampering the cross-disciplinary exchange of quantitative data and the execution of policies aimed at minimizing the major soil hazards. As a result, the increased need for high-resolution soil data covering huge regions is challenging to provide [5].

Traditional laboratory methods for analyzing soil parameters are often unworkable since they are time-consuming, costly, and occasionally inaccurate [6]. Often, these approaches frequently necessitate extensive sample preparation, the use of potentially dangerous chemicals, and complex apparatus that is unsuitable when several measurements are required, such as for soil mapping, monitoring, and modeling. Historically, this kind of laboratory examination has helped us understand the soil system and evaluate its quality and activity. We need to improve our analytical skills in order to better comprehend the soil as an overall system as well as a resource that we can utilize with greater efficiency while also preserving it for future generations. This is more crucial than ever before since acquiring bigger volumes of precise soil data is needed in order to manage our main resources sustainably and meet the needs of the next generations for nutrition and fiber

[7]. Furthermore, there are still doubts and disagreements about present procedures and their outcomes, which can lead to misinterpretations and misleading information. These difficulties necessitated study into other ways for optimizing or assisting these previously significant wet approaches [8].

Spectroscopic techniques such as visible (VIS), near-infrared (NIR), and mid-infrared (MIR) spectroscopy are being examined as potential solutions to supplement or substitute conventional laboratory methods of examining soil. Most of these approaches are non-destructive, allowing the essential integrity of the soil system to be preserved, and they may efficiently describe soil [9]. Spectroscopic measurements are quick, accurate, and affordable. The spectra contain data pertaining to the soil's fundamental makeup, which includes minerals, organic molecules, and water. In reaction to its surroundings and human intervention, soil acquired properties throughout its formation from its parent material, including minerals and tightly bound water [4]. All of these encodings become apparent in the spectra as an intake at certain frequencies of electromagnetic light, and their assessments may be used to subjectively and quantitatively characterize soil. Spectroscopy can also offer data regarding the dimension of soil particles, and consequently the soil matrix. Another appealing aspect of spectroscopy is that spectra may be obtained at different locations or by scanning various platforms, such as proximate sensing in the field, in the laboratory utilizing collected material, or from platforms for imaging with multi and hyperspectral features[10].

As we progress through the pages of this thesis, we will embark on an adventure to discover various aspects of soil spectroscopy. In the next chapters, we will look at the fundamental ideas and procedures that underpin this technology. We will investigate the methodology, data analysis, and practical applications made available by soil spectroscopy. These studies seek to address critical concerns about the ability to identify characteristics of soil, predict soil quality, and have an influence on a wide range of disciplines, from precision agriculture to environmental sustainability and mineral resource management. We will examine the landscape of soil spectroscopy, looking at its potential, limitations, and future prospects.

2. Literature review

2.1 Visible and infrared spectroscopy

Spectroscopy, the art of deciphering the secrets contained within light, is a fundamental instrument for unraveling the mysteries of matter. We will look at the detailed procedures of visible and infrared spectroscopy in this chapter. These techniques operate similarly to expert detectives, retrieving crucial information from substances' particular spectral fingerprints. We equip ourselves with these exact instruments as we begin our study through the nuances of visible and IR spectroscopy in our narrowed examination. We unravel Earth's well-guarded secrets with these powerful approaches, providing significant insights with wide-ranging implications in agriculture, ecology, and geology [11].

2.1.1 The visible

Only a small fraction of the electromagnetic range is covered by visible light. It has a wavelength between 0.4 to 0.7 μm. When visible light interacts with soil, it causes energy shifts in the atoms, typically via electron interactions such as the crystal field effect and charge transfer. There is also a scattering effect that takes place over the visible range. The broad spectrum of absorption produced by these electron processes in soils determines the color of the soil at visible wavelengths while scattering effects alter the spectrum's albedo sequence's baseline. Even though the spectral response in the Vis area appears imprecise, it is possible to infer quantitative and qualitative information from this data that is not visible to the human eye. [12].

2.1.2 The infrared

Infrared radiation has a wavelength range of 0.7 μm to 1 mm. When it comes to molecules, their vibration energy transitions normally need a frequency in the infrared region. As a result, molecular interatomic vibrations can potentially be accomplished by using infrared radiation which serves as the foundation of the infrared spectroscopy method. An IR spectrum, in essence, offers a chemical characterization of the given sample [13]. Both electric and magnetic components are contained in electromagnetic radiation, however, the electrical vector of infrared rays interacts with bonds between atoms inside molecules to initiate distinct vibrations, which leads to the absorption of infrared light. When infrared radiation is absorbed, a variety of molecular vibrations take place, such as bending, stretching, and even wagging motions among the atoms that make up the molecule [14]. To exhibit IR activity, a compound must have covalent bonds and

be subjected to an electric field that oscillates while atomic bonds vibrate. IR radiation is absorbed owing to molecular vibrations in particular types of chemical bonds such as O-H, C-H, and C-N. In contrast, symmetric bonds with equal electron sharing do not produce IR-active vibrations. The spectrum of infrared radiation, shown in Figure 1 [15], represents the IR three primary regions: near-infrared (NIR) spanning 0.70-2.5 μm, mid-infrared (MIR) spanning 2.5-25 μm, and far-infrared (FIR) spanning 25-1000 μm [16]. The implications and combined modes of basic atom vibrations that are active in the MIR and FIR areas are mostly seen in the NIR region. NIR spectroscopy is employed extensively in a variety of sectors, including food science, semiconductor electronics, pharmaceutics, chemical identification, and soil quality study. The fundamental vibration of the functional groups that make up a material can be used to explain its absorption in MIR spectra. [17]. The MIR predominantly reflects atom vibrational modes such as wagging, stretching, and twisting. Furthermore, the MIR spectral region, like the NIR spectral region, has lately acquired prominence in soil applications.

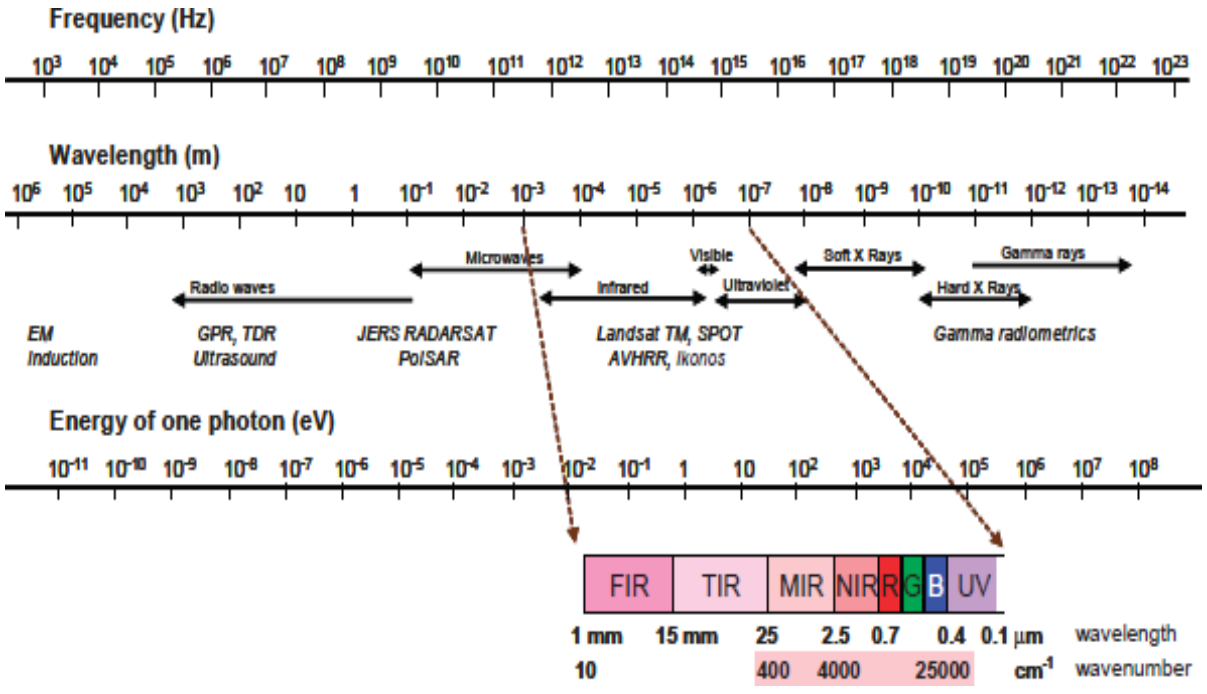

Figure 1: The electromagnetic spectrum

2.2 Soil visible and infrared spectroscopy

Understanding soil spectroscopic techniques is important for discovering the numerous mysteries lying under the Earth's surface. Soil spectroscopy is a useful instrument for deciphering the complexity of soil composition, structure, and condition. This chapter takes you on a scientific and technological heading into the foundations of soil spectroscopy, delving into the methods and concepts that allow us to discover the hidden characteristics of soil [18]. We are able to discover more about it by examining its spectral attributes. We provide the basis in this framework for addressing the different uses of soil spectroscopy and its relevance across various sectors, developing our understanding of Mother Earth beneath our feet. Figure 2 [19] depicts a flowchart of the technique's key steps, from preparation of the soil sample to prediction.

