Mid-Infrared spectroscopy calibration models for base cation concentration prediction in soils of the North-West Province, South Africa

BR Raath

orcid.org/0000-0002-6097-4064

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

Supervisor:	Prof GM van Zijl
Co-supervisor:	Dr PD Ramphisa
Co-supervisor:	Dr DE Elephant

Graduation May 2023

23495545

ACKNOWLEDGEMENTS

Thank you to everyone involved in this study which aims to create prediction models for all soil properties involved in plant growth. This work is based on the research supported by the National Research Foundation of South Africa (Grant Numbers: 121302).

I thank the following people contributing to the completion of this study:

- My supervisor Dr GM van Zijl, who believes in guidance as opposed to spoon feeding his students. Also, for being supportive and never unreasonable.
- Martiens Du Plessis at NWK Limited & Dailena Pienaar NviroTek Laboratories for providing soil samples and soil property data.
- Dup Haarhoff at GWK & Dries Bloem for providing soil samples and soil property data.
- Godfrey Lekhuleni at Bruker South Africa (Pty) Ltd for the lease of the ALPHA II DRIFTS MIR spectrometer.
- Ruan Gagiano at the North-West University Soil Laboratory for the use of laboratory equipment.
- Anru Kock as fellow student for his help and support.

A special thanks to the following:

- My mother and father for their love and financial support.
- All my friends for always being there for me.
- My dog, Alfie.

ABSTRACT

Fertilizers are essential for plant nutrition to sustain global food demands. Fertilizer recommendations requires soil analysis. Conventional laboratory analysis for soil chemistry is often slow and expensive. Mid-Infrared (MIR) spectroscopy may be a promising solution to overcome the limitations of conventional soil analysis, but these require soil specific calibration algorithms. Insufficient MIR analysis calibration algorithms exist for South African soils. The aim of this study was to create calibration algorithms for the prediction of exchangeable base cations (calcium (Ca^{2+}), magnesium (Mg^{2+}), potassium (K^{+}), and sodium (Na^{+})) concentrations for soils from North-West Province, South Africa. Soil analysis data was received from Noordwes Kooperasie (NWK) and Griekwaland Wes Korporatief (GWK) which included 4393 and 175 soil samples, respectively. Conditioned Latin Hypercube Sampling (cLHS) was used to select a total of 1000 samples (900 from NWK, 100 from GWK), of which 979 were deemed fit and represented the soil spectral database (SSD). The samples were crushed and sieved (53 micron) before being scanned at 4000 – 600 cm^{-1} spectral range at 2 cm^{-1} resolution. The data was captured by OPUS Base software, exported with Spectrograph 1.2 software to R Studio. A spectral library was created by combining the SSD and the spectra of the samples from the SSD in R Studio using the R programming language. The spectral library was divided into a training and validation datasets at a 75:25 split. Calibration algorithms were created from the training dataset using Cubist, Partial Least Squared Regression (PLSR) and Random Forest (RF) calibration models. The calibration algorithms were used to predict values of the validation dataset from the spectral library. The accuracy of the models was tested with the independent validation dataset with statistical analysis including coefficient of determination (R^2), root mean square error (RMSE) and ratio of performance to deviation (RPD). Cubist showed the best overall performance with order of declining performance accuracy of the base cations as follows: Ca (R^2 = 0.77; RMSE = 129; RPD = 2.09), Mg (R^2 = 0.75; RMSE = 40; RPD = 1.89), K (R^2 = 0.41; RMSE = 59; RPD = 1.28), and Na (R^2 = 0.29; RMSE = 6.45; RPD = 1.14), followed by PLSR, then RF. Base cations are not active in the MIR band. Prediction algorithms use soil properties which are active in the MIR band that correlate with exchangeable base cations, to predict the concentrations of the exchangeable base cations. To improve the accuracy of the models, it is recommended to increase sample numbers; using additional calibration models; using different scanning methods; and to include spectral processing before calibration.

Key terms

Soil spectroscopy, exchangeable base cations, mid-infrared, calibration models.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

Ca	Calcium
Mg	Magnesium
K	Potassium
Na	Sodium
NPK	Nitrogen, phosphorus, potassium
NWK	Noordwes Kooperasie
GWK	Griekwaland Wes Korporatief
cLHS	Conditioned Latin Hypercube Sampling
SPD	Soil property database
SSD	Soil spectral database
PLSR	Partial least squared regression
RF	Random Forest
R^2	Coefficient of determination
RMSE	Root mean square error
RPD	Ration of performance to deviation
EBC	Exchangeable base cations
PA	Precision agriculture
N	Nitrogen
P	Phosphorus
IR	Infrared
MIR	Mid-infrared
VIR	Visible infrared

NIR	Near-infrared
SWIR	Short wave infrared
C	Carbon
SOM	Soil organic matter
EC	Electrical conductivity
CEC	Cation exchange capacity
SOC	Soil organic carbon
Mn	Manganese
Fe	Iron
Cu	Copper
S	Sulphur
pm	Picometer
ICP	Inductively coupled plasma
AAS	Atomic absorption spectrometry

