Creation of mid-infrared spectroscopy calibration algorithms for soil property predictions

A Kock

orcid.org/0000-0001-5349-9798

Dissertation accepted in fulfilment of the requirements for the degree *Master of Science in Environmental Sciences* at the North-West University

Supervisor:	Dr GM van Zijl
Co-supervisor:	Dr PD Ramphisa

Graduation May 2022

25901966

ACKNOWLEDGMENTS

This thesis is part of a bigger project which aims to create soil spectroscopy calibration models that cover the entirety of South Africa. This project in conjunction with another project also conducted within the North West province focuses on the pH, Effective CEC, P and base cations which covers all soil properties that are important for precision agriculture.

This work is based on the research supported wholly / in part by the National Research Foundation of South Africa (Grant Numbers: 121302)

Psalm 100: 4-5 (KJV): [4] Enter into his gates with thanksgiving, and into his courts with praise: be thankful unto him and bless his name. [5] For the Lord is good; his mercy is everlasting; and his truth endureth to all generations.

I sincerely thank my supervisor Dr GM van Zijl and Co-supervisor Dr PD Ramphisa, for guidance and support throughout the project.

I also want to thank the people and organisations without whom many aspects of this research would not have been achieved.

- Martiens Du Plessis at NWK Limited in conjunction with Dailena Pienaar NviroTek Laboratories for providing access to seasonal samples and soil property data and support during this study.
- Dup Haarhoff at GWK for samples and soil property data used in the study.
- Dries Bloem for providing samples and data.
- Bruker South Africa (Pty) Ltd for the lease of the ALPHA II DRIFTS MIR spectrometer and Godfrey Lekhuleni for providing excellent service and support.
- Ruan Gagiano at the North-West University Soil Laboratory for assistance and the use of laboratory equipment.

Finally, I would like to thank all my family and friends there was nothing random about your acts of kindness, support and care. Thank you for all you have done.

ABSTRACT

Precision agriculture (PA) has been named as a cultivation method which could help alleviate food shortages in the future. However, PA relies on knowing the spatial distribution of soil properties, which requires repeated, quick, accurate, and cost-effective soil analysis. Conventional methods of soil analysis are slow and costly, reducing the application of PA in South Africa. Soil spectroscopy can fulfil the need for quick cost-effective soil analysis, but depend on robust calibration curves, of which none exist openly for South Africa. It is expected that mid-infrared soil spectroscopy can be used to analyse soil samples for soil properties pH, Effective Cation Exchange Capacity and Phosphorus within the Western Highveld summer grain area. Data for this study was provided by Noord Wes Kooperasie and Griekwaland Wes Kooperasie which resulted in a selection pool of ± 5 180 samples already analysed for soil properties. Conditioned Latin Hypercube sampling (cLHS) was then used to select 1 000 samples based on the pH, Effective CEC and P values of the dataset. The samples were then prepared and scanned with Mid Infrared spectroscopy (4 000 to 400 cm^{-1}) using a Bruker Alpha II with DRIFTS module attached to create a spectral library. The soil property database and spectral library was combined and the R programming language was used to create calibration models using Cubist, Partial Least Square regression (PLSR) and Random Forest (RF). Each calibration algorithm was also applied onto two types of spectral datasets which includes spectra with pre-processing and spectra with minimal to no pre-processing applied. These models were validated using statistical performance measures including root mean square error (RMSE), squared correlation coefficient (r^2), standard deviation, bias and ratio of performance to deviation (RPD). Results show that models created for pH with Cubist and pre-processed spectral data had the best performance ($\mathrm{R}^2 = 0.86,\ \mathrm{RMSE} = 0.3, \mathrm{RPD} = 2.66$) along with ECEC with Cubist ($\mathrm{R}^2 = 0.86,\ \mathrm{RMSE} = 0.3, \mathrm{RPD} = 2.66$) and RF ($\mathrm{R}^2 = 0.85,\ \mathrm{RMSE} = 0.72$) and then P with some success using RF ($\mathrm{R}^2 = 0.57,\ \mathrm{RMSE} = 13.48, \mathrm{RPD} = 1.51$). Overall performance increase was observed with Cubist, PLSR and RF models using pre-processed spectral data compared to spectra with no pre-processing and produced acceptable prediction models to be able to predict pH and ECEC from soil spectra. Findings are consistent with other studies conducted worldwide but with little to no data to compare from South Africa more research and data is needed to create models that include all soil properties used for PA and that is representative of the whole of South Africa.

Keywords: Soil spectroscopy, Infrared spectroscopy, Soil properties, Calibration models, Mid-infrared

TABLE OF CONTENTS

LIST OF ABREVIATIONS

Al	Aluminium
ANN	Artificial neural network
C	Carbon
Ca	Calcium
CEC	Cation exchange capacity
cLHS	Conditioned Latin hypercube sampling
CNN	Convolutional neural network
CR	Continuum removal
Cu	Coper
CV	Coefficient of variability
DFFE	Department of forestry, fisheries and the environment
DRIFTS	Diffuse reflectance Infrared Fourier transform spectroscopy
DSM	Digital soil mapping
EC	Electrical conductivity
ECEC	Effective Cation exchange capacity
EMR	Electromagnetic radiation
Fe	Iron
FIR	Far-Infrared
GWK	Griekwaland Wes Kooperasie
GUI	Graphical user interface
IR	Infrared
K	Potassium
KCl	Potassium chloride
LHS	Latin hypercube sampling
MARS	Multivariate adaptive regression splines
MBL	Memory based learning
Mg	Magnesium
MIR	Mid-Infrared
MLR	Multiple linear regression
Mn	Manganese
Mo	Molybdenum
MSC	Multiplicative scatter correction
MWA	Moving window average
N	Nitrogen

Na	Sodium
NIR	Near-Infrared
NWK	Noord Wes Kooperasie
OC	Organic carbon
OM	Organic matter
P	Phosphorus
PA	Precision agriculture
PCA	Principle component analysis
PLSR	Partial least square regression
RF	Random forest
RMSE	Root mean square error
RPD	Ratio of performance to deviation
RPIQ	Ratio of performance to inter quartile difference
SAM-PLS	Self-adapted partial least squares
SG	Savitzky - Golay filter
SNV	Standard normal variate
SPD	Soil property database
SSD	Soil spectral library
SVM	Support vector machines

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 INTRODUCTION

1.1 Introduction

It is well known that soil is a heterogenous system with complex mechanisms. Understanding these mechanisms is impossible without the use of analysis methods which gives insight into the processes that govern soil (Viscarra Rossel *et al*., 2006). Soil analysis also aid in determining the quality of soil which in return enables us to better manage soil for better crop production and yields. Spectroscopic soil analysis techniques are increasingly being used as alternatives for conventional soil analyses methods and one such method is mid-infrared spectroscopy, which is the focus of this study. Details regarding the use of mid-infrared spectrometry will be discussed below as well as the aim and proposed procedures for this study.

1.2 Background to study

The world population is dramatically increasing from the reported 7 billion in 2012 to a predicted 8.5 billion in 2030 and 10.1 billion in 2050 (United Nations, 2019). South Africa represents 0.74% of the world population (Daw and Molefe, 2017) and our population is predicted to increase from the current 56 million to 72 million by the end of 2050 (UN-DESA, 2017). Population growth has a great effect on a country's ability to sustain food security. This is as a result of an increase in population which leads to lower food security (Tilman *et al.*, 2011). It will be difficult for South Africa to maintain food security due to the population growth. Currently, if South Africa is affected by a drought, it will not be able to provide sufficient food during such periods (Le Roux *et al.*, 2016).

Precision Agriculture (PA) could enable farmers to increase productivity to counter the effects of increasing population (Van Vuuren *et al*., 2006). Precision agriculture as defined by Pierce and Nowak, (1999) is "*the application of technologies and principles to manage spatial and temporal variability associated with all aspects of agricultural production for the purpose of improving crop performance and environmental quality".* The fundamental purpose of PA, as defined by the above-mentioned definition, is to monitor and manage soils to increase crop yield and promote sustainable agricultural practices, using technology, such as spectrometry. Soil properties are known to vary over short distances due to natural soil forming processes as well as man-made production management practices (Mills and Fey, 2004). Precision agriculture methods requires the monitoring and analysis of spatially variable soils on a more frequent and affordable manner which also enables more farmers in developing countries to cultivate for increased crop yields. Traditional laboratory methods that are used to analyse soil properties are time consuming, costly

and use chemicals that are hazardous to the environment (Van Vuuren *et al.*, 2006). A study conducted by Cantarella *et al.* (2006) found that traditional soil analysis conducted by different laboratories yields variable results for soil property analysis with measurement errors and low precision in some cases.

Soil property predictions using spectroscopic methods in contrast to conventional analysis methods are cheaper to run, reduce analysis time, require less personnel and does not use environmentally hazardous reactants during the analysis (Van Vuuren *et al.*, 2006). These spectroscopic measurements which include Visible, Near- and Mid-Infrared (MIR) are able to predict important soil properties like nitrogen (N), organic carbon (OC), Organic matter (OM), pH, electrical conductivity (EC), sand, silt, exchangeable calcium (Ca), potassium (K), Aluminium (Al), iron (Fe), copper (Cu), phosphorus (P) and cation exchange capacity (CEC) with great accuracy (McCarty and Reeves, 2006). Mid-infrared spectroscopy (4 000 – 400 cm^{-1}) has been found to produce better predictions for soil properties compared to visible and near-infrared spectroscopy (McCarty and Reeves, 2006). For this method to be successfully implemented, robust soil property specific calibration models for the mid-infrared region need to be developed and calibrated for South African soils.

Soil spectroscopy has emerged as a new technology by which soil properties could be measured quickly and cost effectively. It has been shown to be applicable internationally (McCarty and Reeves, 2006; Dangal *et al.*, 2019; Coblinski *et al.*, 2020), and its use is also increasing in developing countries (Shepherd and Walsh, 2007). However, to be able to apply soil spectroscopy for soil property analysis, calibration curves for each soil property must first be created (Waruru *et al.*, 2015). This calibration relies on a robust soil spectral library, i.e., a conventionally analysed soil property database, being correlated with spectroscopic scans of the same soils (Waruru et al., 2015).

1.3 Problem statement

Precision agriculture has been named as a cultivation method which could help alleviate food shortages in the future. However, PA relies on knowing the spatial distribution of soil properties, which requires repeated, quick, accurate, and cost-effective soil analysis. Conventional methods of soil analysis are slow and costly, reducing the application of PA in South Africa. Soil spectroscopy can fulfil the need for soil quick cost-effective soil analysis, but is dependant on robust calibration curves, of which none exist openly for South Africa. To date, such a database is not openly accessible for any region in South Africa, rendering soil spectroscopy unreliable. This study will address that need by creating mid-infrared calibration curves for the soil properties:

soil pH, P and effective cation exchange capacity (ECEC) for the Western Highveld summer grain producing region.

1.4 Hypothesis, research aim and objectives

1.4.1 Hypothesis

The hypothesis tested during this study is that mid-infrared soil spectroscopy can be used to analyse soil samples for soil pH, P and ECEC within the Western Highveld summer grain area.

1.4.2 Research aim

The aim of the study is to create mid-infrared soil property specific calibration models for soil pH, P and ECEC for soils in the Western Highveld summer grain area of South Africa.

1.4.3 Research objectives

To test the hypothesis and meet the research aim, the following objectives must be met:

1. To establish a soil database for the Western Highveld summer grains area which includes soil property data.
2. Determine, collect, analyse and scan 1 000 samples with a MIR instrument which covers the entire soil attribute space of the Western Highveld summer grain area to create a soil property database and soil spectral library.
3. Create and evaluate soil property specific calibrations algorithms for soil pH, P and ECEC for use throughout the Western Highveld summer grain area.

CHAPTER 2 LITERATURE REVIEW

2.1 Introduction

In this chapter, a literature review was conducted to explain and present research on soil spectroscopy and how it is used for soil property prediction. The focus of this study revolves around the concept of spectrometry and the use thereof to predict soil properties, thus all aspects that are important and relating to soil spectroscopy will be discussed in detail. Spectroscopy, the creation of a soil spectral library, the calibration used for soil property prediction and validation of the calibration models are important sections that will be presented.

2.2 Infrared spectroscopy

Electromagnetic radiation (EMR) is defined in physics as waves of electromagnetic fields traveling through space with electromagnetic energy, EMR includes radio waves, microwaves, infrared (IR), visible radiation (light), ultraviolet radiation, X-rays, and gamma rays (Atkins and De Paula, 2010). When molecules interact with EMR it causes rotational, vibrational, electronic or ionization processes to occur with increasing energy exposure from microwave, infrared and ultraviolet radiation, respectively (Hollas, 2004). The infrared portion of the electromagnetic spectrum ranges from 10 to 14 790 cm^{-1} and is divided into three regions namely near- (NIR) (14 290 $-4\,000\ \mathrm{cm}^{-1}$), mid- (MIR) (4 000 $-\ 400\ \mathrm{cm}^{-1}$) and far infrared (FIR) (400 – 10 cm^{-1}) as depicted in Figure 2.2.1. Infrared spectroscopy on soils have predominantly been conducted within the NIR region with most soil spectroscopy studies using NIR to predict soil properties (Viscarra Rossel *et al.,* 2006).

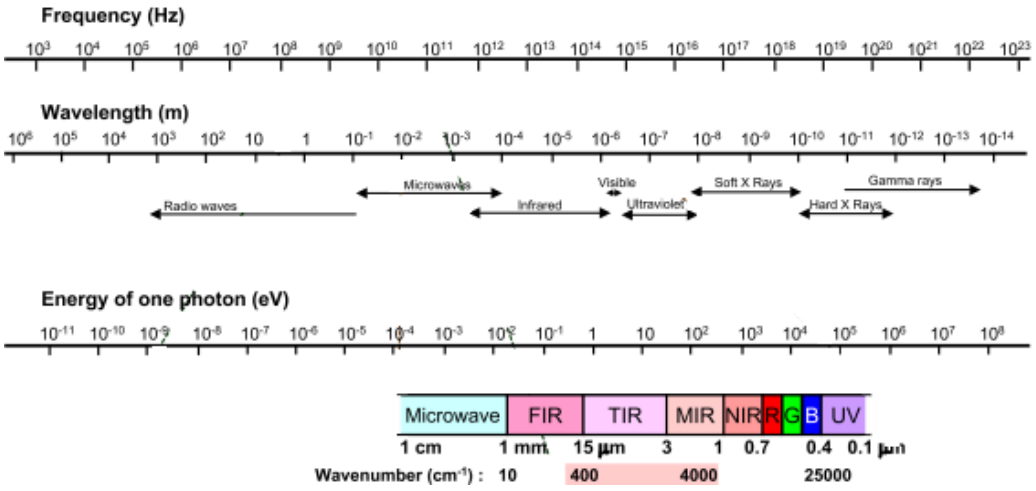

Figure 2.2.1:Different parts of the electromagnetic spectrum as presented by McBratney *et al*., (2003) and Viscarra Rossel *et al.*, (2006).

The interaction of IR radiation with bonds of molecules that are naturally non polar causes an electric dipole moment and the molecule becomes electromagnetically interactive, such molecules are called infrared-active molecules (Atkins and de Paula, 2010). This interaction between IR radiation and molecular bonds causes inter-atomic bonds to stretch and contract resulting in the formation of an electric dipole moment (Atkins and de Paula, 2010). The stretching of a bond between two atoms causes vibrational features and is portrayed in Figure 2.2.2.

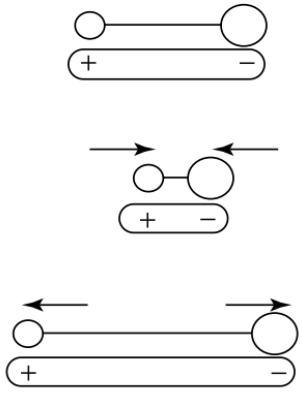

Figure 2.2.2: An infrared-active heteronuclear diatomic molecule that undergoes vibrational stretching of the bond between the atoms causes an electric dipole moment when IR wavelengths of radiation interacts with the molecule (Stuart, 2005).

Infrared spectrometry is therefore defined as the study of the molecular vibrations upon the exposure to radiation. Such vibrational information can be used to investigate bulk or microscopic amounts of samples that can have a wide range of temperatures as well as different physical states (Khoshhesab, 2012). To collect this information or IR spectrum, IR radiation is radiated onto or thorough a sample. The fraction of energy that is absorbed by the sample at a specific energy is then determined and related to the vibrational features of a molecule (Khoshhesab, 2012). IR spectroscopy combined with reflectance theory use the reflected radiation from samples to determine the absorption properties of the sample and is described by Khoshhesab (2012) as IR reflectance techniques.

There are two main categories of IR reflectance methods, namely internal and external reflection. Internal reflection is the interaction of EMR on the surface of the sample and a medium with a high refraction index shown by Figure 2.2.3 (Khoshhesab, 2012). External reflection can be subdivided into specular and diffuse reflection, as can be seen in

Figure 2.2.4. Specular reflection occurs on smooth and polished surfaces such as mirrors, while diffuse reflections are associated with rough surface samples, such as soil.