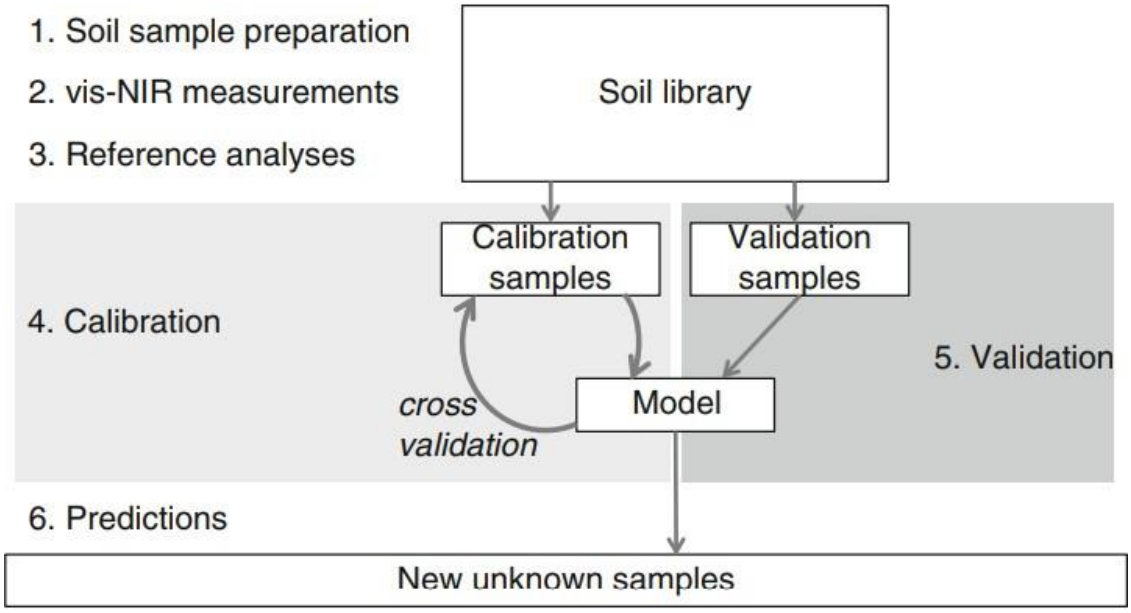

Figure 2: Procedure for developing vis-NIR prediction models.

2.2.1 Instruments

It might be difficult to select a suitable vis-NIR spectrophotometer. Various manufacturers provide a variety of alternatives with varying benefits. The selection procedure is dependent on the particular application, with a cost-performance trade-off. In scientific contexts, high-resolution equipment with a resolution of 10 nm or greater is desirable. It should be emphasized, however, that resolution and noise are inversely related. It is advised for scientific purposes to have a spectral

range that includes both the visible and the full NIR region in order to capture critical wavelength bands. However, for certain specialized applications, the entire vis-NIR spectrum may not be required.

Flexibility is critical in the choosing of instruments. Several criteria, such as the desired usage environment, must be addressed. Instruments used in laboratories have different needs than those used for outside measurements. Outdoor applications require toughness and ease of handling, whereas laboratory devices may prioritize other attributes. Furthermore, the nature of the samples being analyzed affects instrument selection, especially when working with multiple sample varieties. For outdoor or online measurements, the flexibility to show samples in a variety of ways becomes critical. Because of its adaptability in these situations, fiber optics are frequently selected for spectral collecting. It is recommended to utilize post-dispersive equipment to assist in reducing the influence of ambient stray light [20].

2.2.2 Methods

Soil sample preparations

Soil spectroscopy requires a two-step sample preparation procedure similar to that used for chemical and physical studies. These methods offer practical benefits and allow for accurate calibrations even with moist soil samples. Some cases have demonstrated the benefits of calibrating field-moist samples, particularly when consistent remoistening processes are used. However, due to improved standardization and less interference from water bands, dry soil calibrations performed in a laboratory environment outperform those performed on field-moist soil.

Soil spectroscopy relies significantly on crushing and screening soil samples. These methods remove stones and plant residues while allowing for representative subsampling [21]. Furthermore, further grinding and screening might result in a more uniform particle size, which affects spectrum outputs. It is worth mentioning that grinding can significantly improve reflectance, especially for clay samples. Yet, this impact can be minimized by doing pre-treatment activities prior to spectral analysis.

Measurements

It is common for each instrument to have a sample display arrangement and appropriate sample containers of its own. To put it simply, the general advice is to adhere to the instructions specific to the instrument. But below are some general things to keep in mind. [22]:

Sample Presentation and Handling

- Take measurements on a representative section of the soil sample because soils are highly heterogeneous.

- A setup that allows for scanning of a significant portion of the sample is preferable.

- Use repeat spectral sampling for small sampled regions.

- Never shake the sample since it might cause the particles to stratify.

- Instead, evenly flatten the sample surface with a tool.

-Maintain uniformity in terms of volume and pressure while packaging soil samples in containers.

- Do not utilize water or alcohol/organic solvents to clean the container between samples.

-Using a dry, dust-free tissue, clean the measurement window that comes into touch with the soil.

White and Dark Reference

- White and dark references can be taken automatically or manually, depending on the instrument.

- It is critical to take white and dark references to ensure high-quality spectra.

- In numerous instruments, white and dark references should be taken every ten minutes.

-When a white reference is taken, certain instruments automatically take a dark reference.

- If utilizing an external white (or dark) reference, make sure it matches the sample measurement settings.

- The white reference should have 100% reflection at all wavelengths between 400 and 25,000 nm.

Minimizing External Light Sources

- During measurements, reduce or regulate any other sources of light to avoid affecting results.

-Controlling sources such as fluorescent light and ambient light from windows that may interfere with readings is a significant task for these measurements.

Pre-treatment of the spectra

Several essential steps in the field of soil analysis utilizing visible-near infrared (vis-NIR) spectroscopy can considerably improve the dependability of the gathered results. To begin, compute the average spectra of many scans on the same soil sample. This method not only eliminates the impact of false duplicates in later studies but also increases the signal-to-noise ratio, making measurements more robust and precise.

Following the averaging of spectra, the observed reflectance data should be transformed, specifically by calculating the logarithm of the reciprocal of reflectance. This transformation contributes to the development of a more linear connection between the measured absorbance and the concentration of the chemical elements of interest. This improves the alignment of the spectral data with the required analytical aims.

Additional spectral pre-processing techniques are frequently advocated in the search for more chemically relevant peaks and the minimization of disruptive variables such as baseline shifts and overall curvature. This is especially important given the variety of soil data sets and the individuality of each analysis endeavor. Several such approaches are widely available in specialist spectroscopic software, but no one-size-fits-all solution exists. As a result, it is recommended to conduct a systematic review by evaluating several transformations on a representative calibration set and adjusting the selection to the project's unique requirements. This method ensures that the soil analysis is both trustworthy and precisely linked with the chemical goals at hand [23].

2.2.3 Calibration and validation

There are several ways available for calibrating soil vis-NIR spectra in order to predict soil characteristics. Multiple linear regression (MLR), principal component regression (PCR), and partial least squares regression (PLS) are examples. They all have pros and limitations, and we will not provide particular suggestions on which to use, although the linear ones are the easiest and

most often utilized. Simultaneously, the usage of data mining is growing, particularly for big diversified data sets, where it is believed to serve somewhat better than linear analysis [23]. Instead, we will provide some fundamental guidance on factors to take into account when choosing calibration samples and verifying the model.

Calibration Set

Creating an effective calibration set for soil spectroscopy is pivotal. The set must encompass the relevant variations found in the data's intended use. If the model is meant for a countrywide application, it should cover all existing soil types. Conversely, for a local-scale model, capturing the local variation is essential, and including irrelevant soil types should be avoided.

The number of calibration samples required is determined by the scope of the variance. In general, more samples result in a more robust model. For extensive geographical coverage and various soil types, 100 to 200 calibration samples are recommended, but farm or field-scale predictions can be made with as few as 25. The needed sample count varies with the desired variable; directly measured qualities such as clay may necessitate fewer samples, whereas indirectly measured features such as vis-NIR absorbing traits may necessitate more samples [24].

Creating a soil spectral library necessitates a thorough soil sampling technique, regardless of scale. It is critical to capture the necessary variety. When working with a significant volume of vis-NIR scanned samples, selecting calibration samples based on their spectra aids in capturing a wide range of dataset variance. The calibration set should ideally have an equal distribution, which may be accomplished through techniques like the Kennard and Stones uniform mapping.