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

1.1 Background

Food insecurity remains a main concern as global hunger is ever increasing (Von Grebmer *et al.*, 2017). Although South Africa was seen as food self-sustainable in 2006 (Van der Berg, 2006), population growth still demands further agricultural development in South Africa (Nyam *et al.*, 2020). With the increase in population, more food will have to be produced, without increasing agricultural land area (Herrero *et al.*, 2010; Sutton *et al.*, 2013). The increase in food production should be done sustainably, as the food supply would need to be increased annually, to keep up with the food demand from population growth (Sutton *et al.*, 2013). Farmers are generally willing to adopt sustainable agricultural farming practices when paid, or when saving cost regarding inputs (Boufous *et al.*, 2023). Bonilla-Cedrez *et al.* (2021) suggest that agriculture can be intensified sustainably in sub-Saharan Africa by improving soil fertility and spatial targeting of fertilizer recommendations, which also reduce fertilizer inputs.

Soil analysis helps determine the soil fertility and allows a farmer to apply fertilizer inputs based on what the plants need (Wall & Plunkett, 2021). Soil testing is a very important part of commercial agriculture (Ray *et al.*, 2010). Soil testing increase yield, save cost and improve fertilizer use efficiency (Yun-peng *et al.*, 2010). For example: peach orchards in China had decreased fertilizer usage when soil testing was implemented (Xiao *et al.*, 2019), while also in China, fertilizer nutrient balancing with the help of soil testing, increased the yield of peas, maize, and peaches; and reduced leaching significantly, whilst using less nitrogen, phosphorus and potassium (NPK) fertilizers (Chen, Hu*, et al.*, 2021).

Cost saving and yield increases are also of interest to farmers through variable rate application (VRA) of fertilizer (Thompson *et al.*, 2019). VRA of N resulted in 0.8 tonne per hectare higher maize yields compared to uniform application in South Africa (Maine *et al.*, 2010). VRA with a high spatial sampling rate, reduced winter wheat N application rates while increasing profit margins (Stamatiadis *et al.*, 2018). However, VRA of fertilizers, requires more soil samples per area of agricultural land (Thompson *et al.*, 2019) and this has resulted in poor adoption by precision agriculture (PA) practitioners (Schimmelpfennig & Ebel, 2016; Griffin *et al.*, 2017; Mitchell *et al.*, 2018). When a large amount of soil samples is required, it involves more field and laboratory work which translates in a higher economic input when using conventional soil analysis.

Conventional soil analysis is the most widely used techniques used at the given time, in this case wet chemistry methods (Ryan *et al.,* 2013). Conventional soil analysis is slow, expensive, hazardous, and destructive to the soil sample (Rossel *et al.*, 2006). Mid-Infrared (MIR)

spectroscopy may be a promising solution to the limitations of conventional soil analysis (Gates, 2018). MIR spectroscopy is an analytical method that that give qualitative and quantitative information of the chemical composition of samples (Zorin *et al.*, 2021). MIR spectroscopy has shown promise as a soil analysis method internationally to predict various soil properties when used with predictive machine learning algorithms (Shepherd & Walsh, 2007; Rossel *et al.*, 2008; Johnson *et al.*, 2019; Gholizadeh *et al.*, 2021; Li, Xu*, et al.*, 2021; Metzger *et al.*, 2021; Parent *et al.*, 2021; Sabetizade *et al.*, 2021; Breure *et al.*, 2022). Examples of machine learning algorithms are partial least squared regression (PLSR), Random Forest (RF) and Cubist.

The agricultural sector in South Africa may benefit largely from MIR spectroscopy analysis due to it being more affordable, faster, requires less training and infrastructure, and measures multiple properties from a single sample (Rossel *et al.*, 2006; Paterson *et al.*, 2015).

1.2 Problem statement

A gap in the spectroscopic application stem particularly from the insufficient calibration algorithms for soils in South Africa (Paterson *et al.*, 2015). The main challenge is building a soil property database for the baseline study which requires large amounts of soil samples. Collaborative research efforts are required to build up soil spectral libraries for South African soils (Paterson *et al.*, 2015). This includes data already collected by: ISCW, National and Provincial Departments of Agriculture and Universities; data held by private companies which are difficult to include due to confidentiality clauses; and lastly data that still needs to be collected from both private and public entities that can contribute to the soil property libraries (Paterson *et al.*, 2015). Exchangeable base cations (EBC) are also seldom the focus of MIR soil spectroscopy studies (Gholizadeh *et al.*, 2021; Li, Feng*, et al.*, 2021; Metzger *et al.*, 2021; Parent *et al.*, 2021; Sabetizade *et al.*, 2021; Breure *et al.*, 2022). The Soil Fertility and Analytical Services laboratories of the KwaZulu-Natal Department of Agriculture and Environmental Affairs have set up confident MIR calibrations for organic carbon, total nitrogen, and clay. However, base cations calibration models are not widely available in South Africa.

1.3 Hypothesis

The hypothesis is that acceptable calibration algorithms can be created whereby the exchangeable base cations (Ca, Mg, K and Na) can be determined using MIR for the North-West Province, using machine learning algorithms.