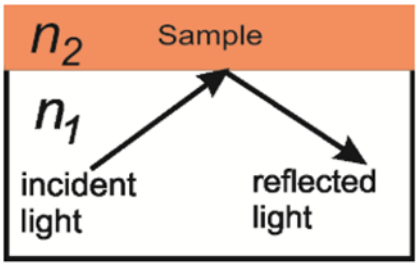

Figure 2.2.3: Internal reflection of incident IR radiation on a surface or medium with high refraction index value (Khoshhesab, 2012).

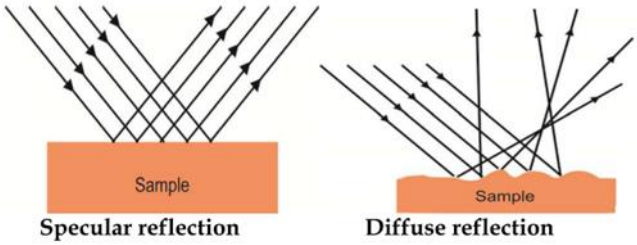

Figure 2.2.4: Specular and diffuse reflection of IR radiation on surfaces that are either smooth or rough as for soil samples (Khoshhesab, 2012).

Diffuse reflectance spectra are obtained from samples with uneven or irregular surfaces, usually finely powdered samples. Incoming IR radiation penetrates these samples at different depths and is reflected at different angels. Some IR radiation will reflect as a specular component, whilst the majority of IR radiation will reflect as a diffuse reflectance component (Thompson, 2018). To accurately use diffuse reflectance for spectroscopy, it is imperative that the specular reflectance component is not detected with the diffuse reflectance component due to the distorting effect of specular reflectance component on the diffuse spectra (Khoshhesab, 2012). Diffuse reflectance infrared Fourier transform spectroscopy (DRIFTS) is used to eliminate this problem by removing the specular reflectance component. When using diffuse reflectance, it is not possible to directly relate the intensity, also known as band intensity, of the spectra to the concentration of constituents in the sample as it is for traditional transmission spectroscopy where band intensity is directly related to the concentration (Khoshhesab, 2012). Soil spectroscopy is an indirect quantitative method that is used to rapidly predict soil properties from spectral

measurements of soil samples (Nocita *et al.*, 2015). This is achieved using multivariate calibration models developed from spectral libraries that contain spectral measurement data and the accompanying soil property measurement from the laboratory (Nocita *et al.*, 2015).

2.3 Mid-infrared spectroscopy

The mid-infrared region (4000 - 400 cm^{-1}) is further divided into sections that can be seen in Figure 2.3.1, which corresponds to the following group frequencies: hydrogen bonds, X-H (4000 - 2500 cm^{-1}), triple-bonds (2500 - 2000 cm^{-1}), double-bonds (2000 - 1500 cm^{-1}) as well as the "fingerprint" region located between 1500 cm^{-1} and 600 cm^{-1} (Stuart, 2005). The "fingerprint" region is associated with intense vibrational features specific to a molecule, which can act to identify or "fingerprint" the molecule (Stuart, 2005).

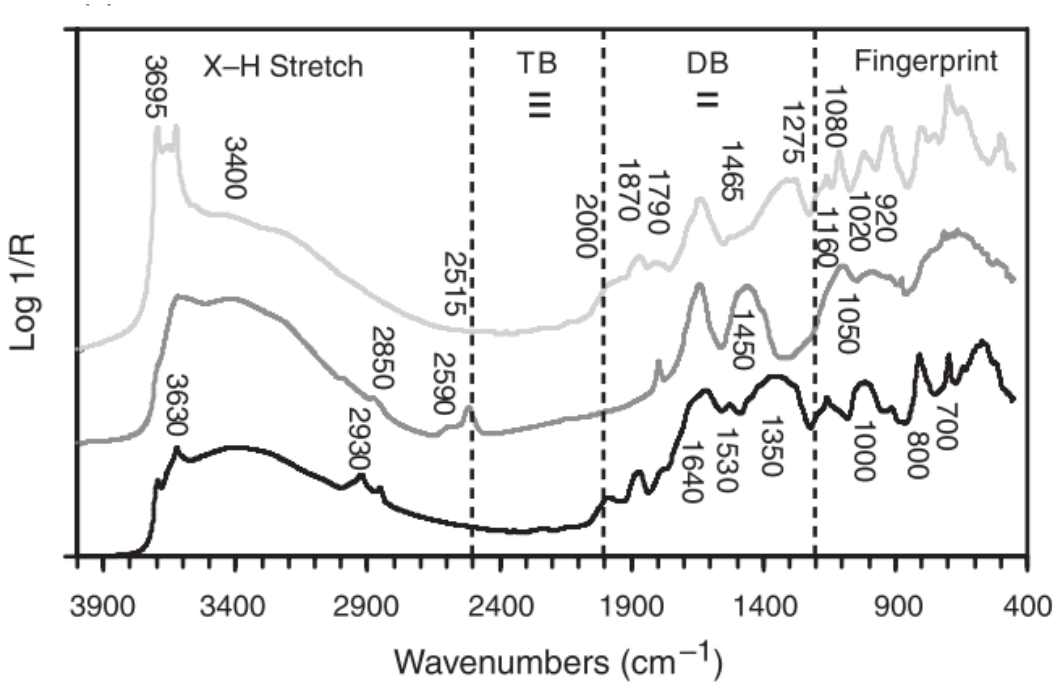

Figure 2.3.1: Different regions within the MIR wavelength that correspond to certain molecular bonds (Rossel *et al.*, 2008).

Studies by Soriano-Disla *et al*. (2014) and Viscarra Rossel *et al*. (2006), which compared the performance of visible, NIR and MIR reflective spectroscopy found that several soil properties can be successfully predicted with great accuracy. Some of these properties include soil water content, soil texture, carbon (C), CEC, calcium (Ca), exchangeable magnesium (Mg), total N, pH

as well as other metals (Soriano-Disla *et al.*, 2014). Soriano-Disla *et al*. (2014) compiled a table (Table 2.3.1) with the regions at which certain soil components react, within the MIR range (Soriano-Disla *et al.*, 2014).

Table 2.3.1: Spectral absorptions associated with soil components in the MIR region. Table adapted from Soriano-Disla *et al.*,(2014).

Soil component	**Wavenumbers cm^{-1}**	**References**
Quartz	1 100 - 1 000	Fischer (1977), Nguyen *et al.* (1991)
Clay minerals	(Kaolinite) 3 690 - 3 620	Fischer (1977), Nguyen *et al.* (1991), Rossel and Behrens (2010)
	(Smectite and illite) 3 620 - 3 630 and 3 400 - 3 300	Fischer (1977), Nguyen *et al.* (1991), Rossel and Behrens (2010)
Carbonates	2 520 - 1 430	Fischer (1977), Nguyen *et al.* (1991), Rossel and Behrens (2010)
Iron oxides	700 - 600	Fischer (1977)
Iron	3 100, 900, 800	Fischer (1977)
Organic matter	1 670 and 1 530 ($\mathrm{OC-NH}$),	Fischer (1977), Nguyen *et al.* (1991), Rossel and Behrens (2010)
	2 930 - 2 850 ($-\mathrm{CH_2}$),	Fischer (1977), Nguyen *et al.* (1991), Rossel and Behrens (2010)
	1 720 (COOH),	Fischer (1977), Nguyen *et al.* (1991), Rossel and Behrens (2010)
	1 630 Water associated,	Fischer (1977), Nguyen *et al.* (1991), Rossel and Behrens (2010)
	1 600 and 1 400 ($-\mathrm{COO^-}$),	Fischer (1977), Nguyen *et al.* (1991)
	1 600 - 1 570 (aromatic groups)	Fischer (1977)

The use of MIR spectroscopy has greatly increased over the years due to several factors that makes the MIR region preferable to the Visible or NIR regions (McCarty and Reeves, 2006). It has been suggested by McCarty and Reeves (2006) that the MIR region contains a greater collection of vibrational signals compared to the NIR region which contains weak broad signals that cause overtones in the spectra. The MIR region has also been found to produce more robust calibration models for soil property predictions compared to the NIR region which can be attributed to the high amounts of spectral information obtainable from the MIR spectra (McCarty and Reeves, 2006).

2.4 Soil spectral library creation

To predict soil properties using MIR DRIFTS spectrometry a sample set needs to be established known as a soil database. This soil database must include soils covering the entire study area to account for the variability of soils (Shepherd and Walsh, 2002). Shepherd and Walsh (2002) gives a detail description on how to use this soil data to create a successful soil spectral library that will support the creation of robust models to predict soil properties from MIR soil reflectance data. This process was first illustrated by Shepherd and Walsh (2002) (Figure 2.4.1) and shows that the first step is to acquire samples into a library that adequately represents the target area. Next, a soil reflectance library is built using soil spectral data of each sample and combined with the accompanying soil property data. A calibration of a model is created with the soil property data to the spectral data and the accuracy is evaluated. If the model performs well a prediction can be made and if the accuracy is not accepted the process is repeated from the spectral data sampling stage. New samples can be added to the library by taking the soil spectral data of the new samples and checking for spectral outliers.

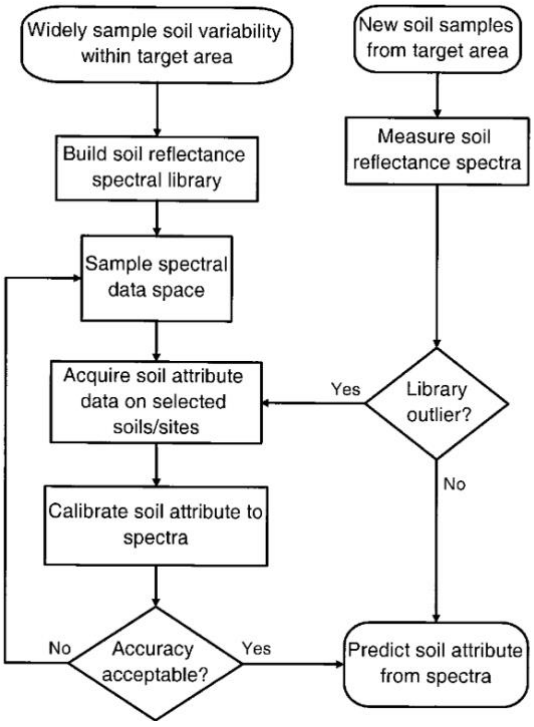

Figure 2.4.1: A scheme developed by Shepherd and Walsh, (2002) that shows the process of building a soil spectral library which includes a soil property database.

A similar process to that of Shepherd and Walsh have been adapted and presented by Rossel et al. (2008) where it is explained how legacy soil samples could be used to create a soil spectral library, and how new soil samples can be added to the spectral library. Soil samples that are selected from a legacy soil sample set or a targeted new study area should contain the maximum number of representative samples that will properly represent the soil spatial variability found in

the study area or dataset (Wadoux *et al.*, 2021). The way samples are selected, and the size of subsamples plays an important role in the accuracy that will be produced for a calibration model (Wadoux *et al.*, 2021). Sample selections in spectroscopic research usually use sample designs that select sample sets based on spectral data (Wadoux *et al.*, 2021). This study will use the soil property data instead of using the soil spectral data to select a subset of samples for use in calibrating the different models. Wadoux *et al.* (2021) confirms that it is theoretically possible to use sample designs in such a way and thus this will only be used in the initial selection of samples when reducing the number of samples to include in the study. Ideally sample designs like conditioned Latin hypercube sampling (cLHS) are well suited for use in expanding and improving soil spectral libraries by using soil spectral data to create a subsample.

2.4.1 Conditioned Latin hypercube sampling

Minasny and McBratney, (2006) developed the conditioned Latin hypercube sampling, based on Latin hypercube sampling (LHS), which is a constrained Monte-Carlo sampling design developed by McKay *et al.* (1979). The main function of cLHS is to equally represent the entire distribution of a set variable within the sample collection. For a single variable (single dimension) the distribution of the variable is divided into a set number of strata equal to the sample size required (Wadoux *et al.*, 2021). For more than one variable, a sample is selected using an optimization algorithm which minimizes an objective function that is a weighted sum of two components (Wadoux *et al.*, 2021). These components are known as O_1 which is the first component and guarantees that each variable contains one unit per stratum in the multi-dimensional feature space, while O_2 correlates between the value of the variable in the sample and selection population (Wadoux *et al.*, 2021). According to Malone *et al.* (2017), cLHS sampling was initially designed for the purpose of digital soil mapping (DSM) and has seen use in DSM around the world with studies performed in China (Sun *et al.*, 2017), South Korea (Jeong *et al.*, 2017), Canada (Scarpone *et al.*, 2016), Australia (Thomas *et al.*, 2015) and South Africa (van Zijl, 2019). There are many ways in which cLHS can be utilized and different tools like MATLAB (MATLAB, 2010), R (R Core Team, 2020) for example the $spsann$ package and the $clhs$ package (Wadoux *et al.*, 2021) as well as a standalone application using Microsoft Windows Command Prompt.

2.4.2 Spectral pre-processing

The use of pre-processing techniques on spectral data has primarily been to improve the spectra in terms of quality for qualitative, calibration and prediction of soil properties (Wadoux *et al.*, 2021). Spectral pre-processing was adopted from chemometrics to soil spectroscopy (Buddenbaum and

Steffens, 2012). There are many pre-processing techniques that are available for use and there is no determined best performing singe method or sequence of methods for pre-processing spectra (Wadoux *et al.*, 2021). Most spectra that are obtained from spectroscopic analysis are disordered, noisy, inadequate, defective or exhibits a combination of these and other problems including noise spikes, non-detections, and unrealistic values (Wehrens, 2020), which leads to pre-processing to be required for analysis. Wadoux *et al*. (2021) present six different categories for spectral pre-processing which includes: noise removal, scatter correction, derivatives, centring, and standardizing, spectral or dimension reduction and other specific transformations that are summarized in Table 2.4.1.

Table 2.4.1: Summary of pre-processing techniques presented by Wadoux *et al.*,(2021) with references to other studies which included these methods.

Pre-Processing category	**Technique**	**Reference**
Noise Removal	Spectral Trimming	Metzger *et al.* (2020),Wadoux *et al.* (2021)
	Moving Window Average (MWA)	Wadoux *et al.* (2021)
	Savitzsky-Golay (SG)	Cambule *et al.* (2012), Peng *et al.* (2016), Metzger *et al.* (2020), Wadoux *et al.* (2021)
Scatter correction	Standard Normal Variate (SNV)	Buddenbaum and Steffens, (2012), Hobley and Prater (2019), Metzger *et al.* (2020), Wadoux *et al.* (2021)
	Multiplicative Scatter Correction (MSC)	Buddenbaum and Steffens, (2012), Cambule *et al.*, (2012), Metzger *et al.* (2020), Peng *et al.*, (2016), Wadoux *et al.*, (2021)
	Detrending	Wadoux *et al.* (2021)
Derivatives	First- and Second Order	Buddenbaum and Steffens (2012), Cambule *et al.* (2012), Peng *et al.* (2016), Wadoux *et al.* (2021)
Centring and Standardization		Metzger *et al.* (2020), Wadoux *et al.* (2021)
Spectral or dimensions reduction	Resampling	Wadoux *et al.* (2021)
	Wavelets	Wadoux *et al.* (2021)
Other	Splice Correction	Wadoux *et al.* (2021)
	Continuum Removal (CR)	Buddenbaum and Steffens (2012), Peng *et al.* (2016), Wadoux *et al.* (2021)

As mentioned before, there is no predetermined sequence of implementing these techniques on mid-infrared spectral data. Cambule *et al.* (2012) first used raw spectral data then MSC, after which they applied the first derivatives and finally the SG filter. Peng *et al.* (2016) also used spectral trimming, SG smoothing, MSC, SNV, CR and the first derivatives. The pre-processing techniques chosen for this study and the MIR spectral data were selected based on a study by Peng *et al.* (2016), where they found that SG filtering combined with the first derivatives were was the best performing pre-processing method for MIR spectral data.

Spectral trimming as the name implies trims the spectra to remove high noise-to-signal wavelength ranges. For vis-NIR this is the 28 571 – 20 040 cm^{-1} and 4 080 – 4 000 cm^{-1} and for MIR the edges of the measured spectral range of 400 – 4000 cm^{-1} is trimmed to remove noise associated with the detector (Wadoux *et al.*, 2021). When calculating the MWA, the average of the neighbouring wavelengths is determined based on the size of the MWA also known as the size of the window (w) (Wadoux *et al.*, 2021). When using MWA, wavelengths are lost at the beginning and end of the wavelength range and the number of these wavelengths lost during the calculation is determined by Equation **2.4.1** (Wadoux *et al.*, 2021).

Equation 2.4.1:

$$\frac{w-1}{2}$$

Where *w* represents the size of the window.

Savitzky-Golay filtering developed by Savitzky and Golay, (1964) applies a polynomial regression on a set of spectral values which produces a smoother wavelength value for a set filter length or window size (Wadoux *et al.*, 2021). Spectral data that is run through a Savitzky-Golay filter algorithm can also be transformed to a smooth first or second derivative spectra which enables the combined use of a Savitzky-Golay and the first or second derivatives pre-processing methods (Peng *et al.*, 2016; Lohninger, 2021).