Validation

To assess calibration accuracy effectively, an independent validation set is vital. It is supposed to be sampled and analyzed separately, ideally not sharing the same sampling time or strategy as the calibration set. If you lack a separate validation set, consider these guidelines:

- For field and farm-scale analyses, include all samples within a soil profile in either the calibration or validation set to avoid dependence.
- For regional or global-scale analyses, avoid using geographically clustered samples that may introduce dependence. If clusters are used, ensure every sample from a given cluster

exhibits the same pattern in either the calibration or validation set to prevent overly optimistic predictions.

Within the calibration process, an internal validation step is employed to refine the model's performance. This internal validation serves various purposes, such as determining the optimal number of components and selecting the most informative wavelengths. It can also aid in optimizing and refining the calibration model. Popular techniques for internal validation include cross-validation and bootstrapping, which systematically test the model's performance on a subset of data while using the remaining data for calibration. When it comes to the ratio of calibration to validation samples, there is no universally fixed rule [25]. Nevertheless, a common practice involves a ratio of 2/3 calibration samples to 1/3 validation samples, which provides a useful benchmark.

Model Assessment

Numerous statistical measures describe the accuracy of the estimations. To account for prediction accuracy and imprecision, we propose using the root mean squared error (RMSE), bias, and standard deviation of the error distribution (SDE), as well as the ratio of performance to deviation (RPD) for evaluations across units [19].

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (\hat{y}_i - y_i)^2}{N}}$$

$$Bias = \sqrt{\frac{\sum_{i=1}^{N} (\hat{y}_i - y_i)}{N}}$$

$$SDE = \sqrt{\frac{\sum_{i=1}^{N} (\hat{y}_i - y_i - Bias)^2}{N - 1}}$$

$$RPD = \frac{Standard\ Deviation}{RMSE}$$

For sample i, y is the measured value with N samples and $\hat{y}_i$ is the forecasted value.

Software

A profusion of specific commercial software programs is easily accessible in the field of spectral data analysis and calibration, intentionally developed to simplify the process and give user-friendly functionality for these jobs. Furthermore, many instrument makers provide specialized software with their hardware, providing easy integration and frequently allowing real-time forecasts. These analytical and calibration techniques are also accessible via a variety of software options, including commercial programs, shareware, and open-source platforms like the R-project [23].

3. Materials and methods

3.1 Methods of spectral-based soil parameter prediction

This part of the study endeavored to find out whether the conducted VIS–NIR spectrum analysis information could potentially be used to predict soil characteristics. The importance of being able to investigate the predictability of cation exchange capacity (CEC), pH, humus content, and clay content for 400 soil samples collected from the soil archives of the laboratory of the Department of Soil Science (MATE). Conducting reference laboratory analysis is an essential requirement for the multivariate calibration in addition to the spectral measurements. Mathematical-statistical chemometric models can be developed based on laboratory reference clay content, humus content, pH, and CEC evaluations on assigned calibration samples covering the whole dataset's spectral variability, as well as on their reflectance spectrum. According to the spectral reflectance of samples whose composition is unknown, the proposed models may be used to forecast these soil attributes. The primary goal of using chemometric approaches for soil research is to partially or completely replace traditional laboratory procedures with an easy-to-use and quick method. This will make it possible to determine the attributes of a given number of samples more quickly and affordably, or to analyze more samples in a given amount of time and within a certain amount of money.

3.1.1 Exploration of Terrain and Soil Sampling

The soil samples were collected by the department staff as a part of a broader project (Agrotechnology National Laboratory project) at The Hungarian University of Agriculture and Life Sciences' Soil Science Department. To guarantee representative sampling, rigorous procedures, and guidelines were followed while collecting soil samples. Samples were gathered, allowed to air dry, and then sieved using a 2 mm sieve in preparation for additional examination. The objectives of this procedure were to preserve sample integrity and provide details about the characteristics of the soil in the research region.

3.1.2 Spectral and Laboratory Reference Analysis

The Analytical Spectral Devices (ASD) LabSpec 4 Hi-res portable spectroradiometer from the Department of Soil Science was utilized for the spectroscopic measurements in the lab. The spectra have been obtained by the device in the 350–2500 nm spectral range. 379 soil samples were collected, and the Muglight probe attachment was employed to gauge the spectral reflectance of each sample. This probe allowed for direct contact with the samples, minimizing environmental factors that might have a detrimental impact on the measurement quality. Additional samples were selected in accordance with the chemometric analysis findings and the visual interpretation of the spectra.

3.2 Data Acquisition and Pre-processing

3.2.1 Spectral Acquisition

R Studio was used to import the unprocessed spectral data and visualize it. Any required data wrangling was done when the data structure and variable identification (reflectance and wavelength) were verified. This could comprise managing missing values, changing data formats, or choosing relevant columns. Lastly, the wavelength values were shown using the x-axis, and the reflectance values on the y-axis. The raw spectra are subsequently shown in the plot created in R Studio, as seen in Figure 3 below.

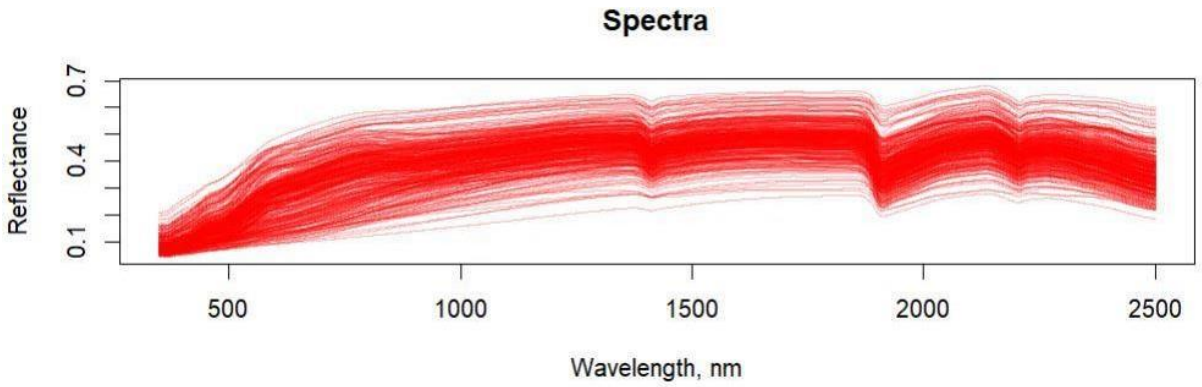

Figure 3: Raw spectra

3.2.2 Pre-processing Methods

Reflectance spectra into absorbance spectra transformation

Initially, as a main step of spectroscopic transformations, we need to transform the spectral reflectance data to spectral absorbance by simply using the formula below. It is an important step for enhancing spectral features and for further preprocessing procedures.

$$Absorbance = log\left(\frac{1}{Reflectance}\right)$$

Signal processing

In this section, signal processing attempts to reduce unwanted spectrum variance to enhance the quality of the spectra before modeling. It enhances both the quantitative and qualitative spectral model outcomes in a comparable manner. Particular precautions must be used while selecting methods to prevent the loss of valuable physical information. The intended outcome and the NIR data are always taken into consideration while choosing the method to be used.

Noise removal

In this section, a digital filter called a Savitzky-Golay (SG) filter was employed to smooth the data and eliminate noise. The method involves fitting a low-degree polynomial to a window of neighboring data points, and then substituting the fitted value for the center data point. The envisioned cutoff frequency of the filter and the polynomial's order define the window size, which equals 17 in this case. Other methods can be opted for noise removal like splice correction and moving average techniques. However, the selection of Savitzky-Golay for the given data was encouraged by the ability to calculate derivatives without making considerable distortion.

Differentiation (First-Derivative)

We can pair smoothing with differentiation, as discussed in the preceding section simply because differentiation increases spectral noise. The Savitzky-Golay function can be employed also to apply differentiation, it smooths out the coefficients of the polynomial in the moving window so that the noise is not significantly amplified. Moreover, considering a fixed band width, the first derivative of a spectrum can likely be calculated using the finite difference technique (difference

between successive data points). This approach improved small spectral absorption characteristics, while contributing to resolving absorption overlaps, and improved the dataset's prediction accuracy.

Scatter correction

Multiplicative Scatter Correction

The multiplicative impact of scattered light on the NIR spectra is lessened by the Multiplicative Scatter Correction (MSC) approach. More specifically, it's a technique that assumes that each spectrum in an identified set of spectra can have its offset and scattering effects eliminated. Regressing each spectrum onto the mean spectrum yields this model. Based on the average spectrum of the collection of spectra, it can additionally be regarded as an offset and slope correction of the spectra, which is considered important for the prediction accuracy.

Standard Normal Variate

This approach is regarded as another straightforward method for spectra normalization that aims to account for light scatter. However, it's implemented in a row-wise manner compared to the Multiplicative Scatter Correction technique. Within SNV's advantages are its capacity to eliminate physical occurrences from spectra, improve model interpretation and forecast accuracy, and enhance spectral linearity correction.