1.4 Research aim

This study aims to create and test calibration algorithms using MIR to predict EBC for the soils of the North-West Province, using machine learning.

1.5 Research objectives

To reach the research aim, the following objectives must be met:

1. Establishing an EBC database of conventionally analysed soil samples of the North-West province.
2. To create an MIR soil spectral database for the same soil samples.
3. Forming a spectral library by merging the soil database with the MIR spectral database.
4. Creating MIR calibration algorithms for the EBC, using machine learning and the created spectral library.
5. To validate the calibration algorithms with an independent soil dataset.

CHAPTER 2 LITERATURE REVIEW

2.1 Introduction

"*Food security exists when all people, at all times, have access to sufficient, safe and nutritious food to meet their dietary needs for an active and healthy life"* (FAO, 2005). The Sustainable Development Goals (UN, 2015) includes a goal to "*end hunger, achieve food security, improve nutrition and promote sustainable agriculture"* by 2030. This will be difficult to achieve (Hawkes & Fanzo, 2017; Cai *et al.*, 2020), as the interim goal to halve hunger by 2015 was met by less than half of all developing countries (Barbosa-Cánovas *et al.*, 2017). Furthermore, it is expected that Africa will contribute to approximately 58% of the world population growth by 2050 (Ciceri & Allanore, 2019), which necessitates the need for increased food production to achieve food security on our continent.

Food insecurity is still a main concern as global hunger was still prevalent at 21.8% in 2017 (Von Grebmer *et al.*, 2017). Poor human and environmental health are the result of poor soil-, crop-, and livestock health in many African rural areas (Shepherd & Walsh, 2007). Coupled with an increase in population, a great threat is posed to ecosystems and the environmental sustainability of developing countries in Africa (Shepherd & Walsh, 2007). Fertilizers are essential for plant nutrition and may increase crop production by up to 90% by achieving higher yields and cropping intensity (Ciceri & Allanore, 2019; Qiao *et al.*, 2019). Fertilization also improves agricultural efficiency and product quality (Savci, 2012). However, the sustainability of extensive use of fertilizers is questioned, as it has been found to lead to groundwater contamination, soil acidification, soil degradation, nutrient imbalance and deterioration of soil fertility; all affecting plant growth (Savci, 2012; Rahman & Zhang, 2018b; AL-Zabee & AL-Maliki, 2019; Naik *et al.*, 2019).

According to the FAO, in the past fifty years, the use of nitrogenous, phosphate, and potash fertilizers were increased by 800%, 300%, and 125% respectively (Tian *et al.*, 2021). Fertilizer over-application is also known to have a negative effect on plant growth (Wei *et al.*, 2018). This is most evident with broadcast fertilizer applications, generally coupled with low-density conventional soil analysis (Rahman & Zhang, 2018a).

2.2 Precision agriculture

Precision agriculture (PA) relies on farm systems designed to increase productivity, profitability, production efficiency; while minimising environmental impact (Whelan & Taylor, 2013). These systems include technologies that automate application of pesticides, irrigation, and fertilizer (Say *et al.*, 2018). Natural soils may be variable in a short distance and could cause variable yields

(Elkateb *et al.*, 2003). Geypens *et al.* (1999) noted higher variable ranges for carbon, pH and calcium and lower variable ranges for K content in a study in Belgium as seen in Figure 2.1 and Figure 2.2.

