



UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

***Evaluating the Use of Mid Infrared Spectroscopy in Predicting Soil
Carbon, Nitrogen Contents, and Soil Texture in Varying South
African Soil Samples***

ABONGA MATA

Student Number: 2694389

A thesis Submitted in partial fulfilment of the degree requirements

Master of Science by Research (Geography, Archaeology, and Environmental Studies)

FACULTY OF SCIENCE

UNIVERSITY OF THE WITWATERSRAND

Supervisors (Professor Elhadi Adam and Doctor Goodman Jezile)

May 2025

DECLARATION

I declare that this dissertation is my own, unaided work. It is being submitted for the Degree of Master of Science at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other University.



(Signature of candidate)

14 day of May 2025 at _____

Abstract

There is a growing need for soil information to help address problems such as soil nutrient depletion and incorrect fertilizer inputs which affect yields and threaten food security and the environment. Low soil fertility is a significant issue in main grain-producing regions in South Africa. Routine soil analysis addresses these issues as farmers are provided with valuable information to enhance agricultural productivity and sustainability. However, lack of efficient and cost effective soil testing infrastructure is a limitation in South Africa.

Soil spectroscopy has proven to be an efficient soil analysis method that complements traditional methods of soil analysis. It is more cost-effective than analytical laboratory procedures since measurements may be completed faster, and multiple soil parameters can be deduced from a single spectral measurement (Nocita *et al.*, 2015). This research aimed at developing a spectral library of South African soils and to test the ability of mid-infrared (MIR; 4000 – 500cm⁻¹) diffuse reflectance spectroscopy coupled with machine learning algorithms to predict total carbon, total nitrogen, and soil texture. Soil samples for this study were selected from the ARC-SCW archive using the Kennard-Stone algorithm to ensure diversity. Chemical analysis of total carbon (%) and nitrogen (%) was conducted through total combustion with a CHNS-O analyser, while physical properties (sand (%), silt (%), and clay (%) content) were determined through the hydrometer analysis technique. This data was then used with corresponding spectra to develop PLSR (partial least squares regression) and MBL (memory-based learner) predictive models. During model assessment, the criteria used to determine a good predictive model was that an R² value had to exceed 0.7, RMSE_p < 10% of the range of the independent validation dataset and RPIQ > 1.7.

Predictive regression models were developed successfully, the results for clay, silt, and sand contents showed that soil texture can be predicted well with diffuse reflectance MIR spectroscopy. The highest R² values were 0.80, 0.78, and 0.81 for clay, silt, and sand contents respectively. Based on these results, these models can be used in real world situations for accurate predictions of sand, silt, and clay content. The best sand and clay content models were developed with UVS-selected variables, selected from a raw spectra dataset. On the other hand, during model assessment, the results were poor for predicting TC (total carbon) and TN (total nitrogen). The highest R² values were 0.51 and 0.61 for TC and TN respectively. This meant that the predictive models developed in this study for predicting TC and TN were unreliable

and may not be applicable for predictions. This poor result was due to highly variable and limited data used during model training.

Compared to results obtained with full-spectrum, Boruta, and CARS-selected variables datasets, the raw spectra UVS-selected variables datasets produced better results. Generally, MBL produced better results than PLSR. Pre-processing the spectra with either Savitzky-Golay first derivative filter (SG) or standard normal variate (SNV) was proved to be ineffective in improving model performance. Soil spectroscopy is a data hungry analysis method, therefore it is recommended to develop predictive models with many samples such that the number of samples is greater or equal to the number of predictor variables especially when the data is diverse.

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Professor Elhadi Adam and Dr Goodman Jezile for their support, guidance, patience, motivation, and sharing of skills and knowledge all of which contributed to the success of this project. I am deeply thankful for the opportunity which allowed me to learn and grow, both as a person and as a researcher; it has truly been an invaluable experience. I extend my heartfelt thankfulness to the Agricultural Research Council Institute for Soil, Climate, and Water (ARC-ISCW) and the Department of Science and Innovation (DSI) for their generous financial support, infrastructure, and technical assistance. I would also like to express my gratitude to my family and friends for their unwavering support, as well as to everyone who contributed, directly or indirectly, to the success of this project.

List of Abbreviations

PLS	Partial Least Squares
PLSR	Partial Least Squares Regression
MBL	Memory-Based Learner
MIR	Mid-Infrared
NIR	Near-Infrared
MIRS	Mid-Infrared Spectroscopy
NIR	Near-Infrared
RMSE	Root Mean Square Error
RMSE _p	Root Mean Square Error of Prediction
R ²	Coefficient of Determination
CARS	Competitive Adaptive Reweighted Sampling
UVS	Univariate Filter Variable Selection
TC	Total Carbon
TN	Total Nitrogen
SG	Savitzky-Golay filter
SNV	Standard Normal Variate
DRIFT	Diffuse Reflectance Infrared Fourier Transform
CV	Coefficient of Variation
RPIQ	Ratio of Performance to the Inter-Quartile Range
ARC-ISCW	Agricultural Research Council-Institute for Soil, Climate, and Water
SOC	Soil Organic Carbon
SOM	Soil Organic Matter

Contents

1.1	General Introduction	1
1.2	Research Problem Statement	2
1.3	Research Aim and Objectives.....	4
2.1	Model Calibration	5
2.1.1	Partial Least Squares Regression (PLSR)	5
2.1.2	Application of PLSR: Global Perspective.....	6
2.1.3	Application of PLSR: Continental (Africa) Perspective	7
2.1.4	Application of PLSR: Local Perspective	8
2.1.5	Memory Based Learner (MBL).....	9
2.1.6	Application of MBL: Global Perspective	10
2.1.7	Application of MBL: Continental & Local Perspective.....	11
2.2	The Dimensionality of Spectroscopy Data	11
2.2.1	Variable Selection Techniques	11
2.3	Model Accuracy Evaluation Metrics	15
2.3.1	Root Mean Square Error (RMSE).....	15
2.3.2	Coefficient of Determination (R^2).....	15
2.3.3	Ratio of Performance to Inter-Quartile Range (RPIQ)	15
2.3.4	Bias.....	16
2.4	Lessons Learned From the Literature	16
3.1	Soil Data Acquisition.....	18
3.2	Soil Data Preparation for Spectral Measurements	18
3.3	Soil Chemical and Physical Analysis	19

3.4	Laboratory Spectra Measurements.....	19
3.5	Outlier Analysis.....	19
3.6	Spectra Pre-processing.....	20
3.7	Variable Selection	21
3.7.1	Boruta.....	21
3.7.2	Competitive Adaptive Reweighted Sampling (CARS).....	22
3.7.3	Univariate Filter Variable Selection (UVS).....	22
3.8	Calibration and Validation Datasets	23
3.9	Modelling Soil Properties.....	23
3.9.1	Partial Least Squares Regression (PLSR).....	24
3.9.2	Memory-Based Learner (MBL)	24
3.10	Model Assessment.....	26
3.10.1	Coefficient of Determination (R^2).....	26
3.10.2	Root Mean Square Error of Prediction ($RMSE_p$).....	26
3.10.3	Ratio of Performance to the Inter-Quartile Range (RPIQ)	27
3.10.4	Bias.....	27
4.1	Descriptive Statistics	28
4.2	Predicting Soil Properties Using Full Spectrum Datasets.....	31
4.2.1	Predicting Sand Content	31
4.2.2	Predicting Silt Content.....	32
4.2.3	Predicting Clay Content	33
4.2.4	Predicting Total Carbon Content.....	33
4.2.5	Predicting Total Nitrogen Content.....	34
4.3	Improving Models Performance Using Variable Selection Methods.....	35
4.3.1	Sand Content	35
4.3.2	Silt Content.....	36
4.3.3	Clay Content.....	38

4.3.4	Total Carbon Content.....	39
4.3.5	Total Nitrogen Content.....	40
4.4	Determination of the Best Model for Each Soil Property	41
4.4.1	Sand Content.....	41
4.4.2	Silt Content.....	41
4.4.3	Clay Content	42
4.5	Variable (Wavelengths) Selection.....	43
4.5.1	Sand Content.....	43
4.5.2	Silt Content.....	44
4.5.3	Clay Content	45
4.5.4	Total Carbon Content.....	46
4.5.5	Total Nitrogen Content	47
4.5.6	Frequently Selected Variables by Each Selection Method	48
5.1	Prediction of Soil Properties Based on MIR Full Spectrum Datasets	50
5.1.1	Sand Content.....	50
5.1.2	Silt Content.....	51
5.1.3	Clay Content	51
5.1.4	Total Carbon and Total Nitrogen Content	51
5.2	Prediction of Soil Properties Based on Selected Key Variables in the MIR Range..	52
5.2.1	Sand Content.....	52
5.2.2	Silt Content.....	53
5.2.3	Clay Content	54
5.2.4	Total Carbon and Total Nitrogen Content	54
5.3	The Effect of the Spectra Pre-processing on the Model Performance	55
5.4	The Role of Variables Selection on the Model Performance	56
5.5	Limitations	57
5.6	Conclusions.....	58

List of Tables

Table 2.1: Summary of the results from the above-reviewed studies.....	14
Table 4.1: Descriptive statistics of the whole data, calibration, and the validation set.	29
Table 4.2: Validation results of the sand content predicted using full spectrum raw spectra and pre-processed spectra datasets respectively.	32
Table 4.3: Validation results of the silt content predicted using full spectrum raw spectra and pre-processed spectra datasets respectively.	32
Table 4.4: Validation results of the clay content predicted using full spectrum raw spectra and pre-processed spectra datasets respectively.	33
Table 4.5: Validation results of the total carbon content predicted using full spectrum raw spectra and pre-processed spectra datasets respectively.	33
Table 4.6: Validation results of the total nitrogen content predicted using full spectrum raw spectra and pre-processed spectra datasets respectively.	34
Table 4.7: PLSR and MBL validation results of the sand content predicted using predictor variables selected using the Boruta, CARS, and UVS algorithms from raw spectra and pre-processed spectra datasets respectively.	36
Table 4.8: PLSR and MBL validation results of the silt content predicted using predictor variables selected using the Boruta, CARS, and UVS algorithms from raw spectra and pre-processed spectra datasets respectively.	37
Table 4.9: PLSR and MBL validation results of the clay content predicted using predictor variables selected using the Boruta, CARS, and UVS algorithms from raw spectra and pre-processed spectra datasets respectively.	38
Table 4.10: PLSR and MBL validation results of the total C content predicted using predictor variables selected using the Boruta, CARS, and UVS algorithms from raw spectra and pre-processed spectra datasets respectively.	39
Table 4.11: PLSR and MBL validation results of the total N content predicted using predictor variables selected using the Boruta, CARS, and UVS algorithms from raw spectra and pre-processed spectra datasets respectively.	40
Table 4.12: Comparison between MBL and PLSR: Results of the best and worst sand content models.	41

Table 4.13: Comparison between MBL and PLSR: Results of the best and worst silt content models.	42
Table 4.14: Comparison between MBL and PLSR: Results of the best and worst clay content models.	42
Table 4.15: Frequently selected key variables by each variable selection method from all datasets (raw and pre-processed spectra) for predicting each studied soil property.	49

List of Figures

Figure 2.1: The red dots represent samples of interest, the dotted lines around them depict a boundary defined by a K-nearest neighbour search, for local modelling. The above figure was adopted from Ramirez-Lopez et al. (2013a).	10
Figure 4.1: Histograms of sand, silt, clay, TC and TN contents showing their respective distribution, respectively (ordered in rows from top to bottom). The histograms in the first column are for the whole dataset, in the second and third columns are for calibration and validation sets respectively.	31
Figure 4.2: Plots showing selected spectral variables for predicting sand content. The variables were selected from both unprocessed spectra and pre-processed spectra through Boruta, CARS, and UVS algorithms.	44
Figure 4.3: Plots showing selected spectral variables for predicting silt content. The variables were selected from both unprocessed spectra and pre-processed spectra through Boruta, CARS, and UVS algorithms.	45
Figure 4.4: Plots showing selected spectral variables for predicting clay content. The variables were selected from both unprocessed spectra and pre-processed spectra through Boruta, CARS, and UVS algorithms.	46
Figure 4.5: Plots showing selected spectral variables for predicting TN content. The variables were selected from both unprocessed spectra and pre-processed spectra through Boruta, CARS, and UVS algorithms.	48

Chapter One

1.1 General Introduction

Soil is an aggregate of organic materials, minerals, microorganisms, gases, and water (Al-Kaisi *et al.*, 2017; Wilpiseski *et al.*, 2019). It is further described as an essential resource responsible for regulating climate conditions, fertility, and ultimately, food production, as well as influencing the distribution of animals, plants, and landforms (Dangal *et al.*, 2019). Therefore, an understanding of the complex processes that govern soil is important. Additionally, it is crucial to monitor soil properties that contribute to its health and fertility in a timely manner in order to ensure that the growing human population and natural biodiversity has access to sufficient food supplies. Soil properties play a vital role in influencing the success of food crops and thus contribute to global food security (FAO and ITPS, 2015). Understanding soil properties is especially critical for farmers because it enables them to apply sustainable soil management procedures ensuring the optimal growth and health of their crops (Hati *et al.*, 2022). Soil is highly crucial in achieving the United Nations Sustainable Development Goals (SDG) 2030 Agenda, as it contributes to SDG 2, SDG 3, SDG 12, and SDG 15. These SDGs address issues such as zero hunger, good health, and well-being, responsible consumption and production as well as life on land respectively.

Various key soil properties influence the fertility of the soil, soil stability, and carbon flux. Such properties include soil organic carbon, nitrogen content, soil texture, and others. Both soil organic carbon and nitrogen have been identified as key soil fertility properties (Stevens *et al.*, 2013; Sun *et al.*, 2014; Nawar and A. Mouazen, 2017). Soil organic carbon influences aggregate stability, and carbon flux (Stevens *et al.*, 2013). One of the most critical soil properties for agriculture is soil texture, which is the relative proportional distribution of different-sized particles in the soil such as sand, silt and clay. Soil texture is essential in regulating water infiltration and erosion control and in influencing nutrient retention. Sandy soils have large particles, easily drain water, and do not hold nutrients well. Clay soils, on the other hand, have smaller particles and hold nutrients well. Previous research has shown that clay content influences the carbon content stored in the soil (Seboko *et al.*, 2023). Understanding these soil properties require regular soil testing.

Regular soil testing of soil properties leads to sustainable soil management practices and can help ensure soil remains healthy and productive for a long time. However, the evaluation of soil properties has previously been assessed through conventional laboratory and field

assessments (Peck, 1990). While these methods are usually accurate, they are often tedious and expensive and may lead to sample contamination (Bonett *et al.*, 2015; Nomngongo *et al.*, 2017).

Alternatively, soil spectroscopy is a faster, more convenient, affordable, environmentally friendly, and accurate soil testing procedure (Baumann *et al.*, 2021). Previous research has successfully demonstrated the application of spectroscopy in predicting a range of soil properties including total carbon, and soil texture (Waruru *et al.*, 2015; Ludwig *et al.*, 2021; Ramírez *et al.*, 2021; Summerauer *et al.*, 2021). Spectroscopy is especially valuable in determining soil properties especially within the MIR region of the electromagnetic spectrum (EMS). The MIR falls in the range of $4000 - 400 \text{ cm}^{-1}$ of the EMS and these wavelengths correspond to the fundamental vibrational modes of soil molecular bonds. Contrastingly, the NIR region ($\sim 14286 - 4000 \text{ cm}^{-1}$) corresponds to weak signals from vibration overtones and a combination of fundamental frequencies (Dangal *et al.*, 2019; Hati *et al.*, 2022). Spectroscopy relies on utilising multivariate statistics and geo-statistics for its accuracy (Bonett *et al.*, 2015).

Spectroscopic soil prediction techniques require the development of robust predictive models. To accurately predict soil properties with spectroscopy, multivariate regression modelling in conjunction with MIR spectral data offers a comprehensive and efficient approach. There are various statistical and machine learning algorithms or regression models that are utilized to correlate the soil spectra and the target soil property. These algorithms include random forest, support vector machine, PLSR, principal component regression, and MBL to name a few. Amongst the listed regression methods, PLSR is arguably the most utilized regression method to estimate and quantify soil properties (Dangal *et al.*, 2019; Ludwig *et al.*, 2019).

1.2 Research Problem Statement

Numerous studies have indicated that in different parts of the African continent, soil nutrient depletion is occurring at alarming rates (Henao and Baanante, 1999; Roberts *et al.*, 2003; Tan *et al.*, 2005). Depleted soil fertility challenges the sustainability of crop production and directly threatens food security (Baumann *et al.*, 2021). The soil fertility decline results from the removal of nutrients from crops through high production levels, declined fertilizer use, or loss due to land degradation (Tan *et al.*, 2005). Roberts *et al.*, (2003) revealed that in South Africa nutrient deficiency and depletion are site-specific and are influenced by site-specific soil management over the years.

It is recommended that before a farmer purchases fertilizer, they submit soil samples to the analytical laboratory for testing. The information provided by the testing laboratory helps farmers establish fertility goals and increase their yields. Unfortunately, as noted by Shepherd and Walsh, (2007), in Africa, national soil testing laboratories are deteriorating. They stated that this occurs at a crucial time when such institutions should be facilitating diagnostic surveillance methods in developing countries. These methods are not only applicable to diagnosing issues related to soil, crops, and livestock health, but also to addressing quality control challenges in agricultural production and marketing. Precision agriculture is a method of cultivation that intends to improve crop yield but relies on timely soil properties testing to make informed decisions regarding soil quality and fertility. The time-consuming and costly conventional soil analysis methods reduce the applicability of collecting thousands of samples, performing analysis, and generating results fast and with significantly reduced costs.

In contrast, soil spectroscopy has proven to be an efficient spectral-based soil analysis method that complements traditional methods of soil analysis. Soil spectroscopy is more cost-effective than analytical laboratory procedure since measurements may be completed faster, and enables the estimation of a range of soil properties from an individual spectrum of a soil sample (Nocita *et al.*, 2015). In underdeveloped countries, where soil information is needed at high spatial resolution, infrared spectroscopy can contribute to help in manage and enhance understanding of the observed soil fertility variation on communal fields where the nutrient depletion status is field specific as stated by Roberts *et al.*, (2003). By leveraging the strengths of spectroscopy and multivariate regression, we can develop highly accurate predictive models that predict soil fertility properties based on spectral data. This would allow farmers and agronomists to make data-driven decisions on how they manage their soils, leading to improved crop yields and reduced environmental impact.

Even though both international and continental research has demonstrated that soil spectroscopy has the potential to dramatically increase the extent of soil sampling activities by increasing both spatial and temporal coverage, research into soil spectroscopic techniques using the infrared region of the spectrum is minimal in South Africa and lags far behind the international surge in this area of research (Paterson *et al.*, 2015; Seboko *et al.*, 2023).

1.3 Research Aim and Objectives

This research aims to develop a spectral library of South African soils and demonstrate the ability of MIR diffuse reflectance spectroscopy coupled with machine learning algorithms to simultaneously predict total carbon, total nitrogen, and soil texture.

The objectives of this study is to;

- Calibrate and validate the PLS regression and MBL models for predicting total carbon, total nitrogen, and soil texture contents by applying soil spectroscopy techniques.
- Determine the key wavelengths and regions to predict soil total carbon, total nitrogen, and soil texture fractions.
- Determine the best calibration algorithm by comparing model performance. Ultimately, choose the one with a high performance-to-interquartile distance (RPIQ) ratio, coefficient of determination (R^2), low prediction error ($RMSE_p$), and bias.

Chapter Two

Literature Review

Soil texture is essential in regulating critical soil processes such as infiltration of water, erosion control, and air diffusion (Soriano-Disla *et al.*, 2014). Additionally, it also influences nutrient retention, which is primarily determined by the clay content in the soil (Nocita *et al.*, 2016). Moreover, soil texture significantly influence decision-making concerning soil degradation, and land use management which includes sustainable agriculture management (Nocita *et al.*, 2016; Shete *et al.*, 2019). Soil texture is spectrally active due to its sensitivity and spectral response to the MIR wavelengths. The spectral response is due to the MIR's wavelength sensitivity to clay and sand mineral vibrations resulting in distinctive absorption features (Fongaro *et al.*, 2018; Coblinski *et al.*, 2020). The distinct spectral features in the spectral signatures of sandy soils are mainly determined by presence of quartz whereas clay is influenced by the abundance of smectite, illite, and kaolinite compounds (Soriano-Disla *et al.*, 2014). Carbon content in the soil influences soil fertility, carbon flux, and soil stability (Stevens *et al.*, 2013). Similarly, nitrogen in the soil has been identified as one of the most important soil fertility properties (along with phosphorus and potassium) with a critical role in crop and pasture production (Sun *et al.*, 2014; Nawar and A. Mouazen, 2017). Carbon and Nitrogen in the soil are also spectrally active soil properties as indicated by Ludwig *et al.*, (2023). Overall, the characteristic MIR spectral regions of functional groups relevant to each soil property such as soil texture fractions, TC, and TN enable accurate predictions of soil properties using spectroscopy (Baumann *et al.*, 2021).