3.3 Calibration sampling

In an effort to solve the multicollinearity issue and to make it easier to identify the ideal sample configuration that accurately captures the spectral variability, sampling methods are often employed in the principal component (PC) space of the spectral variables. Because it performs better with fewer variables, as in this study, it is thus required to calculate the Primary components (PCs) of the spectra before performing the calibration sampling to decrease error prediction, represent underlying variability, and reduce the number of variables.

3.3.1 Sample Size Selection

For the purpose of calibration set size, we can use different methods, which are: Kennard-Stone sampling (KSS), K-means sampling (KMS), and Conditioned Latin hypercube sampling (cLHS).

Each of these approaches has advantages and disadvantages. In this study, it is remarkably appropriate to use the KSS method since it provides a reliable means of selecting calibration samples that evenly cover the distribution of predicted probabilities, despite its simplicity compared to alternative methods.

Based on these steps, the ideal sample size for calibrating the vis-NIR models was determined:

1. On the principal components' space, the vis-NIR spectra were projected. By applying the resemble package in R studio, the singular value decomposition technique was used to determine the PCs.

2. Subsets of varying sizes were sampled from the collection of possible calibration samples. In increments of 10, we began from thirty to three hundred samples. The Kennard-Stone sampling technique was used to sample each subset from the PCs of the vis-NIR data.

3. We calculated the mean squared Euclidean distance (msd) between estimates of the probability density functions (pdfs) of the samples in the subset and the pdfs of the samples in the entire set of samples for each calibration subset. The msd is calculated as follows [18]:

$$msd = \frac{1}{k}\sum_{j=1}^{k} d^2[P_s(x_j \in cs),\ P_P(x_j)]$$

Where cs: a given subset of samples, $Ps(xj \in cs)$: the estimated pdf of the jth PC of cs, $Pp(xj)$: is the pdf of the jth PC for the whole population, $d2$: reflects the two single distributions under comparison's squared Euclidean distance, and k: the total amount of kept PCs.

The estimates for both the Pp and the Ps were estimated for the same values inside the PCp ranges, by employing the same bandwidth and Gaussian kernel. The optimal calibration set size was indicated by a notable decrease in the msd, which did not alter further when more samples were included; that is, the Ps was comparable to the Pp. This was determined visually by comparing the calibration sample size adjusted to the mean msd. To achieve trustworthy predictions of msd as a function of sample set size, The first two steps were repeated ten times, and the average of those repetitions was used to determine the final msd.

3.3.2 Evaluation and visualization of sample size selection

The visualization and analysis of Kennard-Stone selected samples is valuable as it provides an understanding of how the dataset looks like. In this regard, a subset of samples is chosen based on their Mahalanobis distance from each other in the principal component space. The selection method ensures that the chosen samples represent the variability contained within the dataset. Patterns and trends in the data are made more obvious when these selected samples are plotted together with the original dataset. It is also helpful because clusters, outliers, and other distinct groupings can be identified through this graphical representation.

3.4 Chemometric analysis

Chemometric approaches based on multivariate mathematical or statistical methodologies are appropriate for determining the statistical connection between dependent and independent variables. This study implemented PLSR with leave-one-out (LOO) cross-validation to calibrate spectral data with reference soil data from the laboratory.

3.4.1 The applied methods – Partial Least Squares Regression

In this study, the calibration of spectral data with laboratory soil data was undertaken using Partial Least Squares Regression (PLSR) and Leave-one-out cross-validation. It is deployed to model predictions in cases where numerous predictor factors exhibit strong correlations with one another. This method is similar to principal components regression (PCR). Nevertheless, unlike PCR, the PLSR approach picks out the next few orthogonal factors that minimize the covariance between predictor (X spectra) and response variables (Y laboratory data). The goal of this model is to pinpoint a select few variables that together explain the majority of the variation in responses and predictions.

Leave-one-out cross-validation was utilized to deduce the amount of factors for retention in the calibration models. The optimal cross-validated calibration model was determined by calculating the root mean squared error of predictions (RMSEP). We usually select the model with the least RMSE but we also checked if it is appropriate to have a more parsimonious model than one that minimizes RMSE. In order to compare all models with fewer components, the model that produced the lowest RMSE was utilized as the point of reference ($RMSE_{CV}$). In order to ensure that the

RMSE of the final model is not appreciably higher than the $RMSE_{CV}$, the model with the fewest factors was sought after.

4. Results and discussion

After preprocessing and eliminating spectral outliers, 355 soil samples remained. Test-set validation was used to confirm the prediction models' stability with approximately 70 % calibration (around 255 samples) and 30 % validation samples (around 255 samples).

4.1 Characteristics of the soil spectral curves

The spectral pre-processing results are shown in the Figures below. Particularly, the Vis-NIR reflectance spectra of all 379 soil samples had a comparable overall shape. The visual study of the spectra revealed noticeable variances caused by soil characteristics. The finding is consistent with previously published studies [5], demonstrating the study findings' congruence with current literature. We did not fall into over-fitting, as we employed this rigorous test-set validation method which helped us to make our prediction models more generalizable. We could examine the model performance properly by splitting it into the calibration-validation process whereby it was possible to create several data subsets out of them; thus allowing us to get more reliable findings. Additionally, this iterative approach assisted us in making alterations and refining our models so that they became easier to comprehend and their forecasting precision improved. We also analyzed performance data with validation steps taken and evaluated model predictability.