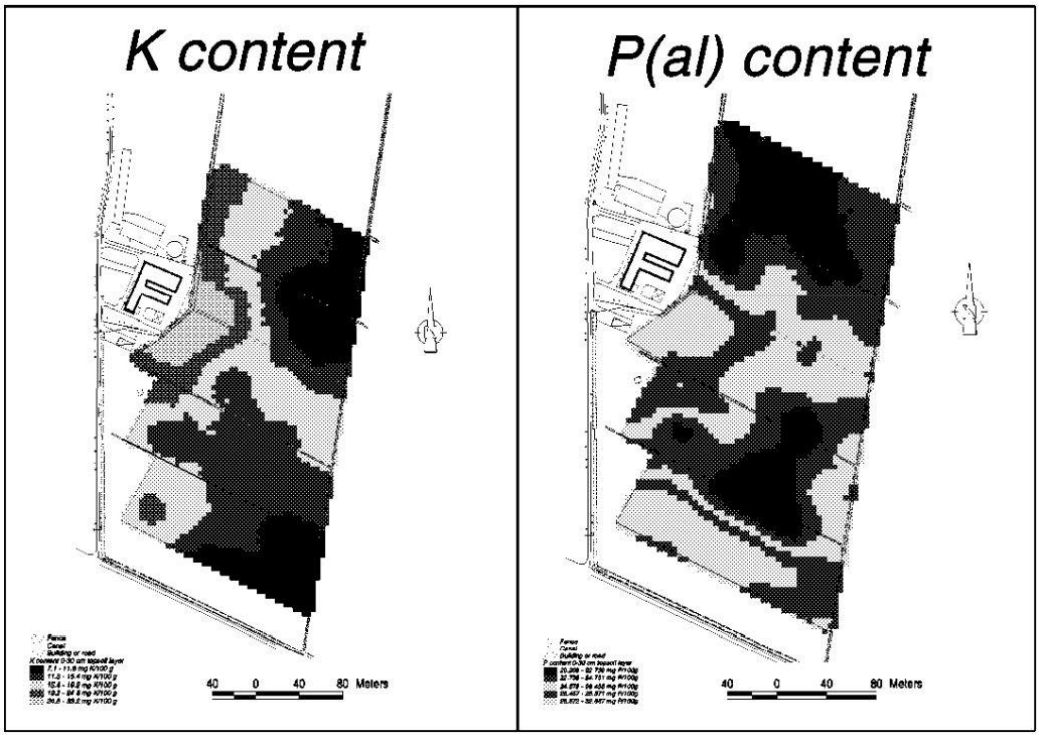


Figure 2.1: Variability of K content and P content in a study by Geypens *et al.* (1999)

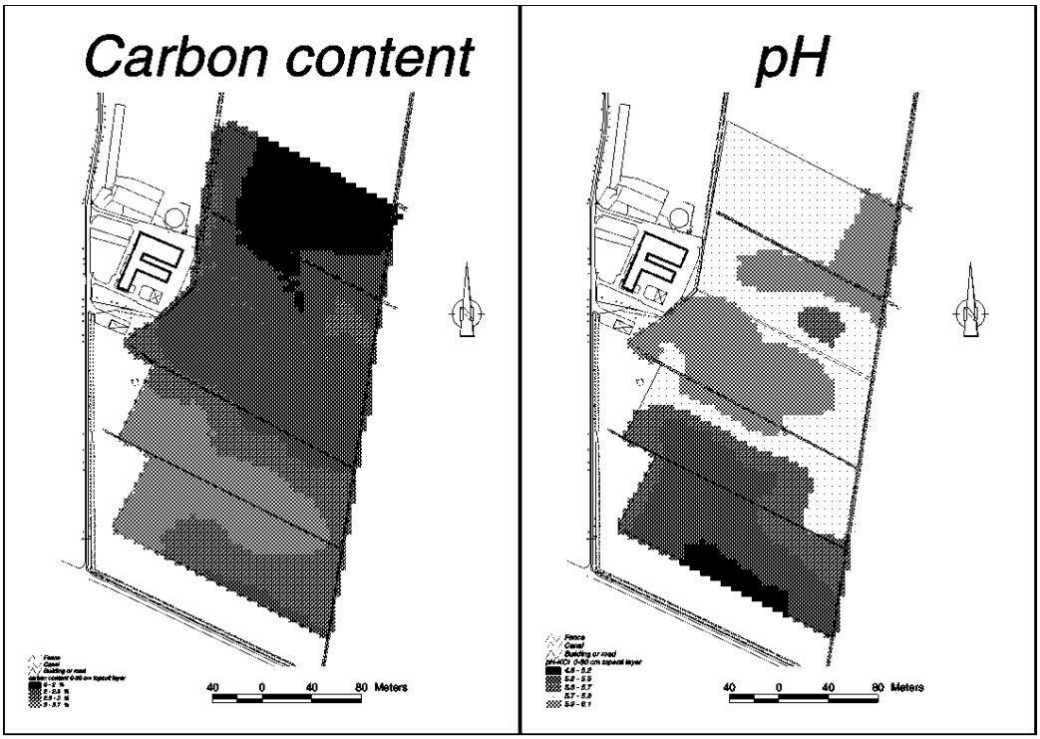


Figure 2.2: Variability of carbon content and pH in a study by Geypens *et al.* (1999)

To homogenise and increase the yields, fertilizer applications need to address this soil variability (Geypens *et al.*, 1999). PA offers economic benefit through reduction in the application of agricultural inputs (Tey *et al.*, 2017; Kendall *et al.*, 2021), increased production by management of variability of crop fields (Schimmelpfennig & Ebel, 2016; Kendall *et al.*, 2021) and environmental benefit through precise application of agrochemical applications (Ma *et al.*, 2014; Kendall *et al.*, 2021). PA offers a solution to food security, while preserving agricultural and environmental sustainability (Kendall *et al.*, 2021), by allowing the farmer to reduce inputs, which also has an economic benefit (Bongiovanni & Lowenberg-DeBoer, 2004).

In Australia, R242 to R520/ha (current exchange rate) annual benefits were recorded when implementing PA. Savings ranging between R17.35 and R381/ha were recorded across six case studies that followed VRA of fertilizers (Robertson *et al.*, 2007). In the Sichuan province of China, soil testing reduced fertilizer application by 23.5% - 28.2% on peach orchards, while also discovering and rectifying soil acidification (Xiao *et al.*, 2019).