2.1 Model Calibration

2.1.1 Partial Least Squares Regression (PLSR)

To quantify various chemical and physical soil properties, multivariate chemometric techniques are utilized to correlate soil reflectance spectra and soil properties of interest (Gholizadeh *et al.*, 2015). These chemometric techniques predict soil properties by extracting qualitative and quantitative information recorded in the soil spectrum. The most commonly utilized method is PLSR (Wold *et al.*, 2001; Gholizadeh *et al.*, 2013; Dangal *et al.*, 2019; Ludwig *et al.*, 2019). PLSR was introduced first by Geladi & Kowalski (1986) for building predictive models from a dataset containing many predictor variables. PLSR considers the correlation between spectral measurements and corresponding soil chemical or physical properties (Ji *et al.*, 2016; Dangal *et al.*, 2019; Hati *et al.*, 2022), and handles collinearity very

well (Rossel and Behrens, 2010; Kiala, Odindi and Mutanga, 2017). PLSR is efficient in processing the large number of spectra produced by the high resolution MIR instruments (Janik *et al.*, 2007). A PLSR model correlates the predictor variables (X) and response variables (Y) simultaneously (Viscarra Rossel *et al.*, 2006). Nawar & Mouazen (2017) stated that in an attempt to obtain maximum covariance between X and Y variables, this chemometric technique selects orthogonal factors known as latent variables. These “*latent variables are simply linear combinations of the original independent variables*” (Bangelesa *et al.*, 2020, page 4). PLSR is similar to principal component regression, however, the significant difference is that PLSR uses both X and Y variables to build vectors with a high predictive ability (Rossel and Behrens, 2010; Bangelesa *et al.*, 2020).

Furthermore, PLSR is able to handle cases where the number of samples is significantly smaller than the number of predictor variables (Nomngongo *et al.*, 2017). This method gained popularity over the years in quantitative chemometric studies of soil spectra analysis (Viscarra Rossel *et al.*, 2006; Nawar and A. Mouazen, 2017) and has successfully estimated various soil properties utilizing the MIR range of the EMS (Xie *et al.*, 2011; Vendrame *et al.*, 2012; Nocita *et al.*, 2014; Sun *et al.*, 2014). However, Bayer *et al.*, (2012) argue that PLS models built based on local observations are not applicable on a regional scale.

2.1.2 Application of PLSR: Global Perspective

Haaland & Thomas (1988) applied PLSR to quantify soil properties by extracting soil information from soil spectra, their study pioneered the application of infrared spectroscopy in conjunction with statistical methods for soil analysis. In their study they used simulated spectra with a range of 0 to 100 data points. To correlate soil chemistry and soil spectral measurements, PLS statistically derives spectral attributes of each soil sample (Bayer *et al.*, 2012). Viscarra Rossel *et al.*, (2006) discovered that in the MIR (4000 – 400 cm⁻¹) range of the spectrum, PLSR produced more accurate results than the visible and NIR when predicting soil organic carbon and soil texture of Kelso NSW soils in Australia. Many other studies around the world reported satisfactory results from PLSR models for varying soil properties (Wold *et al.*, 2001; Xie *et al.*, 2011; Ji *et al.*, 2016; Dangal *et al.*, 2019). In 2016 another study conducted by Gholizadeh *et al.* (2016) in dumpsites in the mines Bílina and Tušimice in the Czech Republic, demonstrated that PLSR has a good predictive ability for soil texture fractions (sand, silt, and clay) with reported R² values ranging from 0.68 to 0.79. Sand, silt and clay ranges were 11.3-63.6%, 23.8-44.9% and 7.5-53.3% respectively. In contrast, Vestergaard *et al.*, (2021) reported that PLSR

performed poor in the prediction of clay (as well as silt and sand) in 13 different farms in Ontario, Canada. They suggested that the poor predictive ability of PLSR for clay was likely due to lower clay content (1.38 to 31.73%) in the soil samples utilized in their study than in those reported in previous studies. However, Viscarra Rossel *et al.*, (2006) stated in their study that, in many studies in literature in which PLSR was utilized to predict soil texture fractions, TN, and TC in the MIR range of the electromagnetic spectrum, PLSR yielded good results.

Studies show that the predictive ability of PLSR model is depends on various factors, amongst which is the selection of a single or combination of different spectral pre-processing method(s) (Stenberg *et al.*, 2010; Gautam *et al.*, 2015; Gholizadeh *et al.*, 2015; Minu, Shetty and Gopal, 2016; Dotto *et al.*, 2019; Vestergaard *et al.*, 2021). Another crucial factor is selecting the optimal number of principal components (Dangal *et al.*, 2019).

2.1.3 Application of PLSR: Continental (Africa) Perspective

Johnson *et al.*, (2019) analysed soil data sampled from multiple rice fields across sub-Saharan Africa using PLSR models in the NIR (7500 – 4000 cm^{-1}), MIR as well as when both regions were combined. From their observations, they demonstrated that soil texture was predicted fairly well, however, they also highlighted that when the measured sand content exceeded 74%, the sand model tended to overestimate. Furthermore, TN (with a range of 0.005 to 0.750 mg/kg) was among other soil fertility properties they assessed and obtained fairly good results and recorded an $R^2 > 0.75$. Baumann *et al.*, (2021) also reported similar findings in the Yam belt soils of West Africa, where PLSR-MIR models were calibrated to predict various soil properties including TC (ranging from 2.4 to 24.7 mg/kg) and TN (ranging from 0.18 to 2.48 mg/kg) and both were predicted with high accuracy with R^2 values of 0.92 and 0.89, respectively.

In contrast, Waruru *et al.*, (2015) tested the applicability of PLSR in the MIR range to predict soil stability properties for civil engineering applications in the western region of Kenya. Soil texture fractions were among the many soil properties they evaluated. According to their observations, PLS could not accurately predict total clay content (0-60%), total silt content (0-60%), and total sand content (~10-90%) as an R^2 value of less than 0.65 was recorded for all soil texture fractions. Takele & Iticha (2020) observed similar results for soil texture fractions which were poorly predicted using MIR-PLSR and were then predicted using conventional laboratory methods. Their data set range was 2.81- 66.25%, 2.69 - 50.14%, and 20.0 - 79.19% for sand, silt, and clay, respectively.

2.1.4 Application of PLSR: Local Perspective

The early scientific records about the use of infrared reflectance in South Africa dates back to 1980s. Research has shown that during that time, there was significant interest in applying this analysis method in plant tissue analysis, for measuring nitrogen (N) levels in leaf material. Meyer (1989) evaluated the use of NIR reflectance for determining organic matter, soil particle distribution, and the nitrogen-mineralizing potential of certain sugar industry soils. They noted that it was possible to simultaneously predict a variety of soil properties (from one individual spectrum) using NIR. Furthermore, they stated that calibrations for various plants such as sugarcane, maize, citrus and other plants including veld grasses had already been established. Paterson *et al.*, (2015) also noted that in addition to measuring important macro and micronutrients (aside from nitrogen) in soil and leaf samples, NIR spectroscopy has long been used in South Africa's sugarcane industry. It has been utilized to quantify soil texture, organic matter, and nitrogen as well as to satisfy the data-driven demands of precision agriculture. Furthermore, they stated that the KwaZulu-Natal Department of Agriculture and Environmental Affairs' Soil Fertility and Analytical Services laboratories have established robust MIR calibrations for predicting organic carbon, total nitrogen, and clay content.

Bayer *et al.*, (2012) conducted a study in the Subtropical Thicket in the Eastern Cape to assess the predictive ability of PLSR compared to multiple linear regression (MLR) in predicting SOC, iron oxides, and clay content demonstrated that SOC and iron oxides can be accurately predicted using PLSR. Furthermore, they concluded that the results produced by PLSR and MLR for predicting studied soil properties were similar. While Seboko *et al.*, (2023) conducted a study within the Johannesburg region in the Gauteng province of South Africa where they assessed the potential of Fourier-transform MIR spectroscopy in predicting soil carbon contents in granitic soils using PLS. In their findings they reported that soil carbon contents were predicted well with PLS and the results correlated with lab measurements. In another study, Nel *et al.*, (2023) assessed the predictive ability of PLSR models developed with NIR and MIR spectra. These models were calibrated with various chemical extraction methods for predicting soil pH and the levels of exchangeable basic cations in soils located in a Mediterranean climate zone in South Africa. Based on their results, they concluded that models developed with NIR-based models were accurate in predicting exchangeable Ca and Mg content, however, the models could not accurately predict other soil properties which include pH, exchangeable Na and K content.

Recently, a study by Sewpersad *et al.*, (2024) analysed the correlation between the SOM and the soil reflectance spectra collected using the Analytical Spectral Device FieldSpec3 instrument in the agricultural, residential, rangeland and eroded classes at the Emakhosaneni area of uThukela district, Kwazulu-Natal Province, South Africa. Their results showed that PLSR models developed with pre-processed data performed better than those developed with raw spectral data. Based on the results, they observed that pre-treating the spectra with the First-Order Derivative Transform (FDT) method and Savitzky-Golay smoothing was more suitable when predicting SOM content.

In addition, no other known studies were conducted in South Africa to evaluate various soil properties using spectroscopy and PLSR. Other authors did not measure soil spectra in the MIR wavelength range and did not use diffuse reflectance spectroscopy (Bayer *et al.*, 2012; Mashimbye *et al.*, 2012; Guo *et al.*, 2017; Nomngongo *et al.*, 2017; Bangelesa *et al.*, 2020a). Other South African-based studies in which PLSR was employed focused on plants (Mzinyane *et al.*, 2015; Kiala *et al.*, 2017; Reddy *et al.*, 2017).

2.1.5 Memory Based Learner (MBL)

Memory-based learning is a family of learning algorithms in machine learning that solves new problem instances by comparing them to previously seen instances stored in memory instead of performing explicit generalization, (Ramirez-Lopez *et al.*, 2013a). Figure 2.1 shows how MBL develops local partial least squares models to predict each sample using nearest neighbour search. Instead of fitting a general function (for prediction) to all samples, this machine-learning approach generates a new and distinct function (f_i) as needed when predicting a particular response variable Dangal *et al.* (2019). To create this new function, the approach only uses a subset of samples that are spectrally similar to the queried reference data,

which is determined through a k-nearest neighbour search (Gholizadeh *et al.*, 2016; Summerauer *et al.*, 2021).

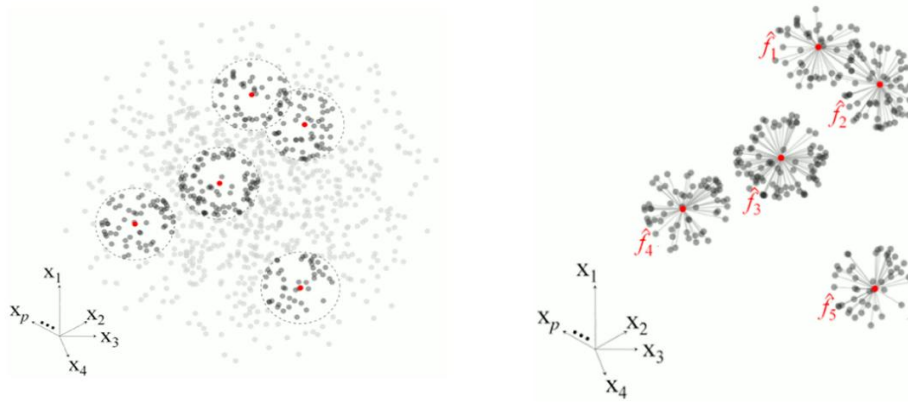


Figure 2.1: The red dots represent samples of interest, the dotted lines around them depict a boundary defined by a K-nearest neighbour search, for local modelling. The above figure was adopted from Ramirez-Lopez *et al.* (2013).

A study conducted by Ramirez-lopez and Stevens, (2013) demonstrated that MBL has the ability to handle complex spectral datasets. These observations were supported by Gholizadeh *et al.*, (2016) as they also reported that MBL resulted in a low prediction error and high R^2 values in many soil properties compared to PLSR and other methods. The MBL calibration procedure involves using a collection of n reference samples in conjunction with a training set composed of m samples.

2.1.6 Application of MBL: Global Perspective

Ramirez-Lopez *et al.* (2013a) conducted a study in which they demonstrated that memory-based learning (MBL) is a promising approach that is able to handle complex non-linear infrared spectral datasets. In their study, MBL was tested on a regional (São Paulo, Brazil) and global scale VNIR soil libraries and their study provides evidence that MBL produced more accurate results than PLSR, locally-weighted PLSR, and support vector machine in predicting clay content (Regional spectral library ranging from 1% to 81% and Global spectral library ranging from 0% to 96.8%), organic carbon (Regional spectral library ranging from 0.06% to 4.0% and Global spectral library ranging from 0.00% to 45.80%), and exchangeable calcium (Regional spectral library ranging from 1 cmol_c/kg to 170 cmol_c/kg and Global spectral library ranging from 0.00 cmol_c/kg to 168.20 cmol_c/kg). Gholizadeh *et al.* (2016) also reported high accuracy in predicting soil texture fractions in the VNIR region with R^2 values of 0.76, 0.81, and 0.89 for sand (9.1%-74.3%), silt (23.8%-48.0%), and clay (6.1%- 68.4%) respectively in

the Czech Republic. Dangal *et al.* (2019) and Tziolas *et al.* (2019) conducted independent studies in different geographic locations (USA & Middle East, respectively) and demonstrated that MBL performs better than global regression methods. Similarly, studies by Sanderman *et al.* (2020) & Ng *et al.* (2022) reported that MBL can accurately predict soil texture fractions, total nitrogen, and total carbon in both VNIR and MIR regions of the electromagnetic spectrum.

2.1.7 Application of MBL: Continental & Local Perspective

Summerauer *et al.* (2021) evaluated the applicability of MBL and MIR spectroscopy in predicting key soil properties such as TC and TN using the soil spectral library for Central Africa (CSSL). In their study, they targeted areas they identified as understudied African regions and demonstrated the usefulness of MBL in finding similar spectra in the soil library, thus reducing the need for spiking when new samples are incorporated into the library provided they cover the same PCA space. Furthermore, they stated that “*augmenting a smaller SSL (soil spectral library) with a continental SSL, even when scanned on different instruments, leads to reasonably accurate predictions for new regions which allows analyses of TC and TN*” (lines 478-479). Local studies evaluating the applicability of MBL and MIR spectroscopy in predicting soil properties are scarce.

2.2 The Dimensionality of Spectroscopy Data

The precision of the rapid estimation of soil properties with non-destructive spectroscopy is often compromised by multicollinearity and noise in the soil spectra (Wang *et al.*, 2022). Various variable selection methods have been developed to optimize model calibrations by reducing the full spectrum to fewer predictor variables (Vohland *et al.*, 2016; Wang *et al.*, 2022). Reducing predictor variables and retaining only informative ones in relation to soil properties of interest directly addresses issues such as redundancy, model complexity, and heavy computation associated with full spectra modelling (Li *et al.*, 2009; Ge, Wadoux and Peng, 2022; Zhang *et al.*, 2023). However, the variable selection in the MIR spectroscopic data has been poorly explored (Vohland *et al.*, 2014).

2.2.1 Variable Selection Techniques

Vohland *et al.* (2016) explored the application of selecting key predictor variables in the vis-NIR region ($\sim 28571\text{ cm}^{-1} - 4000\text{ cm}^{-1}$) for the estimation of organic carbon and total nitrogen from four sub-regions in Germany. A total of 172 soil samples were sampled at 0–10 cm depth with organic carbon ranging from 0.98–4.01 mg/kg and total nitrogen ranging from 0.085–0.347 mg/kg for the whole data set. They tested competitive adaptive reweighted sampling

(CARS) (Li *et al.*, 2009). They combined PLSR with CARS to improve prediction accuracy and their results show that when two or three of the studied four sub-regions dataset were combined, the results improved. The R^2 value of the organic carbon's PLSR model calibrated from the full spectrum was improved from 0.72 to $0.82 \leq R^2 \leq 0.85$. However, when a single model was calibrated from the combination of all four regions' datasets, they stated that “*estimates for OC could not even slightly be improved with CARS*” (page 261). They suggested that the poor organic carbon results may be due to high variability in the large dataset they used. It was also reported that total nitrogen was poorly predicted in all cases with an R^2 range of $0.63 \leq R^2 \leq 0.73$, these observations are similar to those reported by Martínez-España *et al.* (2019) who reported $R^2 = 0.71$ on 457 samples with a total nitrogen range of 0-0.38%.

Wang *et al.* (2022), investigated four variable selection algorithms; stability competitive reweighted sampling (sCARS) (Zheng *et al.*, 2012), bootstrapping soft shrinkage (BOSS) (Deng *et al.*, (2016), interval combination optimization (ICO) (Song *et al.*, 2016), and the combination of ICO and successive projections algorithm (SPA) (Araújo *et al.*, 2001). They developed PLSR model for predicting soil organic carbon in the MIR range with full spectra and then with selected variables. The data ranged from 0.22 - 2.97 mg/kg and the reported results show that PLSR models calibrated with selected variables outperformed the full-spectrum PLSR models. They also reported that the combination of ICO-SPA-PLSR was the most accurate model with an R^2 value of 0.93 and the authors further suggested that ICO-SPA variable selection “*may be a good alternative variables selection method for MIR spectroscopic data*” (Wang *et al.*, 2022, page 1).

Cheng *et al.* (2021) evaluated three variable selection methods, namely CARS, SPA and genetic algorithm (GA) (Babatunde *et al.*, 2014) combined with four multivariate techniques for predicting total nitrogen (range: 0.32-1.98) in the Vis-NIR region and they reported that CARS was the best variable selection technique for all four models even though all their models underperformed compared to other studies (Xie *et al.*, 2011; Sanderman *et al.*, 2020; Summerauer *et al.*, 2021). Additionally, similar results were also observed by Nawar and Mouazen (2017) without selecting informative key spectral variables, their data set range was 0.03-0.35.

In contrast, Zhang *et al.* (2023) stated that the results of their study “*showed that there was no silver bullet for the optimal variable selection algorithm among different predictive algorithms*” (page 1) and further stated that CARS produced the best results when combined

with PLS, while the recursive feature elimination (RFE) (Granitto *et al.*, 2006) performed best for random forest models and the Cubist algorithm produced best results with genetic algorithm. Furthermore, they reported that each multivariate method performed better with a matching variable selection algorithm, i.e., linear variable selection technique matched with a linear algorithm, and non-linear variable selection matched with a nonlinear algorithm. These observations are supported by Li *et al.* (2022). They also highlighted that their study focused on global modelling algorithms and suggested that variable selection for local modelling algorithms such as MBL should be investigated in future studies. Below is Table 2.1 with a tabulated summary of the results from the above-reviewed studies.

Table 2.1: Summary of the results from the above-reviewed studies.

Soil Property	Multivariate method	R ² (full spectrum)	R ² (selected variables)	Spectrum Range	Variable Selection Methods	Reference	Country
Organic Carbon	PLSR	0.90	0.93	MIR	sCARS, ICO-SPA & ICO	(Wang <i>et al.</i> , 2022)	China
Organic Carbon	PLSR	0.90		MIR	BOSS	(Wang <i>et al.</i> , 2022)	China
Total Nitrogen	PLSR	0.61	0.68	Vis-NIR	CARS	(Cheng, Wang and Du, 2021)	China
Total Nitrogen	PLSR	0.61	0.64	Vis-NIR	GA	(Cheng, Wang and Du, 2021)	China
Total Nitrogen	PLSR	0.61	0.46	Vis-NIR	SPA	(Cheng, Wang and Du, 2021)	China
Total Nitrogen	SVMR	0.61	0.75	Vis-NIR	CARS	(Cheng, Wang and Du, 2021)	China
Total Nitrogen	PLSR	0.72	$0.82 \leq R^2 \leq 0.85$	Vis-NIR	CARS	(Vohland <i>et al.</i> , 2016)	Germany

2.3 Model Accuracy Evaluation Metrics

In quantitative studies, such as those based on diffuse reflectance infrared spectroscopy, accuracy assessment metrics are used to evaluate model accuracy after calibration (Wadoux *et al.*, 2021a). A model's performance is assessed based on the comparison of laboratory-measured values versus the predicted values (Ge, Wadoux and Peng, 2022). There are five key model accuracy assessment parameters used in most soil spectroscopy studies, these include RMSE (root mean square error), R^2 (coefficient of determination), RPD (ratio of performance deviation), RPIQ (ratio of performance to interquartile range), and bias (Janik *et al.*, 2007; Terhoeven-Urselmans *et al.*, 2010; Dotto *et al.*, 2019; Ng *et al.*, 2022).