2.5 Model calibration for soil property prediction

There are many different approaches that are used for soil property prediction and include: Cubist, memory-based learning (MBL), Random Forest (RF), Partial least square (PLSR), principal component analysis (PCA), artificial neural networks (ANN), multivariate adaptive regression splines (MARS), support vector machines (SVM), multiple linear regression (MLR) and convolutional neural network (CNN). Dangal *et al.* (2019) used four of these calibration methods, three of which are machine learning algorithms Cubist, MBL and RF. Partial least square regression is the most widely used and was also included by the study of Dangal *et al.* (2019). This study chose three calibration methods based on the findings of Dangal *et al.* (2019) and includes Cubist, PLSR and RF.

2.5.1 Cubist

Cubist is a machine learning model that was first used for DSM by McBratney *et al.* (2003), which introduced Cubist for DSM for predicting continuous soil attributes. Cubist is regarded as a data mining tool that was developed from early model trees C4.5 and M5 (Peng *et al.*, 2015). Data that is imported into Cubist is divided into sets of data that have the same features as the predictor and the variables, Cubist uses a series of rules like that of a decision tree to divide the data into a hierarchy made out of nodes, using “if [condition is true]”, “then [regress]” and “else [apply the next rule]” statements (Peng *et al.*, 2015). If a node has a true result, a prediction is made for the soil property value using least-squares regression and the variable values in that partition (Peng *et al.*, 2015). When a condition is false a new node in the tree is created by the rule, which repeats the if, then and else statements (Peng *et al.*, 2015). The final tree is trimmed to include all the rules that were accepted and a path is created that runs from top to bottom, the model is then adjusted and simplified which reduces the absolute error (Peng *et al.*, 2015). Cubist models are created using the $cubist$ package (Kuhn, 2021) in R (R Core Team, 2020). Cubist was used by Minasny and McBratney, (2008) for total carbon, clay and CEC model calibration and prediction with MIR diffuse reflectance spectra. Dangal *et al.* (2019) used Cubist to predict nine soil properties including CEC and pH with r^2 values of 0.99 and 0.88 respectively using a large MIR

spectral library. Few studies have used Cubist for soil property prediction, the increased prediction accuracy will lead to more studies including this algorithm (Dangal *et al.*, 2019).

2.5.2 Partial least square regression

Partial least square regression has been determined to be useful when working with MIR spectroscopy because of the amount of spectral data that is produced with MIR (Janik *et al.*, 2007). This method enables MIR spectral data to be processed with speed and ease (Janik *et al.*, 2007). Partial least square regression can accomplish this by combining principal component analysis and multiple regression (Wadoux et al., 2021). Soil properties are predicted using PLSR by using latent variables created from the spectral data, the soil property data and spectral data is divided into smaller sets called orthogonal loadings and scores (Janik *et al.*, 2007). The loadings and scores created by PLSR is equivalent to the principal components of PCA but are instead more focussed on the soil property data than just the spectra (Bramley and Janik, 2005). When calibrating the PLSR model a principal component calculation is run with the aim of determining the optimal amount of principle components that will create a model with the lowest root mean square error (RMSE) without overfitting (Bramley and Janik, 2005; Kassambara, 2018). The method of combining the spectral- and soil property data leads to the creation of a calibration model, which is then used to make soil property predictions from spectral data with unknown soil property values (Janik *et al.*, 2007). Partial least square regression can also be implemented using R (R Core Team, 2020) with the $plsr$ function in the pls package created by Mevik and Wehrens, (2015). Partial least square regression has been widely used for soil property predictions with numerous studies including this calibration algorithm. Most recent studies include: (Wold *et al.*, 2001), Dangal *et al.* (2019), Mura *et al.* (2019), Comstock *et al.* (2019), Metzger *et al.* (2020) and Sleep *et al.* (2021) which all used PLSR to predict different soil properties using NIR and MIR spectra. The best PLSR results for pH and CEC was produced by Dangal *et al.* (2019) with r^2 values of 0.96 for CEC and 0.74 for pH.

2.5.3 Random Forest

Random forest is a newer non-linear modelling algorithm compared to Cubist and PLSR (Knox *et al.*, 2015). Random forest functions as an ensemble of decision trees which are able to perform classifications and regressions (Breiman, 2001; Yiu, 2019). This data mining algorithm uses multiple decision trees that are constructed from different perturbations of the data being used (Wadoux *et al.*, 2021). Data is split up in two thirds that are used to create or grow the regression tree while the rest is used for cross-validation in conjunction with the calibration step, these are known as out-of-bag samples and gives an idea about how well the models performs (Shahrayini *et al.*, 2020; Wadoux *et al.*, 2021). When RF is used for regression a final prediction is determined by the average of the outputs for the individual trees, for classification the trees are voted on by count of majority (Breiman, 2001; Yiu, 2019; Wadoux *et al.*, 2021). Random forest is ideally suited to work with soil spectral data, especially MIR, with the high dimension of wavelengths. It is also optimized to work with large amounts of spectra with small sample sizes and is powerful enough to handle noise and features from the MIR that is irrelevant (Rossel and Behrens, 2010). Random forest is easy to use and tune for optimal results which will in return reduce the tendency of RF to over-fit for calibration models (Rossel and Behrens, 2010). Cubist and PLSR has a usage advantage over RF for soil property calibrations as new data is sometimes added to the dataset to improve the model. Cubist and PLSR can accept this data, while RF cannot (Wadoux *et al.*, 2021). Random forest has seen implementation in soil property prediction studies from 2010 and is increasingly gaining popularity as a prediction algorithm (Rossel and Behrens, 2010). However, minimal studies exist where RF was used for pH, ECEC and P calibration and predictions with MIR. Rossel and Behrens, (2010) produced results for Organic C, clay and pH (water) with a r^2 value of 0.63, while Dangal *et al.* (2019) produced r^2 values for pH and CEC using RF of 0.82 and 0.97 respectively.

2.6 Model validation

Model validation or goodness of fit plays an important role in model calibration and validation, it is performed to assess the models created by calibration by comparing the predicted soil property values with that of the actual laboratory measured values (Wadoux *et al.*, 2021). There are several

indicators used in predictive soil spectroscopy described in Wadoux *et al.* (2021) which include: RMSE (Terhoeven-Urselmans *et al.*, 2010), Mean error (bias) (Terhoeven-Urselmans *et al.*, 2010), squared correlation coefficient (r^2), coefficient of determination (R2) (Terhoeven-Urselmans *et al.*, 2010), Lin's concordance coefficient (p_c), Ratio of performance to deviation (RPD) (Terhoeven-Urselmans *et al.*, 2010) and Ratio of performance to inter-quartile distance (RPIQ). Focus will be on RMSE, bias, r^2 and RPD as they were commonly used in soil spectroscopy calibration evaluations, coefficient of variability (CV) used by Janik *et al.* (2009) will also be calculated to express the RMSE as an percentage. Dangal *et al.* (2019) like many other studies used all four performance parameters in evaluating the prediction performances of Cubist, PLSR and RF models for pH and CEC.

2.6.1 Root mean square error

The root mean square error is calculated from the standard deviation of the measured soil property values subtracted from the predicted soil property values (residuals), which gives the accuracy of the prediction (Viscarra Rossel *et al.*, 2006) or the dispersion of the residuals around a 1:1 line of best fit (Wadoux *et al.*, 2021). The following Equation 2.6.1 is used to calculate RMSE, where n is the sample size and obs_i and $pred_i$ are vectors of the measured and predicted soil property values respectivily (Wadoux *et al.*, 2021).

Equation 2.6.1: Root mean square error

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

2.6.2 Bias

Mean error is used to quantify bias and is calculated along with RMSE to evaluate the bias of the predictions made (Viscarra Rossel et al., 2006; Wadoux et al., 2021). A good prediction model will have a very low bias and ideally it should be zero (Hobley and Prater, 2019; Wadoux *et al.*,

2021). Bias is calculated by Equation 2.6.2 as the average of errors between measured and predicted values.

Equation 2.6.2: Bias

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

where n is the sample size and obs_i and $pred_i$ are vectors of the measured and predicted soil property values respectivily (Wadoux *et al.*, 2021).

2.6.3 Coefficient of determination

Coefficient of determination is the sum of squared errors divided by the total sum of squares given by Equation 2.6.3 (Peng *et al.*, 2015). The optimal value for r^2 according to Wadoux et al. (2021) should be one and in some cases negtive when the RMSE is greater tha the standard deviation of the data.

Equation 2.6.3: Coefficient of determination

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

where n is the sample size and obs_i and $pred_i$ are vectors of the measured and predicted soil property values respectivily, $\overline{obs}$ is the total measured soil property values (Wadoux *et al.*, 2021).

2.6.4 Ratio of performance to deviation

The RPD is given “as the ratio of standard error in prediction to the standard deviation” by Wadoux *et al.* (2021). Chang *et al.* (2001) established categories using RPD values to classify the performance of the prediction models regarding soil properties: Category A (RPD > 2.0 and r^2 = 0.80-1.00), Category B (RPD = 1.4~2.0 and $r^2 = 0.50 - 0.80$), Category C (RPD < 1.4).

Category A models are able to accurately predict soil properties, Category B models is regarded as an intermediate class where improvement is possible using different models and Category C models have no prediction abilities (Nocita *et al.*, 2011). Ratio of performance to deviation is calculated using Equation 2.6.4 (Wadoux *et al.*, 2021).

Equation 2.6.4: Ratio of performance to deviation

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

where n is the sample size and obs_i and $pred_i$ are vectors of the measured and predicted soil property values respectivily, $\overline{obs}$ is the total measured soil property values (Wadoux *et al.*, 2021).

2.6.5 Coefficient of variability

Coefficient of variability is used to perform an informed evaluation on the prediction performance between two or more prediction models prediction models (Janik *et al.*, 2009). Coefficient of variability is expressed as a percentage and is calculated with Equation **2.6**.**5**.

Equation 2.6.5: Coefficient of variability.

$$CV = \frac{RMSE}{Mean}\ x\ 100$$

where $RMSE$ is the root mean square error calculated of the calibration and prediction model with Equation 2.6.1 and $Mean$ the average of the soil property dataset used for the specific model.

2.7 MIR soil spectroscopy for soil property prediction

2.7.1 pH

Soil pH plays an important role with regard to the availability of soil nutrients, soils with low pH will reduce the molybdenum (Mo), P, Mg and Ca available to plants (Maher *et al.*, 2003). Other elements like Al, Fe and manganese (Mn) are more available with low pH and can become toxic

at certain concentrations (Maher *et al.*, 2003). Some natural soils are acidic by nature, which is caused by poor buffering in high rainfall areas. South Africa has an increasing soil acidification problem mostly caused by anthropogenic activities like cultivation which reduces organic matter and over fertilization of N (Barnard and du Preez, 2004). Soil laboratories currently measure soil pH using a pH electrode, this method uses a pH-sensitive glass electrode and a standard electrode as reference (a combination of both in one electrode is also used) that is inserted into a soil water mixture of 1:1 or 1:2.5 known as pH_{water} or pH_{KCl} when potassium chloride (KCl) is used instead of water (Brady and Weil, 2017). More detail about the laboratory measurement of pH is discussed in Appendix A. The prediction and calibration of soil pH using soil spectroscopy has widely been implemented with NIR and MIR wavelengths, some studies include Shepherd and Walsh (2002); Viscarra Rossel *et al.* (2006); Rossel and Behrens (2010); Soriano-Disla *et al.* (2014); Wijewardane *et al.*, 2018) ; Sleep *et al.* (2021) are also summarized in

Table 2.7.1. pH is best predicted within the 4 000 - 25 000 cm^{-1}range (Sleep *et al.*, 2021).

Table 2.7.1: Studies that produced results for the MIR prediction of soil pH.

Reference	**pH**	r^2	**RMSE**	**bias**	**Location**
Shepherd and Walsh (2002)	-	0.83	0.34	-0.02	Malawi, Kenya, Rwanda, Tanzania, Uganda, Zambia, and Zimbabwe
Viscarra Rossel *et al*. (2006)	pH_{Ca}	0.86	0.10	0.00	Australia
	pH_w	0.75	0.13	-0.01	
	pH_b	0.64	0.05	0.00	
Rossel and Behrens (2010)	pH_w	0.73	-	-	Australia
Wills and Libohova (2018)	pH_w	0.80	0.57	-	United States of America
Sleep *et al.* (2021)	-	0.4	0.78	-	Australia
Rossel *et al.* (2008)	pH_{Ca}	0.7	0.41	-0.05	Australia
Dangal *et al.* (2019)	pH	0.88,0.74,0.2	<0.5	-	United States of America

2.7.2 Phosphorus

Phosphorus is one of the most important elements that plants require and enough P needs to be available for crops to sustain the fundamental processes of photosynthesis, flowering, fruiting and seed production and to fully mature (Brady and Weil, 2017:662). An excess amount of P in soil can lead to negative environmental impacts (Abdi *et al.*, 2012). Major P and mostly insoluble deposits are found in South Africa originating from the igneous and marine sedimentary geologic environments including the Phalaborwa Complex (Barnard and du Preez, 2004). Mass balance calculations in agriculture and environmental studies generally use total P concentration, whilst this method is useful for such applications it is not adequate for the analysis of P in soils where it is preferred to calculate the available fractions of P (Soriano-Disla *et al.*, 2014). Phosphorus is calculated as extractable P and there are eight methods that include Bray-1, Bray-2, Resin-bag, Resin, $NaHCO_3$(Oslen), Modified ISFEI (Hunter), Citric acid method and Modified Truog method (The non-affiliated Soil Analysis Work Commitee, 1990), Bray-1 is most commonly used in South Africa and is discussed fully in the Appendix A. Maleki, *et al.*(2006) in Soriano-Disla *et al.* (2014) states that minerals and organic matter control the availability of P in soil due to the effect of absorption. Soil MIR spectroscopy is able to detect phosphorus compounds with the MIR spectra between 3 700 cm^{-1} and 1 500 cm^{-1} (Barnard and du Preez, 2004). Investigations by Soriano-Disla *et al.* (2014) showed that predictions using MIR for extractable P generally produced low R^2-values compared the predicting the P sorption and the total P in soil and can be seen in

Table 2.7.2.

Table 2.7.2: Studies that produced results for the MIR prediction of P.

Reference	**P**	r^2	**RMSE**	**bias**	**Location**
Soriano-Disla *et al.* (2014a)	Extractable P	0.35	-	-	-
	Sorption P	0.83	-	-	-
	Total P	0.58	-	-	-
Janik *et al.* (1998)	Available P	0.07	-	-	Australia
Daniel *et al.* (2003)	Available P	0.81	-	-	Thailand
Janik *et al.* (2009)	Extractable P	0.28	-	-	Australia
Rossel *et al.* (2008)	Extractable P	0.34	20.7	2.4	Australia

2.7.3 Effective cation exchange capacity

Effective CEC is calculated as the sum of the exchangeable acidity and the exchangeable cations (Shepherd and Walsh, 2002; Van Zijl *et al.*, 2014). Cation exchange capacity depends on the pH , ionic strength, nature and amount of absorbed anions and can be influenced to manage soil cation retainability (Edmeades, 1982). Cation exchange capacity is determined by the relative amounts of different colloids in the soil, sandy soils with low colloidal material have low CEC compared to silt loam and clay loams (Brady and Weil, 2017). Soils with high CEC are attributed to their hummus content which causes higher CEC values compared to that of inorganic clays (Brady and Weil, 2017). Extractable cations are determined using an Ammonium acetate solution method and the extractable acid with KCl that is given in Appendix A (The non-affiliated Soil Analysis Work Commitee, 1990). Effective CEC and CEC is highly correlated to soil texture and organic matter, this enables good prediction results when using MIR soil spectroscopy for ECEC and CEC prediction (Van Groenigen *et al.*, 2003). As presented in Table **2.3**.**1**, soil texture and organic matter is well represented with the MIR region. Wijewardane *et al.* (2018) showed diagnostic absorption bands associated with the regression coefficients of PLSR models used in his study. These bands include 3 500 – 3 800 cm^{-1}, 2 900 cm^{-1} and 1 500 – 1 800 cm^{-1} also presented in Table **2.3**.**1** for organic matter absorption bands. A few studies have predicted ECEC with soil spectroscopy and is presented in

Table 2.7.3.

Table 2.7.3: Studies that produced results for the MIR prediction of ECEC and CEC.

Reference	ECEC / CEC	r^2	RMSE	bias	Location
Shepherd and Walsh (2002)	ECEC	0.95	2.6	-0.11	Malawi, Kenya, Rwanda, Tanzania, Uganda, Zambia, and Zimbabwe
Van Groenigen *et al.* (2003)	ECEC	0.56	-	-	California, USA
Janik *et al.* (1998)	CEC	0.88	-	-	Australia
Rossel *et al.* (2008)	CEC	0.89	32.5	0.03	Australia
Dangal *et al.* (2019)	CEC	>0.96	2.3 - 4.02		United States of America

2.8 Conclusion

This section concludes that research in this field has been widely conducted outside of South Africa and shows that soil spectroscopy is a valid and accurate means of soil property determination. This also shows that more research is needed to be able to adopt this method for use in South Africa.