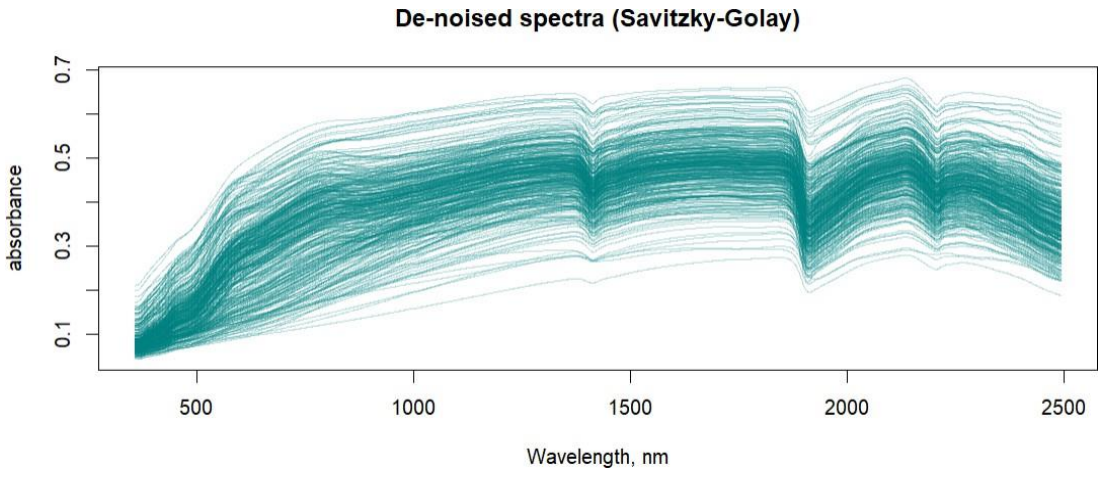

Figure 4: De-noised spectra by Savitzky-Golay

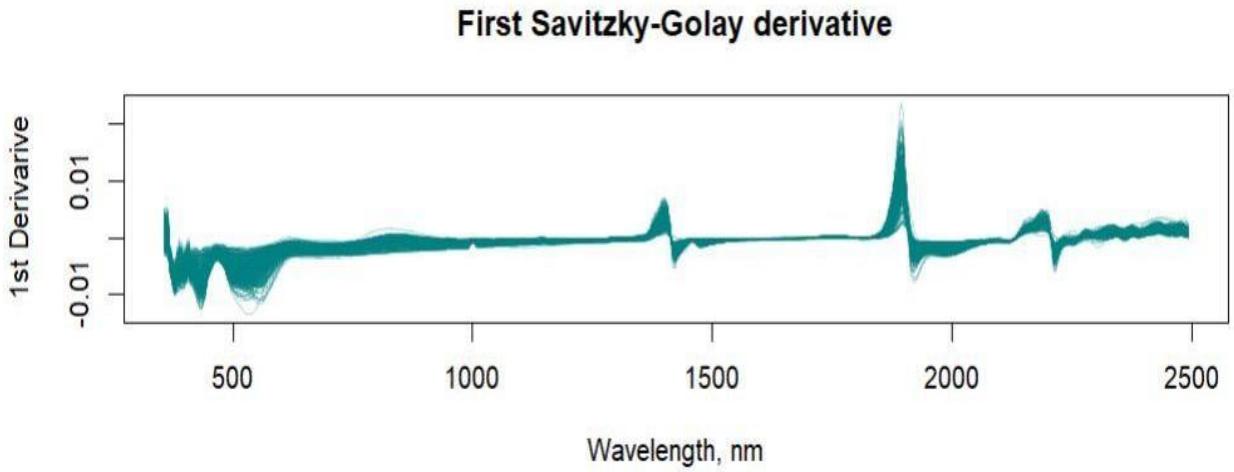

Figure 5: First Savitzky-Golay derivative

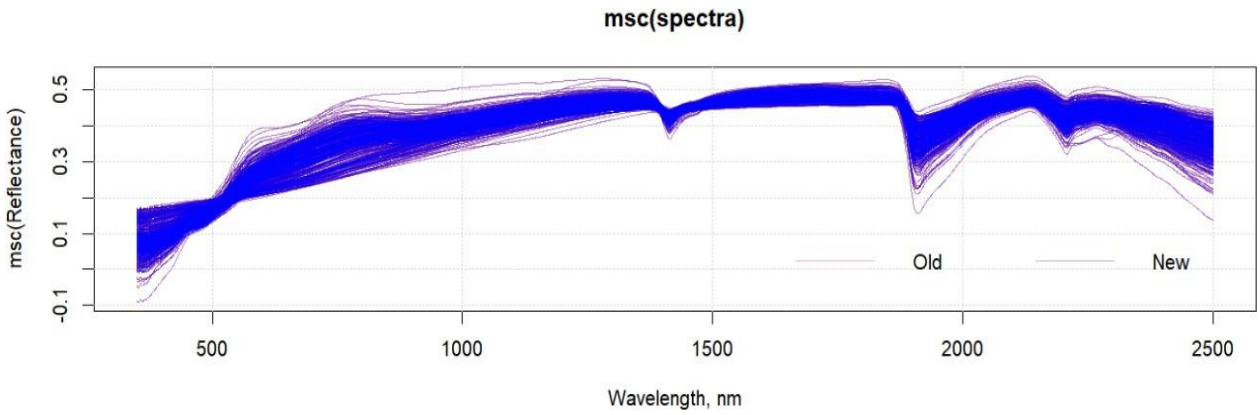

Figure 6: Multiplicative Scatter Correction

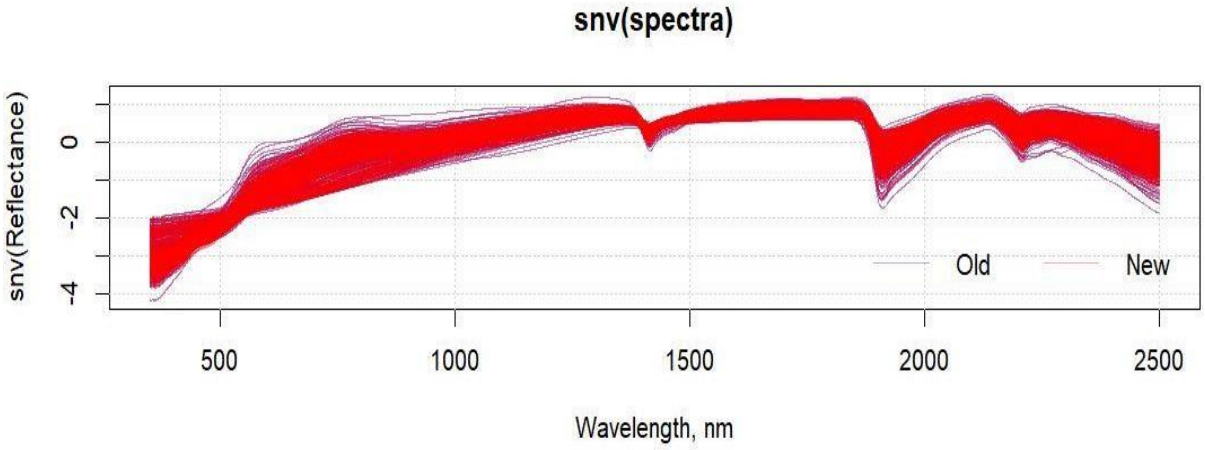

Figure 7: Standard Normal Variate

4.2 Calibration Sampling Assessment

4.2.1 Determining the ideal calibration set size

The first three PCs accounted for 99% of the spectrum fluctuation and were used to determine the ideal calibration set size. When comparing estimates of the probability density functions (pdfs) of the complete set and the pdfs of the samples in each calibration set, Figure 8 shows the mean squared Euclidean distance (msd) values that are related to those comparisons. The msd values declined as the sample set size increased. Notwithstanding this, there were only minor changes in msds between the calibration sets after 100 samples. As a result, we chose this number of samples as the ideal size for calibrating the vis-NIR models of the desired soil characteristics.

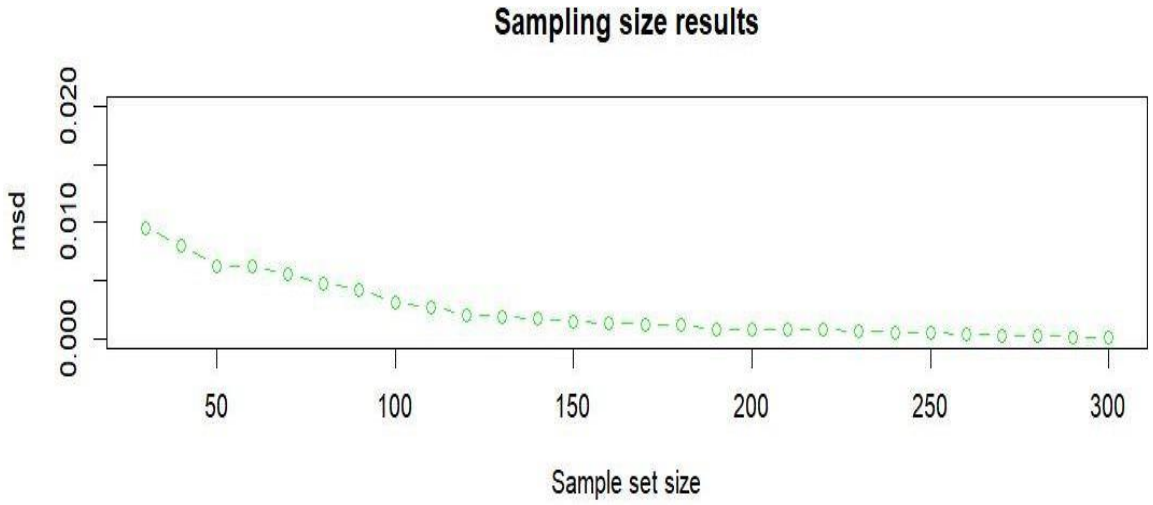

Figure 8: The msd computed for the different calibration dataset sizes.

4.2.2 Calibration sample selection

The Kennard-Stone approach was also used to choose a representative subset of calibration samples from the principal components scores for our chemometric model. It prioritizes samples that are evenly distributed throughout the data set and hence has good population coverage. For a visual representation of the selection process, we have plotted the first two principal components (PCs) in a scatter plot. Each data point represents one soil sample. The blue translucent circles show all samples while the larger red circles depict the calibration samples chosen by the Kennard-Stone algorithm. From Figures 9 and 10 below, it can be seen that these selected calibration samples are well positioned across the PC space, highlighting the variability

contained in the original dataset. This results in a more reliable and representative calibration set for our chemometric model. Figure 9 shows all soil samples (blue circles) in the first two principal components (PCs), while Figure 10 highlights the calibration samples chosen by Kennard-Stone (red circles) on the same PC space.