In Beijing, China; nutrient balancing by soil testing was used to test the growth of peas, maize, and peaches; whilst also monitoring leaching of excess fertilizer (Chen, Hu*, et al.*, 2021). Nitrogen (N) input was reduced by 60%; 34% and 87% for peas, maize, and peaches; respectively. Phosphorus (P) inputs decreased by 77%; 69% and 93%; for peas, maize, and peaches; respectively. Potassium (K) inputs for maize and peaches were decreased by 8% and 83% respectively; whilst a deficient soil, hosting the peas, could be rectified by adding 83% more P. These nutrient balancing practises also increased yields for peas, maize, and peaches by 2.7%; 3.3%; and 49%; respectively. Furthermore, leaching of N reduced for peas, maize, and peaches by 61%; 43%, 88%; respectively, whilst leaching of P reduced for peas, maize, and peaches by 78%; 69%, and 93%. Variable rate application resulted in 38% less total N use; and an increase of 14% N use efficiency in winter wheat in Larissa, Greece (Stamatiadis *et al.*, 2018). This related to a return of R53/ha per unit of N, compared to R31/ha return by farmer practise. PA is therefore one way to use less fertilizers, but is dependent on high sample density, which can be timely and costly when conventional soil analysis methods are used (Gates, 2018).

Conventional soil laboratory analysis is often slow and expensive, makes use of hazardous chemicals and is destructive to the soil sample (Palm *et al.*, 2007; Paterson *et al.*, 2015). When different soil properties are tested, samples are split into sub-samples because they will be mixed with different chemicals to test different soil properties. With MIR analysis the same samples are used to simultaneously test the various soil, properties. With multiple sub-samples needed to be able to analyse multiple soil properties, deviations in subsamples may influence the correlations drawn between the various soil properties, as chemical extraction may disrupt the soil equilibrium (Rossel *et al.*, 2006; Stenberg *et al.*, 2010; Gates, 2018). The number of observations to be made on a specific site is also limited with sampling associated with conventional soil analysis, which

become problematic when soils vary significantly over a short distance (Paterson *et al.*, 2015). Infrared (IR) spectroscopy may be a promising solution to the drawbacks posed by conventional soil analysis (Janik *et al.*, 1998; Bramley & Janik, 2005; Nocita *et al.*, 2015; Gates, 2018) information on soil properties or soil conditions are require a high sample amount per unit for the are being analysed, due to its fast and cost effective soil analysis (Shepherd & Walsh, 2007).

2.3 Infrared spectroscopy

Ng and Simmons (1999) states: "*Infrared (IR) spectroscopy measures the absorption of infrared radiation by chemical bonds in a material*". Infrared radiation is absorbed by functional groups in molecules in a frequency range regardless of the rest of the structure of the molecule. Identification of these functional groups and sequentially the unknown molecule is then possible due to the connection to the frequency at which it absorbs IR radiation (Ng & Simmons, 1999).

Spectrally active soil properties can be predicted by the use of infrared spectroscopy (Gates, 2018). This can be done with wavelengths ranging from mid-infrared (MIR) to visible (VIR) and near-infrared (NIR) spectroscopy (McCarty & Reeves, 2006; Reeves III, 2010; Hutengs *et al.*, 2019; Johnson *et al.*, 2019). VIR includes the range of 350 to 780 nm and NIR ranges from 780 to 2,500 nm (Fang *et al.*, 2018). In remote sensing, the 350 to 1,000 wavelength range can also be referred to as VNIR whilst short-wave infrared (SWIR) then refers to the 1,000 to 2,500 range (Figure 2.3).

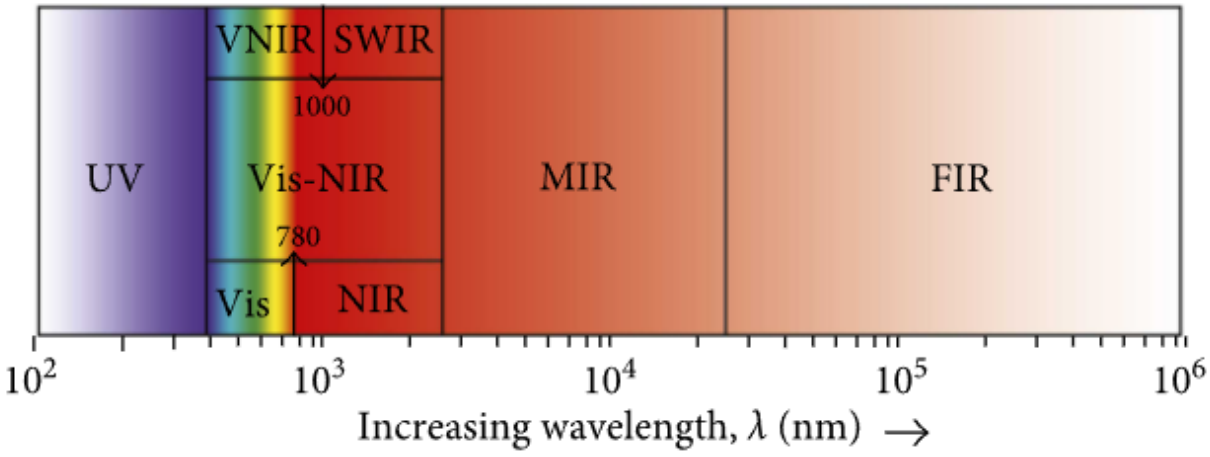


Figure 2.3: Graph displaying wavelengths of the different IR ranges in nanometres (nm) by (Fang et al., 2018).