CHAPTER 3 MATERIALS AND METHODS

3.1 Introduction

The main components that are fundamentally important towards the successful implementation of this study includes the establishment of a soil property database, soil spectral database and the combination of both databases to create a soil spectral library which is then used for calibrating prediction models for the soil properties pH, P and ECEC.

3.2 Sample acquisition and study Area

3.2.1 Sample acquisition

A vast collection of soil samples is required to construct a soil property database that will provide information regarding the chemical properties of the soil samples required for this study. Soil samples donated from North West Kooperasie (NWK) and Griekwaland Wes Kooperasie (GWK) facilitated in creating a soil property database without the need to collect samples from the field. Both NWK and GWK provided access to their seasonal soil samples that they processed and analysed along with the chemical data. North West Kooperasie provided a selection of 4 393 soil samples and GWK had 175 samples for us to select from. A total of 1 500 soil samples provided by NviroTek Central (Hartbeespoort, North West) were selected from the 4 393 soil samples available using cLHS developed by Minasny and McBratney (2006) with the soil properties pH, P and T value. The samples provided by NviroTek Central were from soil samples sent to the laboratory for analysis of which they had completed during each two-week interval. In addition, 100 soil samples provided by GWK was also selected from the 175-soil sample collection mentioned above.

3.2.2 Study area

A map of the study area in Figure 3.2.1 was created by plotting the locations of samples received from NWK with accompanying GPS co-ordinate data. All samples are located within the Western

Highveld grain producing area which are situated in the North-West province of South Africa and is located at 26°27'24.3", S 26°04'58.0"E. The study area is located at an average height of 1 769 m above sea-level (USGS,2021) with an average annual temperature of 19.7°C and an average annual precipitation of 567 mm (Malherbe *et al.*, 2016), making it part of the arid to semi-arid regions.

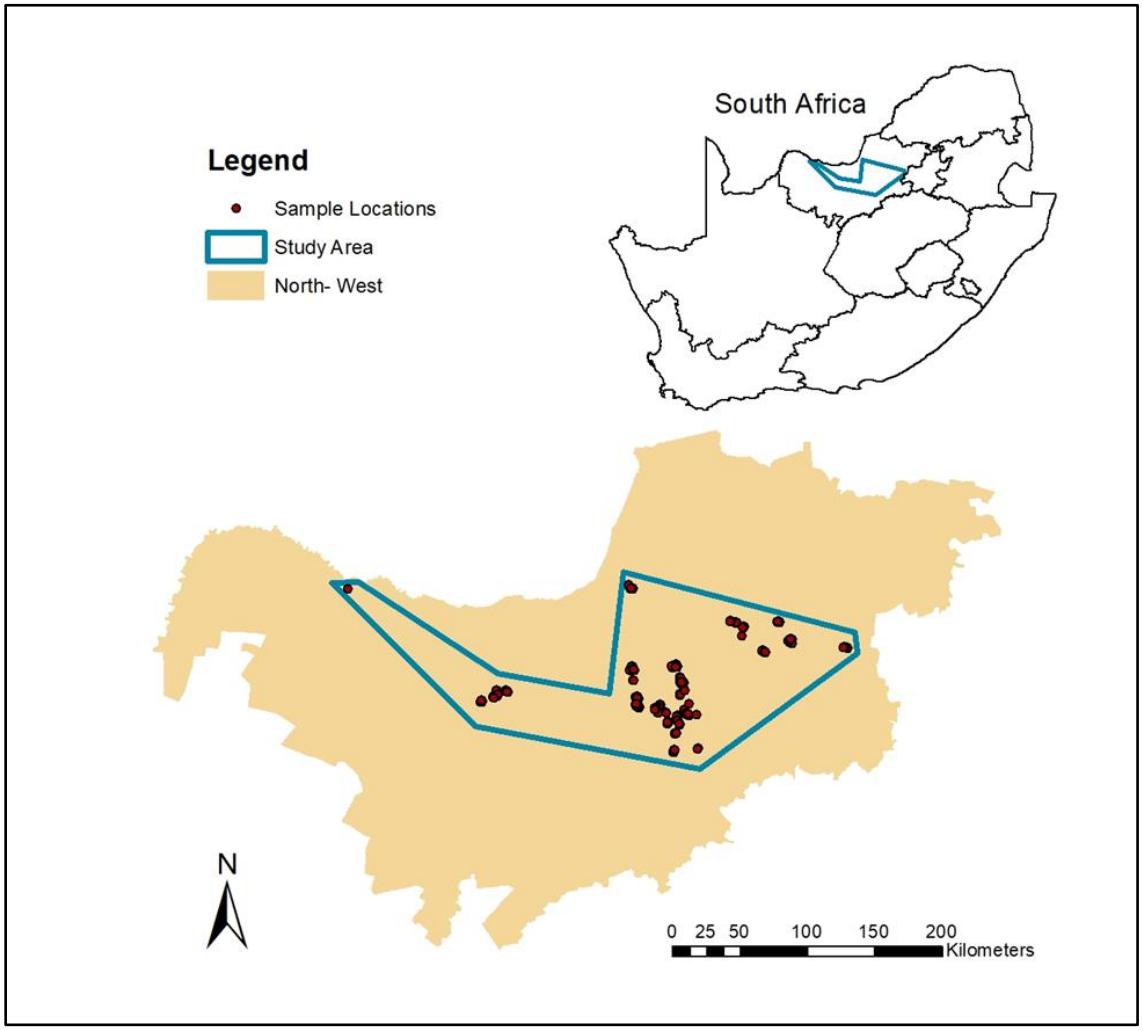

Figure 3.2.1: A map showing various sample locations within the North-West province of South Africa.

The geology of the study area consists out of several formations: the Kalahari group which covers the westerns parts of the study area and mostly consists out of calc-conglomerate, mudstone, gritstone, siliceous/calcareous sandstone, silcrete, diatomaceous limestone and calcrete. While the eastern parts of the study area include the Klipriviersberg group, Malmani subgroup and Rietgat formations which comprises out of dolomite, chert, shale, limestone, quartzite, basalts, andesites, tuff, agglomerate, gneiss, and granites (Council for Geoscience, 2019). Several broad land types are present which consist of A - red and yellow redoximorfic, B - plintic, E - black or red clays, F - shallow soils and I - alluvia and rock outcrops (Land Type Survey Staff, 2006). Landcover data from Department of Forestry, Fisheries and the Environment (DFFE, 2018) indicate that all-sample locations occur on commercial farmland that is used for the cultivation of annual crops, including *Zea Mays L. and Helianthus annus L.*

3.3 Establishment of soil property database

Due to capacity and time constraints, we were limited to a total of 1 000 samples of which 900 were selected from the 1 500 samples of NWK and 100 from the 175 collected of GWK as described in **Error! Reference source not found.**. The final selection was made using cLHS with the soil properties pH, P and ECEC (T value). Only 854 NWK and 95 GWK soil samples were finally added to the soil spectral library, 46 soil samples needed for a total of 900 NWK soil samples and 5 for a total of 100 GWK soil samples were lost due to unintentional human errors during mid-infrared scanning process and were excluded from the final soil spectral library. All soil samples selected and collected had accompanying soil property data provided by NWK and GWK. The soil property data supplied by NWK and GWK both contained measurements for pH and P. Effective CEC however, was only obtained for samples analysed by NviroTek Central (Hartbeespoort, North West). Both NWK and GWK use a potentiometric method to measure pH(KCl) (The Non-affiliated Soil Analysis Work Commitee, 1990). Available/extractable P for all soil samples were determined using the Bray-1 method, which is preferable for acidic soils like those found in some parts of South Africa (Ali and Narayan, 2009). Effective CEC is calculated by adding the titratable acidity and base cation concentrations (Schwertfeger and Hendershot, 2009). Soil chemical data received from NWK which contains data for ECEC is calculated the same way by adding the sum of extractable cations which is denoted as the S-value and the

titratable acidity to obtain a value for ECEC. The extractable cations are determined using the ammonium acetate (NH_4OAc) method (The Non-affiliated Soil Analysis Work Committee, 1990). These methods are described in detail in Appendix A.

3.4 Soil spectral database

A soil spectral library was established by combining all the MIR spectral data with the laboratory analysis data from the soil property database. Samples that were collected from NWK and GWK were already dried and sieved to < 2 mm that was suitable for the wet laboratory analysis. Samples were scanned to obtain the MIR spectral data. To scan each sample, additional preparation was necessary. Each sample was ground using a Retsch Mortar Grinder RM 200 (Retsch GmbH, 2021) to obtain a grain size < 100 $\mu\mathbf{m}$ (Terhoeven-Urselmans *et al*., 2010). Each soil sample was then scanned using a Bruker Alpha II with DRIFT module (Figure 3.4.1) (Bruker, 2021) with a spectral range of 3996 - 398 $\mathbf{cm^{-1}}$ and a resolution of 2 $\mathbf{cm^{-1}}$. Background measurements were conducted at the start of each scanning session and continued at one-hour intervals during the scanning procedure. This was to ensure that the spectral measurements were adjusted for temperature and humidity changes. About 2 g of soil from each sample was loaded into the sample cup and levelled to provide a consistent scan for all soil samples. The cup was then placed into the sample holder and carefully pressed into the DRIFT module to scan. OPUS Base package software provided by Bruker (2021) was used to integrate with the spectrometer and automatically scanned each soil sample 36 times to obtain spectral data for each sample. Soil spectra obtained from the OPUS software were then converted from the .opus file format to comma-separated values (.csv) using Spectrograph 1.2 software which enables easy to use graphical user interface (GUI) to convert the spectra to a usable format. Spectrograph 1.2 was also used to inspect the individual spectra of each sample to ensure that all spectral data with any anomalies were removed or corrected before conversion.

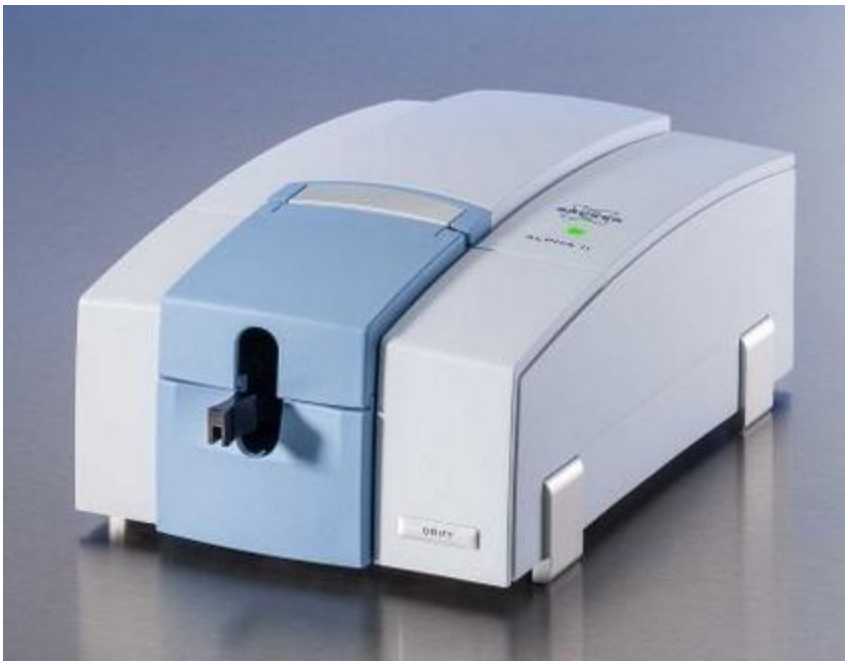

Figure 3.4.1:Bruker Alpha II with FT-IR DRIFT module attached (Bruker,2021).

3.5 Soil spectral library

3.5.1 Calibration and prediction models for pH, P and ECEC

Both soil property and spectral datasets were successfully combined using R (R Core Team, 2020), running in R studio and used to create calibration and prediction models to predict soil chemical properties using soil spectral data. The spectral library was used to develop pH, P and ECEC calibration models from which the chosen soil properties can be predicted from soil spectra, using three different approaches, namely Cubist, PLSR and RF. The following functions were used in the R program: for Cubist, the $cubist$ function in the $Cubist$ package (Kuhn, 2021) was used; for PLSR the $plsr$ function included in the pls package (Mevik and Wherens, 2015) was used and the $randomforest$ function from the $randomForest$ package (Breiman and Cutler, 2018) was used for RF. The soil spectral library was divided into calibration and validation datasets set with a 70:30 ratio respectively, using the inbuilt $sample$ function of the R programming language (Malone *et al.*, 2017).The model calibrations were performed in two steps.

Firstly, the raw spectral data with no pre-processing, followed by calibration algorithms with spectral data that underwent pre-processing treatment.

Pre-processing included spectral trimming to remove noise from the spectra, applying a SG filter and using the first derivative of the spectral data (Wadoux *et al.*, 2021). This was done for the Cubist, PLSR and RF methods. In order to reduce overfitting when using the PLSR algorithm, a pre-processing technique was used to determine the optimal number of PLSR factors (principal components) for the calibration samples alone (Wadoux *et al.*, 2021). This was performed using the $plsr$ function from the pls package created by Mevik and Wehrens (2015).

3.6 Prediction model validation

Each of the calibration algorithms were validated by calculating the following statistical parameters for the independent validation dataset as well as the prediction model calibration: r^2 (Wadoux *et al.*, 2021), RMSE, standard deviation, RPD (Wadoux *et al.*, 2021), coefficient of variability (CV) and bias (Dangal *et al.*, 2019). More detail and how the calculations are performed is given in Chapter 2. Based on these parameters the algorithms were evaluated to determine how well they represent the real data.

CHAPTER 4 RESULTS AND DISCUSSION

This section focusses on the results that were obtained as the outcomes of this study which included: the creation of a soil property database, creation of a soil spectral database with spectral data obtained from the samples from the soil property database and finally the creation of a soil spectral library which is a combination of the previously mentioned databases, and with accompanying calibration and prediction models. Each soil property prediction model was validated with an independent dataset.

4.1 Soil property database

The descriptive statistics for the 4 568 soil samples included in the soil property database (SPD) obtained from NWK and GWK for the three soil properties used for calibration in this study are given in Table 4.1.1.

Table 4.1.1: Basic descriptive statistic calculation results for the initial 4 567 soil sample database from which a final soil property database was constructed.

	Mean	**Median**	**σ**	**Min**	**Max**	**Range**	**Count**
pH (KCl)	5.08	4.97	0.79	3.54	7.75	4.21	4567
P	25.2	19	21.21	0.83	255.68	254.85	4567
ECEC	3.58	3.12	1.96	0.93	33.06	32.13	4392

σ = Standard deviation, Min = Minimum, Max = Maximum. Units for Phosphorus = P $(\mathrm{mg.kg^{-1}})$ and Effective cation exchange capacity = ECEC $(\mathrm{cmol(+).kg^{-1}})$

The average pH of the soil samples was calculated at 5.08, with minimum and maximum values of 3.54 and 7.75, respectively. Therefore, the pH ranged from very strongly acidic to slightly alkaline (Fertilizer Society of South Africa, 2016; Hazelton and Murphy, 2017). Most pH values falls within the recommended pH(KCl) range for *Zea Mays L.* of between 4.5 and 8 according to the Fertilizer Society of South Africa (2016) or between 5.5 and 7 according to Hazelton and Murphy (2017) with some values at the lower end being below the optimal values.

The P (Bray1) values of all the soil samples has an average value of 25.20 $\mathrm{mg.kg^{-1}}$ with a minimum of 0.83 $\mathrm{mg.kg^{-1}}$, maximum value of 255.68 $\mathrm{mg.kg^{-1}}$and standard deviation of 21.21 $\mathrm{mg.kg^{-1}}$. The average P (Bray1) value falls into the medium to high bracket specified for *Zea Mays L.* and is regarded as a normal P level (The Fertilizer Society of South Africa, 2007). Many samples exceed the higher limit (>30 $\mathrm{mg.kg^{-1}}$ for maize) and indicates over fertilization, which is common for cultivated soils.

Effective CEC had an average value of 3.58 $\mathrm{cmol(+).kg^{-1}}$ which is regarded as a soil with a low CEC rating (< 5 $\mathrm{cmol(+).kg^{-1}}$) (Hazelton and Murphy, 2017). The standard deviation is 1.96 $\mathrm{cmol(+).kg^{-1}}$and the minimum and maximum was 0.93 $\mathrm{cmol(+).kg^{-1}}$and 33.06 $\mathrm{cmol(+).kg^{-1}}$respectively.

4.2 Soil spectral database

4.2.1 Soil laboratory measurements

Descriptive statistics performed on the selected 948 samples included into the soil spectral database (SSD) are presented in Table 4.2.1: which contains the same statistical parameters as for the 4 568 SPD sample set. Conditioned Latin Hypercube sampling was used to select the 948 samples from the SPD.

Table 4.2.1: Basic descriptive statistic calculation results for the final 948 soil samples selected for the soil property database.

	Mean	Median	σ	Min	Max	Range	Count
pH (KCl)	5.2	5.12	0.79	3.68	7.62	3.94	948
P	25.16	19	22.59	0.83	255.68	254.9	948
ECEC	3.63	3.17	1.8	1.02	13.09	12.09	853

σ = Standard deviation, Min = Minimum, Max = Maximum. Units for Phosphorus = P $(\mathrm{mg.kg^{-1}})$ and Effective cation exchange capacity = ECEC $(\mathrm{cmol(+).kg^{-1}})$

The final dataset had an average pH value of 5.20 and minimum of 3.68, maximum of 7.62, a range of 3.94 with a standard deviation of 0.79. These results for pH were almost matching to the initial dataset provided. Box-and-whiskers plots for pH values in the SPD and the SSD are displayed in Figure **4.2**.**1** and clearly shows that the selection of pH values for the SSD almost perfectly represented the properties and range of the pH values from the SPD.