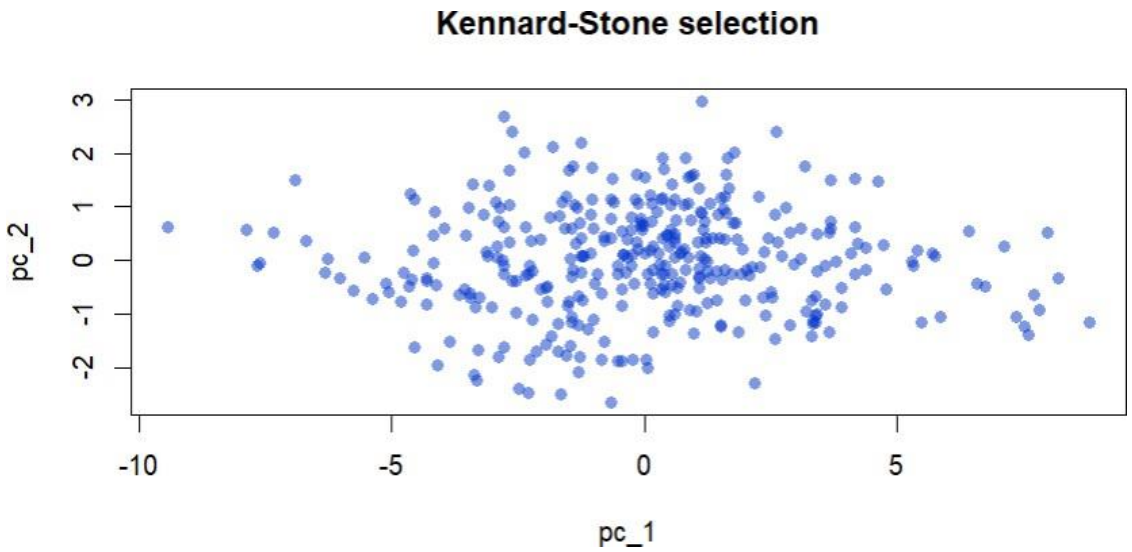

Figure 9: All soil samples

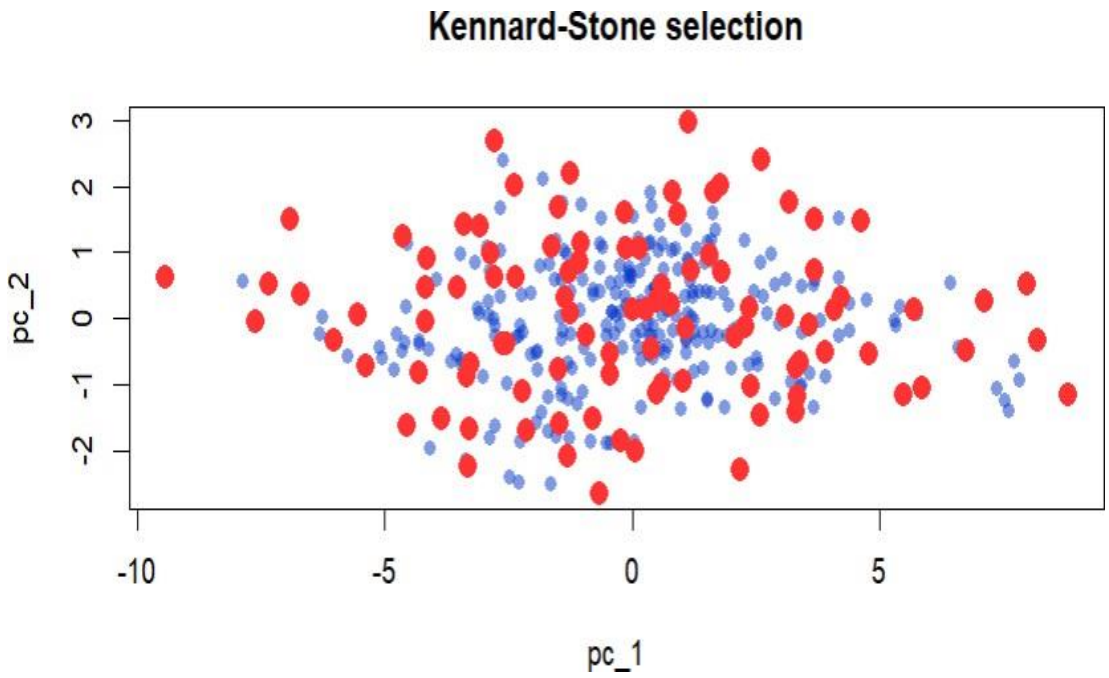

Figure 10: Calibration samples selected by Kennard-Stone Sampling

4.3 Multivariate calibration

Partial Least Squares Regression (PLSR) is used as a very powerful multivariate calibration method in the field of spectroscopy. After preprocessing steps and sample calibration have taken place, we will now look at the application of PLSR for creating robust calibration models. In modeling complex relationships between spectral data and analyte concentrations, PLSR takes an advantage over other regression methods by using latent variables that contain most of the variance present while minimizing overfitting. The purpose of this study is to assess how effective PLSR is based on various performance indicators such as Root Mean Squared Error of Prediction (RMSEP), measured versus predicted plots indicating predictive accuracy, and wavelength plotted against loading values explaining spectral components influencing predictions made from the model. These analyses are aimed at demonstrating whether or not it is possible to achieve calibration objectives through interpretation when using the PLSR approach for solving the given spectroscopic measurement problems.

Additionally, the models were constructed using the training datasets, and their correctness was evaluated using the testing datasets. The Root Mean Squared Error of Prediction (RMSEP) and the coefficient of determination (R^2) are the two performance measurements taken into consideration. The optimal combination of spectral processing and investigated regression models was selected for each soil attribute and dataset by considering the highest R^2 and the lowest RMSEP.

4.3.1 Utilizing the Number of Components versus RMSEP for Enhanced Model Performance

To optimize model performance in component selection, we use RMSEP as a way of evaluating the performance of models across different number of components. RMSEP measures the difference between actual data and model predictions, with the lowest values representing the best fit as mentioned above. By plotting the number of components versus RMSEP, we can identify an elbow point where the reduction in RMSEP starts to flatten. This point indicates that there is no longer much significance associated with adding more components when it comes to model improvements. The optimal choice for the number of components is just before this inflection since it strikes a balance between maintaining complexities inherent in data and diminishing overfitting.

Therefore, this technique helps determine the trade-off between model complexity and generalization.

Moreover, by using graphs presenting RMSEP at different numbers for clay, humus, pH, and CEC, we can directly establish which principal component number has minimum RMSEPs. Besides that, we can use the R studio to determine it directly from the graph. Consequently, as seen in Figure 11 below, the optimal number of components of clay, humus, pH, and CEC are 18, 16, 15, and 12 respectively.

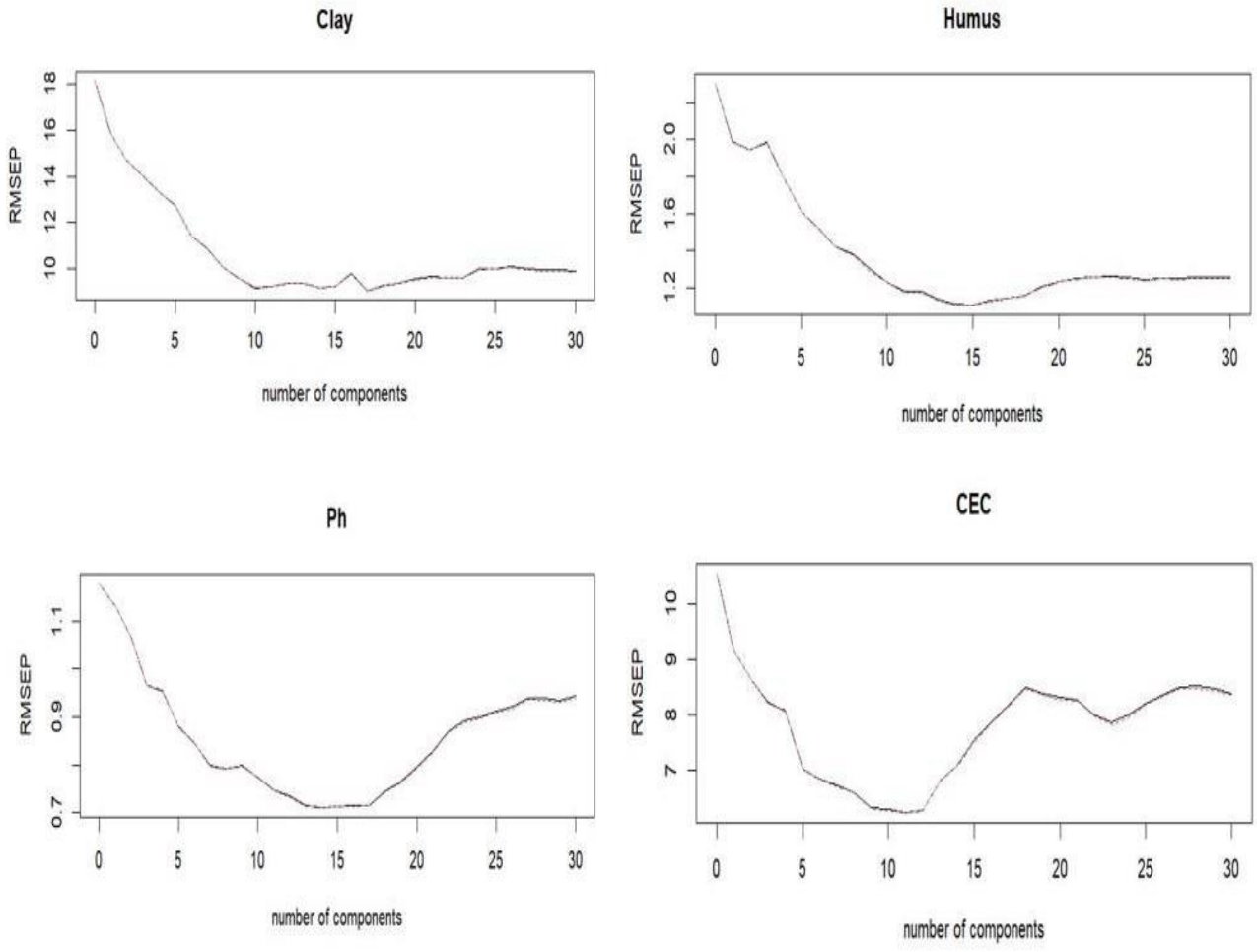

Figure 11: The number of components versus Root Mean Squared Error of Prediction (RMSEP) for clay, humus, pH, and CEC

4.3.2 Cross-Validation of Predictive Models for Soil Properties

The comparative analysis of clay content, humus content, pH, and cation exchange capacity (CEC) depicted in Figure 12 below is achieved through the measured versus predicted plots. Predicted values on the y-axis that were generated by PLSR models are shown while measured values that were obtained after laboratory analyses are plotted on the x-axis. It is expected that points should lie close to the diagonal line which indicates accurate predictions. Results from these plots indicate that Clay content and pH can be predicted well by the PLSR models as most of the points are concentrated around the diagonal line. The calibration step resulted in a higher R2 value and lower RMSE than the validation step for each parameter. This implies that after calibration the same dataset in which a model is built is used to test it out. However, during the validation stage, we can observe leave-one-out cross-validation results which approximates how well a model may perform in reality. For the given dataset, humus was predicted best using the PLSR model with the obtained calibration sample size (R^2 = 0.72694), followed by clay content (R^2 = 0.65917), pH (R^2 = 0.52588), and CEC (R^2 = 0.51824).