IR spectroscopy soil analysis have been applied to different soil properties such as carbon (C) content (Bellon-Maurel & McBratney, 2011a; Deiss *et al.*, 2020), soil organic matter (SOM) (Chen *et al.*, 2016; Dangal *et al.*, 2019), total N (Sanderman *et al.*, 2020), extractable P (Forrester *et al.*, 2015), pH (Dangal *et al.*, 2019; Sanderman *et al.*, 2020), electrical conductivity (EC) (Sanderman *et al.*, 2020), exchangeable base cations (Ca, Mg, Na, K) (Gates, 2018; Sanderman *et al.*, 2020),

cation exchange capacity (CEC) (Dangal *et al.*, 2019; Sanderman *et al.*, 2020), heavy metals (Jean-Philippe *et al.*, 2012; Wang *et al.*, 2017; Dangal *et al.*, 2019), moisture content (Janik *et al.*, 2007), particle size distribution (Tümsavaş *et al.*, 2019; Janik *et al.*, 2020), clay mineralogy (Reeves III, 2012; Kasprzhitskii *et al.*, 2018); soil respiration (Meyer *et al.*, 2018; Liu *et al.*, 2021), microbial biomass (Kamnev *et al.*, 2021), atmospheric components (Liu *et al.*, 2020), leachability of pesticides (Li *et al.*, 2012; García-Jaramillo *et al.*, 2014); soil salinity (Triki Fourati *et al.*, 2015; Peng *et al.*, 2016), soil acidity and alkalinity (Merry & Sabljic, 2009), and soil type (Linker *et al.*, 2005; Linker, 2007; Du *et al.*, 2008).

IR spectroscopy is a low cost, repeatable soil analysis method being more widely used in present times (McClure, 2003; Workman Jr & Shenk, 2004; Shepherd & Walsh, 2007). Infrared (MIR and NIR) spectroscopy instrumentation can aid in soil analysis from a single sample to analyse multiple soil properties by using no chemicals inexpensively and fast (Shepherd & Walsh, 2007).

2.3.1 NIR spectroscopy

NIR spectroscopy is based on vibration of molecules that exist in the wavelength range from 13,333 to 4000 cm^{-1} (Pasquini, 2018). Near infrared technologies have previously been reported successful in applications ranging from material science, food, environment, medicine, pharmaceutics, agriculture and archaeology (Agelet & Hurburgh Jr, 2010). NIR spectrometers are less expensive than MIR spectrometers and when rapid in field analysis is needed, VNIR may be a better option than MIR (Gates, 2018). NIR spectrometers are being utilized by tillage systems for in field measurements (Adamchuk *et al.*, 2004; Shepherd & Walsh, 2007).

Previous studies on soil with the use of NIR spectroscopy was mainly used to predict organic carbon in soils (Reeves III, 2010; Bellon-Maurel & McBratney, 2011b; Branco de Freitas Maia *et al.*, 2013; Gholizadeh *et al.*, 2013; Gobrecht *et al.*, 2014; Johns *et al.*, 2015). There have also been successful reports when predicting CEC, EBC, Total P, SOC, Total N, pH (Gates, 2018; Haghi *et al.*, 2021) and heavy metal contamination in soils (Shi *et al.*, 2014). However, NIR models are not as robust as MIR models and MIR should be investigated when looking for more accurate prediction models (Brown *et al.*, 2006; Gates, 2018).

2.3.2 MIR Spectroscopy

MIR spectroscopy exist in the range from 2,500 – 25,000nm (4,000 cm^{-1} to 400 cm^{-1}) (Haas & Mizaikoff, 2016). It can provide quantitative information on molecules in any phase (liquid, solid, gas), that can support a variety of use cases (Haas & Mizaikoff, 2016). Applications range from manufacturing monitoring, materials science, medicine, environmental analysis and biotechnology (Haas & Mizaikoff, 2016). The bonds able to identify with MIR spectroscopy can be seen in Figure 2.4.

MIR spectroscopy gives stronger soil properties peaks due to the detection of fundamental vibrations of mineral and organic compounds, compared to NIR which offer more overtone features (Shepherd & Walsh, 2007). Generally, MIR calibrations also outperform NIR based calibrations on the same sample sets (Pirie *et al.*, 2005; McCarty & Reeves, 2006; Rossel *et al.*, 2006). MIR is also better suited for organic matter and may provide more robust calibrations when different soil types are being analysed (Janik *et al.*, 1998; Shepherd & Walsh, 2007). In this case study, as predictive accuracy of models is the preferred outcome, MIR spectroscopy analysis will be used. MIR spectra, which contain large amounts of spectral information (McCarty & Reeves, 2006) is investigated.