Calculations preformed on the SSD for P produced an average value of 25.61 $\mathrm{mg.kg^{-1}}$ with a minimum and maximum of 0.83 and 255.68, respectively, a range of 254.9 and a standard deviation of 22.59 $\mathrm{mg.kg^{-1}}$. These results are also very close to the values calculated for the SPD. When comparing the box-and-whisker plots for SPD and SSD in Figure **4.2**.**2** the SSD plot is almost identical to the SPD plot for P and no difference can be distinguished.

Only 853 ECEC samples were used for the soil property database as GWK do not measure or calculate T values for their routine soil analysis. The average ECEC was 3.63 $\mathrm{cmol(+).kg^{-1}}$ with a minimum and maximum of 1.02 $\mathrm{cmol(+).kg^{-1}}$ and 13.09 $\mathrm{cmol(+).kg^{-1}}$respectively, a range of 12.09 while the standard deviation of the dataset is 1.80. The major difference between the SPD and the SSD for ECEC is the reduction in range from 2.13 to 12.09. This was caused by the cLHS method and the minimum value increased with 0.09 $\mathrm{cmol(+).kg^{-1}}$, which is a trivial increase while the maximum was reduced with 19.97 $\mathrm{cmol(+).kg^{-1}}$. This reduction in range is well presented in the box-and-whiskers plot for ECEC in Figure **4.2**.**3**. The plot also indicates that all other features of the dataset are almost identical to that of the SPD.

4.2.2 Soil mid infrared spectra

The mid-infrared spectra of the 949 samples scanned as described in Chapter 3 mostly exhibit the same spectral features compared to one another as seen in Figure 4.2.4. Strong spectral clusters of peaks occur at regions between 3 700 – 3 000 $\mathrm{cm^{-1}}$, 2 400 – 2 100 $\mathrm{cm^{-1}}$ and 2 000 – 400 $\mathrm{cm^{-1}}$ respectively.

While scanning minor variations were observed at the region 2 400 – 2 300 $\mathrm{cm^{-1}}$ with deep valleys becoming progressively clearer as scanning progressed (Figure **4.2**.**5**). This is due to moisture in the air saturating the instrument as well as the silicon moisture traps, therefore increasing the

humidity in the scanning and detection chamber of the spectrometer. The anomalies were easily rectified in scans done after they were observed by completely drying the silicone moisture traps located within the detection chamber of the Bruker MIR instrument by baking them in an oven at 105 °C degrees for four hours, as specified by Bruker. However, as this does not influence the soil properties investigated in this study, and due to limited access to the Bruker MIR instrument, the affected samples were not rescanned.

Some samples were scanned at a shorter spectral range of 4 000 – 600 cm^{-1}, caused by transferring software to a different computer during the scanning process. The spectra scanned at this range was removed from the spectral dataset to keep a constant overall spectral range of 4 000 – 400 cm^{-1}.

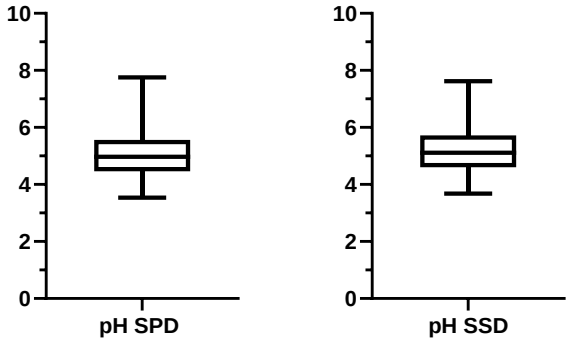

Figure 4.2.1: Box-and-whisker plots for pH values of the soil property database (SPD) and the soil spectral database (SSD) selected with conditioned Latin Hypercube sampling (cLHS).

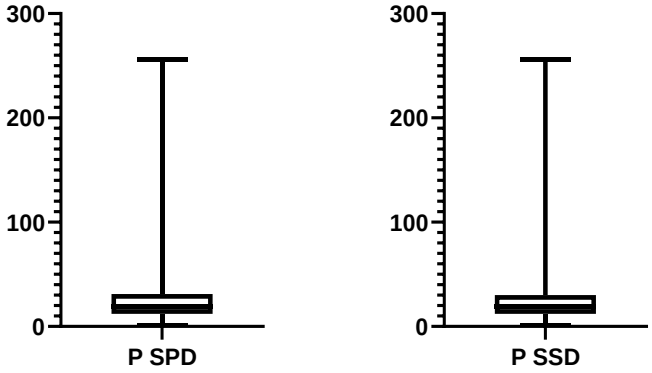

Figure 4.2.2: Box-and-whisker plots for P(Bray-1) values of the soil property database (SPD) and the soil spectral database (SSD) selected with conditioned Latin Hypercube sampling (cLHS).

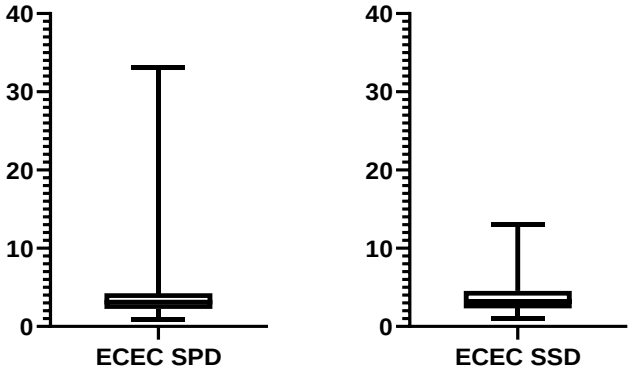

Figure 4.2.3: Box-and-whisker plots for ECEC values of the soil property database (SPD) and the soil spectral database (SSD) selected with conditioned Latin Hypercube sampling (cLHS).

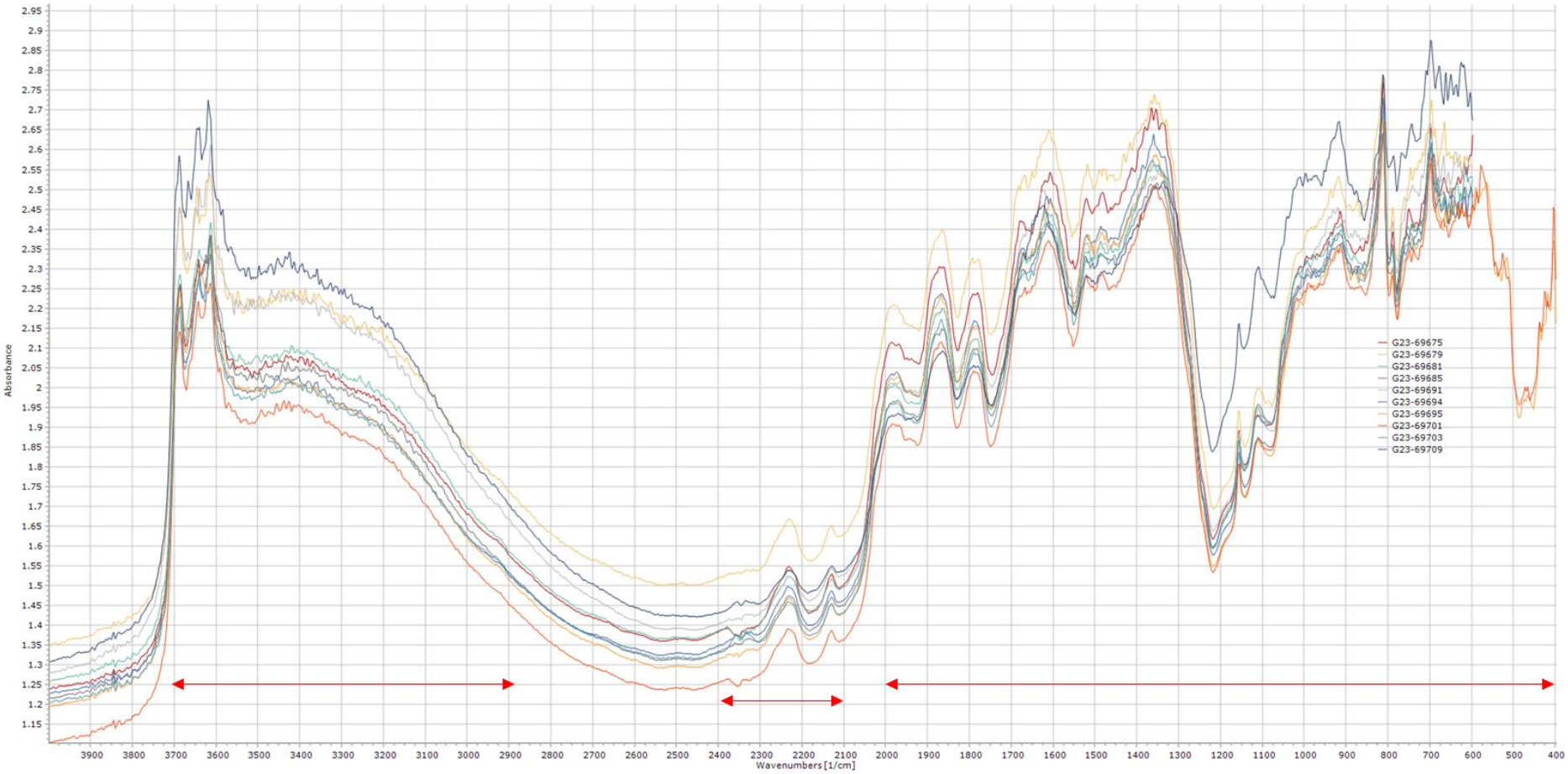

Figure 4.2.4: A selection of 10 spectra obtained from scanning the soil samples using the Bruker Alpha II with MIR DRIFT Module and red arrows indicating the major regions that contain important spectral information about the soil samples.

Figure 4.2.5: Six spectra selected from the spectra dataset scanned using the Bruker Alpha II with MIR DRIFT Module indicating variation between 2 400 – 2 300 $\mathbf{cm^{-1}}$ (indicated with a red circle).

4.3 Soil spectral library

The soil spectral library was created by combing the soil property database created in **Error! Reference source not found.** and the soil spectral database in 4.2. A total of 949 samples with soil property data were merged with the corresponding soil spectral data measured in Chapter 3.

4.4 Model creation

Two separate sets of Cubist, PLSR and RF calibration models were successfully created for each soil property one using unprocessed and the other using pre-processed spectral data from the soil spectral library. This yielded a total of 18 individual models (three soil properties, three machine learning algorithms and two sets of data), as described in Chapter 3. The calibrated models were then used to make soil property predictions with the test dataset also described in Chapter 3 as well as the training data used to calibrate the model, to validate and calculate the statistical performance parameters for comparison.

The calibration and validation results for pH, P and ECEC models with unprocessed and pre-processed are given in Table **4.4.1** – 4.4.3, along with scatterplots in Figure 4.4.1 – 4.4.6.

Table 4.4.1: The calibration and validation results for pH unprocessed and pre-processed spectral data from Cubist, PLSR and RF calibration models.

	Calibration					**Validation**				
					pH unprocessed					
Model	***r^2***	***RMSE***	***CV %***	***RPD***	***Bias***	***r^2***	***RMSE***	***CV %***	***RPD***	***Bias***
Cubist	0.90	0.25	4.81	3.16	-0.0128	0.85	0.31	5.96	2.56	0.0057
PLSR	0.91	0.24	4.62	3.29	1.0263	0.83	0.34	6.54	2.38	0.0391
RF	0.97	0.22	4.23	3.6	-0.0022	0.61	0.51	9.81	1.55	-0.0046
					pH pre-processed					
Cubist	0.91	0.24	4.62	3.29	-0.0107	0.86	0.3	5.77	2.66	-0.0054
PLSR	0.88	0.27	5.19	2.92	1.6689	0.84	0.33	6.35	2.43	0.0433
RF	0.97	0.17	3.27	4.64	-3.9330	0.74	0.42	8.08	1.88	-0.0187

Coefficient of determination = r^2, root mean square error = RMSE, coefficient of variability = CV, ration of performance to deviation = RPD, partial least square regression = PLSR and random forest = RF

Table 4.4.2: The calibration and validation results for P unprocessed and pre-processed spectral data from Cubist, PLSR and RF calibration models.

	Calibration					Validation				
					P unprocessed					
Model	***r^2***	***RMSE***	***CV %***	***RPD***	***Bias***	***r^2***	***RMSE***	***CV %***	***RPD***	***Bias***
Cubist	0.81	11.13	44.24	2.07	-1.5167	0.43	15.32	60.81	1.33	-0.0613
PLSR	0.66	13.42	53.34	1.72	1.0263	0.42	16.28	64.71	1.25	1.458
RF	0.93	8.13	32.31	2.84	0.2419	0.53	14.32	56.92	1.43	1.779
					P pre-processed					
Cubist	0.11	20.52	80.56	0.23	96.1772	0.04	22.96	91.26	0.89	4.9004
PLSR	0.54	15.62	62.08	1.48	-1.5509	0.51	14.55	57.83	1.4	0.5984
RF	0.94	6.82	27.11	3.39	1.7765	0.57	13.48	53.58	1.51	1.7765

Coefficient of determination = r^2, root mean square error = RMSE, coefficient of variability = CV, ration of performance to deviation = RPD, partial least square regression = PLSR and random forest = RF

Table 4.4.3: The calibration and validation results for ECEC unprocessed and pre-processed spectral data from Cubist, PLSR and RF calibration models.

	Calibration					Validation				
					ECEC unprocessed					
Model	***r^2***	***RMSE***	***CV %***	***RPD***	***Bias***	***r^2***	***RMSE***	***CV %***	***RPD***	***Bias***
Cubist	0.91	0.55	15.15	3.16	-0.0035	0.86	0.81	22.31	2.45	-0.1988
PLSR	0.87	0.64	17.63	2.72	7.9547	0.80	0.91	25.07	2.21	-0.1081
RF	0.97	0.39	10.74	4.59	-0.0069	0.73	0.99	27.27	1.88	-0.0372
					ECEC pre-processed					
Cubist	0.92	0.25	6.89	3.2	-0.0136	0.86	0.3	8.26	2.66	-0.0054
PLSR	0.85	0.68	18.73	2.56	-2.6263	0.81	0.88	24.24	2.28	-0.067
RF	0.97	0.30	8.26	5.97	4.5813	0.85	0.72	19.83	2.56	0.0116

Coefficient of determination = r^2, root mean square error = RMSE, coefficient of variability = CV, ration of performance to deviation = RPD, partial least square regression = PLSR and random forest = RF

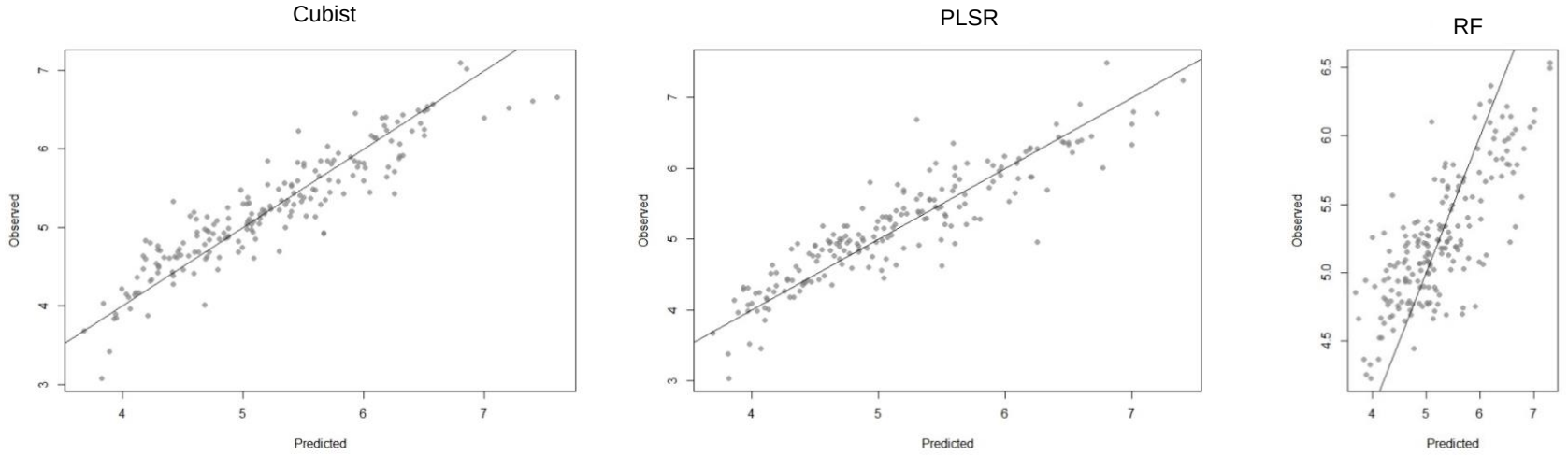

Figure 4.4.1: Scatter plots for validation model predictions for soil pH from the Cubist, PLSR and RF calibration models respectively using unprocessed spectral data. The black line represents the 1:1