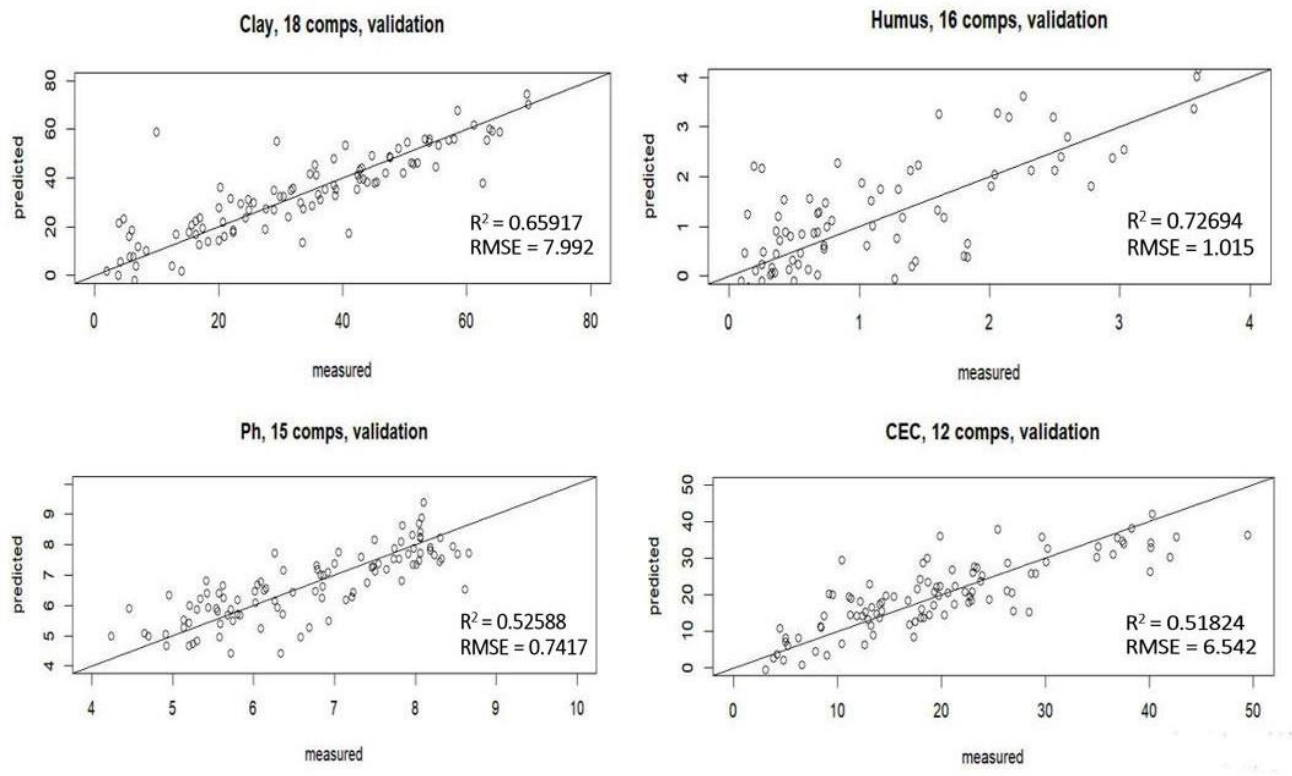

Figure 12: Measured versus predicted values of clay content, humus content, pH, and CEC

4.3.3 Unveiling the informative wavelengths using loading plots

The multivariate data analysis was also conducted to investigate the patterns and internal arrangement of the spectral information using loading plots. These helped establish how important different wavelengths of electromagnetic radiation were in terms of patterns exhibited by each principal component among other aspects. Most frequently, we carry out these analyses to examine the loadings of each wavelength variable; thus, determining which wavelengths emitted more energy and thus contributed most to variations in the datasets and therefore allowing for the selection of proper features. More specifically, a loading plot used in multivariate calibration models gives us a view of how original variables are related to the principal components extracted from the data set. Each part on these plots represents a variable and its position indicates both the magnitude and direction of its contribution to each principal component. Moreover, looking at the pattern of variables on these plots can unravel possible groupings or hidden structures within data that could be suggestive for further investigation. It is aimed at identifying significant factors necessary for model performance as well as refining the calibration model by eliminating less useful ones. In our studies, the significant wavelengths will be assigned based on the 3 first PCA loading values.

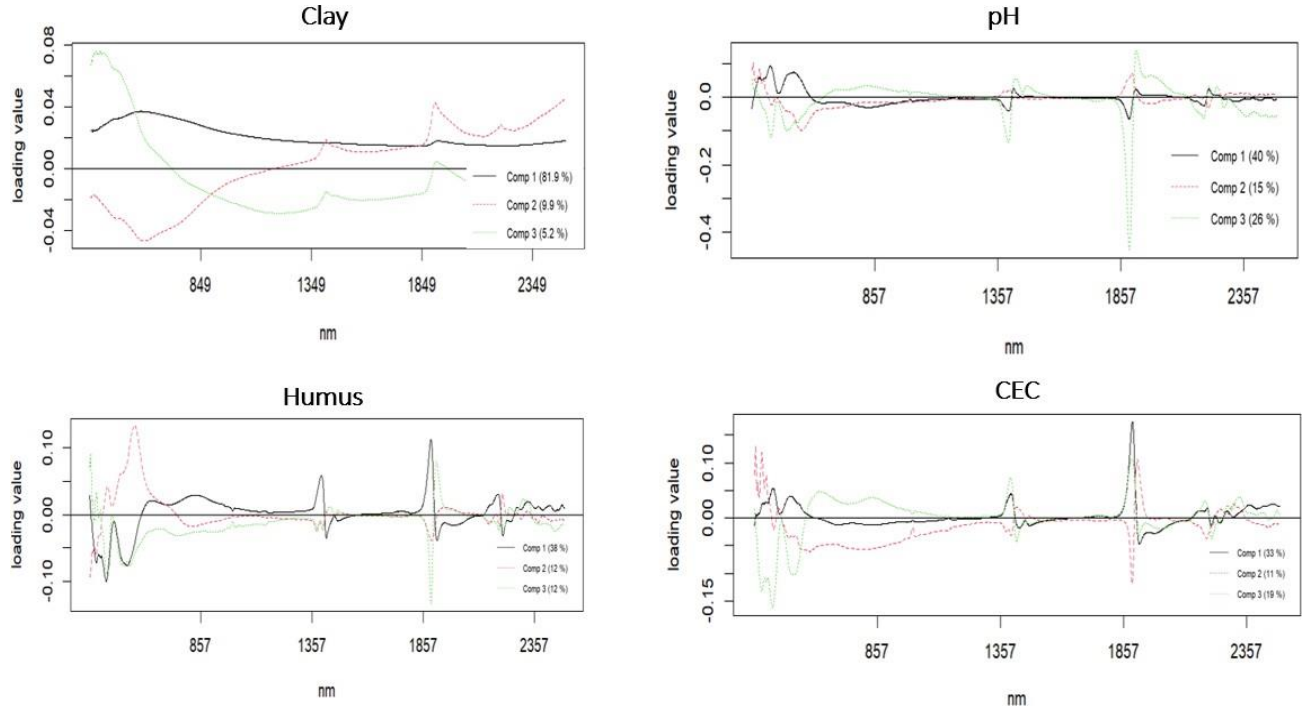

Figure 13: Loading plots of clay, humus, pH, and CEC.

The above figure 13 shows the loading plots for clay, humus, pH and CEC which are based on PLS regression from the first three principal components Comp 1, Comp2 and Comp3. The wavelength (nm) of each variable is denoted on the x-axis while the vertical axis represents loading values, where greater absolute values indicate more influence on a particular component. Furthermore, component 1 in all of the plots, has the highest percentage of explained variance, unveils significant insights. Other revelations about spectral information and measured soil properties can be found in components 2 and 3 depending on their explained variance. For example, clay has a considerable positive loading between 350 nm and 500 nm, while humus, and CEC exhibit strong positive loadings (picks) at approximately 1850 nm suggesting that these wavelengths are pivotal in capturing information related to their content.