2.3.1 Root Mean Square Error (RMSE)

$RMSE_p$ is commonly used in spectroscopic literature to describe the model's predictive ability based on prediction error (Bellon-Maurel *et al.*, 2010) and is calculated as the square root of the sum of the squared difference of predicted and measured values of a validation set (Ludwig *et al.*, 2021).

Ludwig *et al.* (2021) highlighted that the value of RMSE is often affected by numerous factors such as sample preparation prior to scanning (grinding and drying) and signal-to-noise ratio which is resultant of the instrument used to scan soil samples. Wadoux *et al.* (2021) stated that the values of $RMSE_p$ of different datasets should not be compared because this accuracy assessment metric depends on the range of values in the data. Furthermore, they stated its values can only be greater or equal to zero, and that $RMSE = 0$ indicates a high-accuracy model and a perfect relationship between measured and predicted values.

2.3.2 Coefficient of Determination (R^2)

R^2 describes how much variance the model can explain and is computed as the “*sum of squared errors divided by the total sum of squares*” (Wadoux *et al.*, 2021). An optimal value of R^2 is positive 1, however, the R^2 value can also be negative. A positive R^2 describes the amount of variance the model explains, and a negative value indicates the model's inability to predict, thus, the mean of the observations is a better predictor (Ge, Wadoux and Peng, 2022).

2.3.3 Ratio of Performance to Inter-Quartile Range (RPIQ)

For several years infrared soil spectroscopy studies utilized RPD as one of the indices to evaluate the accuracy of the models, however, its limitation is that it assumes normal distribution of the data and therefore does not represent the spread of the population accurately

when the data is skewed (Bellon-Maurel *et al.*, 2010; Soriano-Disla *et al.*, 2014; Guo *et al.*, 2017). Bellon-Maurel *et al.*, (2010) proposed an alternative index called RPIQ, this index utilizes the quartiles to determine equivalent ranges of population spread. An RPIQ value greater or equal to 2.7 demonstrates good model estimation (Ludwig *et al.*, 2021).

2.3.4 Bias

Bias is a metric that quantifies the extent to which the model tends to overestimate or underestimate soil properties of interest (Ge, Wadoux and Peng, 2022). To quantify bias, mean square error is calculated. The optimal value of mean error is 0 which indicates that there is no bias which means that the model does not overestimate or underestimate, not that there is no error (Wadoux *et al.*, 2021).

2.4 Lessons Learned from the Literature

Based on the observations from the literature described in this study, IR spectroscopy in combination with PLSR is a commonly used statistical method for assessing and quantifying the concentration of various soil properties worldwide. On the other hand, even though MBL has proved to be superior to PLSR according to the studies mentioned in previous sections, this method is still underutilized thus far as compared to PLSR. Generally, the results from the studied literature indicate that MIR produces more precise estimation compared to NIR irrespective of the choice of calibration algorithm, however, in a study conducted by Johnson *et al.* (2019) the combination of both produced quite good results. Furthermore, it was also noted that in most studies, soil spectra were measured more in the NIR than in the MIR.

As noted in the literature review above, when predicting sand, silt, clay, TC, and TN, each multivariate method performs better with a matching variable selection algorithm as described by Zhang *et al.* (2023). The absorption features at 1100 - 1000, 1598–1600, 1862–1870, 2044–2058 cm^{-1} were found to be relevant for predicting sand content (Soriano-Disla *et al.*, 2014; Hutengs *et al.*, 2018). For predicting clay content, Hutengs *et al.*, (2018) reported that clay minerals such as kaolinite, smectite, and illite absorb mostly in the 3690–3300 cm^{-1} region. In addition, Le Guillou *et al.*, (2015) reported that the spectrum region at 1950–1750 cm^{-1} also corresponds to absorptions related to both sand and clay minerals. Soil carbon and TN can be predicted accurately with spectroscopy because of their association with vibrations related to functional groups such as hydroxyl, alkyl, protein, amides and aromatic groups (Ludwig *et al.*, 2023). These vibrations occur around 2900–2940, 2850, 1800–1570 cm^{-1} (Ludwig *et al.*, 2023).

Generally, soil texture, TC and TN were predicted well with both PLSR and MBL models at different scales i.e. regional, national and continental scale (Xie *et al.*, 2011; Johnson *et al.*, 2019; Summerauer *et al.*, 2021). This was true for models developed with full spectrum as well as models developed with selected variables datasets respectively (Soriano-Disla *et al.*, 2014; Nocita *et al.*, 2016a; Hutengs *et al.*, 2018; Sanderman *et al.*, 2020; Baumann *et al.*, 2021; Li *et al.*, 2022). The MBL models generally outperformed PLSR models in predicting both physical and chemical properties (Sanderman *et al.*, 2020). Currently, there is limited information on the application of variable selection in the MIR range particularly in conjunction with local modelling techniques such as MBL (Li *et al.*, 2022). Research indicate that there is also limited research on the use of MIR spectroscopy soil analysis technique on South African soils (Seboko *et al.*, 2023).

Chapter Three

Methodology

3.1 Soil Data Acquisition

In this research study, soil samples from various landscape positions and parent materials were selected from the archives of the ARC-Soil, Climate & Water Analytical Laboratory for analysis. These samples were submitted by farmers from across South Africa for fertility assessment. Before measuring soil spectra, soil samples were oven-dried (34-41 °C) and sieved to pass through a 2 mm sieve then stored in 300ml plastic containers at room temperature on the shelves. The soil samples are stored in chronological order in the archives. A total of 716 samples for soil texture fractions and 300 samples for total carbon and total nitrogen were selected from the archive using the Kennard Stone selection to ensure that the selected samples were diverse and not correlated with each other (Ferreira *et al.*, 2022). This also helped ensure that the selected samples were suitable for developing the calibration models to predict TC, TN, clay, silt, and sand contents.

During the selection of the soil samples from the archives, it was ensured that they had complete reference (soil property measurements from the laboratory) data. Otherwise, sub-samples of the selected group of samples were sent to the laboratory to be analyzed for total carbon, total nitrogen, and soil texture fractions adhering to standard laboratory procedures as outlined by The Non-Affiliated Soil Analysis Working Committee (1990). This data was then used as reference data during model training.

3.2 Soil Data Preparation for Spectral Measurements

In preparation for spectral measurements, soil samples were sub-sampled from the 300ml plastic containers, finely ground (<0.5 mm) using a Mortar Grinder Pulverisette 2, then transferred into 30ml plastic containers and labeled with corresponding sample IDs to avoid compromising the integrity of the data. In the last step of preparation, these finely ground sub-samples were mixed thoroughly with a spatula to maximize homogeneity before loading them in duplicates into alpha cups (instrument sample cups). The soil samples were duplicated to account for in-sample variation. After loading into each sample cup, the soil surface was flattened to avoid specular reflection (Xie *et al.*, 2011; Sommerauer *et al.*, 2021).

3.3 Soil Chemical and Physical Analysis

During soil chemical analysis, the total carbon and total nitrogen were both determined in the laboratory through total combustion using the CHNS-O analyzer. The soil physical analysis was also performed in the laboratory to determine soil texture fractions contents through the hydrometer method. During measurements, the laboratory personnel adhered to standard laboratory procedures as outlined by The Non-Affiliated Soil Analysis Working Committee (1990).

3.4 Laboratory Spectra Measurements

The spectral response of soil samples was measured in the MIR range of 4000 - 500 cm^{-1} with the Bruker Alpha II MIR DRIFT (Diffuse Reflectance Infrared Fourier Transform) spectrometer with the diffuse-reflectance accessory. Before acquiring the spectral response of soil samples, the temperature in the laboratory was set to 25°C and kept constant and the spectrometer was switched on and left to stabilize for at least 30 minutes. Background readings were measured at the start of each scanning session (eight samples per session) to ensure accurate spectral measurements and account for possible temperature and humidity fluctuations.

For every soil sample, the spectral measurements were replicated, and the average of these duplicate spectra was computed in R Studio. This approach was employed to consider the variations within each sample, as well as the differences in sample particle size and packing density (Terhoeven-Urselmans *et al.*, 2010). Each point of the spectral response resulted from 32 scans and the spectra were measured in the MIR range from 4000 to 500 cm^{-1} at 8 cm^{-1} resolution.

3.5 Outlier Analysis

Prior to PLSR and MBL model development, the spectra were assessed to screen for spectra that were not consistent with the rest of the spectra (spectral outliers). First, the spectra were plotted for visual analysis, and three spectral signatures that were clearly not similar in shape to the rest of the spectra were identified as outliers (noted their sample IDs) but were not removed at this stage. The region between 2500 cm^{-1} and 2000 cm^{-1} was where the outlier spectra were clearly distinct from the rest of the spectra.

Furthermore, since not all spectral outliers could be visualized from the plots of the spectra, the spectra were also plotted in the principal component space and outliers were determined

through Mahalanobis distance. This method tries to normalize the distribution of the data around its mean in the form of Gaussian distribution. The application of Mahalanobis distance ensured that features in the dataset were extracted such that the extremes in feature values indicate outliers (von Eye and Schuster, 1998). The Mahalanobis distance was computed based on the first and second principal components and outliers were identified based on the significance level of 0.95. A total of sixteen samples, including the three samples identified through a visual analysis were identified as outliers. These samples were removed and were not part of model development.

Another method of identifying and removing outliers that was used in this study was adopted from a study by de Santana *et al.*, (2018). Soil samples were considered to be outliers if they had residual values that were three times greater than the $RMSE_p$. The identified outliers were removed for each model and the final models were then redeveloped. This procedure was performed for both PLSR and MBL models.

3.6 Spectra Pre-processing

In IR spectroscopy, several algorithms are used to pre-process soil spectra. In this study, two pre-processing algorithms were used. The standard normal variate (SNV) as proposed by Barnes *et al.*, (1989) was used in this study to normalize spectral data and correct light scattering. To achieve this, this method subtracts each row of the spectral matrix from its mean and then divides it by its standard deviation.

Another pre-processing method that was used in this study is the Savitzky-Golay first derivative (SG) first proposed by Savitzky and Golay (1964). This method produces smoother wavelength values based on set window size and by applying a polynomial regression on the spectral matrix. After numerous tests, the optimal window size of the SG first derivative second-order polynomial for sand, silt, and clay contents was found to be 27, and 17 for total nitrogen and total carbon contents.

These preprocessing methods were chosen for this study based on their association with improved model performance as reported in previous studies (Rinnan *et al.*, 2009; Gholizadeh *et al.*, 2015; Dotto *et al.*, 2019; Barra *et al.*, 2021a; Vestergaard *et al.*, 2021).

Another important part of pre-processing spectra was the removal of spectral bands showing artifacts such as atmospheric carbon dioxide (CO_2) ($2350\text{ cm}^{-1} - 2380\text{ cm}^{-1}$). This region

contains information that is not relevant to changes in soil samples as suggested by Dangal *et al.*, (2019), and has a negative impact on the model performance.

3.7 Variable Selection

Variable selection algorithms were employed to reduce the dimensionality of the data and improve model computation and performance. In this study, Boruta, competitive adaptive reweighted sampling (CARS), and univariate variable selection (UVS) were tested. These three variable selection procedures were applied to raw and pre-processed spectral data for PLSR and MBL regression models. All variable selection methods were implemented in R (R Core Team 2023).

3.7.1 Boruta

The Boruta algorithm (Kursa and Rudnicki, 2010) emulate the random forest algorithm proposed by Breiman (2001), which adds randomness and compute results from the ensemble of randomized samples to minimize the misleading effects of random fluctuations and correlations. The Boruta algorithm measured the importance of each variable based on the Z score since it accounts for fluctuations of the mean accuracy loss among trees in the forest. However, since the Z score alone cannot directly determine the significance of the importance of each variable, then shadow attributes corresponding to each variable were created. In simple terms, shadow attributes are copies of variables.

The importance of all variables was then computed after performing classification using all variables including shadow attributes as well. The importance of each shadow attribute was used to decide whether its corresponding variable was truly important.

As described by Kursa and Rudnicki (2010), this algorithm works in the following sequence:

1. Adds shadow attributes
2. Shuffle shadow attributes to minimize their correlation to the target soil property
3. Run a random forest classifier to compute Z scores
4. Determine the maximum Z score among shadow attributes (MZSA)
5. For variables with undetermined importance, perform a two-sided equality test for each variable.
6. Remove variables with importance significantly lower than MZSA
7. Retain variables with importance significantly higher than MZSA.

This variable selection method was implemented using the Boruta package in R. During parameter tuning, the maxRuns parameter was set to 500 iterations. The Tentative Rough Fix function, which performed step 5 was applied to determine the importance of tentative variables. On average, the computational time of this algorithm was nine minutes for all soil properties in this study.

3.7.2 Competitive Adaptive Reweighted Sampling (CARS)

CARS is a variable selection method developed by Li *et al.*, (2009). This method uses PLS to establish the linear relationship between the matrix of predictor variables (X) and the vector containing corresponding values of the target soil property (y) based on latent variables (LVs), then selects a subset number of predictor variables from the full spectrum. The variables are selected based on four phases. In the first phase, the CARS algorithm performs the Monte Carlo strategy for model sampling in each sampling run. Secondly, it performs an enforced wavelength selection by utilizing the exponentially decreasing function (EDF). During the third phase, CARS performs the adaptive reweighted sampling (ARS) to exclude unimportant variables similar to Darwin's survival of the fittest evolution theory. Lastly, to evaluate the selected variables, internal cross-validation is performed. These phases are repeated N times to obtain an optimal subset of variables. The weights of the variables are therefore determined as follows:

$$y = Tc + e = XWc + e = Xb + e$$

Where T is the scores matrix, e is the prediction error, and b is the vector whose absolute value of each element reflects the contribution of a corresponding variable to y . The larger the element's value, the more important the corresponding variable is. After N iterations, the optimal subset of variables is retained based on the lowest $RMSE_{CV}$ which is computed based on the PLS model initially developed.

In this study, CARS was implemented using matlab and pls packages developed by Roebuck P (2022) and Liland K *et al.*, (2023) respectively. During parameter tuning, the number of iterations was set to 100 while the number of latent variables ($LV = 10$) and number of folds (fold = 10) were left on default. For all soil properties studied, the optimal number of iterations was <80.

3.7.3 Univariate Filter Variable Selection (UVS)

Interestingly, the caret package has the selection by filter (sbf) function (Kuhn, 2023) which was used in this study to select informative spectral variables. Each variable was passed to a

score function to establish the relationship between each predictor variable (x) and the target variable (y). The scores of each variable were passed to a logical vector to filter variables to be used to fit the subsequent random forest model. Only columns (variables) with the value TRUE were retained for model fitting. This random forest model was used to establish the relationship between x and y and subsequently rank the variables based on their contribution in predicting y. Variables with high scores were then retained while variables with low scores were not considered important. The general syntax for the *sbfc* function was given as follows:

$$sbfc(x, y, sbfControl = sbfControl(), \dots)$$

The internal repeated cross-validation resampling method was performed with 10 folds and 5 repetitions. This method is known as univariate filter or univariate variable selection (UVS).

3.8 Calibration and Validation Datasets

In this study, the dataset was split into two groups: the calibration dataset and the validation dataset. The former was utilized to train the model, the latter was used as an independent set of samples that were put aside during model training and used later to validate the model. The dataset was split such that 70% of the samples were allocated for model training, and the remaining 30% was for validation.

For clay, silt, and sand contents, the dataset was split using stratified random sampling. This method of partitioning the data ensured that the calibration dataset and validation dataset covered the diversity in the whole dataset and maintained the same proportion. Stratified random sampling stratifies the data and then randomly selects samples from each stratum based on the values of a single soil property at a time. On the contrary, this method was found to be not as effective when applied to the total nitrogen and total carbon contents dataset. Therefore, to achieve a similar result, the total nitrogen and total carbon contents data was chronologically ordered and then every third sample was selected for the validation dataset while the remaining samples were assigned to the calibration dataset.

3.9 Modelling Soil Properties

To predict each of the studied soil properties, PLSR and MBL regression models were developed, first with full spectrum datasets and then with selected variable datasets. The full spectrum dataset comprised raw spectra, spectra pre-processed with SG, and spectra pre-processed with SNV, and the same was true for the selected variable dataset as well. These regression models were developed in R.

3.9.1 Partial Least Squares Regression (PLSR)

PLSR was introduced first by Geladi & Kowalski (1986) for building predictive models from a dataset containing many predictor variables by considering the correlation between spectral measurements and soil chemical or physical properties (Ji *et al.*, 2016; Dangal *et al.*, 2019; Hati *et al.*, 2022). A PLSR model establishes the relationship between the predictor variables (X) and response variables (Y) simultaneously (Viscarra Rossel *et al.*, 2006). Nawar & Mouazen (2017) stated this chemometric technique selects orthogonal factors known as latent variables to maximise covariance between X and Y variables. These “*latent variables are simply linear combinations of the original independent variables*” (Bangelesa *et al.*, 2020, page 4). According to (Geladi and Kowalski, 1986; Nocita *et al.*, 2016b). The model can be expressed as follows:

$$U = TB + E_d$$

Where B is a diagonal regression coefficient matrix and E_d denotes a random error. Thus, the chemical concentration of unknown samples can only be determined through their spectral measurements using the formula below.

$$Y = x(UX)'BQ$$

Interestingly, in situations where the number of variables significantly surpasses the number of available samples, PLSR is able to handle such cases (Nomngongo *et al.*, 2017). The *pls* package in R developed by Liland K *et al.*, (2023) was used in this study to develop PLSR models. For the *pls()* function tuning parameters, the maximum number of components was set to 25 (for all datasets with a number of predictor variables >25), and the number of folds for the internal cross-validation was set to 10. This regression method was chosen for this study for its ability to handle collinearity very well (Rossel and Behrens, 2010; Kiala, Odindi and Mutanga, 2017) and its ability to predict soil properties with high accuracy.

3.9.2 Memory-Based Learner (MBL)

Memory-based learning is a machine learning algorithm developed by Ramirez-Lopez *et al.*, (2013a), it solves new problem instances by comparing them to previously seen instances stored in memory instead of performing explicit generalization. Instead of fitting a general function (for prediction) to all samples, this machine-learning approach generates a new and distinct function (*fi*) as needed when predicting a particular response variable (Dangal *et al.*, 2019). To create this new function, the approach only uses a subset of samples that are

spectrally similar to the queried reference data, which is determined through a k-nearest neighbor search (Gholizadeh *et al.*, 2016; Summerauer *et al.*, 2021).

The MBL calibration procedure involves using a collection of n reference samples in conjunction with a training set composed of m samples:

$$(X_r, Y_r) = \{X_{ri}, Y_{ri}\}_{i=1}^n$$

$$(X_u, Y_u) = \{X_{uj}, Y_{uj}\}_{j=1}^m$$

Where X_r is a matrix of predictor variables corresponding to the reference data of the calibration dataset, and Y_r is a vector that contains numeric values of response variables corresponding to the reference dataset of the calibration dataset. X_u is a matrix of the prediction dataset, and Y_u is an optional matrix containing the values of the response variables corresponding to the data to be predicted (independent validation dataset). Before modelling, each data point in the prediction dataset must have an established k-nearest neighbour. Using these referenced neighbours, a local model is trained to estimate Y_u 's corresponding value based on X_u . The weighted partial least square average was then used to fit the local models (at each neighbourhood). This weighted average was calculated from the set of predicted values generated from several partial least square models using a range between minimum and maximum PLS components (Ramirez-lopez and Stevens, 2013; Ramirez-Lopez *et al.*, 2013b).

During the development of the MBL regression model for this study, the resemble package developed by Ramirez-Lopez *et al.*, (2023) was used. The *mbl()* function included in the resemble package has a range of methods that fit a local regression model at k-nearest neighbours. For this study, the local fit method was locally weighted average partial least squares (*local_fit_wapls*) with a maximum of 19 components, the k-nearest neighbor search was sequenced such that the minimum k-neighbor is 20. The maximum was 500 for sand, silt, and clay contents models at a 20 neighbour interval, whereas due to a limited number of samples for total nitrogen and total carbon, the maximum number of nearest neighbors was set to 200 at a 20 neighbors interval as well. The validation type was set to nearest neighbor validation (*NNv*) for all datasets. This modeling approach was adopted in this study for its feasibility of handling complex spectral datasets and high predictive ability.

3.10 Model Assessment

This study evaluated model performance using the R^2 , RMSE, bias, and RPIQ and utilized a range of values as described in the literature for each metric to assess the goodness of fit. All the PLSR and MBL regression models developed in this study had to meet all the requirements described below to be categorized as good models.