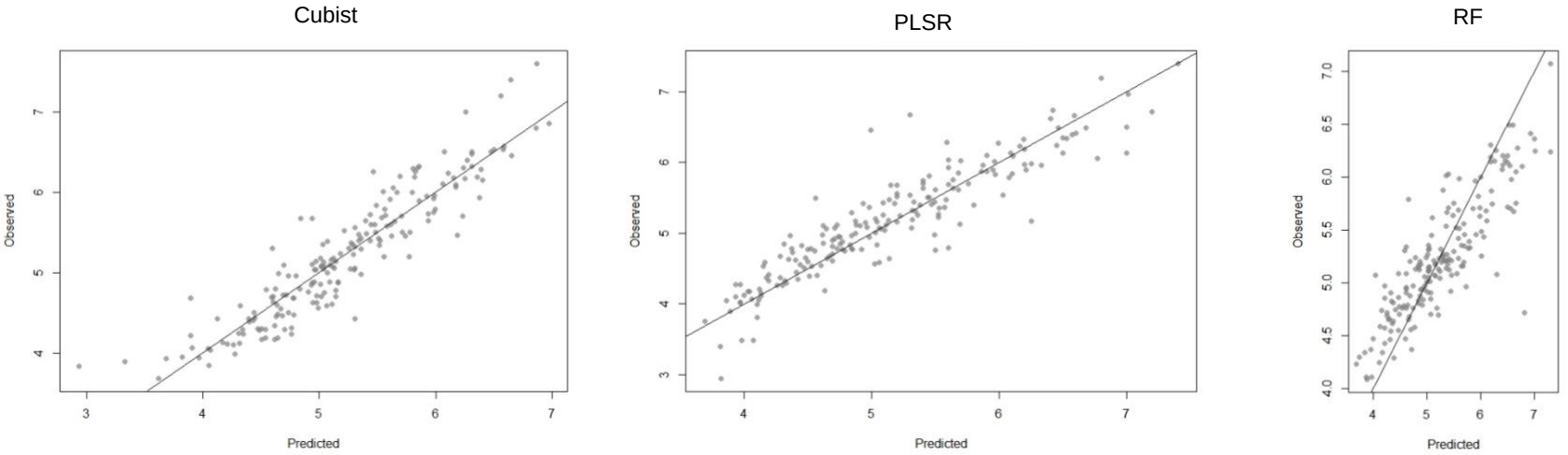

Figure 4.4.2: Scatter plots for validation model predictions for soil pH from the Cubist, PLSR and RF respectively using pre-processed spectral data. The black line represents the 1:1

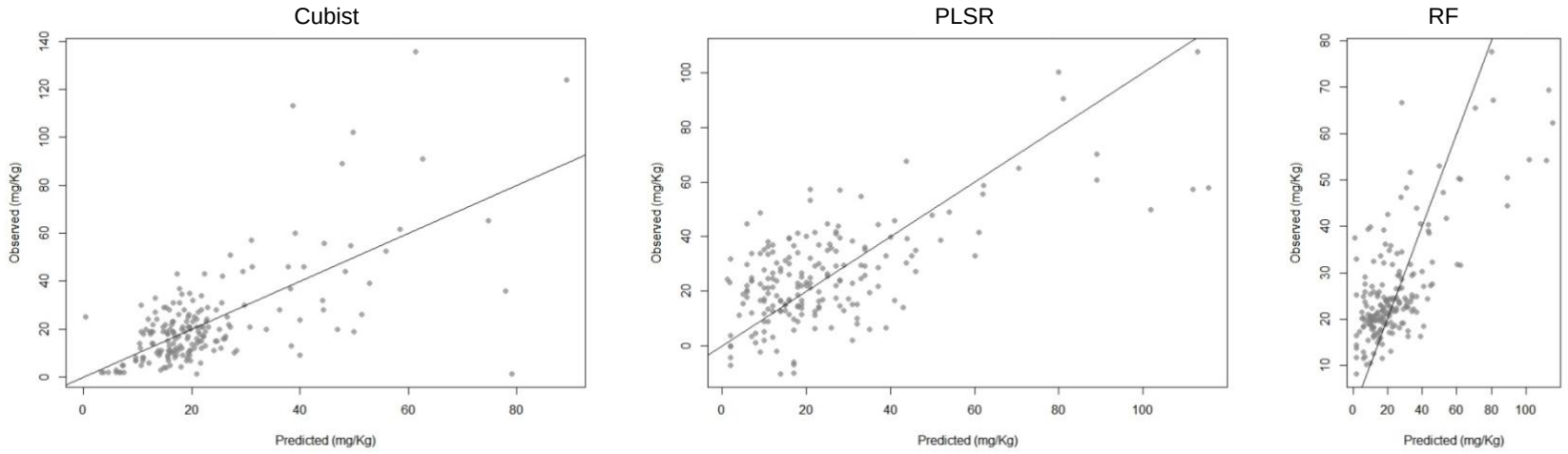

Figure 4.4.3: Scatter plots for validation models predictions for soil P from the Cubist, PLSR and RF respectively using unprocessed spectral data. The black line represents the 1:1

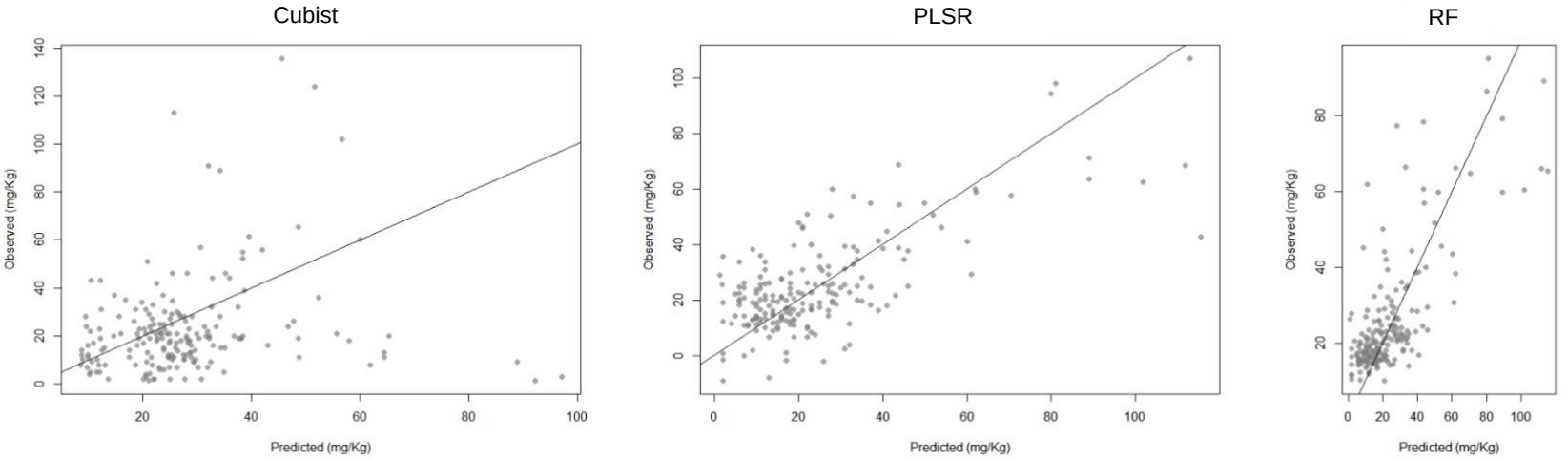

Figure 4.4.4: Scatter plots for validation model predictions of soil P from the Cubist, PLSR and RF respectively using pre-processed spectral data. The black line represents the 1:1

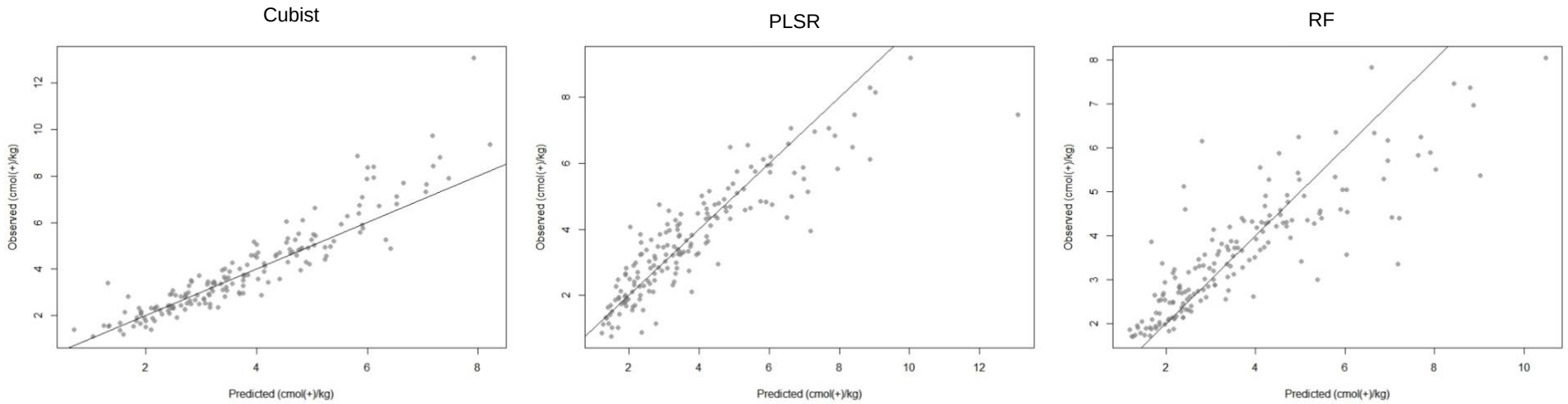

Figure 4.4.5: Scatter plots for validation model predictions for soil ECEC from the Cubist, PLSR and RF respectively using unprocessed spectral data. The black line represents the 1:1

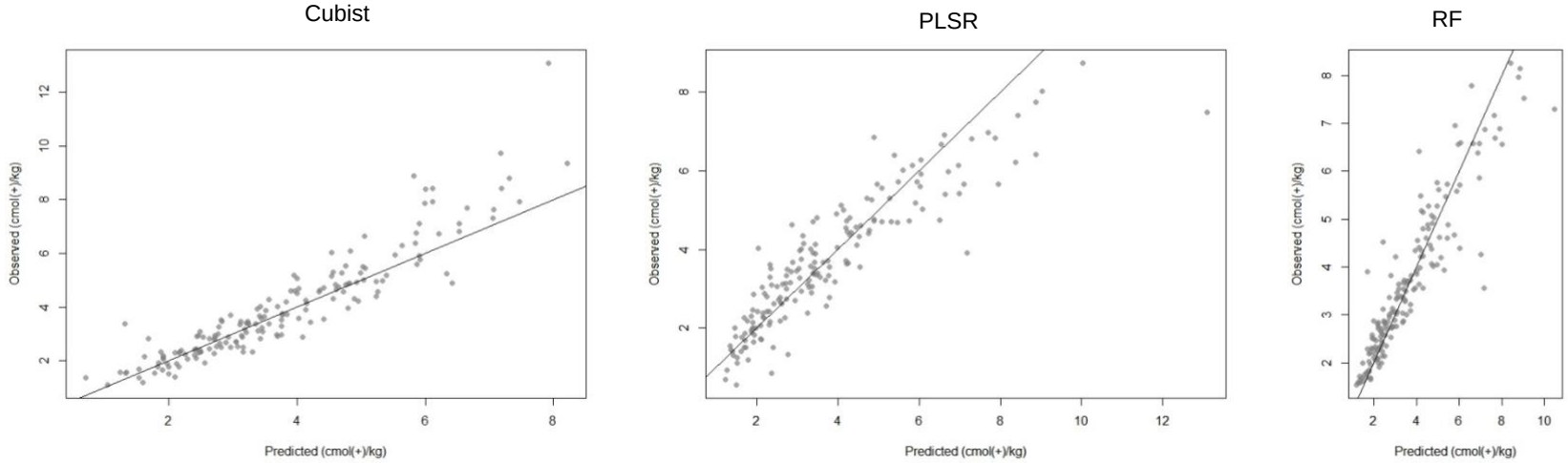

Figure 4.4.6: Scatter plots for validation model predictions for soil ECEC from the Cubist, PLSR and RF respectively using pre-processed spectral data. The black line represents the 1:1

4.4.1 pH

Results show that the Cubist model was the best performing prediction model for pH with the highest r^2 value of 0.86 and lowest RMSE of 0.30 and lowest CV of 5.77% for pre-processed spectral data, this result is shown in Table **4.4**.**1** and highlighted in grey. The RPD value of 2.66 for the Cubist model also indicated that Cubist was the most reliable model for predicting pH from pre-processed spectral data, this model is graded under category A according to the RPD ranges set out by Dangal *et al.* (2019) presented in Chapter 2. The Cubist model created with unprocessed spectral data was the second-best prediction model for pH with slightly inferior values. A lower value of 0.01 was calculated for r^2, increase of 0.01 for RMSE, 0.19 % increase for CV and 0.1 decrease for RPD.

Both PLSR models for pH using pre-processed and unprocessed spectra had similar results with r^2 values of 0.84 and 0.83, RMSE of 0.33 and 0.34, CV of 6.35% and 6.54% as well as RPD values of 2.43 and 2.38, respectively. Both the Cubist and PLSR models changed in performance in prediction validation due to pre-processing was negligible. Comparison of the Cubist and PLSR calibration models using the calibration statistics show that the PLSR model with pre-processing had the smallest difference between calibration and validation performance results compared to all other models. This indicates that while Cubist performed the best, PLSR mentioned above resulted in a prediction model that was better fitted for the pH dataset with minimal overfitting taking place. Scatter plots for the Cubist and PLSR models in Figure 4.4.1 and Figure 4.4.2 show little difference when comparing the unprocessed and pre-processed models, as expected from the above-mentioned results the Cubist model had a greater concentration of predictions around the 1:1 line compared to the PLSR models, attributed to the lower RMSE values in both the Cubist models compared to the RMSE values of the PLSR models.

The RF model had the most inferior predictions for pH compared to the Cubist and PLSR models. The best RF predictions were achieved for pre-processed spectral data with r^2 = 0.74, RMSE = 0.42, CV = 8.08% and RPD of 1.88. The RF model that used unprocessed spectral data had 0.13 lower values for r^2 , a 0.11 higher RMSE value and a lower RPD value of 1.73. Unlike Cubist and PLSR models, RF showed more improvement with the use of pre-processing when calibrating for pH. Figure **4.2**.**1** and Figure **4.2**.**2** shows the effects of the increased RMSE of the RF models for both pre-processing and unprocessed data, predictions are distributed further from the 1:1 line indicating far less accurate predictions.

According to the Soil and Plant Analysis Council (2000), the necessary precision to which pH is read for routine soil analysis is a unit of 0.1. This is 0.2 units lower than the best performing model mentioned previously. Cubist and PLSR models for both raw and pre-processed data showed similar results as for pH predictions made with PLSR in Wijewardane *et al.* (2018) and Janik *et*

al. (2009) with MIR (r^2 > 0.80 and RMSE < 0.5). A study conducted by Dangal *et al.* (2019) proves that the Cubist model is able to produce even better pH predictions. Results for the Cubist model from Dangal *et al.* (2019) in Table 2.7.1 (r^2 = 0.88, RMSE = 0.36 and RPD = 2.9) outperformed the Cubist, PLSR and RF models produced in this study. Partial least square regression and the RF models created by Dangal *et al.* (2019) also performed well with RF following Cubist as best prediction models trailed by the PLSR model which had a r^2 value of 0.74, RMSE of 0.54 and RPD of 1.91 which had lower performance compared to both PLSR models created for pH in this study. A self-adapted partial least squares model (SAM–PLS) using 1 456 samples from Nanjing, China resulted in a better performing model than Dangal *et al.* (2019) and all other studies mentioned with r^2 = 0.89, RMSE = 0.32 and RPD = 3.03 (Ma *et al.*, 2019). This study and all other studies mentioned ultimately proves that pH can be constantly predicted with similar variance using the Cubist, RF, PLSR or SAM–PLS calibration models.

4.4.2 Phosphorus

Phosphorus calibration models performed the worst. Calibration model results for P in

Table **4.4.2** show that both unprocessed and pre-processed model calibrations for P had lower r^2 values compared to other soil property calibrations in Table **4.4.1** and Table **4.4.3**. Models seemingly trained or calibrated well for the Cubist model (r^2 = 0.81, RMSE = 11.13, CV = 44.24%, RPD = 2.07, bias = -1.5167) and RF (r^2 = 0.93, RMSE = 8.13, CV = 32.31%, RPD = 2.84, bias = 0.2419) for unprocessed spectra as well as the RF model with pre-processed spectra (r^2 = 0.94, RMSE = 6.82, CV = 27.11%, RPD = 3.39, bias = 1.7765), but failed to produce any acceptable models when validated with a test dataset for prediction. The best performing prediction model was the RF model with pre-processed spectra (r^2 = 0.57, RMSE = 13.48, CV = 53.58%, RPD = 1.51, bias = 1.7765) and is graded as a category B model followed by the RF model for unprocessed spectra (r^2 = 0.53, RMSE = 14.32, CV = 56.92%, RPD = 1.43, bias = 1.779), the PLSR model with pre-processing (r^2 = 0.51, RMSE = 14.55, CV = 57.83%, RPD = 1.4, bias = 0.5984), the Cubist model with no pre-processing (r^2 = 0.43, RMSE = 15.32, CV = 60.81%, RPD = 1.33, bias = -0.0613), the PLSR model with no processing (r^2 = 0.42, RMSE = 16.28, CV = 64.71%, RPD = 1.25, bias = 1.458) and finally the Cubist model with pre-processed spectra (r^2 = 0.04, RMSE = 22.96, CV = 91.26%, RPD = 0.89, bias = 4.9004) that produced results that were lower than expected. This decrease in performance from calibration to validation, especially for the RF model indicates extreme overfitting for predicting the P values from the test dataset.