5. Conclusion

The soil science profession is dealing with an increase in demand for regional, continental, and international databases to monitor soil situations. Nevertheless, such information is lacking. Low-cost equipment for measuring soil characteristics across large areas is needed. Soil spectroscopy has been proven to be an efficient, affordable, safe, and consistent analytical technique. As a result, we consider that the fundamental purpose of this research is to put into words the current status of soil spectroscopy along with its future application for soil monitoring. The limits of soil spectroscopy as a substitute for conventional laboratory testing are explored, as are the restraints. The study also underlines the importance of developing a standard for collecting laboratory soil spectra, as well as exchanging spectral collections and digitizing existing soil archives, in order to eliminate the requirement for pricey sample campaigns. subsequently, frequent soil analysis employing soil spectroscopy would benefit people using it by minimizing analytical costs and boosting laboratory results comparability.

This thesis thoroughly investigates various soil spectroscopy techniques to improve understanding and application in soil analysis. The first step involved meticulous preprocessing that included noise removal using Savitzky-Golay, first derivative and scatter correction through MSC and SNV methods. These steps of data preparation are important because they ensure quality control and subsequent data analysis. Following this, calibration sampling was addressed with attention to details. The sample size determination was done by Kennard-Stone sampling (KSS) method whereas sample robustness and selection visualization ensured reliable calibration samples.

Lastly, chemometric analysis using Partial Least Squares Regression (PLSR) was performed towards developing predictive models for the target soil properties. The optimal number of latent variables in each model was determined by analyzing the relationship between Root Mean Square Error of Prediction (RMSEP) and number of components. Model generalizability as well as its performance during cross-validation were evaluated. Furthermore, loading plots identified the spectral wavelengths which provided most informative contributions to prediction of each soil property.

This thesis’s main point is how crucial systematic preprocessing, careful sampling and insightful chemometric analysis are in advancing soil spectroscopy as a strong tool for the analysis and management of soil properties. This work's overall conclusion is that there could be limitless opportunities for practical soil science with the widespread application of VIS-NIR reflectance spectroscopy. However, the requirement for its success is developing such a spectral library that represents a pedological diversity. The development and continuous improvement of a database of this sort would improve the efficiency of predicting soil characteristics and make the process of classifying soil more data-driven and attainable, considering that it is based on standard measurements.

6. References

[1] Soil Spectroscopy: An alternative to wet chemistry- ScienceDirect. Available at: https://www.sciencedirect.com/science/article/pii/S0065211315000425.

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

[3] Sanchez, P.A., Ahamed, S., Carre, F., Hartemink, A.E., Hempel, J., Huising, J., Lagacherie, P., McBratney, A.B., McKenzie, N.J., Mendoca-Santos, M.L., Minasny, B., Montanarella, L., Okoth, P., Palm, C.A., Sachs, J.D., Sheperd, K.S., Vagen, T.G., Vanlauwe, B., Walsh, M.G., Winowiecki, L.A., Zhang, G.L., 2009. Digital soil map of the world. Science 325, 680e681.

[4] Viscarra Rossel a et al. (2016) A global spectral library to characterize the world's Soil, Earth-Science Reviews. Available at: https://www.sciencedirect.com/science/article/pii/S0012825216300113.

[5] Grunwald, S., Thompson, J.A., Boettinger, J.L., 2011. Digital soil mapping and modeling at

continental scalese finding solutions for global issues. Soil Sci. Soc. Am. J. 75, 1201e1213.

[6] Lyons, D., Rayment, G., Hill, R., Daly, B., Marsh, J., Ingram, C., 2011. Aspac soil proficiency testing program report 2007–08. Tech. rep.ASPAC, Melbourne, Victoria (URL http://www.aspac-australasia.com/index.php/documents/upload-documents/doc_download/232-annual-review-soil-07-08)

[7] Viscarra Rossel, R.A., McBratney, A.B., 1998a. Soil chemical analytical accuracy and costs: implications from precision agriculture. Australian Journal of Experimental Agriculture 38, 765–775.

[8] M. Demattê a et al. (2019) The Brazilian Soil Spectral Library (BSSL): A general view, application, and challenges, Geoderma. Available at: https://www.sciencedirect.com/science/article/pii/S0016706118318548.

[9] Janik, L.J., Skjemstad, J.O., Raven, M.D., 1995. Characterization and analysis of soils using mid infrared partial least-squares. I. Correlations with XRF-determined major-element composition. Australian Journal of Soil Research 33, 621–636.

[10] Characteristic variations in reflectance of surface soils. Available at: https://acsess.onlinelibrary.wiley.com/doi/10.2136/sssaj1981.03615995004500060031x.

[11] Zhao a et al. (2023) Quantifying soil properties relevant to soil organic carbon biogeochemical cycles by infrared spectroscopy: The importance of compositional data analysis, Soil and Tillage Research. Available at: https://www.sciencedirect.com/science/article/pii/S0167198723000855.

[12] Owen, A., 2000. Fundamentals of Modern UV-visible Spectroscopy. Agilent Technologies.

Pimstein, A., Notesco, G., Ben Dor, E., 2011. Performance of three identical spectrometers

in retrieving soil reflectance under laboratory conditions. Soil Sci. Soc. Am. J. 75, 746e759.

[13] Herzberg, G., 1945. Molecular Spectra and Molecular Structure: Infrared and Raman Spectra of Polyatomic Molecules. D Van Nostrand, New York.

[14] Alpert, N.L., Keiser, W.E., Szymanski, H.A., 1970. Theory and Practice of Infrared

Spectroscopy. Plenum Press, New York.

[15] Viscarra Rossel a et al. (2005) Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties, Geoderma. Available at: https://www.sciencedirect.com/science/article/pii/S0016706105000728#bib18.

[16] Aenugu, H.P.R., Sathis Kumar, D., Srisudharson, P.N., Ghosh, S.S., Banji, D., 2011. Near

infrared spectroscopy- an overview. Int. J. ChemTech Res. 3, 825e836.

[17] Davis, T., 1998. The history of near infrared spectroscopic analysis: past, present and

future - from sleeping technique to the morning star of spectroscopy. Anal. Mag. 26, 17e19.

[18] Wartini Ng a et al. (2023) Near and mid infrared soil spectroscopy, Encyclopedia of Soils in the Environment (Second Edition). Available at: https://www.sciencedirect.com/science/article/abs/pii/B9780128229743000227.

[19] RA;, W.J.B. Soil analysis using visible and near infrared spectroscopy, Methods in molecular biology (Clifton, N.J.). Available at: https://pubmed.ncbi.nlm.nih.gov/23073878/.

[20] Stenberg B, Rossel RAV, Mouazen AM, Wetterlind J (2010) Visible and near infrared spectroscopy in soil science. In: Sparks DL (ed) Advances in agronomy, vol 107. pp 163–215

[21] Fystro G (2002) The prediction of C and N content and their potential mineralisation in heterogeneous soil samples using Vis-NIR spectroscopy and comparative methods. Plant Soil 246:139–149

[22] Williams PC, Norris K (2001) Variables affecting near-infrared spectroscopic analysis. In: Williams P, Norris K (eds) Near-infrared technology in the agricultural and food industries, 2nd edn. American Association of Cerial Chemists Inc., Minesota, pp 171–185

[23] Viscarra Rossel RA, Behrens T (2010) Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma 158:46–54

[24] Wetterlind J, Stenberg B, Soderstrom M (2010) Increased sample point density in farm soil mapping by local calibration of visible and near infrared prediction models. Geoderma 156:152–160

[25] Brown DJ, Bricklemyer RS, Miller PR (2005) Validation requirements for diffuse re fl ectance soil characterization models with a case study of VNIR soil C prediction in Montana. Geoderma 129:251–267

DECLARATION

on authenticity and public assess of final mater's thesis

Student's name:	Mohamed Yassine Zaghdoudi
Student's Neptun ID:	PC2R1K
Title of the document:	Soil spectroscopy
Year of publication:	2024
Department:	Department of Soil Science

I declare that the submitted final master's thesis is my own, original individual creation. Any parts taken from an another author's work are clearly marked, and listed in the table of contents.

If the statements above are not true, I acknowledge that the Final examination board excludes me from participation in the final exam, and I am only allowed to take final exam if I submit another final essay/thesis/master's thesis/portfolio.

Viewing and printing my submitted work in a PDF format is permitted. However, the modification of my submitted work shall not be permitted.

I acknowledge that the rules on Intellectual Property Management of Hungarian University of Agriculture and Life Sciences shall apply to my work as an intellectulal property.

I acknowledge that the electric version of my work is uploaded to the repository sytem of the Hungarian University of Agriculture and Life Sciences.

Place and date: 2024/04/20 Hungary.

Student's signature