3.10.1 Coefficient of Determination (R^2)

R^2 describes total variance the model can explain, and is computed as “*the sum of squared errors divided by the total sum of squares*” (Wadoux *et al.*, 2021). An optimal value of R^2 is positive 1. R^2 is computed through the equation below:

$$R^2 = 1 - \frac{\sum_{i=1}^n (obs_i - pred_i)^2}{\sum_{i=1}^n (obs_i - \overline{obs})^2}$$

Where n is the total number of samples, obs_i is the vector of measured values, $pred_i$ is the vector of estimated values and $\overline{obs}$ is the mean of measured values (Wadoux *et al.*, 2021a). In this study, when each of the regression models was assessed based on the R^2 value, it was considered to be good if it produced an R^2 value exceeding 0.7 (irrespective of the soil property being predicted) as stated by Nel *et al.*, (2023). They claimed that models with an R^2 value greater than 0.7 demonstrate a strong correlation between the predictor variables and the target soil property.

3.10.2 Root Mean Square Error of Prediction (RMSE_p)

RMSE_p describes the model's precision based on prediction error (Bellon-Maurel *et al.*, 2010) and is calculated as the square root of the sum of the squared difference of predicted and measured values of a validation set (Ludwig *et al.*, 2021). The equation is expressed as follows:

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

Where N is the number of the validation dataset samples, $\hat{y}_i$, and y_i are vectors of predicted and measured values, respectively. A RMSE_p value that is equal to zero indicates a high-accuracy model and a perfect relationship between measured and predicted values (Wadoux *et al.*, 2021).

In addition to the R^2 value, Nel *et al.*, (2022) stated that good models must also have a $RMSE_p$ value that is less than 10% of the range of the target soil property. This was added to the criteria used in this study to assess PLSR and MBL based on the $RMSE_p$ value.

3.10.3 Ratio of Performance to the Inter-Quartile Range (RPIQ)

For several years, infrared soil spectroscopy studies utilized RPD as one of the indices to evaluate the accuracy of the models. However, its limitation is that it assumes normal distribution of the data, and therefore does not represent the actual distribution of the population accurately when the data is skewed (Bellon-Maurel *et al.*, 2010; Soriano-Disla *et al.*, 2014; Guo *et al.*, 2017). Bellon-Maurel *et al.*, (2010) proposed an index called RPIQ, this index utilizes the quartiles to identify equivalent ranges of population distribution. RPIQ is calculated as follows:

$$RPIQ = \frac{IQ}{RMSE}$$

Where IQ is the interquartile range calculated from the predicted values ($IQ = Q3_{pred} - Q1_{pred}$). The RPIQ range of values given in a study by Nawar and Mouazen (2017) was adopted in this study to assess both PLSR and MBL performances. In their study, they stated that models with RPIQ values greater than 2.5 were excellent, those that had RPIQ values ranging between 1.7 and 2.5 were identified as good models, and those with an RPIQ value less than 1.4 as very poor models.

3.10.4 Bias

Bias is a metric that quantifies the extent to which the model tends to overestimate or underestimate soil properties of interest (Ge, Wadoux and Peng, 2022). To quantify bias, the mean square error was calculated (Viscarra Rossel *et al.*, 2006; Ge, Wadoux and Peng, 2022) using the equation is as follows:

$$MSE = \sum_{i=1}^n \frac{(obs_i - pred_i)^2}{n}$$

The optimal value of mean error is 0 which indicates that there is no bias (the model does not overestimate or underestimate), not that there is no error. A negative value reveal that the model tend to underestimate, and a positive value indicates a model's tendency to overestimate (Wadoux *et al.*, 2021). This metric was used as an indicative measure of both PLSR and MBL's tendency to overestimate or underestimate.

Chapter Four

Results

4.1 Descriptive Statistics

The analysis revealed that numerous studied soil properties significantly diverged from the normal distribution, as demonstrated by their high skewness values, reflecting asymmetry, and kurtosis values, which measure the extent of light or heavy tails compared to the normal distribution. Table 4.1 presents the descriptive statistics for sand, silt, clay, TC, and TN contents. The skewness and kurtosis values of all studied soil properties indicate that both calibration and validation datasets were skewed (and heavy-tailed). The similarity between the means and median values indicated a harmonious division of respective datasets.

The sand content exhibited a broad range, with values ranging between 8% and 92%, a mean of 67.58% and a median was 72%. The slightly high median value suggested that most of the samples had a high sand composition. The standard deviation for sand content was calculated at 15.76%, indicating a moderate spread around the mean. The skewness of -0.98 reflected a left-skewed distribution, implying that many samples were concentrated towards the higher end. The coefficient of variation (CV) was determined to be 0.23 (23%), suggesting that the variability of sand within the dataset was relatively low.

In contrast, the silt content was found to be much lower, ranging from 0% to 38%, with a mean and median of 8.43% and 8.00%, respectively. The value of the median was almost equal to the mean indicating an almost symmetric distribution around the lower silt content range. The standard deviation and skewness for silt were calculated at 5.1 and 1.16 respectively. The skewness value was indicative of a right-skewed distribution, suggesting that a smaller number of samples had higher silt content. The coefficient of variation (CV) for silt was 0.61 (61%), suggesting a higher relative variability than sand.

Clay content ranged between 6% and 72%, with a mean of 23.99% and a median of 20%, which implied that the data was slightly skewed towards the lower end. The standard deviation for clay was 9.78, and the CV of 0.51 indicates a moderately high variability level. Similar to silt content, the skewness value was calculated at 1.18 for clay content indicating a right-skewed distribution, suggesting that fewer samples had a higher percentage of clay.

Both the TC and TN content datasets exhibited the highest CV values, calculated at 0.81 (81%) and 0.87 (87%), respectively. These datasets were highly skewed, based on the calculated skewness values of 2.45 for TC and 4.43 for TN, indicating a right-skewed distribution.

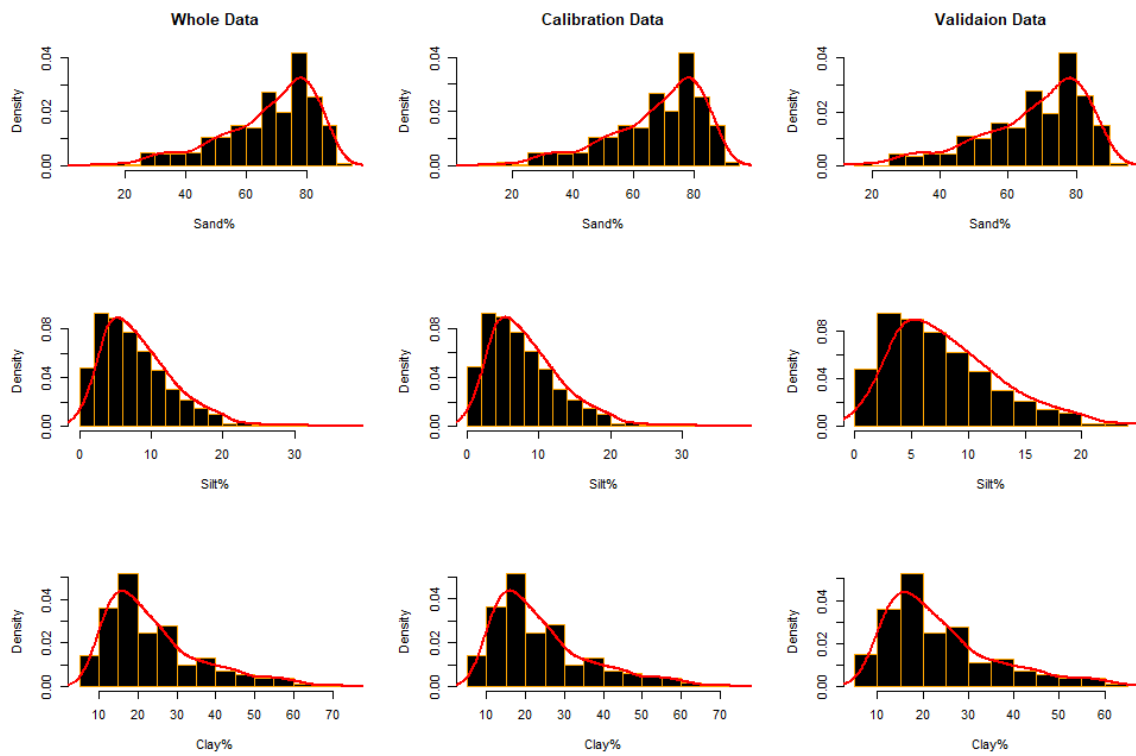
Overall, the data suggested that the sand content in the samples was generally high with low variability, while silt and clay contents are low with moderate to high variability. Silt, clay, TC, and TN contents datasets tended to be right-skewed. Based on the observation by Xiao *et al.*, (2016), this observed variability in the (dataset) could negatively influence the accuracy of the respective subsequent models. The equal means and median values were indicative of a balanced partitioning of the data.

Table 4.1: Descriptive statistics of sand, silt, clay, total carbon, and total nitrogen contents.

Soil Property	Dataset	Number of Samples (n)	Min	Max	Mean	Median	Standard Deviation	CV	Skewness	Kurtosis
Sand	Whole data	716	8.00	92.00	67.58	72.00	15.76	0.23	-0.98	0.47
	Calibration	500	8.00	92.00	67.44	72.00	15.98	0.24	-1.00	0.53
	Validation	216	18.00	92.00	67.93	72.00	15.24	0.22	-0.91	0.22
Silt	Whole data	716	0.00	38.00	8.43	8.00	5.11	0.61	1.16	2.40
	Calibration	501	0.00	38.00	8.50	8.00	5.25	0.62	1.28	2.94
	Validation	215	0.00	24.00	8.27	8.00	4.77	0.58	0.76	0.17
Clay	Whole data	716	6.00	72.00	23.99	20.00	9.78	0.51	1.18	1.01
	Calibration	501	6.00	72.00	23.99	20.00	9.78	0.51	1.18	1.01
	Validation	215	6.00	72.00	24.12	20.00	12.43	0.52	1.20	1.07
Total Carbon	Whole data	300	0.04	7.07	1.30	1.01	1.05	0.81	2.45	7.96
	Calibration	200	0.04	6.49	1.28	1.01	1.01	0.79	2.41	7.88

	Validation	100	0.06	7.07	1.32	1.02	1.11	0.84	2.57	8.61
Total Nitrogen	Whole data	300	0.00	0.83	0.11	0.08	0.09	0.87	4.43	26.77
	Calibration	200	0.00	0.82	0.11	0.08	0.09	0.85	4.28	26.56
	Validation	100	0.02	0.83	0.11	0.08	0.10	0.91	4.36	26.07

For visualization, the statistics from Table 4.1 were represented in histograms in Figure 4.1 to show the distribution of the data.



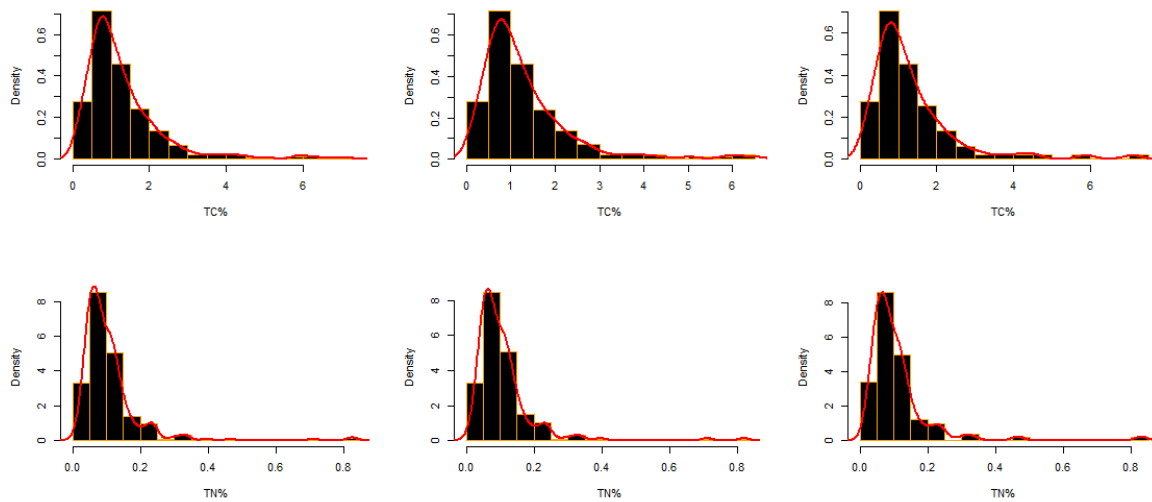


Figure 4.1: Histograms of sand, silt, clay, TC and TN contents showing their respective distribution, respectively (ordered in rows from top to bottom). The histograms in the first column are for the whole dataset, in the second and third columns are for calibration and validation sets respectively.

4.2 Predicting Soil Properties Using Full Spectrum Datasets

The current study focused on evaluating several spectral pre-processing methods along with both linear (PLSR) and nonlinear (MBL) regression approaches for the estimation of three physical soil parameters (sand, silt, clay) and two chemical properties (total carbon and total nitrogen). MBL models had better performance than PLSR models in predicting studied soil properties.

4.2.1 Predicting Sand Content

The predictive performance of full spectrum PLSR and MBL models, both for raw and pre-processed spectra, is summarized in Table 4.2.

The PLSR models of sand content prediction showed R^2 and $RMSEP$ values ranging from 0.57 and 10.05% to 0.67 and 8.59%, respectively. Similarly, MBL models of sand content prediction showed R^2 and $RMSEP$ values ranging from 0.67 and 8.86% to 0.74 and 7.65%, respectively. MBL models developed with full spectrum raw spectra demonstrated the best predictive ability. Among the six sand content predictive models, two presented an R^2 value higher than 0.70. Models developed with SG pre-processed data (PLSR-SG and MBL-SG), and models developed with raw spectra dataset (PLSR-Raw and MBL-Raw) tended to perform better than those developed with SNV pre-processed dataset (PLSR-SNV and MBL-SNV)

Table 4.2: Validation results of the sand content predicted using full spectrum raw spectra and pre-processed spectra datasets respectively.

Soil Property	Model	Pre-Processing Method	R ²	RMSE _p	Bias	RPIQ
Sand	PLSR	Raw Spectra	0.65	9.04	-1.16	2.21
		SG	0.67	8.59	-1.02	2.33
		SNV	0.57	10.05	-0.72	1.99
	MBL	Raw Spectra	0.74	7.65	0.01	2.61
		SG	0.73	7.63	0.18	2.62
		SNV	0.67	8.86	0.21	2.26

4.2.2 Predicting Silt Content

The PLSR models of silt content prediction showed R² and RMSE_p values ranging from 0.57 and 3.13% to 0.67 and 2.69%, respectively. Similarly, for MBL models of silt content prediction showed R² and RMSE_p values ranging from 0.65 and 2.78% to 0.69 and 2.53%, respectively. Among the six silt content predictive models, none presented an R² value higher than 0.70 (Table 4.3).

The trend in silt content models was analogous to that observed for the sand content. Developing predictive models for silt content with SNV pre-processed dataset yielded poor results than when developed with raw and SG pre-processed full spectrum datasets. It was also noted that PLSR models tended to slightly over-predict compared to MBL models with bias values varying between 0.14 and 0.29 for PLSR and 0.02 and 0.18 for MBL models respectively.

Table 4.3: Validation results of the silt content predicted using full spectrum raw spectra and pre-processed spectra datasets respectively.

Soil Property	Model	Pre-Processing Method	R ²	RMSE _p	Bias	RPIQ
Silt	PLSR	Raw Spectra	0.67	2.69	0.14	2.97
		SG	0.65	2.77	0.29	2.53
		SNV	0.57	3.13	0.21	2.56
	MBL	Raw Spectra	0.69	2.53	0.02	2.77
		SG	0.69	2.63	0.14	3.04

		SNV	0.65	2.78	0.18	2.69
--	--	-----	------	------	------	------

4.2.3 Predicting Clay Content

The PLSR models of clay content prediction showed R^2 and $RMSE_p$ values ranging from 0.66 and 6.61% to 0.69 and 6.34%, respectively. Similarly, for MBL models of clay content, prediction showed R^2 and $RMSE_p$ values ranging from 0.73 and 5.87% to 0.78 and 5.22%, respectively. Among the six clay content predictive models, three (all MBL models) presented an R^2 value higher than 0.70 (Table 4.4). This suggests MBL models for predicting clay content had a better accuracy than PLSR and showed stronger correlation between the spectra and clay content.

Table 4.4: Validation results of the clay content predicted using full spectrum raw spectra and pre-processed spectra datasets respectively.

Soil Property	Model	Pre-Processing Method	R^2	$RMSE_p$	Bias	RPIQ
Clay	PLSR	Raw Spectra	0.67	6.54	-0.34	1.83
		SG	0.69	6.34	-0.38	1.89
		SNV	0.66	6.61	-0.55	1.82
	MBL	Raw Spectra	0.78	5.22	-0.05	2.3
		SG	0.75	5.65	0.35	2.21
		SNV	0.73	5.87	-0.43	2.04

4.2.4 Predicting Total Carbon Content

The PLSR models of TC content prediction showed R^2 and $RMSE_p$ values ranging from 0.44 and 0.52% to 0.51 and 0.58%, respectively. Similarly, for MBL models of TC content, prediction showed R^2 and $RMSE_p$ values ranging from 0.31 and 0.61% to 0.49 and 0.75%, respectively. Among the six TC content predictive models, all presented an R^2 value below 0.70 (Table 4.5), which suggests that the performance of both PLSR and MBL models developed with full spectrum dataset, either raw or pre-processed datasets for predicting TC content produced inadequate results.

Table 4.5: Validation results of the total carbon content predicted using full spectrum raw spectra and pre-processed spectra datasets respectively.

Soil Property	Model	Pre-Processing Method	R ²	RMSE _p	Bias	RPIQ
TC	PLSR	Raw Spectra	0.51	0.52	0.15	1.59
		SG	0.44	0.58	0.0	1.43
		SNV	0.47	0.57	-0.02	1.47
	MBL	Raw Spectra	0.49	0.75	0.18	1.16
		SG	0.31	0.74	0.11	1.16
		SNV	0.43	0.61	0.17	1.37

4.2.5 Predicting Total Nitrogen Content

Similar to the results observed above for TC content, the performance metrics suggested that both PLSR and MBL could not predict TN accurately with full spectrum datasets whether raw or pre-processed.

The PLSR models of TN content prediction showed R² and RMSE_p values ranging from 0.22 and 0.05% to 0.42 and 0.06%, respectively. Similarly, for MBL models of TN content, prediction showed R² ranged from 0.35 to 0.61% with RMSE_p values of 0.05%, respectively. Among the six TN content predictive models, none presented an R² value higher than 0.70 (Table 4.6).

Table 4.6: Validation results of the total nitrogen content predicted using full spectrum raw spectra and pre-processed spectra datasets respectively.

Soil Property	Model	Pre-Processing Method	R ²	RMSE _p	Bias	RPIQ
TN	PLSR	Raw Spectra	0.39	0.05	0	1.2
		SG	0.22	0.06	0	1.07
		SNV	0.42	0.05	0	1.23
	MBL	Raw Spectra	0.61	0.05	0	1.40
		SG	0.35	0.05	0.01	1.20
		SNV	0.50	0.05	0.02	1.24

4.3 Improving Models Performance Using Variable Selection Methods

In an attempt to improve PLSR and MBL models' predictive ability based on the results observed with full-spectrum raw and pre-processed datasets, variable selection algorithms (Boruta, CARS, and UVS) were applied to reduce the dimensionality of the datasets. The variable selection algorithms reduced dimensionality by selecting fewer predictor variables that contain useful information in relation to the target soil property. The new datasets with fewer predictor variables were then used to calibrate predictive models for sand, silt, clay, TC, and TN contents. MBL models developed with reduced spectral features were generally better compared to PLSR models for predicting all studied soil properties even though in some cases there was only a slight difference.

4.3.1 Sand Content

The results for predicting sand content produced with Boruta and CARS selected variables were comparable and both PLSR and MBL could not predict sand content well with these datasets. As shown in Table 4.7, the R^2 values ranged from 0.58 to 0.69 and 0.65 to 0.70 for PLSR and MBL respectively. The bias values were below zero for both PLSR and MBL and this was an indication that both these models tended to under predict sand content. When sand content was predicted using the UVS-selected variables from a raw spectra dataset, both PLSR and MBL models, had the highest R^2 values ($R^2 \geq 0.76$) and lowest prediction errors respectively. The MBL model developed with predictor variables selected using UVS from raw spectra had excellent predictive ability.

Table 4.7: PLSR and MBL validation results of the sand content predicted using predictor variables selected using the Boruta, CARS, and UVS algorithms from raw spectra and pre-processed spectra datasets respectively.