None of the P validation models using either unprocessed or pre-processing spectral data performed well enough to be accepted for soil P prediction according to the RPD ranges set out

by Dangal *et al.* (2019) in Chapter 2. The standard precision for P (Bray-1) measured in a laboratory given by the Soil and Plant Analysis Council (2000), results in a CV of 5% which is well below the lowest mentioned for the RF model for pre-processed spectra. These results are in-line with findings from Soriano-Disla *et al.* (2014) which showed that extractable P for MIR had low r^2 values of 0.5 – 0.7. Wijewardane *et al.* (2018) also included P for MIR spectroscopy and found similar results with r^2 = 0.14 and RMSE = 45.17 for P values ranging from 0 $mg.kg^{-1}$ to 595 $mg.kg^{-1}$. All scatter plots in

Figure 4.4.3 and Figure 4.4.4 show a cluster of predictions that are close to the 1:1 within the 0 – 40 $mg.kg^{-1}$ P range and a greater distribution as the P increases from 40 $mg.kg^{-1}$ to > 100 mg/Kg. This observation should indicate that better predictions are expected if a validation set of soils with P values < 40 $mg.kg^{-1}$ are used. This however could be misleading as P is determined form an extraction method whereas the MIR measures spectra from a soil matrix (Soriano-Disla *et al.*, 2014). Better predictions for P have been achieved with MIR when determining P sorption (r^2 = 0.83) and total P (r^2 = 0.58) but with almost no studies providing data (Soriano-Disla *et al.*, 2014).

4.4.3 Effective CEC

The Cubist model using pre-processing was the best performing model for ECEC, which is also a category A model (r^2 = 0.86, RMSE = 0.3, CV = 8.26%, RPD of 2.66, bias = -0.0054) followed by the Cubist model using unprocessed spectra (r^2 = 0.86, RMSE = 0.81, CV = 22.31%, RPD of 2.45, bias = -0.1988), the RF model with pre-processing (r^2 = 0.85, RMSE = 0.72, CV = 19.83%, RPD of 2.56, bias = 0.0116), the PLSR model with pre-processing (r^2 = 0.81, RMSE = 0.88, CV = 24.24%, RPD of 2.28, bias = -0.067), the PLSR model with unprocessed spectra (r^2 = 0.80, RMSE = 0.91, CV = 25.07%, RPD of 2.21, bias = -0.1081) and the RF model with unprocessed spectra (r^2 = 0.73, RMSE = 0.99, CV = 27.27%, RPD of 1.88, bias = - 0.0372). These results did not perform as well as the calibration models created by Shepherd and Walsh, (2002) which produced a r^2 value of 0.95 and RMSE of 2.6 for a MARS model, but did produce prediction models with lower RMSE values. The use of pre-processed spectral data for calibration reduced the RMSE and CV values and in return improved on the RPD, this is also observed by the scatter plot of the predicted vs observed values in Figure 4.4.5 and Figure 4.4.6 for the Cubist model which shows an overall narrower concentration of data around the 1:1 resulting from lower RMSE. The scatter plot for the RF model with pre-processed spectra shows that the model has very good predictions with a highly concentrated distribution of prediction points on the 1:1 within the range 0 $cmol(+).kg^{-1}$ to 4 $cmol(+).kg^{-1}$. This shows that the specific model should produce the most

accurate predictions if ECEC values are within that range of prediction compared to all other models even with the inferior performance statistical results. This could be explained by the decrease in range of the SSD from that of the SPD when cLHS was used for ECEC.

The PLSR model also showed that spectral pre-processing reduces RMSE, CV as well as bias values. The PLSR calibration for both unprocessed and pre-processed spectral data showed similar results for the models during validation, unlike that of the Cubist and RF model which showed a significant reduction in performance calculations. While the PLSR model seemingly did not perform as well as the Cubist and RF models, Janik *et al.* (1998) and Wijewardane *et al.* (2018) shows that PLSR is able to produce good results with r^2 values of 0.88 and 0.86 respectively. Laboratory determination of ECEC relies on the precision and accuracy at which the cations are extracted using an ammonium acetate solution. This method of extraction from which ECEC is calculated will produce coefficients of variations of 5 – 10% (Soil and Plant Analysis Council, 2000). This criterion allows for the Cubist model that used pre-processed spectral data to be used for the prediction of ECEC values that are within the same margins as for laboratory ECEC analysis.

4.5 Conclusion

All results obtained from this study were presented in this chapter for each soil property calibration model and validation model. Results for pH, ECEC and P were evaluated and compared with statistical performance parameters and included. It was found that certain models performed better for each soil property. According to the results, a conclusion can be made that pre-processing is an essential step for soil spectroscopy although the improvement is slight for certain soil properties and calibration algorithms.

CHAPTER 5 CONCLUSION AND RECCOMENDATIONS

This research study aimed to create MIR soil property specific calibration models using the Cubist, PLSR and RF calibration models for soil pH, P and ECEC for soils in the Western Highveld region of South Africa. Results presented from this study indicated that different calibration algorithms performed differently for each soil property when trained and used for soil property predictions. The first and second objective of the study was met successfully with the creation of a SPD with 4 567 soil samples and a SSD with 948 samples that was created using cLHS and combined with spectral data and calibration models to produce a soil spectral library. Eighteen calibration models in total were successfully calibrated and validated to evaluate the performance of the models with both unprocessed and pre-processed spectral data to accomplish the third and final objective of the study.

The soil spectral library contains a Cubist calibrated model with pre-processed spectra that was the best performing algorithm for predicting soil pH as well as a Cubist model calibrated with pre-processes spectra that can accurately predict ECEC soil values. A RF model with pre-processing was the best performing model for extractable P but did not provide accurate enough P values during validation to be used as an extractable P prediction model. The study hypothesis was accepted for both pH and ECEC but rejected for P. Both prediction models for pH and ECEC produced accurate predictions that fall within the laboratory accuracy ranges mentioned for pH and ECEC respectively. Results showed that pre-processing is an important step before model calibration and will produce the best calibration and prediction models. Calibration and prediction models produced similar and comparable results to that of other studies mentioned in the paper for all three soil properties and calibration algorithms. The best prediction models for soil pH and ECEC created in this study were well calibrated with low RMSE values compared to the models presented for pH and ECEC from other studies.

The MIR spectroscopic region is very well suited for determining soil pH and ECEC due to the relationships of these soil properties to soil OC and even soil texture absorptions and reflections that are strongly represented in the MIR spectral range. Other forms of soil P like total P, absorbed P and available P also mentioned in the paper that were used in other studies can also be accurately predicted with MIR prediction models.

This study used a relatively small sample size in the soil spectral database to calibrate the respective prediction models for each soil property. This resulted in slightly lower r^2 values than expected when calibrated for pH and ECEC. Other studies that used larger sample sizes or large spectral MIR libraries (Dangal *et al.*, 2019) produced better r^2 values for their respective prediction models. Increasing the sample size could further improve the prediction model calibrations of this study. It is also recommended that other statistical calibration methods be used for MIR soil spectroscopy including Artificial Neural Network (ANN) and Machine Based Learning (MBL) algorithms.

The conducted study only included soil samples that originated from areas within the North West Province, South Africa. To expect acceptable results when using these calibration models on soil samples that originate from outside the study area, additional samples must be included from different regions of the country and added to the soil spectral library. Additional soil properties are also required for routine soil analysis and new calibration models need to be created. Only after the successful calibration of prediction models with sufficient accuracy can MIR soil spectroscopy be used as a standard soil analysis method.

BIBLIOGRAPHY

Abdi, D., Tremblay, G.F., Ziadi, N., Bélanger, G., and Parent, L.-É. 2012. Predicting Soil Phosphorus-Related Properties Using Near-Infrared Reflectance Spectroscopy. *Soil Science Society of America Journal*. 76(6):2318–2326.

Ali, T. and Narayan, S. 2009. Manual of Soil, Plant and Water Analysis. Daya Publishing House: Delhi.

Atkins, P. and de Paula, J. 2010. Atkins' physical chemistry. 9th ed. Oxford University Press: London, England.

Barnard, R.O. and du Preez, C.C. 2004. Soil fertility in South Africa: The last twenty five years. *South African Journal of Plant and Soil*. 21(5):301–315.

Weil, R. and Brady, N. 2017. The Nature and Properties of Soils. 15th ed. Pearson Education: Essex.

Bramley, R.G.V. and Janik, L.J. 2005. Precision agriculture demands a new approach to soil and plant sampling and analysis - Examples from Australia. *Communications in Soil Science and Plant Analysis*. 36(1–3):9–22.

Breiman, L. 2001. Random Forests. *Machine Learning*. 45(1):5–32.

Breiman, L. and Cutler, A. 2018. Breiman and Cutler's Random Forests for Classification and Regression. *CRAN*.

Bruker. 2021. https://www.bruker.com/en/products-and-solutions/infrared-and-raman/ft-ir-routine-spectrometer/alpha-ii-compact-ft-ir-spectrometer.html

Buddenbaum, H. and Steffens, M. 2012. The effects of spectral pretreatments on chemometric analyses of soil profiles using laboratory imaging spectroscopy. *Applied and Environmental Soil Science*. 2012.

Cambule, A.H., Rossiter, D.G., Stoorvogel, J.J., and Smaling, E.M.A. 2012. Building a near

infrared spectral library for soil organic carbon estimation in the Limpopo National Park, Mozambique. *Geoderma*. *Elsevier B.V.* 183–184:41–48.

Cantarella, H., Quaggio, J.A., Van Raij, B., and De Abreu, M.F. 2006. Variability of soil analysis in commercial laboratories: Implications for lime and fertilizer recommendations. *Communications in Soil Science and Plant Analysis*. 37(15–20):2213–2225.

Chang, C.-W., Laird, D.A., Mausbach, M.J., and Hurburgh, C.R. 2001. Near-Infrared Reflectance Spectroscopy-Principal Components Regression Analyses of Soil Properties. *Soil Science Society of America Journal*. 65(2):480–490.

Coblinski, J.A., Giasson, É., Demattê, J.A.M., Dotto, A.C., Costa, J.J.F., and Vašát, R. 2020. Prediction of soil texture classes through different wavelength regions of reflectance spectroscopy at various soil depths. *Catena*. 189(January).

Comstock, J.P., Sherpa, S.R., Ferguson, R., Bailey, S., Beem-Miller, J.P., Lin, F., Lehmann, J., and Wolfe, D.W. 2019. Carbonate determination in soils by mid-IR spectroscopy with regional and continental scale models. *PLoS ONE*. 14(2):1–19.

Council for Geoscience. 2019. *Geological Data 1: 1 000 000*. South Africa, PTA: Council for Geoscience.

Dangal, S.R.S., Sanderman, J., Wills, S., and Ramirez-Lopez, L. 2019. Accurate and precise prediction of soil properties from a large mid-infrared spectral library. *Soil Systems*. 3(1):1–23.

Daniel, K.W., Tripathi, N.K., and Honda, K. 2003. Artificial neural network analysis of laboratory and in situ spectra for the estimation of macronutrients in soils of Lop Buri (Thailand). *Australian Journal of Soil Research*. 41(1):47–59.

Daw, O.D. and Molefe, E.K. 2017. Empirical analysis of population growth and economic performance in South Africa. *Gender & Behaviour.* 15(4):10117–10125.

DFFE (Department of Forestry, Fisheries and the Environment). 2018. South African Nation Land Cover (SANLC) maps. Pretoria: Department of Forestry, Fisheries and the Environment.

Edmeades, D.C. 1982. Effects of lime on effective cation exchange capacity and exchangeable cations on a range of New Zealand soils. *New Zealand Journal of Agricultural Research*. 25(1):27–33.

Fischer, W.R. 1977. H.W. van der Marel und H. Beutelspacher (Herausg.): Atlas of infrared spectroscopy of clay minerals and their admixtures. Elsevier, *Zeitschrift für Pflanzenernährung und Bodenkunde*. 140(2):247–248.

Hazelton, P.A. and Murphy, B.W. 2017. Interpreting Soil Test Results: What do all the Numbers mean? 3rd ed. , European Journal of Soil Science. CSIRO PUBLISHING: Clayton South.

Hobley, E.U. and Prater, I. 2019. Estimating soil texture from vis–NIR spectra. *European Journal of Soil Science*. 70(1):83–95.

Hollas, M.J. 2004. Modern Spectroscopy. 4th ed. John Wiley and Sons Ltd: West Sussex.

Janik, L.J., Forrester, S.T., and Rawson, A. 2009. The prediction of soil chemical and physical properties from mid-infrared spectroscopy and combined partial least-squares regression and neural networks (PLS-NN) analysis. *Chemometrics and Intelligent Laboratory Systems*. *Elsevier B.V.* 97(2):179–188.

Janik, L.J., Merry, R.H., and Skjemstad, J.O. 1998. Can mid infrared diffuse reflectance analysis replace soil extractions? *Australian Journal of Experimental Agriculture*. 38(7):681.

Janik, L.J., Skjemstad, J.O., Shepherd, K.D., and Spouncer, L.R. 2007. The prediction of soil carbon fractions using mid-infrared-partial least square analysis. *Australian Journal of Soil Research*. 45(2):73–81.

Jeong, G., Choi, K., Spohn, M., Park, S.J., Huwe, B., and Ließ, M. 2017. Environmental drivers of spatial patterns of topsoil nitrogen and phosphorus under monsoon conditions in a complex terrain of South Korea. *PLoS ONE*. 12(8):1–19.

Kassambara, A. 2018. Machine Learning Essentials: Practical Guide in R. 1st ed. Statistical tools for high-throughput data analysis.

Khoshhesab, Z.M. 2012. Reflectance IR Spectroscopy., Infrared Spectroscopy - Materials Science, Engineering and Technology.

Knox, N.M., Grunwald, S., McDowell, M.L., Bruland, G.L., Myers, D.B., and Harris, W.G. 2015. Modelling soil carbon fractions with visible near-infrared (VNIR) and mid-infrared (MIR) spectroscopy. *Geoderma*. *Elsevier B.V.* 239–240:229–239.

Land Type Survey Staff. 1972-2006. Land Types of South Africa: Digital Map (1: 250 000 Scale) and Soil Inventory Datasets. South Africa, PTA: Agriculture Research Council Institute for Soil, Climate and Water.

Le Roux, P.A.L., Hensley, M., van Rensburg, L.D., and Botha, J.J. 2016. The cropping potential of South Africa: land evaluation results obtained during the last 50 years. *South African Journal of Plant and Soil*. 33(2):83–88.

Lohninger, H. 2021. Fundamentals of Statistics. http://www.statistics4u.info/fundstat_eng/cc_filter_savgolay.html Date of access: 25 Aug. 2021

Ma, F., Du, C.W., Zhou, J.M., and Shen, Y.Z. 2019. Investigation of soil properties using different techniques of mid-infrared spectroscopy. *European Journal of Soil Science*. 70(1):96–106.

Maher, N.M., Richter, D.D., and Markewitz, D. 2003. Understanding Soil Change: Soil Sustainability over Millennia, Centuries, and Decades. *Environmental History*. 8(1):145.

Malherbe, J., Dieppois, B., Maluleke, P., Staden, M., and Pillay, D. 2016. South African droughts and decadal variability. *Natural Hazards*. 80.

Malone, B., Minasny, B., and Mcbratney, A.B. 2017. Progress in Soil Science Using R for Digital Soil Mapping.

MATLAB. 2010. version 7.10.0 (R2010a). The MathWorks Inc.: Natick, Massachusetts.

Kuhn, M. 2021. Package “Cubist” Title Rule-And Instance-Based Regression Modeling.

McBratney, A.B., Mendonça Santos, M.L., and Minasny, B. 2003. On digital soil mapping., *Geoderma*. 117 (1-2):3-52.

McCarty, G.W. and Reeves, J.B. 2006. Comparison of near infrared and mid infrared diffuse reflectance spectroscopy for field-scale measurement of soil fertility parameters. *Soil Science*. 171(2):94–102.

McKay, M.D., Beckman, R.J., and Conover, W.J. 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*. 21:239–245.

Metzger, K., Zhang, C., Ward, M., and Daly, K. 2020. Mid-infrared spectroscopy as an alternative to laboratory extraction for the determination of lime requirement in tillage soils. *Geoderma*. *Elsevier B.V.* 364.