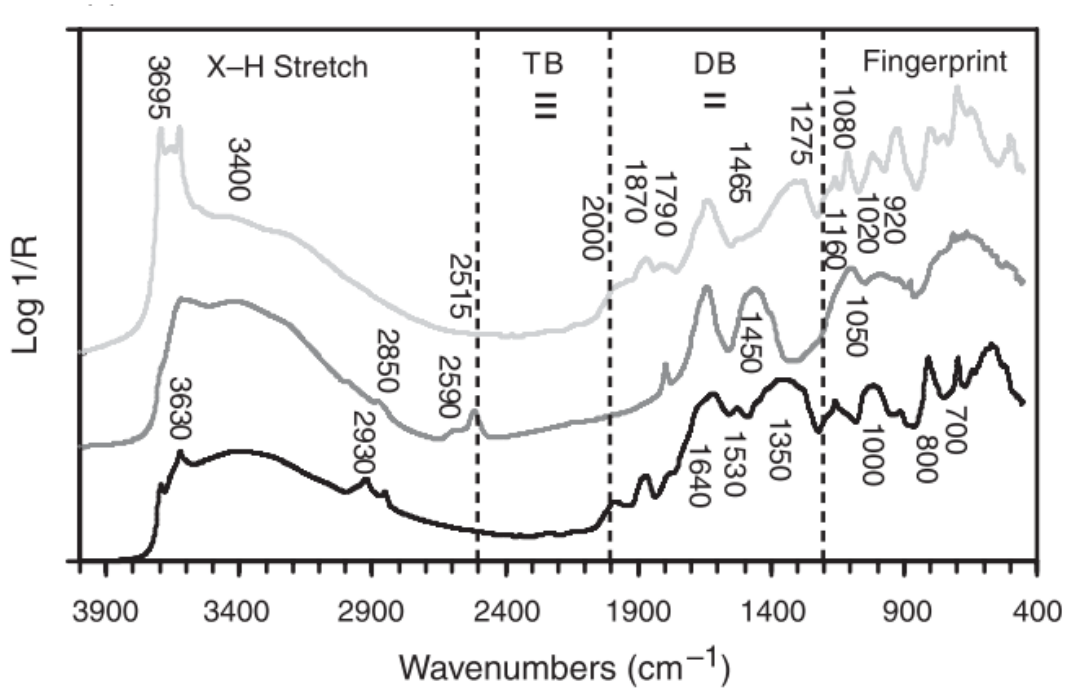


Figure 2.4: Comprehensive display of wavenumbers in cm-1 associated with the bonds of molecules showing active vibration the mid infrared region by Rossel et al. (2008).

2.4 Soil Spectral Library

The creation of a soil spectral library is described in Figure 2.5. The process by Shepherd and Walsh (2002) explains the need for a soil property database, of high density, in conjunction with soil spectra from the same database to create a soil spectral library. According to Shepherd and Walsh (2002), samples should be collected covering the entire study area, using a representative sampling technique and size. A soil spectral library is then built by gathering spectral data for each sample and combining it with the conventional laboratory analysis data of each sample. The calibration accuracy of the spectral and soil data is tested. If the model accuracy is accepted, a prediction is made, but if unacceptable spectral information is received, new spectral information should be gathered for the specific sample. New samples added to the database should be sampled with the same technique and sample size and scanned. If the sample does not test as an outlier, it can be added to the database. Powerful machine learning software is often used to solve the complex formulas presented in calibration algorithms, especially for MIR spectroscopy (Reeves III, 2010).

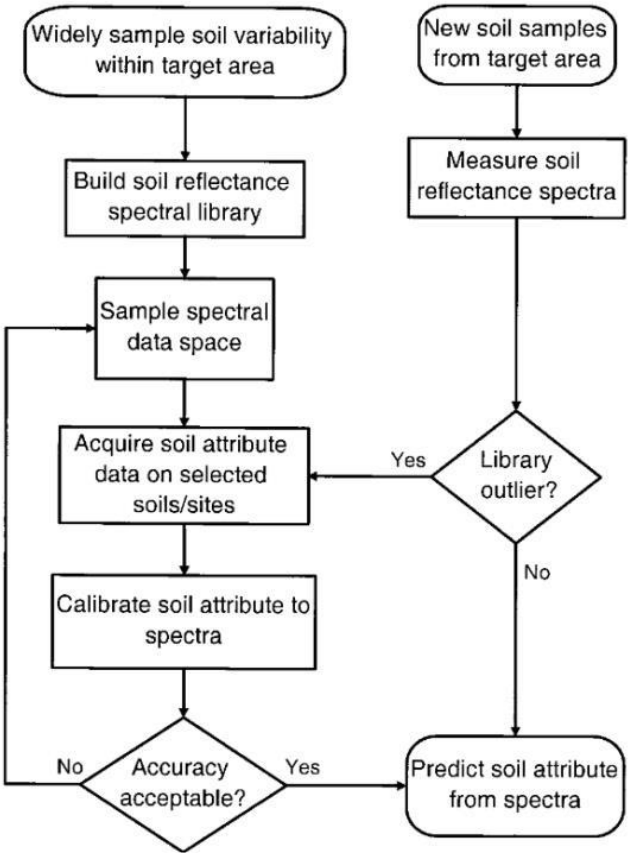


Figure 2.5: The detailed process of building a soil spectral library as proposed by Shepherd and Walsh (2002).