			Boruta				CARS				UVS			
Soil Property	Model	Pre-Processing Method	R ²	RMSE _p	Bias	RPIQ	R ²	RMSE _p	Bias	RPIQ	R ²	RMSE _p	Bias	RPIQ
Sand	PLSR	Raw Spectra	0.66	8.72	-0.44	2.29	0.68	8.48	-0.66	2.36	0.76	7.47	-0.77	2.68
		SG	0.69	8.28	-0.09	2.42	0.68	8.66	-0.95	2.31	0.67	8.60	-1.01	2.32
		SNV	0.58	9.89	-0.96	2.02	0.63	9.10	-0.2	2.20	0.60	9.66	-1.15	2.07
	MBL	Raw Spectra	0.67	8.61	-0.46	2.32	0.70	8.30	-0.37	2.41	0.81	6.59	0.12	3.03
		SG	0.70	8.21	-0.61	2.44	0.68	8.37	-0.05	2.39	0.72	7.89	0.10	2.54
		SNV	0.65	9.04	-0.64	2.21	0.67	8.57	-0.55	2.33	0.68	8.71	0.40	2.30

4.3.2 Silt Content

Models developed with variables selected with CARS algorithm for predicting silt content were generally superior. The results shown in Table 4.8 show that amongst all silt content models developed with variables selected using CARS algorithm, the models developed with SG pre-processed dataset were superior with the highest $R^2 = 0.78$ for MBL, while PLSR recorded the highest $R^2 = 0.77$ with the same dataset. Silt content models developed with SNV pre-processed datasets were the worst. The PLSR models developed with Boruta and UVS selected variables produced relatively similar results for predicting silt content with R^2 values < 0.70 . MBL on the other hand demonstrated a stronger correlation between the soil spectra and silt content when developed with raw spectra (Boruta and UVS selected variables) and SG pre-processed (Boruta selected variables) dataset producing R^2 values of 0.73.

Table 4.8: PLSR and MBL validation results of the silt content predicted using predictor variables selected using the Boruta, CARS, and UVS algorithms from raw spectra and pre-processed spectra datasets respectively.

			Boruta				CARS				UVS			
Soil Property	Model	Pre-Processing Method	R ²	RMSE _p	Bias	RPIQ	R ²	RMSE _p	Bias	RPIQ	R ²	RMSE _p	Bias	RPIQ
Silt	PLSR	Raw Spectra	0.68	2.65	0.27	3.02	0.71	2.56	0.27	3.13	0.67	2.72	0.18	2.20
		SG	0.67	2.70	0.08	2.96	0.77	2.26	0.27	3.32	0.66	2.76	0.02	2.90
		SNV	0.63	2.80	0.29	2.50	0.61	2.87	0.29	2.44	0.59	3.05	0.11	2.62
	MBL	Raw Spectra	0.73	2.39	0.33	2.93	0.76	2.31	0.04	3.46	0.69	2.63	-0.02	2.28
		SG	0.73	2.41	0.09	2.91	0.78	2.23	0.03	3.37	0.73	2.45	0.24	3.06
		SNV	0.67	2.65	0.12	2.83	0.67	2.64	0.20	2.65	0.63	2.88	0.06	2.78

4.3.3 Clay Content

As shown in 4.9, all MBL clay content models developed with selected variables demonstrated a relatively strong predictive ability ($R^2 > 0.71$) and were considered good except for one model developed with SNV pre-processed (Boruta selected variables) dataset which had an R^2 value < 0.71 . The best MBL model for predicting clay content produced an R^2 value of 0.8 and was considered excellent. The R^2 values for PLSR models for predicting clay content with selected variables varied from 0.63 to 0.77 and the best result was produced with UVS selected variables dataset. In addition, the highest $RMSE_p$ and lowest R^2 recorded by both PLSR and MBL were observed when these models were developed with SNV pre-processed dataset.

Table 4.9: PLSR and MBL validation results of the clay content predicted using predictor variables selected using the Boruta, CARS, and UVS algorithms from raw spectra and pre-processed spectra datasets respectively.

			Boruta				CARS				UVS			
Soil Property	Model	Pre-Processing Method	R^2	$RMSE_p$	Bias	RPIQ	R^2	$RMSE_p$	Bias	RPIQ	R^2	$RMSE_p$	Bias	RPIQ
Clay	PLSR	Raw Spectra	0.71	6.07	-0.36	1.98	0.75	5.63	0.30	2.13	0.77	5.59	0.41	2.50
		SG	0.65	6.65	-0.41	1.80	0.71	6.16	-0.07	1.95	0.69	6.31	-0.19	1.90
		SNV	0.63	6.84	-0.22	1.76	0.67	6.44	-0.34	1.86	0.64	6.72	-0.34	1.79
	MBL	Raw Spectra	0.75	5.58	-0.10	2.15	0.76	5.41	0.34	2.22	0.80	5.25	0.08	2.67
		SG	0.72	5.86	-0.20	2.13	0.78	5.33	0.42	2.25	0.76	5.47	0.44	2.28
		SNV	0.69	6.56	-0.30	1.83	0.72	5.92	-0.17	2.03	0.74	5.73	-0.48	2.09

4.3.4 Total Carbon Content

The results in Table 4.10 show that TC was poorly predicted by both PLSR and MBL models when developed with selected variables datasets. The former produced a maximum R^2 value of 0.51 and the latter produced 0.56 for all datasets (raw and pre-processed spectra datasets).

Table 4.10: PLSR and MBL validation results of the TC content predicted using predictor variables selected using the Boruta, CARS, and UVS algorithms from raw spectra and pre-processed spectra datasets respectively.

			Boruta				CARS				UVS			
Soil Property	Model	Pre-Processing Method	R^2	RMSE _p	Bias	RPIQ	R^2	RMSE _p	Bias	RPIQ	R^2	RMSE _p	Bias	RPIQ
TC	PLSR	Raw Spectra	0.43	0.60	0.15	1.39	0.48	0.58	0.13	1.44	0.46	0.57	0.12	1.45
		SG	0.51	0.54	-0.02	1.55	0.48	0.65	-0.01	1.31	0.46	0.56	-0.02	1.48
		SNV	0.43	0.58	0.01	1.44	0.40	0.60	0.03	1.40	0.41	0.59	-0.03	1.41
	MBL	Raw Spectra	0.39	0.67	0.21	1.25	0.56	0.64	0.10	1.32	0.35	0.74	0.29	1.14
		SG	0.44	0.62	0.13	1.36	0.36	0.68	0.16	1.22	0.38	0.63	0.14	1.32
		SNV	0.39	0.63	0.18	1.32	0.39	0.64	0.17	1.30	0.33	0.65	0.10	1.29

4.3.5 Total Nitrogen Content

Similar to TC, the validation results in Table 4.11 show that both PLSR and MBL models developed with selected variables datasets for predicting TN content performed poorly with both models producing maximum R^2 values <0.6 .

Table 4.11: PLSR and MBL validation results of the TN content predicted using predictor variables selected using the Boruta, CARS, and UVS algorithms from raw spectra and pre-processed spectra datasets respectively.

			Boruta				CARS				UVS			
Soil Property	Model	Pre-Processing Method	R^2	RMSE _p	Bias	RPIQ	R^2	RMSE _p	Bias	RPIQ	R^2	RMSE _p	Bias	RPIQ
TN	PLSR	Raw Spectra	0.31	0.06	0	1.13	0.41	0.05	0	1.21	0.40	0.05	0	1.21
		SG	0.32	0.06	0	1.13	0.34	0.06	0	1.16	0.30	0.06	0	1.12
		SNV	0.42	0.05	0	1.22	0.47	0.05	0	1.28	0.37	0.06	0	1.18
	MBL	Raw Spectra	0.44	0.05	0.01	1.22	0.54	0.05	0.01	1.37	0.59	0.04	0.01	1.44
		SG	0.46	0.05	0.01	1.21	0.55	0.05	0.01	1.36	0.41	0.05	0	1.19
		SNV	0.46	0.05	0.01	1.24	0.30	0.06	0	1.11	0.47	0.06	0.02	1.14

4.4 Determination of the Best Model for Each Soil Property

Based on the criteria used in this study to determine a good predictive model as described in the previous chapter, R^2 value has to exceed 0.7, $RMSE_p$ has to be less than 10% of the range of the independent validation dataset and the RPIQ value must be greater than 1.7. Out of all the good predictive models for each soil property studied, the best models (one PLSR and one MBL model) were selected including the worst performing models and their results were displayed in tables. However, since all models for predicting TC and TN were poor their results were not reported below.

4.4.1 Sand Content

With the full-range spectrum datasets, raw or pre-processed, MBL models for predicting sand content performed better than PLSR. The worst results produced by an MBL model (MBL-SNV) were the best PLSR (PLSR-SG) could produce with the same dataset. The RPIQ values were relatively similar for both methods, the difference in performance was more evident with R^2 values. Both MBL and PLSR underperformed and had similar results when developed with selected variables from a raw spectra dataset using Boruta and CARS. However, the use of UVS to select important variables from the raw spectra improved the performance of both modeling algorithms. In addition, PLSR could not predict sand content better than MBL. Generally, the performance of PLSR models was either similar to or worse than MBL models for predicting sand content. The best predictive model for sand content was the MBL model developed with selected variables from raw spectra using UVS. In Table 4.12 are the results of the best and worst-performing models for each modeling algorithm. The best model for predicting sand content was highlighted.

Table 4.12: Comparison between MBL and PLSR: Results of the best and worst sand content models.

Best Models	Best Results	Worst Models	Worst Results
MBL-RAW (UVS)	$R^2 = 0.81$, $RMSE_p = 6.59$, bias = 0.12, and RPIQ = 3.03	MBL-SNV (Boruta)	$R^2 = 0.65$, $RMSE_p = 9.04$, bias = -0.64, and RPIQ = 2.21
PLSR-RAW (UVS)	$R^2 = 0.76$, $RMSE_p = 7.47$, bias = -0.77, and RPIQ = 2.68	PLSR-SNV	$R^2 = 0.57$, $RMSE_p = 10.05$, bias = -0.72, and RPIQ = 1.99

4.4.2 Silt Content

Both MBL and PLSR underperformed in predicting silt content with full-spectrum data (raw and pre-processed), however, MBL predicted silt content better than PLSR with the same data. MBL performed better with raw full-range spectra than with pre-processed full-range spectra

whereas PLSR performed better with SG pre-treated full-spectrum data. MBL models developed with selected variables (raw and pre-processed), generally performed better than PLSR with the same data. However, MBL-SG (CARS) and PLSR-SG (CARS) had similar results and these were the best models for predicting silt content in this study. The results from the two aforementioned models were inseparable, thus both models had an equal ability to predict silt content. The results of the best and worst-performing models for each modeling algorithm are shown in Table 4.13, the best models for predicting silt content were highlighted.

Table 4.13: Comparison between MBL and PLSR: Results of the best and worst silt content models.

Best Models	Best Results	Worst Models	Worst Results
MBL-SG (CARS)	$R^2 = 0.77$, $RMSE_p = 2.25$, bias = 0.06, and RPIQ = 3.33	MBL-SNV (UVS)	$R^2 = 0.63$, $RMSE_p = 2.88$, bias = 0.06, and RPIQ = 2.78
PLSR-SG (CARS)	$R^2 = 0.78$, $RMSE_p = 2.26$, bias = 0.27, and RPIQ = 3.32	PLSR-SNV	$R^2 = 0.57$, $RMSE_p = 3.13$, bias = 0.21, and RPIQ = 2.56

4.4.3 Clay Content

MBL models were better than PLSR models for predicting clay content with all datasets (full-spectrum raw, full-spectrum pre-processed, and selected variables). The best model for predicting clay content in the current study was an MBL model developed with selected variables from raw spectra using UVS. The most accurate PLSR model in predicting clay content was also developed with the same dataset and the performance difference between the best MBL model and the best PLSR model for predicting clay content was marginal. It was also noted that the worst models for both MBL and PLSR were developed with the same dataset (variables selected from an SNV pre-processed data using Boruta). The results of the best and worst-performing models for each modeling algorithm are shown in Table 4.14, the best models for predicting silt content were highlighted.

Table 4.14: Comparison between MBL and PLSR: Results of the best and worst clay content models.

Best Models	Best Results	Worst Models	Worst Results
MBL-RAW (UVS)	$R^2 = 0.80$, $RMSE_p = 5.25$, bias = 0.08, and RPIQ = 2.67	MBL-SNV (Boruta)	$R^2 = 0.69$, $RMSE_p = 6.61$, bias = 0.01, and RPIQ = 1.82
PLSR-RAW (UVS)	$R^2 = 0.78$, $RMSE_p = 5.59$, bias = 0.41, and RPIQ = 2.51	PLSR-SNV (Boruta)	$R^2 = 0.63$, $RMSE_p = 6.84$, bias = -0.22, and RPIQ = 1.76

4.5 Variable (Wavelengths) Selection

4.5.1 Sand Content

For sand contents, the Boruta variable selection method tended to generally select most key wavelengths from the double bonds and fingerprint regions ($\leq 2000\text{ cm}^{-1}$) as shown in Figure 4.2. This region is influenced mostly by quartz and silicate minerals. From the raw spectral data, apart from the double bonds and fingerprint regions, fewer spectral variables were selected in the X-H single bond region where clay minerals ($3671\text{ cm}^{-1} - 3540\text{ cm}^{-1}$) are mostly dominant. It was also noted that in the same region, between 3542 cm^{-1} and 2344 cm^{-1} , there were no spectral variables selected. This trend was witnessed in the SNV pre-processed data as well. From a total of 1703 spectral variables (full spectrum), 157 (9.2%), 157 (9.2%), and 123 (7.2%) variables were selected from raw spectral data, SG pre-processed, and SNV pre-processed spectral data respectively.

Interestingly, CARS selected critical wavelengths across the spectrum, predominantly from the double bonds and fingerprint regions, similar to Boruta. All selected variables were representative of all four sub-regions (X-H single bond ($>2500\text{ cm}^{-1}$), triple bonds ($2500 - 2000\text{ cm}^{-1}$), double bonds ($2000 - 1500\text{ cm}^{-1}$), and fingerprint ($<1500\text{ cm}^{-1}$)) of the MIR range of the spectrum. Only 27 (1.6%) spectral variables were selected from the SNV pre-processed spectral data, while from raw spectra and SG pre-processed spectra, a total of 189 (11%) and 64 (3.8%) were selected respectively.

In contrast, the UVS algorithm eliminated fewer unimportant spectral variables compared to Boruta and CARS. However, even though UVS had selected 981 (57.6%) variables from the raw spectra, the predictive models (MBL and PLSR) developed with those variables were the best sand content models compared to full-spectrum models in this study.

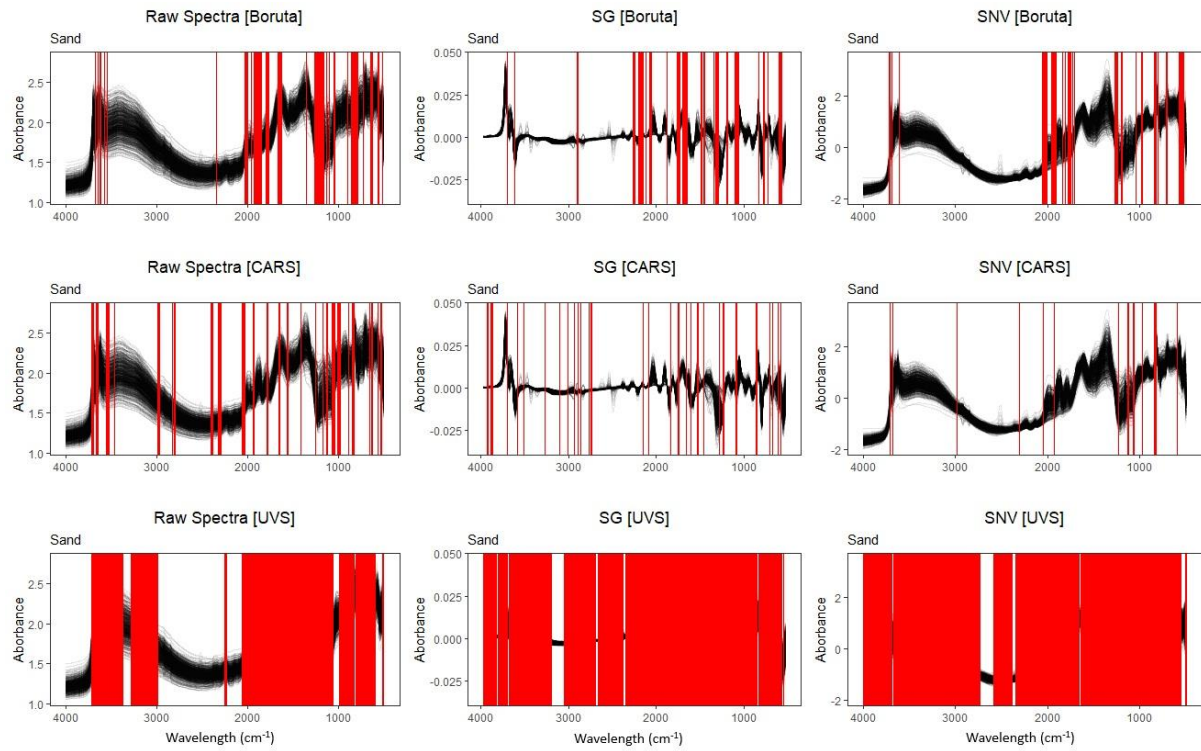


Figure 4.2: Plots showing selected spectral variables for predicting sand content. The variables were selected from both unprocessed spectra and pre-processed spectra through Boruta, CARS, and UVS algorithms.

4.5.2 Silt Content

Boruta predominantly identified important spectral variables for predicting silt content between 2276cm^{-1} and 1033cm^{-1} for both unprocessed and pre-processed datasets. The peaks (3745cm^{-1} - 3609cm^{-1}) associated with kaolinite minerals in the X-H single bond region were commonly identified as important variables for silt content from all datasets. It was also observed that Boruta tended to select few variables or none at all at $\sim 1000\text{cm}^{-1}$ or less (Figure 4.3). The total number of important variables selected by Boruta for predicting silt content was less than 10% of the total variables in the full spectrum, for each dataset (raw and pre-processed) respectively.

On the other hand, CARS and UVS exhibited a similar trend as seen with sand content on Figure 4.2 where CARS identified important variables across the spectrum and UVS eliminated a few variables identified as unimportant. CARS selected less than 5% of the total variables in the full spectrum, while UVS selected over 50% for each dataset respectively.

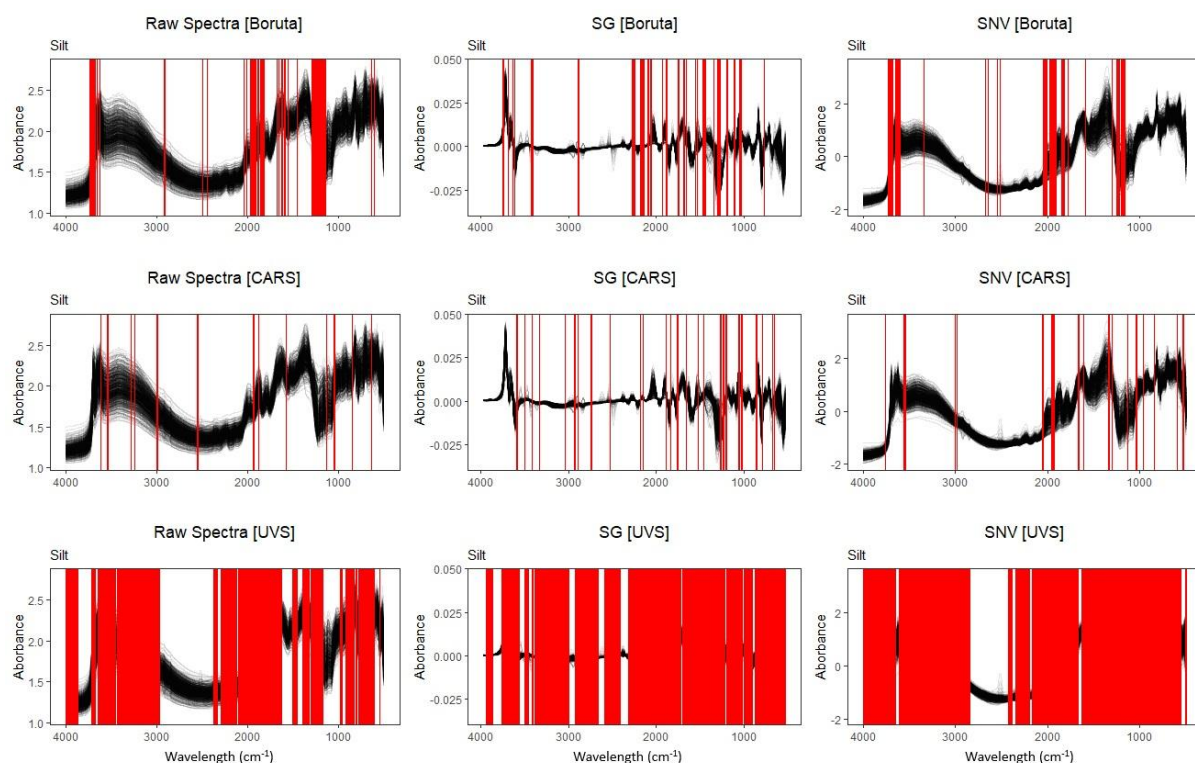


Figure 4.3: Plots showing selected spectral variables for predicting silt content. The variables were selected from both unprocessed spectra and pre-processed spectra through Boruta, CARS, and UVS algorithms.