Mevik, B.-H & Wehrens, R. 2015. Introduction to the pls Package. *Help section of the “pls” package of RStudio software*. (Section 7):1–23.

Mills, A.J. and Fey, M. V. 2004. Declining soil quality in South Africa: Effects of land use on soil organic matter and surface crusting. *South African Journal of Plant and Soil*. 21(5):388–398.

Minasny, B. and McBratney, A.B. 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. *Computers and Geosciences*. 32(9):1378–1388.

Minasny, B. and McBratney, A.B. 2008. Regression rules as a tool for predicting soil properties from infrared reflectance spectroscopy. *Chemometrics and Intelligent Laboratory Systems*. 94(1):72–79.

Mura, S., Cappai, C., Greppi, G.F., Barzaghi, S., Stellari, A., and Cattaneo, T.M.P. 2019. Vibrational spectroscopy and Aquaphotomics holistic approach to determine chemical compounds related to sustainability in soil profiles. *Computers and Electronics in Agriculture*. *Elsevier*. 159(March):92–96.

Nguyen, T., Janik, L., and Raupach, M. 1991. Diffuse reflectance infrared fourier transform (DRIFT) spectroscopy in soil studies. *Soil Research*. 29(1):49.

Nocita, M., Kooistra, L., Bachmann, M., Müller, A., Powell, M., and Weel, S. 2011. Predictions of soil surface and topsoil organic carbon content through the use of laboratory and field spectroscopy in the Albany Thicket Biome of Eastern Cape Province of South Africa. *Geoderma*. *Elsevier B.V.* 167–168:295–302.

Nocita, M., Stevens, A., van Wesemael, B., Brown, D.J., Shepherd, K.D., Towett, E., Vargas, R., and Montanarella, L. 2015. Soil spectroscopy: An opportunity to be seized. *Global Change Biology*. 21(1):10–11.

Peng, J., Ji, W., Ma, Z., Li, S., Chen, S., Zhou, L., and Shi, Z. 2016. Predicting total dissolved salts and soluble ion concentrations in agricultural soils using portable visible near-infrared and mid-infrared spectrometers. *Biosystems Engineering*. 152:94–103.

Peng, Y., Xiong, X., Adhikari, K., Knadel, M., Grunwald, S., and Greve, M.H. 2015. Modeling soil organic carbon at regional scale by combining multi-spectral images with laboratory spectra. *PLoS ONE*. 10(11).

Pierce, F.J. and Nowak, P. 1999. Aspects of Precision Agriculture. *Advances in Agronomy*. 67(C):1–85.

R Core Team. 2020. R: A Language and Environment for Statistical Computing. Vienna, Austria.

Retsch GmbH. 2021. https://www.retsch.com/products/milling/mortar-grinders/rm-200/function-features/

Rossel, R.A.V. and Behrens, T. 2010. Using data mining to model and interpret soil diffuse

reflectance spectra. *Geoderma*. *Elsevier B.V.* 158(1–2):46–54.

Rossel, R.A.V., Jeon, Y.S., Odeh, I.O.A., and McBratney, A.B. 2008. Using a legacy soil sample to develop a mid-IR spectral library. *Australian Journal of Soil Research*. 46(1):1–16.

Savitzky, A. and Golay, M.J.E. 1964. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry*. 36(8):1627–1639.

Scarpone, C., Schmidt, M.G., Bulmer, C.E., and Knudby, A. 2016. Modelling soil thickness in the critical zone for Southern British Columbia. *Geoderma*. *Elsevier B.V.* 282:59–69.

Schwertfeger, D.M. and Hendershot, W.H. 2009. Determination of Effective Cation Exchange Capacity and Exchange Acidity by a One-Step BaCl 2 Method. *Soil Science Society of America Journal*. 73(3):737–743.

Shahrayini, E., Noroozi, A.A., and Eghbal, M.K. 2020. Prediction of Soil Properties by Visible and Near-Infrared Reflectance Spectroscopy. *Eurasian Soil Science*. 53(12):1760–1772.

Shepherd, K.D. and Walsh, M.G. 2002. Development of Reflectance Spectral Libraries for Characterization of Soil Properties. *Soil Science Society of America Journal*. 66(3):988–998.

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):1–19.

Sleep, B., Mason, S., Janik, L., and Mosley, L. 2021. Application of visible near-infrared absorbance spectroscopy for the determination of Soil pH and liming requirements for broad-acre agriculture. *Precision Agriculture*. *Springer US*. (0123456789).

Soil and Plant Analysis Council, 2000. Soil analysis Handbook of reference methods. 4th ed. CRC Press LLC: Boca Raton.

Soriano-Disla, J.M., Janik, L.J., Viscarra Rossel, R.A., Macdonald, L.M., and McLaughlin, M.J. 2014. The Performance of Visible, Near-, and Mid-Infrared Reflectance Spectroscopy for Prediction of Soil Physical, Chemical, and Biological Properties. *Applied Spectroscopy Reviews*. *Elsevier*. 49(2):139–186.

Stuart, B.H. 2005. Infrared Spectroscopy: Fundamentals and Applications., Infrared Spectroscopy: Fundamentals and Applications.

Sun, X.L., Wang, H.L., Zhao, Y.G., Zhang, C., and Zhang, G.L. 2017. Digital soil mapping based on wavelet decomposed components of environmental covariates. *Geoderma*. *Elsevier*. 303(May):118–132.

Terhoeven-Urselmans, T., Vagen, T.-G., Spaargaren, O., and Shepherd, K.D. 2010. Prediction of Soil Fertility Properties from a Globally Distributed Soil Mid-Infrared Spectral Library. *Soil Science Society of America Journal*. 74(5):1792–1799.

The Fertilizer Society of South Africa. 2007. FSSA fertilizer handbook of South Africa. 6th ed. Pretoria.

The Non-affiliated Soil Analysis Work Commitee. 1990. Handbook of standard soil testing methods for advisory purposes. Soil Science Society of South Africa: Pretoria.

Thomas, M., Clifford, D., Bartley, R., Philip, S., Brough, D., Gregory, L., Willis, R., and Glover, M. 2015. Putting regional digital soil mapping into practice in Tropical Northern Australia. *Geoderma*. *Elsevier B.V.* 241–242:145–157.

Thompson, J.M. 2018. Infrared spectroscopy. 1st ed. Jenny Stanford Publishing.

Tilman, D., Balzer, C., Hill, J., and Befort, B.L. 2011. Global food demand and the sustainable intensification of agriculture. *Proceedings of the National Academy of Sciences of the United States of America*. 108(50):20260–20264.

United Nations. 2017. World Population Prospects 2019: Methodology of the United Nations population estimates and projections. Departmenr of Economic and Social Affairs, Population Division. :61.

USGS (United States Geological Survey). 2021. Earth Explorer: Landsat images. https://earthexplorer.usgs.gov/ Date of access: 20 Jun. 2021.

Van Groenigen, J.W., Mutters, C.S., Horwath, W.R., and Van Kessel, C. 2003. NIR and DRIFT-MIR spectrometry of soils for predicting soil and crop parameters in a flooded field. *Plant and Soil*. 250(1):155–165.

Van Vuuren, J.A.J., Meyer, J.H., and Claassens, A.S. 2006. Potential use of near infrared reflectance monitoring in precision agriculture. *Communications in Soil Science and Plant Analysis*. 37(15–20):2171–2184.

van Zijl, G.M. 2019. Digital soil mapping approaches to address real world problems in southern Africa. *Geoderma*. *Elsevier*. 337(May 2018):1301–1308.

Van Zijl, G.M., Ellis, F., and Rozanov, A. 2014. Understanding the combined effect of soil properties on gully erosion using quantile regression. *South African Journal of Plant and Soil*. 31(3):163–172.

Viscarra Rossel, R.A., Walvoort, D.J.J., McBratney, A.B., Janik, L.J., and Skjemstad, J.O. 2006. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. *Geoderma*. 131(1–2):59–75.

Wadoux, A., Malone, B., Minasny, B., Fajardo, M., and McBratney, A. 2021. Soil Spectral Inference with R: Analysing Digital Soil Spectra Using the R Programming Environment., Springer International Publishing AG, Cham.

Waruru, B.K., Shepherd, K.D., Ndegwa, G.M., Sila, A., and Kamoni, P.T. 2015. Application of mid-infrared spectroscopy for rapid characterization of key soil properties for engineering land use. *Soils and Foundations*. *Elsevier*. 55(5):1181–1195.

Wehrens, R. 2020. Chemometrics with R., Use R! Springer Berlin Heidelberg: Berlin, Heidelberg.

Wijewardane, N.K., Ge, Y., Wills, S., and Libohova, Z. 2018. Predicting Physical and Chemical Properties of US Soils with a Mid-Infrared Reflectance Spectral Library. *Soil Science Society of America Journal*. 82(3):722–731.

Wills, S., & Libohova, Z. (2018). Predicting Physical and Chemical Properties of US Soils with a Mid-Infrared Reflectance Spectral Library. *Soil Science Society of America Journal*, *82*(3), 722–731. https://doi.org/10.2136/sssaj2017.10.0361

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

Yiu, T. 2019. Understanding Random Forest. How the Algorithm Works and Why it Is https://towardsdatascience.com/understanding-random-forest-58381e0602d2 Date of access: 25 Aug 2021

ANNEXURE: METHODS USED FOR SOIL PROPERTY DETERMINATION

pH

pH is measured as pH (KCl) and 10 g of the soil sample is mixed with 25 cm^3 of KCl solution at 1 $\mathrm{mol\,dm}^{-3}$ and stirred into suspension. The mixture is then left for 50 minutes and stirred again and left to stand for 10 minutes. The electrode of the pH meter is placed into a buffer solution with known pH and used to calibrate the pH meter. After calibration the electrode is removed from the buffer solution and placed into the soil suspension which contains the soil sample, the pH is then measured and recorded as pH(KCl) (The Non-affiliated Soil Analysis Work Commitee, 1990).

Phosphorus

The soil is extracted with a solution of $\mathrm{NH_4F}$ and HCl called Bray-1 extracting solution and then treated with ammonium molybdate and the blue colour obtained is then further enhanced with stannous chloride. Colour intensity is measured with a photoelectric colorimeter at 660 nm wavelength (Ali and Narayan, 2009). For each sample 6.67g of soil is added to a 50 cm^3 Bray-1 solution and stirred for one minute after which two drops of flocculant is added. The mixture is the filtered immediately into a separate container and treated with ammonium molybdate and stannous chloride. The solution is then measured with a photoelectric colorimeter (The Non-affiliated Soil Analysis Work Commitee, 1990).

Effective Cation Exchange Capacity

Extractable cations are determined using the ammonium acetate ($\mathrm{NH_4OAc}$) method according to the Non-affiliated Soil Analysis Work Committee (1990). A 20g soil sample that was air-dried and sieved to < 2 mm is placed into a 125 cm^3 beaker with 50 cm^3 0.2 $mol\ dm^{-3}\ NH_4OAc$ solution, stirred and left for one hour. The solution was filtered using a Buchner funnel using a vacuum and the leftover soil was leached for one hour using aliquots of the $\mathrm{NH_4OAc}$ solution into 500 cm^3 volumetric flasks (the leachate should not exceed 500 cm^3) and weighed after the leaching prosses. The leachate was made up to the 500 cm^3mark with 0.2 $\mathrm{mol\,dm}^{-3}$ NH_4OAc solution and mixed. The soil contained in the funnel was leached for one hour using 50 cm^3 of 0.2 $\mathrm{mol\,dm}^{-3}$ $\mathrm{K_2SO_4}$ solution into clean 500 cm^3 volumetric flasks and the leachate were tested for $\mathrm{NH_4}^+$ using Nessler's solution. If no $\mathrm{NH_4}^+$ were detected the 0.2 $\mathrm{mol\,dm}^{-3}$ $\mathrm{K_2SO_4}$ solution can be used to

make up the volume to the 500 cm^3 mark and mixed. Sodium and K were then determined by placing 10 cm^3 of Na and K standards of 1000 $\mathrm{mg\ dm}^{-3}$ into a 100 cm^3 volumetric flask and using de-ionized water to fill up to the mark yielding a concentration of 100 $mg\ dm^{-3}$ Na and K. The solution is then further diluted with 100 cm^3 Caesium (Cs) solution with 0, 2, 4 and 8 cm^3 of the previously mentioned solution. The $\mathrm{NH_4OAc}$ leachates for the samples were then diluted with the Cs solution to guarantee a final solution with 800 $\mathrm{mg\ dm}^{-3}$ Cs (1-part leachate: 4 parts Cs) after which the Na and K concentrations were determined using an atomic absorption spectrophotometer. Calcium was determined using the same principle and 10 cm^3 of a Ca standard of 1 000 $\mathrm{mg\ dm}^{-3}$ was placed into a 100 cm^3 volumetric flask a filled to the mark with deionised water. Working standards were again made by pipetting 0, 5, 10, and 15 cm^3 of the above made solution into 100 cm^3 volumetric flasks and adding 20 cm^3 of 6 250 $\mathrm{mg\ dm}^{-3}$ Latium (La) solution and filled to the mark with 0.1 $\mathrm{mol\ dm}^{-3}$ HCl. One part of the NH_4OAc solution with four parts of the 1250 $mg\ \mathrm{dm}^{-3}$ La solution was added to get a final concertation of 1 000 $\mathrm{mg\ dm}^{-3}$ La. Calcium was then determined using the same spectrophotometer as for the Na and K determinations. Finally, Mg was determined in the same manner as for Ca but with a Mg standard solution and dilutions of 0, 2, 4, and 8 $\mathrm{mg\ dm}^{-3}$ Mg that has been prepared using the same steps for the previous cations. Again, La solution was used, and the amount of Mg was determined using the same spectrophotometer. The following equations were used to determine the amount of K, Na, Ca and Mg in $cmol\ (+)\ kg^{-1}$ in each soil sample.

Equation 1

$$\mathrm{cmol\ (+)\ kg^{-1}\ K\ in\ soil} = \frac{(\mathrm{mg\ dm^{-3}\ K\ in\ sample} - \mathrm{mg\ dm^{-3}\ K\ in\ blank}) \times \mathrm{dillution}}{390}$$

Equation 2

$$\mathrm{cmol\ (+)\ kg^{-1}\ Na\ in\ soil} = \frac{(\mathrm{mg\ dm^{-3}\ Na\ in\ sample} - \mathrm{mg\ dm^{-3}\ Na\ in\ blank}) \times \mathrm{dillution}}{230}$$

Equation 3

$$\mathrm{cmol\ (+)\ kg^{-1}\ Ca\ in\ soil} = \frac{(\mathrm{mg\ dm^{-3}\ Ca\ in\ sample} - \mathrm{mg\ dm^{-3}\ Ca\ in\ blank}) \times \mathrm{dillution}}{200}$$

Equation 4

$$\text{cmol}\,(+)\,\text{kg}^{-1}\ \text{Mg in soil}\ = \frac{(\text{mg dm}^{-3}\text{Mg in sample} - \text{mg dm}^{-3}\ \text{Mg in blank}\ ) \times \text{dillution}}{122}$$

Equation 5

$$\text{Dilution} = \frac{\text{Volume leachate } (500\ \text{cm}^3) \times 5}{\text{soil mass } (20\ \text{g})}$$

If some soil samples contain no soluble salts it was assumed that the extracted cations are equal to the exchangeable cations and was reported as extractable cations (The Non-affiliated Soil Analysis Work Commitee, 1990). However, if soluble salts are contained in the soil sample the concentration of salts were determined by multiplying the cation concentration in the extract by the saturation percentage and divided by 1000 to convert to $\text{cmol}\,(+)\,\text{kg}^{-1}$ soil. This value was then subtracted from the concentrations $\text{cmol}\,(+)\,\text{kg}^{-1}$ value calculated for the extracted cation. To determine the extractable acidity (KCl) used to calculate ECEC 10 g of the soil sample and 70 cm^3 1 mol dm^{-3} KCl was placed into a beaker and stirred for one hour after which it was filtered into a 500 cm^3 Erlenmeyer flask. The remaining soil in the filter paper was then washed with three portions of 30 cm^3 KCl solution until AA VOLUME OF 100 cm^3 was reached. A titration was performed with he filtered solution, 0.1 mol dm^{-3} NaOH and phenolphthalein as indicator. One drop of 0.1 mol dm^{-3} HCL was added to remove the pink colour and 10cm^3 NaF solution was added while stirring and further titrating with 0.1 mol dm^{-3} HCl to determine the exchangeable Al (The Non-affiliated Soil Analysis Work Committee, 1990). The following equations are used to calculate the extractable acidity:

Equation 6

$$\text{cmol}\,(+)\,\text{kg}^{-1}\text{KCl acidity} = \frac{\text{cm}^3\ \text{NaOH}\ \times \text{M}\ \times 100}{\text{mass of soil (10g)}}$$

Equation 7

$$\text{mol } (+)\ \text{kg}^{-1} \text{ exchangable Al} = \frac{\text{cm}^3\ \text{HCl} \times \text{M} \times 100}{\text{mass of soil (10g)}}$$

Equation 8

$$\text{Where M} = \text{concentration of the NaOH or HCl}$$

Equation 9

$$\text{mol } (+)\ \text{kg}^{-1}\ \text{H} = \text{KCl acidity} - \text{KCl exchangable Al}$$