4.5.3 Clay Content

As shown in Figure 4.4, the Boruta algorithm selected key variables for predicting clay content predominantly in the double bonds and fingerprint regions of the MIR spectrum. Some variables near the peaks and absorption features highly correlated with clay minerals between the ranges of 3700 and 3400 cm^{-1} and $\sim 1200 - 900\text{ cm}^{-1}$ were selected from all datasets (raw and pre-processed). Boruta tended to select few or no variables in the region between $\sim 3500 - 2300\text{ cm}^{-1}$.

Similar to Boruta, the CARS algorithm predominantly selected key variables for predicting clay content in the double bonds and fingerprint regions and included the ranges associated with clay minerals as indicated above. Noticeably, CARS included few variables in the region between $\sim 3500 - 2300\text{ cm}^{-1}$ while Boruta did not select any variables in that range.

In contrast, the UVS algorithm selected many variables for predicting clay content, more than any other variable selection method used in this study. Since there were very few variables that UVS did not select, therefore most of the variables selected by both Boruta and CARS were also selected by UVS, and the regions associated with clay minerals were also included.

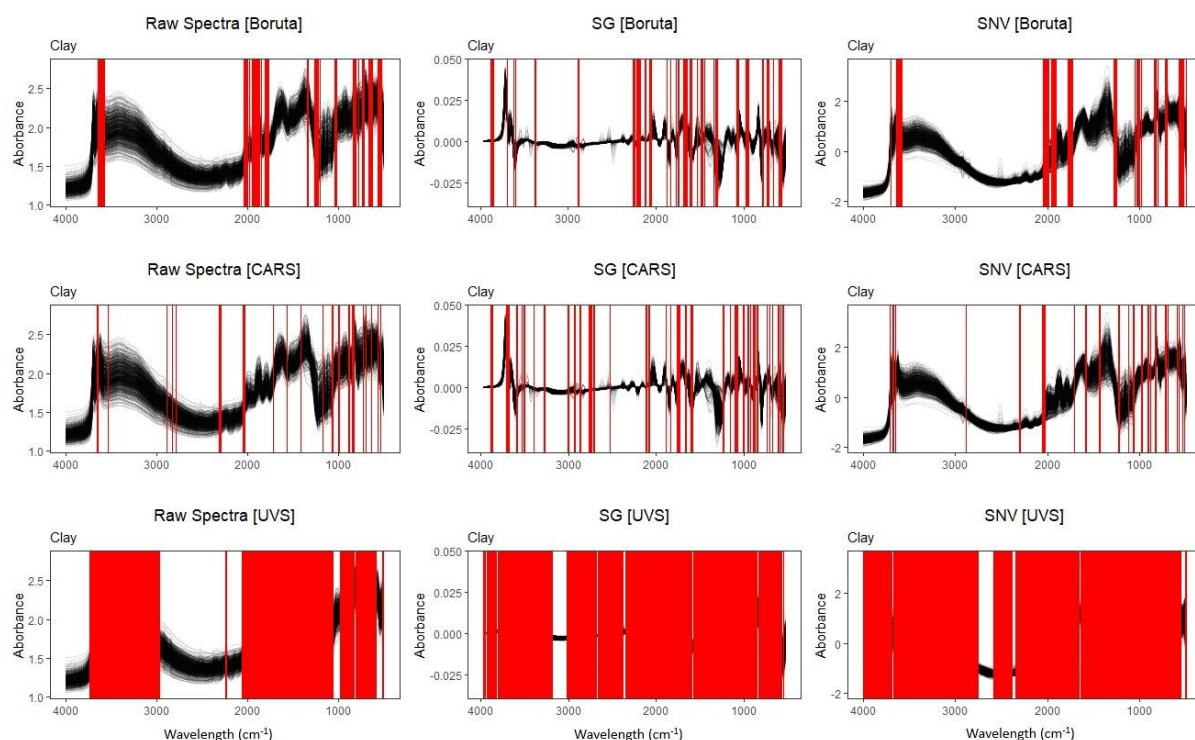


Figure 4.4: Plots showing selected spectral variables for predicting clay content. The variables were selected from both unprocessed spectra and pre-processed spectra through Boruta, CARS, and UVS algorithms.

4.5.4 Total Carbon Content

The Boruta algorithm selected important variables for predicting TC content between 3618 cm^{-1} and 1370 cm^{-1} . This algorithm also managed to select variables in the region $\sim 2900\text{ cm}^{-1} - 2800\text{ cm}^{-1}$ where there are absorption peaks highly associated with organic matter. As presented in Figure 4.5, only a single variable was selected by Boruta from the pre-processed datasets in the region $>3000\text{ cm}^{-1}$, whereas none were selected from the raw spectra dataset.

In contrast, the CARS algorithm did not select any variables between $\sim 3689\text{ cm}^{-1}$ and 2268 cm^{-1} from the raw spectra dataset and the SNV processed dataset. This region encompasses the previously discussed organic matter-sensitive region ($\sim 2900\text{ cm}^{-1} - 2800\text{ cm}^{-1}$). The UVS algorithm noticeably selected a bigger number of variables compared to Boruta and CARS. These variables were selected across the spectrum. Similar to Boruta, UVS selected variables in the organic matter-sensitive region as well.

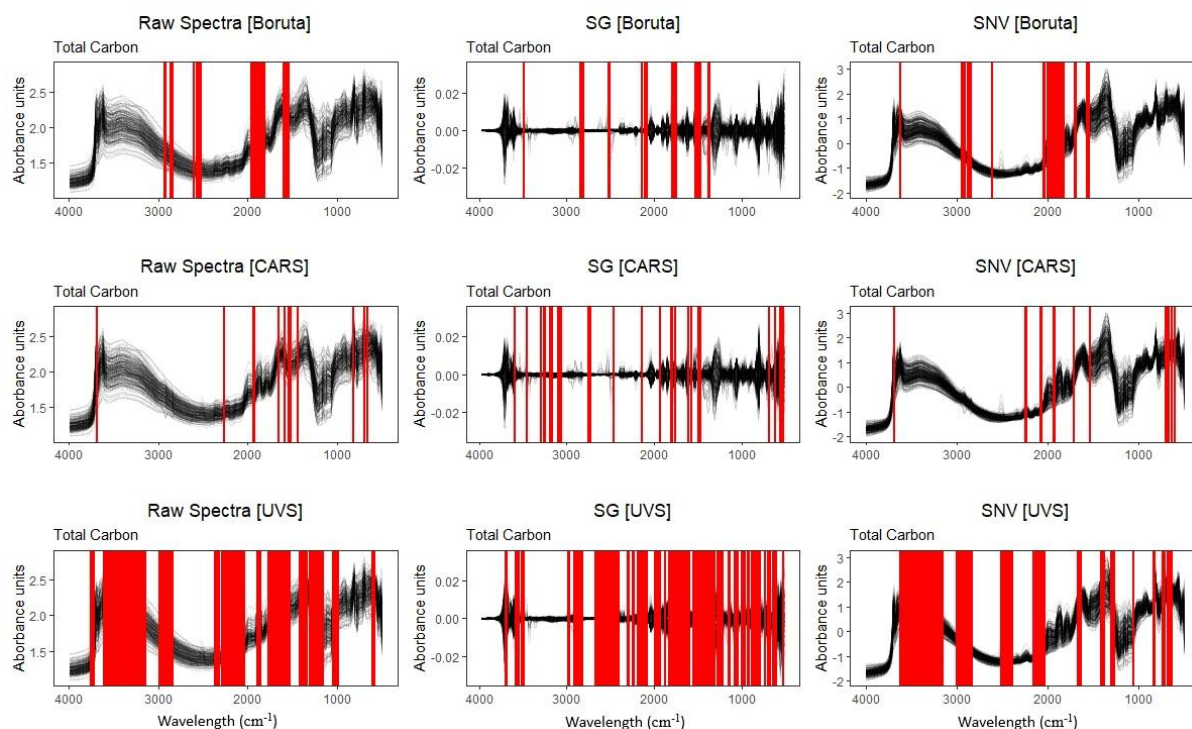


Figure 4.5: Plots showing selected spectral variables for predicting TC content. The variables were selected from both unprocessed spectra and pre-processed spectra through Boruta, CARS, and UVS algorithms.

4.5.5 Total Nitrogen Content

For predicting TN content, as shown in Figure 4.6, the Boruta algorithm selected variables in the range between 2935 and 1078 cm^{-1} while CARS generally selected individual variables across the spectrum. The UVS algorithm selected the most number of variables. It selected over 70% of the total number of variables in the full spectrum.

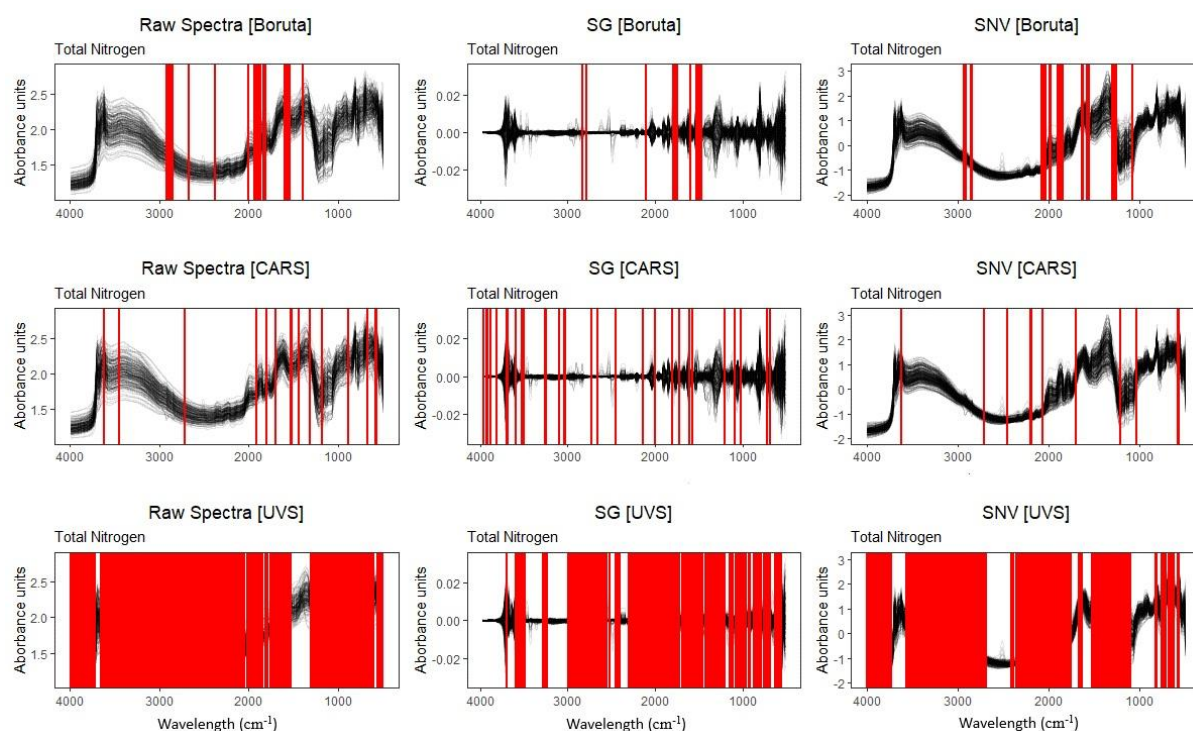


Figure 4.6: Plots showing selected spectral variables for predicting TN content. The variables were selected from both unprocessed spectra and pre-processed spectra through Boruta, CARS, and UVS algorithms.

4.5.6 Frequently Selected Variables by Each Selection Method

The selected specific key variables for predicting each specific studied soil property varied depending on the variable selection algorithms and datasets used in this study. Table 4.15 presents variables that were frequently selected by each variable selection algorithm from raw and pre-processed datasets to predict each of the studied soil properties. The number of most frequently selected variables by the Boruta algorithm were 3 for sand and clay content, 8 for silt content, 1 for TC content, and none for TN content. CARS did not select any common variables from the raw and pre-processed spectra datasets for predicting sand, silt, TC, and TN. Only 2 variables were frequently selected by CARS for predicting clay content. On the other hand, UVS selected at least 116 common variables from all datasets. Single variables or ranges that were selected by at least two variable selection algorithms were highlighted in red as shown in Table 4.15

Table 4.15: Frequently selected key variables by each variable selection method from all datasets (raw and pre-processed spectra) for predicting each studied soil property.

Soil Property	Variable Selection Method	Number of Variables (n)	Frequent Variables (cm⁻¹)
Sand	Boruta	3	1191, 827, 830
	CARS	0	None
	UVS	833	3716-3630, 3628-3385, 3371, 3281-3198, 3053-2991, 2245-2231, 2058-1738, 1731-1060, 982-795, 786-721, 704-638, 626-618, 612-595
Silt	Boruta	8	3685-3687, 1931, 1195-1187
	CARS	0	None
	UVS	786	3934-3868, 3712-3673, 3611-3564, 3500-3457, 3414-3406, 3385-2999, 2288-2194, 2167-2122, 2097-1725, 1703-1670, 1631-1618, 1505-1497, 1489-1452, 1395-1302, 1294-1224, 1201-1173, 977-957, 914-906, 879-832, 809-749, 743-731, 692-601
Clay	Boruta	3	3602, 1769, 721
	CARS	2	881,879
	UVS	891	3737-3184, 3020-2972, 2243-2233, 2056-1937, 1931-1660, 1637-1592, 1573-1571, 1565-1060, 980-879, 873-782, 772-721, 704-638, 626-585
TC	Boruta	1	2849
	CARS	0	None
	UVS	116	3583-3578, 3558-3550, 3513-2983, 2977-2974, 2917-2876, 2853-2841, 2161-2114, 2109-2085, 1670-1662, 1651-1633, 1413-1382, 1304
TN	Boruta	0	None
	CARS	0	None
	UVS	439	3568-3554, 3531-3521, 3498-3488, 3283-3274, 3250-3233, 2991-2929, 2921-2689, 2407-2403, 2305-2192, 2169-2116, 2109-2052, 2013-1941, 1918-1805, 1795-1754, 1670-1658, 1649-1635, 1286-1107, 830, 749-745, 723-710, 682, 680, 624

Chapter Five

Discussion

This research focused on developing a spectral library based on South African soils and demonstrating the ability of MIR diffuse reflectance spectroscopy coupled with PLSR and MBL to predict soil texture, and TC, TN. The main objectives were to identify key spectral variables and regions for predicting each of the studied properties, as well as to establish good predictive models for each soil property based on performance metrics as described in the methodology chapter of this study. The results demonstrated that sand and clay content were predicted accurately with MBL models developed with a full spectrum dataset. TC and TN were poorly predicted with both PLSR and MBL models developed with either a full spectrum dataset or a selected variables dataset. Pre-processing the spectra did not improve model performance in predicting any of the studied soil properties. However, selecting important spectral variables for predicting soil texture proved to be effective as models developed with selected variables were superior to those developed with full spectrum.

5.1 Estimation of Soil Properties Using MIR Full Spectrum Datasets

5.1.1 Sand Content

Based on R^2 , all PLSR models developed with full spectrum datasets (raw or pre-processed spectra) for predicting sand content performed relatively poor, producing an R^2 value < 0.7 . These findings were similar to reports by Johnson *et al.*, (2019) and Terhoeven-Urselmans *et al.*, (2010) in separate studies, one conducted in Sub-Saharan Africa and the other conducted using a globally distributed soil spectral library. On the other hand, MBL models developed in this study with the raw and SG pre-processed full spectrum datasets showed a stronger correlation between the spectra and sand content, with R^2 values exceeding 0.70. However, when the MBL sand content model was developed with the SNV pre-processed full spectrum dataset, the results were similar to those produced by the PLSR models.

These results differ from the findings observed by Sanderman *et al.*, (2020). They conducted a study with over 50 000 soil samples from the soil library of United States of America. In their findings, they reported 0.77 and 0.96 R^2 values produced by PLSR and MBL models respectively. Their models were validated with 6 982 (independent dataset) soil samples which underwent the square root transformation prior to model development for all soil properties they studied. The differences in performance could be due to the different data transformation

methods used and the smaller number of samples (716 in total) in this study, considering that the spatial sampling scale was similar.

5.1.2 Silt Content

Similar to sand content, silt content was also predicted poorly in this study when PLSR and MBL models were developed with full-spectrum datasets. Even though the results were poor, MBL models for predicting silt content performed slightly better than PLSR models developed with the same datasets. The results produced by PLSR and MBL models for predicting silt content in this study were in contrast with observations reported by Hati *et al.*, (2022), Johnson *et al.*, (2019) and Sanderman *et al.*, (2020) however, they were better than results reported by Viscarra Rossel *et al.*, (2006). The poor silt content prediction in the current study may have been caused by the fact that silt content is not associated with any particular compound absorptions in the MIR spectrum (Le Guillou *et al.*, 2015). Additionally, the soil samples utilized in this study had small amounts of silt (compared to clay and sand content) with an average silt content of 8.43% and a narrow range of 38%. Pirie *et al.*, (2005) who obtained an R^2 value of 0.34 from their silt content prediction stated that the poor results of silt was caused by the narrow range of the silt content in their dataset.

5.1.3 Clay Content

Similar to sand and silt content, PLSR could not accurately predict clay content with full spectrum datasets producing R^2 values less than 0.7. These findings aligned with the results observed by Viscarra Rossel *et al.*, (2006) while in contrast with other previous literature (Terhoeven-Urselmans *et al.*, 2010; Dangal *et al.*, 2019; Johnson *et al.*, 2019; Hati *et al.*, 2022).

According to the criteria established in this study to determine a good model, clay content was well predicted by the MBL models developed with full spectrum datasets with R^2 ranging from 0.73 to 0.78. The MBL model developed with raw spectra for predicting clay content was superior to other clay content models (including PLSR models) developed with full spectrum datasets. The findings of this study were consistent to observations reported by Hati *et al.*, (2022) and Terhoeven-Urselmans *et al.*, (2010) obtained using PLSR. However, these results were poorer than the findings reported by Ng *et al.*, (2022) and Sanderman *et al.*, (2020) who observed R^2 values > 0.8 using MBL in their respective studies.

5.1.4 Total Carbon and Total Nitrogen Content

TC and TN content were poorly predicted by PLSR and MBL models developed with full spectrum datasets, and R^2 values hardly exceeded 0.6. Generally, PLSR predicted TC better

and MBL predicted TN content better in this study. Noticeably, the better results were produced with raw spectra and the SNV pre-processed full spectrum datasets yielded second-best results for both TC and TN content models. This suggests that for predicting TC and TN it may not be necessary to pre-process the spectra, otherwise SNV pre-processing may be chosen. The PLSR and MBL models' prediction shortcomings could be associated with the limited sample size (300 samples) in relation to the spatial sampling scale (national). The prediction results produced in this study were contrasting to previous findings (Xie *et al.*, 2011; Johnson *et al.*, 2019; Sanderman *et al.*, 2020; Baumann *et al.*, 2021).

5.2 Estimation of Soil Properties Based on Selected Key Variables in the MIR Range

For enhancing estimation of soil properties, variable selection algorithms such as Boruta, CARS, and UVS were utilized to reduce dimensionality by selecting important variables for each soil property. However, it was difficult to find studies that developed predictive models for predicting soil texture with selected variables datasets especially the ones mentioned above. Hence the results obtained with selected variables datasets were not contrasted with previous studies. Previous studies that utilized variable selection algorithms to determine important variables for predicting soil properties used the NIR range rather than the MIR range of the spectrum.

5.2.1 Sand Content

Table 4.7 presents the outcomes of PLSR and MBL models for predicting sand content, utilizing three algorithms (Boruta, CARS, and UVS) to select informative variables from both raw and pre-processed spectral datasets. Notably, the PLSR models constructed with the Boruta, CARS, and UVS variable selection techniques (as shown in Table) generally yielded comparable or enhanced RMSE_p, R², and RPIQ values when contrasted with full spectrum models (Table 4.2). In contrast, the trends observed in the MBL models were inconsistent. The MBL models that were constructed with SNV pre-processed variables selected by UVS and CARS, yielded slightly improved results that were comparable to those of the full spectrum MBL models. Furthermore, other MBL models developed with selected variables had relatively poor results compared to full spectrum MBL models.

The improved PLSR model performance when utilizing variables selected with CARS algorithm was consistent with observations by Zhang *et al.*, (2023) who reported that PLSR model performance improved when utilizing variables selected with CARS. They further stated

that this was due to the fact that CARS uses PLSR function to calculate variable importance which obeys the linear relationship during computation. However, the improved model performance observed with models developed with Boruta and UVS which fit a nonlinear random forest function to calculate variable importance contradicts this finding. Additionally, Vohland *et al.*, (2014) also demonstrated that PLSR in conjunction with CARS yields improved results compared to full spectrum.

Both Boruta and CARS variable selection algorithms significantly reduced the number of variables to less than 12% of the full spectrum datasets while keeping similar or better results for predicting sand content with PLSR. Meanwhile, UVS retained 57.6% of important variables from full spectrum raw spectra and produced the best sand content models for both PLSR and MBL. A detailed examination of all the results indicated that despite the noted improvements in the PLSR models, both the PLSR and MBL models utilizing variables selected by Boruta and CARS algorithms struggled to accurately predict sand content for this particular dataset. This means none of the models developed with the aforementioned datasets met the criteria outlined in this study for determining a good model.

5.2.2 Silt Content

The analysis of silt content revealed that the models constructed using PLSR and memory-MBL models, which incorporated variable selection methods such as Boruta, CARS, and UVS (Table 4.8), generally yielded comparable or enhanced $RMSE_p$, R^2 , and RPIQ values when contrasted with full spectrum models (Table 4.3). This observation indicate that, although there are no mineral absorptions in the MIR spectrum that can be directly linked to silt according to Le Guillou *et al.*, (2015), the variable selection algorithms used in this study managed to identify important variables that contribute to minimal error for silt prediction. Although the prediction results improved with variable selection, however, not all models developed with selected variables could predict silt content accurately.

Furthermore, the models that utilized the informative variables selected by the CARS algorithm demonstrated a notable superiority over those derived from other variable selection approaches especially when the data was pre-processed with SG. The data presented in Table 4.8 illustrates that both MBL and PLSR models achieved the highest R^2 and RPIQ values and low $RMSE_p$ values when employing the CARS algorithm. This result was achieved with $< 10\%$ of the number of full spectrum variables and these models can accurately estimate silt content in the real world.

5.2.3 Clay Content

The models developed for clay content, specifically the PLSR models, which utilized the CARS variable selection algorithm (Table 4.9), generally demonstrated a consistent improvement in R^2 , RPIQ and $RMSE_p$ values when compared to full spectrum models (Table 4.4). The PLSR developed with UVS-selected variables produced inconsistent results. The MBL models consistently produced better results when important variables were selected using the UVS algorithm. Only one MBL model developed with raw spectral variables selected by CARS showed improved results. Variables selected by Boruta produced relatively poor results compared to full spectrum models, except for one PLSR model developed using raw spectral variables.

Compared to the Boruta variable selection algorithm, CARS and UVS were more consistent in improving predictions for clay content. Furthermore, Boruta and CARS selected < 12% of important variables from the full spectrum datasets while UVS selected a minimum of 61%. Interestingly, both PLSR and MBL models produced the best respective performances with UVS-selected variables from the raw spectra dataset for predicting clay content. A similar result was observed with sand content. Even though the UVS algorithm selected over 61% of variables from the raw spectra, PLSR and MBL models when paired with UVS produced the best clay content predictive models for this study.

5.2.4 Total Carbon and Total Nitrogen Content

The PLSR and MBL models created for TC content and TN prediction which employed the Boruta, CARS, and UVS variable selection methods (as outlined in Table 10 and Table 11), typically exhibited no significant impact, and the observed trends were inconsistent when contrasted with full spectrum models (see Table 5 and Table 6). The highest R^2 value for predicting TC with full spectrum datasets was 0.5. This value was improved to the highest value of 0.56 when TC was predicted with the MBL model developed with variables selected by the CARS algorithm from a raw spectra dataset. The prediction results for TC were relatively similar across all models developed with selected variables datasets ($R^2 < 0.70$). Thus none of the models developed with selected variables datasets managed to predict TC with accuracy.

Similarly, selecting key variables for predicting TN did not improve the result as well, the highest R^2 (0.61) was produced with a full spectrum dataset. As stated by Zhang *et al.*, (2023) spectral modeling is scale dependent, the poor TC and TN prediction results can be associated

with highly variable and limited data (number of samples) which was used in this study. Pirie *et al.*, (2005) also highlighted that a narrow range (as it is the case with TC and TN) of the data may have a negative effect on model accuracy. Additionally, the attempt to select important variables for predicting both TC and TN did not improve model performances. This was caused by the smaller number of samples in comparison to the number of variables in the full spectrum, and frequently, variable selection algorithms struggle to select “*the optimal subset of variables*” (page 103) as stated by Yun *et al.*, (2019).

The results reported in this study were differ from reports in the literature where TC and TN were predicted with much higher accuracy even with full spectrum datasets (Vohland *et al.*, 2014; Baumann *et al.*, 2021; Summerauer *et al.*, 2021; Li *et al.*, 2022).

5.3 The Effect of the Spectra Pre-processing on the Model Performance

Generally, PLSR sand content models had a better predictive ability when the dataset was pre-processed with SG. However, when a PLSR model was developed with UVS-selected variables, the model produced better prediction results with raw spectra and this was the best PLSR model for predicting clay content in this study. On the contrary, MBL models for predicting sand content showed no improvement after pre-processing the data except for the MBL sand content model developed with Boruta-selected variables from the SG pre-processed dataset. Moreover, PLSR and MBL models for predicting silt content had similar results when developed with raw and SG pre-processed datasets while pre-processing with SNV produced the worst outcome. This meant that pre-processing the spectra did not improve model performance when predicting silt content.

For clay content, PLSR models developed with pre-processed spectral data were poorer than those developed with raw spectral data. This was the case except for the PLSR model developed with full spectrum SG pre-processed spectral data. Similarly, MBL models for predicting clay content performed better when developed with raw spectral data than when developed with spectral data that underwent pre-processing. This outcome suggested that spectral pre-processing did not help improve PLSR and MBL predictive ability in predicting clay content in this study. Generally, both PLSR and MBL models performed better when developed with raw spectral data (full spectrum or selected variables) when predicting TC and TN.

Spectral pre-processing did not improve the predictive ability of PLSR and MBL for predicting all the studied soil properties except for silt content, and this outcome suggested that transforming the spectral data based on the SG first derivative and SNV may not help improve

PLSR and MBL in predicting physical and chemical soil properties as suggested by Barra *et al.*, (2021). The findings of this study contradict previous suggestions that pre-processing spectra enhances the precision of predictive models (Rinnan *et al.*, 2009; Gholizadeh *et al.*, 2015; Dotto *et al.*, 2019; Barra *et al.*, 2021a; Vestergaard *et al.*, 2021).

5.4 The Role of Variables Selection on the Model Performance

Generally, a single spectrum is made up of hundreds or thousands of spectral variables. Spectral resolution of the spectrometer used, determines the number of these spectral variables. Many statistical methods struggle when dealing with this high dimensional data which includes noise and interfering spectral variables (Yun *et al.*, 2019). It has been found in the literature that adjacent spectral variables tend to be highly collinear and redundant (Wang *et al.*, 2022), thus contributing to the prediction error and hindering the robustness of the predictive model. Moreover, it is crucial to reduce the dimensionality of spectral datasets and select fewer key variables carrying relevant information to the target soil property such as soil texture, TC, and TN to achieve possible optimal results.

From the three datasets used in this study (Raw spectra dataset, SG, and SNV pre-processed datasets), the Boruta algorithm frequently selected only 3 variables from the three datasets (3602, 1769, 721 cm^{-1} for predicting clay content and 1191, 827, and 830 cm^{-1} for predicting sand content). These variables overlapped with those selected by the UVS algorithm while the CARS algorithm was not consistent in selecting variables from all datasets. The key variables that were identified in a study by Hutengs *et al.*, (2018) for predicting clay and sand content respectively, overlapped with the variables deemed important by the UVS algorithm which subsequently resulted in the best models for estimating both sand and clay content in this study.

The range of variables consistently selected by the UVS algorithm to predict clay and sand content included variables that correspond to O-Si-O absorptions found at 1080 cm^{-1} , 800–780 and 700 cm^{-1} and hydroxyl absorptions found at 3695, 3620, 914, and 936 cm^{-1} which are highly associated with kaolinite minerals as indicated by Baumann *et al.*, (2021) and Le Guillou *et al.*, (2015). Other variables corresponding to absorption features of silicate minerals between ~ 1100 and ~ 600 cm^{-1} were also selected.

The spectrum ranges of variables that were selected for predicting TC and TN were generally similar even though more variables were selected for TN than TC. Based on the frequently selected variables by the UVS algorithm, variables that were also selected in a study by Li *et al.*, (2022) for predicting soil organic carbon (SOC). In their study they also stated that other

studies also reported that the absorption features between 3600 and 2700 cm^{-1} correspond to SOC, which were consistent with the findings from the current study. Some individual variables and ranges described as important for predicting SOC by Ng *et al.*, (2022) were selected for predicting TC in this study. There was a noticeable variability in the variables selected for predicting TN in this study and those that were selected in a study by Li *et al.*, (2022), however, there were overlapping regions or similar single variables. The differences may have been caused by the different variable selection methods they used in their study compared to the present study.

5.5 Limitations

Soil spectroscopy is a data-hungry soil analysis procedure. As stated by Zhang *et al.*, (2023), modeling spectral data is highly depended on the spatial scale of the data amongst other factors. The spatial sampling scale influences the variability in the soil samples. The limited number of samples (300 samples) and high variability in the TC and TN data may have resulted in the poor reported results. Therefore, it is essential to undertake further investigations into the TC and TN content, given that the existing calibrations possess restricted practical importance.

The selection of important variables is a crucial step in the pre-processing of spectral data to maximize the accuracy of the models and minimize computational costs. However, some of the existing variable selection algorithms require high computer processing power. Some of the tested variable selection methods were computationally costly and could not be used due to the limited processing power of the computers used in this study. Since this study intended to pave the way for building the South African soil spectral library, therefore the aforementioned observation was an indication that to manage big spectroscopy data a more powerful computer may be needed.

Chapter Six

6.1 Conclusions

In the problem statement of this study, it was highlighted that farmers rely on timely soil testing for them to be able to make data-driven decisions on which inputs to purchase to maximize their yield. However, the expensive and time-consuming conventional soil analysis methods were a limitation. Furthermore, SOC is essential for maintaining soil health, providing ecosystem services, and mitigating climate change. Due to the significance of SOC, there is a growing need for soil data to assess and track soil carbon stocks. However, traditional soil analysis methods such as wet chemistry and combustion methods, tend to be costly and time-consuming, which can hinder extensive monitoring and limit informed decision-making. In recent decades, infrared spectroscopy proved to be an efficient alternative method for the swift and economical quantitative assessment of SOC and various soil properties, but the investigation of infrared methodologies in South Africa is significantly trailing the global momentum observed in soil spectroscopic research.

Prior studies have indicated that the application of infrared soil analysis methods in South Africa significantly trails behind the advancements observed globally in soil spectroscopic research. Therefore, it is essential to initiate collaborative research initiatives aimed at establishing comprehensive libraries for South African soils. It was further suggested that the collection of samples and the corresponding laboratory analyses conducted at ARC-ISCW can provide a valuable foundation for comparing near-infrared and mid-infrared measurements with conventional chemical analyses.

Consequently, a selection of samples and their related laboratory analyses conducted at ARC-ISCW was identified as a suitable basis for contrasting NIR/MIR measurements against traditional chemical analyses. The objective of the current study was to use a selection of samples and their related laboratory analyses conducted at ARC-ISCW to showcase the capability of MIR diffuse reflectance spectroscopy, in conjunction with machine learning algorithms, to concurrently estimate total carbon, total nitrogen, and soil texture. A further goal was to identify the critical wavelengths and regions that play a significant role in predicting TC, TN, and the particle size fractions. This was achieved using an open-source software application, R Studio.

Results derived from the descriptive analysis showed that the sand content covered a substantial range, with percentages extending from 8% to 92%. The overall mean was computed at 67.58%, in contrast to a median value of 72%. In contrast, the silt content displayed a relatively lesser variability, with percentages varying from 0% to 38%. The average silt content was calculated to be 8.43%, while the median value stood at 8%. The calculated coefficient of variability was found to be 0.61, which signifies a higher level of dispersion relative to the mean. Similarly, to sand, the findings from the descriptive analysis also revealed a wide variability in clay content, which spanned from 6% to 72%. The average clay content was recorded at 24%, in contrast to a median value of 20%.

Furthermore, the descriptive analysis results demonstrated that the TC content was variable, albeit with a narrow range (0.04 - 7.07). The coefficient of variability was calculated to be 0.87 (indicating a greater dispersion around the mean). As expected, the results for TN content followed a similar trend to that for TC content.

Based on the criteria that was used in this study to assess predictive model performance, a good model had to meet *all* the following requirements i.e. $R^2 \geq 0.71$, $RMSEP \leq 10\%$ of the validation dataset range, and $RPIQ \geq 1.7$. The predictive regression models were successfully created, and the evaluation process demonstrated subpar results in forecasting total C and total N. This suggests that the spectral models possess limited predictive strength ($R^2 < 0.71$ and $RPIQ < 1.7$), as highlighted by the established criteria. Therefore, it is imperative to conduct further analyses of additional samples through wet chemistry laboratory methods to improve the predictive efficacy of the TC and TN spectral models.

Conversely, observations for clay, silt, and sand contents showed that soil texture can be predicted well with diffuse reflectance MIR spectroscopy, as indicated by the established criteria. Interestingly, the results showed that (for this dataset), SG and SNV pre-processing of the spectra may not necessarily improve the model accuracy. This was supported by the fact that most of the models that were developed with pre-processed spectra generally underperformed compared to those developed with raw spectra. Instead, it was noted that selecting key variables and the removal of samples with high residual values showed great potential in improving model precision.

The Boruta and CARS variables selection algorithms selected key variables predominantly in the double bonds and fingerprint regions of the spectrum for predicting sand content. Some of the known key variables that are sensitive to minerals highly associated with sand content were

selected by both algorithms. Both algorithms selected less than 15% of the full spectrum variables. UVS on the other hand selected over 55% of the full spectrum variables and also included the known key variables sensitive to sand content minerals such as quartz. PLSR and MBL models developed with Boruta and CARS-selected variables datasets produced similar results (modest) for predicting sand content. Interestingly, the UVS algorithm proved to be a more effective variable selection method when selecting key variables from a raw spectra dataset. Models developed with raw spectra of UVS-selected variables produced the best results for predicting sand content.

For silt content, the CARS algorithm selected less than 5% of the total number of full spectrum variables and predictive models developed with CARS-selected variables selected from the SG pre-processed dataset were superior. This outcome suggests that the CARS algorithm was more suitable for selecting key variables for predicting silt content in this study as opposed to Boruta and UVS. The results obtained with models developed with the Boruta and CARS selected variables datasets for predicting clay content were similar to those produced with full spectrum datasets. In contrast, the UVS algorithm produced similar or better results than the full spectrum datasets models and established the best clay content. A similar result was observed with sand content and this outcome suggests that for both sand and clay contents, the UVS algorithm was a better variable selection method compared to Boruta and CARS.

The goal of using the variable selection algorithms was to improve the precision of regression models for predicting studied soil properties. Unfortunately, it was not achieved for TC and TN, therefore none of the variable selection algorithms can applied in this study can be recommended to determine key predictor variables for TC and TN.

References

- Al-Kaisi, M.M. *et al.* (2017) ‘Fundamentals and Functions of Soil Environment’, *Soil Health and Intensification of Agroecosystems*, pp. 1–23. Available at: <https://doi.org/10.1016/B978-0-12-805317-1.00001-4>.
- Araújo, M.C.U. *et al.* (2001) ‘The successive projections algorithm for variable selection in spectroscopic multicomponent analysis’, *Chemometrics and Intelligent Laboratory Systems*, 57(2), pp. 65–73. Available at: [https://doi.org/10.1016/S0169-7439\(01\)00119-8](https://doi.org/10.1016/S0169-7439(01)00119-8).
- Babatunde, O.H. *et al.* (2014) ‘Available here This Journal Article is posted at Research Online’, *International Journal of Electronics Communication and Computer Engineering*, 5(4), pp. 899–905.
- Bangelesa, F. *et al.* (2020) ‘Predicting Soil Organic Carbon Content Using Hyperspectral Remote Sensing in a Degraded Mountain Landscape in Lesotho’, *Applied and Environmental Soil Science*, 2020(April). Available at: <https://doi.org/10.1155/2020/2158573>.
- Barnes, R.J. *et al.* (1989) ‘Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra’, *Applied Spectroscopy*, 43(5), pp. 772–777. Available at: <https://doi.org/10.1366/0003702894202201>.
- Barra, I. *et al.* (2021) ‘Soil spectroscopy with the use of chemometrics, machine learning and pre-processing techniques in soil diagnosis: Recent advances—A review’, *TrAC - Trends in Analytical Chemistry*, 135(116166). Available at: <https://doi.org/10.1016/j.trac.2020.116166>.
- Baumann, P. *et al.* (2021) ‘Estimation of soil properties with mid-infrared soil spectroscopy across yam production landscapes in West Africa’, *Soil*, 7(2), pp. 717–731. Available at: <https://doi.org/10.5194/soil-7-717-2021>.
- Bayer, A. *et al.* (2012) ‘A Comparison of feature-based MLR and PLS regression techniques for the prediction of three soil constituents in a degraded South African Ecosystem’, *Applied and Environmental Soil Science*, 2012(1). Available at: <https://doi.org/10.1155/2012/971252>.
- Bellon-Maurel, V. *et al.* (2010) ‘Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy’, *TrAC - Trends in Analytical Chemistry*, 29(9), pp. 1073–1081. Available at: <https://doi.org/10.1016/j.trac.2010.05.006>.
- Bonett, J.P. *et al.* (2015) ‘Espectroscopia de infrarrojo medio para la estimación de algunas propiedades

del suelo', *Agronomia Colombiana*, 33(1), pp. 99–106. Available at: <https://doi.org/10.15446/agron.colomb.v33n1.49245>.

Breiman, L. (2001) 'Random Forests', *Springer*, Volume 45, pp. 5–32. Available at: <https://doi.org/https://doi.org/10.1023/A:1010933404324>.

Cheng, H. *et al.* (2021) 'Combining multivariate method and spectral variable selection for soil total nitrogen estimation by Vis–NIR spectroscopy', *Archives of Agronomy and Soil Science*, 67(12), pp. 1665–1678. Available at: <https://doi.org/10.1080/03650340.2020.1802013>.

Coblinski, J.A. *et al.* (2020) 'Prediction of soil texture classes through different wavelength regions of reflectance spectroscopy at various soil depths', *Catena*, 189(July 2019). Available at: <https://doi.org/10.1016/j.catena.2020.104485>.

Dangal, S.R.S. *et al.* (2019) 'Accurate and precise prediction of soil properties from a large mid-infrared spectral library', *Soil Systems*, 3(1), pp. 1–23. Available at: <https://doi.org/10.3390/soilsystems3010011>.

Deng, B.C. *et al.* (2016) 'A bootstrapping soft shrinkage approach for variable selection in chemical modeling', *Analytica Chimica Acta*, 908, pp. 63–74. Available at: <https://doi.org/10.1016/j.aca.2016.01.001>.

Dotto, A.C. *et al.* (2019) 'Alradspectra: A quantification tool for soil properties using spectroscopic data in R', *Revista Brasileira de Ciencia do Solo*, 43, pp. 1–15. Available at: <https://doi.org/10.1590/18069657rbcs20180263>.

Von Eye, A. and Schuster, C. (1998) *Outlier Analysis, Regression Analysis for Social Sciences*. Available at: <https://doi.org/10.1016/b978-012724955-1/50180-7>.

Ferreira, R. de A. *et al.* (2022) 'Kennard-Stone method outperforms the Random Sampling in the selection of calibration samples in SNPs and NIR data', *Ciência Rural*, 52(5), pp. 1–11. Available at: <https://doi.org/10.1590/0103-8478cr20201072>.

Fongaro, C.T. *et al.* (2018) 'Improvement of clay and sand quantification based on a novel approach with a focus on multispectral satellite images', *Remote Sensing*, 10(10). Available at: <https://doi.org/10.3390/rs10101555>.

Gautam, R. *et al.* (2015) 'Review of multidimensional data processing approaches for Raman and infrared spectroscopy', *EPJ Techniques and Instrumentation*, 2(1). Available at: <https://doi.org/10.1140/epjti/s40485-015-0018-6>.

- Ge, Y. *et al.* (2022) ‘Soil spectroscopy Training Material’.
- Geladi, P. and Kowalski, B.R. (1986) ‘Partial least-squares regression: a tutorial’, *ScienceDirect*, 185(Anal. Chim. Acta), pp. 1–17.
- Gholizadeh, A. *et al.* (2013) ‘Visible, near-infrared, and mid-infrared spectroscopy applications for soil assessment with emphasis on soil organic matter content and quality: State-of-the-art and key issues’, *Applied Spectroscopy*, 67(12), pp. 1349–1362. Available at: <https://doi.org/10.1366/13-07288>.
- Gholizadeh, A. *et al.* (2015) ‘Comparing different data preprocessing methods for monitoring soil heavy metals based on soil spectral features’, *Soil and Water Research*, 10(4), pp. 218–227. Available at: <https://doi.org/10.17221/113/2015-SWR>.
- Gholizadeh, A. *et al.* (2016) ‘A memory-based learning approach as compared to other data mining algorithms for the prediction of soil texture using diffuse reflectance spectra’, *Remote Sensing*, 8(4). Available at: <https://doi.org/10.3390/rs8040341>.
- Granitto, P.M. *et al.* (2006) ‘Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products’, *Chemometrics and Intelligent Laboratory Systems*, 83(2), pp. 83–90. Available at: <https://doi.org/10.1016/j.chemolab.2006.01.007>.
- Le Guillou, F. *et al.* (2015) ‘How does grinding affect the mid-infrared spectra of soil and their multivariate calibrations to texture and organic carbon?’, *Soil Research*, 53(8), pp. 913–921. Available at: <https://doi.org/10.1071/SR15019>.
- Guo, L. *et al.* (2017) ‘Comparisons of spatial and non-spatial models for predicting soil carbon content based on visible and near-infrared spectral technology’, *Geoderma*, 285, pp. 280–292. Available at: <https://doi.org/10.1016/j.geoderma.2016.10.010>.
- Haaland, D.M. and Thomas, E. V. (1988) ‘Partial Least-Squares Methods for Spectral Analyses. 1. Relation to Other Quantitative Calibration Methods and the Extraction of Qualitative Information’, *Analytical Chemistry*, 60(11), pp. 1193–1202. Available at: <https://doi.org/10.1021/ac00162a020>.
- Hati, K.M. *et al.* (2022) ‘Mid-Infrared Reflectance Spectroscopy for Estimation of Soil Properties of Alfisols from Eastern India’, *Sustainability (Switzerland)*, 14(9). Available at: <https://doi.org/10.3390/su14094883>.
- Henao, J. and Baanante, C. (1999) *Estimating Rates of Nutrient Depletion in Soils of Agricultural Lands*

of Africa, International Fertilizer Development Center.

- Hutengs, C. *et al.* (2018) ‘Comparison of portable and bench-top spectrometers for mid-infrared diffuse reflectance measurements of soils’, *Sensors (Switzerland)*, 18(4), pp. 1–17. Available at: <https://doi.org/10.3390/s18040993>.
- Janik, L.J. *et al.* (2007) ‘The prediction of soil carbon fractions using mid-infrared-partial least square analysis’, *Australian Journal of Soil Research*, 45(2), pp. 73–81. Available at: <https://doi.org/10.1071/SR06083>.
- Ji, W. *et al.* (2016) ‘Assessment of soil properties in situ using a prototype portable MIR spectrometer in two agricultural fields’, *Biosystems Engineering*, 152, pp. 14–27. Available at: <https://doi.org/10.1016/j.biosystemseng.2016.06.005>.
- Johnson, J.M. *et al.* (2019) ‘Near-infrared, mid-infrared or combined diffuse reflectance spectroscopy for assessing soil fertility in rice fields in sub-Saharan Africa’, *Geoderma*, 354(February). Available at: <https://doi.org/10.1016/j.geoderma.2019.06.043>.
- Kiala, Z. *et al.* (2017) ‘Potential of interval partial least square regression in estimating leaf area index’, *South African Journal of Science*, 113(9–10), pp. 1–9. Available at: <https://doi.org/10.17159/sajs.2017/20160277>.
- Kuhn, M. (2023) *Package ‘caret’ March, Predictive Analytics with KNIME*.
- Kursa, M.B. and Rudnicki, W.R. (2010) ‘Feature selection with the boruta package’, *Journal of Statistical Software*, 36(11), pp. 1–13. Available at: <https://doi.org/10.18637/jss.v036.i11>.
- Li, H. *et al.* (2009) ‘Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration’, *Analytica Chimica Acta*, 648(1), pp. 77–84. Available at: <https://doi.org/10.1016/j.aca.2009.06.046>.
- Li, H. *et al.* (2022) ‘Combining Variable Selection and Multiple Linear Regression for Soil Organic Matter and Total Nitrogen Estimation by DRIFT-MIR Spectroscopy’, *Agronomy*, 12(3). Available at: <https://doi.org/10.3390/agronomy12030638>.
- Ludwig, B. *et al.* (2019) ‘Accuracy of Estimating Soil Properties with Mid-Infrared Spectroscopy: Implications of Different Chemometric Approaches and Software Packages Related to Calibration Sample Size’, *Soil Science Society of America Journal*, 83(5), pp. 1542–1552. Available at: <https://doi.org/10.2136/sssaj2018.11.0413>.

- Ludwig, B. *et al.* (2021) ‘Diffuse reflectance infrared spectroscopy estimates for soil properties using multiple partitions: Effects of the range of contents, sample size, and algorithms’, *Soil Science Society of America Journal*, 85(3), pp. 546–559. Available at: <https://doi.org/10.1002/saj2.20205>.
- Ludwig, B. *et al.* (2023) ‘Optimised use of data fusion and memory-based learning with an Austrian soil library for predictions with infrared data’, *European Journal of Soil Science*, 74(4), pp. 1–17. Available at: <https://doi.org/10.1111/ejss.13394>.
- Martínez-España, R. *et al.* (2019) ‘Developing an intelligent system for the prediction of soil properties with a portable mid-infrared instrument’, *Biosystems Engineering*, 177, pp. 101–108. Available at: <https://doi.org/10.1016/j.biosystemseng.2018.09.013>.
- Mashimbye, Z.E. *et al.* (2012) ‘Model-Based Integrated Methods for Quantitative Estimation of Soil Salinity from Hyperspectral Remote Sensing Data: A Case Study of Selected South African Soils’, *Pedosphere*, 22(5), pp. 640–649. Available at: [https://doi.org/10.1016/S1002-0160\(12\)60049-6](https://doi.org/10.1016/S1002-0160(12)60049-6).
- Meyer, J.H. (1989) ‘Rapid simultaneous rating of soil texture, organic matter, total nitrogen and nitrogen mineralization potential by near infra-red reflectance’, *South African Journal of Plant and Soil*, 6(1), pp. 59–63. Available at: <https://doi.org/10.1080/02571862.1989.10634481>.
- Minu, S. *et al.* (2016) ‘Review of preprocessing techniques used in soil property prediction from hyperspectral data’, *Cogent Geoscience*, 2(1), p. 1145878. Available at: <https://doi.org/10.1080/23312041.2016.1145878>.
- Mzinyane, T. *et al.* (2015) ‘Soil carbon estimation from *eucalyptus grandis* using canopy spectra’, *South African Journal of Geomatics*, 4(4), p. 548. Available at: <https://doi.org/10.4314/sajg.v4i4.12>.
- Nawar, S. and Mouazen, A. (2017) ‘Comparison between random forest and partial least square regression of on-line vis-NIR spectroscopy measurements of soil total nitrogen and organic carbon’, pp. 1–19.
- Nawar, S. and Mouazen, A.M. (2017) ‘Predictive performance of mobile vis-near infrared spectroscopy for key soil properties at different geographical scales by using spiking and data mining techniques’, *Catena*, 151, pp. 118–129. Available at: <https://doi.org/10.1016/j.catena.2016.12.014>.
- Nel, T. *et al.* (2022) ‘Evaluation of simple and multivariate linear regression models for exchangeable

base cation conversion between seven measurement techniques on South African soils’, *Geoderma Regional*, 30(August), p. e00571. Available at: <https://doi.org/10.1016/j.geodrs.2022.e00571>.

Nel, T. *et al.* (2023) ‘Comparison of Soil pH and Exchangeable Cation Quantification by Various Wet Methods with Near- and Mid-Infrared Spectroscopy Prediction’, *Communications in Soil Science and Plant Analysis*, 54(17), pp. 2425–2438. Available at: <https://doi.org/10.1080/00103624.2023.2223657>.

Ng, W. *et al.* (2022) ‘Mid-infrared spectroscopy for accurate measurement of an extensive set of soil properties for assessing soil functions’, *Soil Security*, 6, p. 100043. Available at: <https://doi.org/10.1016/j.soisec.2022.100043>.

Nocita, M. *et al.* (2014) ‘Prediction of soil organic carbon content by diffuse reflectance spectroscopy using a local partial least square regression approach’, *Soil Biology and Biochemistry*, 68, pp. 337–347. Available at: <https://doi.org/10.1016/j.soilbio.2013.10.022>.

Nocita, M. *et al.* (2015) ‘Soil Spectroscopy: An Alternative to Wet Chemistry for Soil Monitoring’, *Advances in Agronomy*, 132, pp. 139–159. Available at: <https://doi.org/10.1016/bs.agron.2015.02.002>.

Nocita, M. *et al.* (2016) ‘Soil Spectroscopy As a Tool To Assess Organic Carbon, Iron Oxides, and Clay Content in the Subtropical Thicket Biome of Eastern Cape Province of South Africa’, pp. 1–23.

Nomngongo, P.N. *et al.* (2017) ‘Near-Infrared Spectroscopy Combined with Multivariate Tools for Analysis of Trace Metals in Environmental Matrices’, *Developments in Near-Infrared Spectroscopy* [Preprint], (March). Available at: <https://doi.org/10.5772/67199>.

Nortjé, G.P. and Laker, M.C. (2021) ‘Soil fertility trends and management in Conservation Agriculture: a South African perspective’, *South African Journal of Plant and Soil*, 38(3), pp. 247–257. Available at: <https://doi.org/10.1080/02571862.2021.1896039>.

Paterson, G. *et al.* (2015) ‘Spatial soil information in South Africa: Situational analysis, limitations and challenges’, *South African Journal of Science*, 111(5–6), pp. 1–7. Available at: <https://doi.org/10.17159/sajs.2015/20140178>.

Peck, T.R. (1990) ‘Soil Testing: Past Present and Future’, *Communications in Soil Science and Plant Analysis*, 21(13–16), pp. 1165–1186. Available at:

<https://doi.org/10.1080/00103629009368297>.

Pirie, A. *et al.* (2005) ‘Spectroscopic Techniques To Predict Several Soil Properties’, *Australian Journal of Soil Research*, 43, pp. 713–721.

Ramirez-Lopez, L. *et al.* (2013) ‘The spectrum-based learner: A new local approach for modeling soil vis-NIR spectra of complex datasets’, *Geoderma*, 195–196, pp. 268–279. Available at: <https://doi.org/10.1016/j.geoderma.2012.12.014>.

Ramirez-lopez, L. and Stevens, A. (2013) ‘Memory based learning methods and tools : towards efficient modelling , predicting and managing tasks in large scale soil spectral libraries Case studies Tools Final remarks Soil vis – NIR libraries Field scale’, (December).

Ramírez, P.B. *et al.* (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(May), pp. 1–13. Available at: <https://doi.org/10.3389/fenvs.2021.634472>.

Reddy, N. *et al.* (2017) ‘A hybrid partial least squares and random forest approach to modelling forest structural attributes using multispectral remote sensing data’, *South African Journal of Geomatics*, 6(3), p. 377. Available at: <https://doi.org/10.4314/sajg.v6i3.8>.

Rinnan, Å. *et al.* (2009) ‘Chapter 2 - Data Pre-processing’, *Infrared Spectroscopy for Food Quality Analysis and Control*, 3, pp. 29–50.

Roberts, V.G. *et al.* (2003) ‘An investigation into soil fertility in two resource-poor farming communities in KwaZulu-Natal (South Africa)’, *South African Journal of Plant and Soil*, 20(3), pp. 146–151. Available at: <https://doi.org/10.1080/02571862.2003.10634924>.

Rossel, R.A.V. and Behrens, T. (2010) ‘Using data mining to model and interpret soil diffuse reflectance spectra’, *Geoderma*, 158(1–2), pp. 46–54. Available at: <https://doi.org/10.1016/j.geoderma.2009.12.025>.

Sanderman, J. *et al.* (2020) ‘Mid-infrared spectroscopy for prediction of soil health indicators in the United States’, *Soil Science Society of America Journal*, 84(1), pp. 251–261. Available at: <https://doi.org/10.1002/saj2.20009>.

De Santana, F.B. *et al.* (2018) ‘Visible and near infrared spectroscopy coupled to random forest to quantify some soil quality parameters’, *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy*, 191, pp. 454–462. Available at:

<https://doi.org/10.1016/j.saa.2017.10.052>.

Savitzky, A. and Golay, M.J.E. (1964) ‘Smoothing and Differentiation of Data by Simplified Least Squares Procedures’, *Analytical Chemistry*, 36(8), pp. 1627–1639. Available at: <https://doi.org/10.1021/ac60214a047>.

Seboko, K.R. *et al.* (2023) ‘Predicting soil carbon in granitic soils using Fourier-transform mid-infrared (FT-MIR) spectroscopy: the value of database disaggregation’, *South African Journal of Plant and Soil*, 40(1), pp. 23–33. Available at: <https://doi.org/10.1080/02571862.2023.2180098>.

Sewpersad, T. *et al.* (2024) ‘The assessment of soil organic matter in the KwaZulu-natal province of South Africa and its relationship to spectroscopy data’, *Geocarto International*, 39(1), p. Available at: <https://doi.org/10.1080/10106049.2024.2361702>.

Shepherd, K.D. and Walsh, M.G. (2007) ‘Review: Infrared spectroscopy - Enabling an evidence-based diagnostic surveillance approach to agricultural and environmental management in developing countries’, *Journal of Near Infrared Spectroscopy*, 15(1), pp. 1–19. Available at: <https://doi.org/10.1255/jnirs.716>.

Shete, P.P. *et al.* (2019) ‘Determination of Soil Texture Distribution (Clay, Sand and Silt) by using Spectral Measurement: A Review’, *Journal of Emerging Technologies and Innovative Research*, 6(2), pp. 625–629.

Song, X. *et al.* (2016) ‘A novel algorithm for spectral interval combination optimization’, *Analytica Chimica Acta*, 948, pp. 19–29. Available at: <https://doi.org/10.1016/j.aca.2016.10.041>.

Soriano-Disla, J.M. *et al.* (2014) ‘The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties’, *Applied Spectroscopy Reviews*, 49(2), pp. 139–186. Available at: <https://doi.org/10.1080/05704928.2013.811081>.

Stenberg, B. *et al.* (2010) ‘Visible and Near Infrared Spectroscopy in Soil Science’, *Advances in Agronomy*, 107(C), pp. 163–215. Available at: [https://doi.org/10.1016/S0065-2113\(10\)07005-7](https://doi.org/10.1016/S0065-2113(10)07005-7).

Stevens, A. *et al.* (2013) ‘Prediction of Soil Organic Carbon at the European Scale by Visible and Near InfraRed Reflectance Spectroscopy’, *PLoS ONE*, 8(6). Available at: <https://doi.org/10.1371/journal.pone.0066409>.

- Sumner, L. *et al.* (2021) 'The central African soil spectral library: A new soil infrared repository and a geographical prediction analysis', *Soil*, 7(2), pp. 693–715. Available at: <https://doi.org/10.5194/soil-7-693-2021>.
- Sun, Z. *et al.* (2014) 'Spectroscopic Determination of Soil Organic Carbon and Total Nitrogen Content in Pasture Soils', *Communications in Soil Science and Plant Analysis*, 45(8), pp. 1037–1048. Available at: <https://doi.org/10.1080/00103624.2014.883628>.
- Takele, C. and Iticha, B. (2020) 'Use of infrared spectroscopy and geospatial techniques for measurement and spatial prediction of soil properties', *Heliyon*, 6(10), p. e05269. Available at: <https://doi.org/10.1016/j.heliyon.2020.e05269>.
- Tan, Z.X. *et al.* (2005) 'Global soil nutrient depletion and yield reduction', *Journal of Sustainable Agriculture*, 26(1), pp. 123–146. Available at: https://doi.org/10.1300/J064v26n01_10.
- Terhoeven-Urselmans, T. *et al.* (2010) 'Prediction of Soil Fertility Properties from a Globally Distributed Soil Mid-Infrared Spectral Library', *Soil Science Society of America Journal*, 74(5), pp. 1792–1799. Available at: <https://doi.org/10.2136/sssaj2009.0218>.
- Tziolas, N. *et al.* (2019) 'A memory-based learning approach utilizing combined spectral sources and geographical proximity for improved VIS-NIR-SWIR soil properties estimation', *Geoderma*, 340(April 2018), pp. 11–24. Available at: <https://doi.org/10.1016/j.geoderma.2018.12.044>.
- Vendrame, P.R.S. *et al.* (2012) 'The potential of NIR spectroscopy to predict soil texture and mineralogy in Cerrado Latosols', *European Journal of Soil Science*, 63(5), pp. 743–753. Available at: <https://doi.org/10.1111/j.1365-2389.2012.01483.x>.
- Vestergaard, R.-J. *et al.* (2021) 'Evaluation of Optimized Preprocessing and Modeling', *Sensors*, 21(20), p. 6745.
- Viscarra Rossel, R.A. *et al.* (2006) 'Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties', *Geoderma*, 131(1–2), pp. 59–75. Available at: <https://doi.org/10.1016/j.geoderma.2005.03.007>.
- Vohland, M. *et al.* (2014) 'Determination of soil properties with visible to near- and mid-infrared spectroscopy: Effects of spectral variable selection', *Geoderma*, 223–225(1), pp. 88–96. Available at: <https://doi.org/10.1016/j.geoderma.2014.01.013>.
- Vohland, M. *et al.* (2016) 'Using variable selection and wavelets to exploit the full potential of visible-near infrared spectra for predicting soil properties', *Journal of Near Infrared Spectroscopy*,

24(3), pp. 255–269. Available at: <https://doi.org/10.1255/jnirs.1233>.

Wadoux, A.M.J.-C. *et al.* (2021) *Soil Spectral Inference with R*, Springer International Publishing AG, Cham.

Wang, J. *et al.* (2022) ‘Spectral variable selection for estimation of soil organic carbon content using mid-infrared spectroscopy’, *European Journal of Soil Science*, 73(4), pp. 1–13. Available at: <https://doi.org/10.1111/ejss.13267>.

Waruru, B.K. *et al.* (2015) ‘Application of mid-infrared spectroscopy for rapid characterization of key soil properties for engineering land use’, *Soils and Foundations*, 55(5), pp. 1181–1195. Available at: <https://doi.org/10.1016/j.sandf.2015.09.018>.

Wilpiseski, R.L. *et al.* (2019) ‘Soil Aggregate Microbial Communities: Towards Understanding Microbiome Interactions at Biologically Relevant Scales’, *Applied and Environmental Microbiology*, 85(14). Available at: <https://doi.org/10.1128/AEM.00324-19>.

Wold, S., Sjöström, M. and Eriksson, L. (2001) ‘PLS-regression: A basic tool of chemometrics’, *Chemometrics and Intelligent Laboratory Systems*, 58(2), pp. 109–130. Available at: [https://doi.org/10.1016/S0169-7439\(01\)00155-1](https://doi.org/10.1016/S0169-7439(01)00155-1).

Xiao, Z., Li, Y. and Feng, H. (2016) ‘Modeling soil cation concentration and sodium adsorption ratio using observed diffuse reflectance spectra’, *Canadian Journal of Soil Science*, 96(4), pp. 372–385. Available at: <https://doi.org/10.1139/cjss-2016-0002>.

Xie, H.T. *et al.* (2011) ‘Predicting soil organic carbon and total nitrogen using mid- and near-infrared spectra for Brookston clay loam soil in Southwestern Ontario, Canada’, *Canadian Journal of Soil Science*, 91(1), pp. 53–63. Available at: <https://doi.org/10.4141/CJSS10029>.

Yun, Y.H. *et al.* (2019) ‘An overview of variable selection methods in multivariate analysis of near-infrared spectra’, *TrAC - Trends in Analytical Chemistry*, 113, pp. 102–115. Available at: <https://doi.org/10.1016/j.trac.2019.01.018>.

Zhang, X. *et al.* (2023) ‘Towards Optimal Variable Selection Methods for Soil Property Prediction Using a Regional Soil Vis-NIR Spectral Library’, *Remote Sensing*, 15(2), pp. 1–16. Available at: <https://doi.org/10.3390/rs15020465>.

Zheng, K. *et al.* (2012) ‘Stability competitive adaptive reweighted sampling (SCARS) and its applications to multivariate calibration of NIR spectra’, *Chemometrics and Intelligent Laboratory Systems*, 112, pp. 48–54. Available at:

<https://doi.org/10.1016/j.chemolab.2012.01.002>.