2.5 Dataset size

Sordo and Zeng (2005) suggest that the size of the training dataset is positively correlated with improved performance in super vector machines and decision tree algorithms. Ng *et al.* (2020) recommends that deep learning for spectral modelling is at a sample size above 2000. The accuracy of PLSR, Cubist and Conventional neural networks plateaued at 4200, 500 and 1800 soil sample number; respectively (Ng *et al.*, 2020).

2.6 Machine learning algorithms

Machine learning makes use of statistical models and algorithms increasing efficiency of applications used in scientific study (Bei *et al.*, 2021). Computer systems use these models to perform tasks without the need for specific instruction, rather relying on patterns and inference (Bei *et al.*, 2021). Predictive machine learning algorithms have been a growing interest for the past quarter century (Janik *et al.*, 1998; Breure *et al.*, 2022). Most recently, numerous calibration algorithms have been used in soil spectroscopy including PLSR (Li, Feng*, et al.*, 2021; Nath *et al.*, 2021; Sabetizade *et al.*, 2021), random forest (Dangal *et al.*, 2019; Chen, Men*, et al.*, 2021), and cubist (Dangal *et al.*, 2019; Ma *et al.*, 2021).

2.6.1 Partial least squared regression

Partial least squared regression (PLSR) is a successfully used calibration model in prediction algorithms for soil spectroscopy analysis (Janik *et al.*, 1998; Reeves III, 2010; Stenberg *et al.*, 2010; Nocita *et al.*, 2011; Nocita *et al.*, 2015; Metzger *et al.*, 2020). PLSR aims to split the collinear variables into latent variables to analyse data with fewer observations, by maximizing the variance between the response and the latent variables (Wold *et al.*, 2001; Gates, 2018). PLS is more robust in handling higher amounts of descriptor variables, increasing accuracy and lowering the risk of chance correlation (Cramer III, 1993). The major limitations are a higher risk of overlooking 'real' correlations and sensitivity to the relative scaling of the descriptor variables.

Janik and Skjemstad (1995) presented PLSR to create soil property prediction models as early as 1995. The study used 300 soils throughout Australia and had excellent prediction accuracy with a R^2 of 0.92. Merry *et al.* (1997) achieved similar precision when predicting values for organic C (R^2 = 0,93) and total nitrogen (R^2 = 0,86) in South Australian soils. Janik *et al.* (1998) followed soon after with a comprehensive study, predicting most of the important soil properties (Appendix 1) with moderate to excellent results. Prediction accuracies vary widely for different properties, and is generally classed as excellent (exchangeable Ca, exchangeable Mg, CEC, Total C, soil organic carbon (SOC), Total N, texture), moderate (exchangeable K, Total P, pH, SOM , Manganese (Mn)) or poor (exchangeable Na, iron (Fe), copper (Cu), extractable P, EC, sulphur

(S)) (Rossel *et al.*, 2008; Ji *et al.*, 2016; Gates, 2018; Sanderman *et al.*, 2020; Takele & Iticha, 2020; Haghi *et al.*, 2021).

PLSR was the first widely used prediction algorithm in soil analysis. However, Sirsat *et al.* (2018) found in his review on over 76 different algorithms that higher performance machine learning methods such as RF, may have a consistent higher performance.

2.6.2 Random Forest

Random Forest (RF) regression tree models combine numerous randomized decision trees and enhance their predictions using averages (Biau & Scornet, 2016). It has been successfully used to predict soil properties such as pH (Chen, Men*, et al.*, 2021; Ma *et al.*, 2021), SOC, clay, and CEC (Ma *et al.*, 2021). Random forest aims to solve missing values by splitting node samples into different weights or filling missing values with an estimate value such as the median then build a random forest on the available data to update the missing values by weighted rules of the estimated values where the closest (Louppe, 2014). However, the mechanisms behind random forest are not fully understood, and a problem arising frequently is the inconsistency of the model related to the learning set (Louppe, 2014).

Soil property prediction by using RF as calibration algorithm is not as widely studied and varies between different studies when it is used. For instance, pH predictions ranged from excellent (Chen, Men*, et al.*, 2021), to moderate (Ma *et al.*, 2021), to poor (Dharumarajan *et al.*, 2017) in these three different studies. Random forest have been used for soil salinity prediction, by remote sensing with satellite imagery, with sufficient accuracy (Fathizad *et al.*, 2020). Predictive properties for SOC, OC and CEC were poor, whilst EC have been predicted with moderate success (Dharumarajan *et al.*, 2017; Ma *et al.*, 2021). Ca could be sufficiently predicted with RF but the other base cations of interest were not tested in the study by Dangal *et al.* (2019). Excellent predictive capabilities were displayed for SOM (Pouladi *et al.*, 2019).

RF cannot accurately explain the correlation between soil spectra and soil properties due to overfitting of data (Gates, 2018). Simplicity can also be a good factor as overfitting of data is a real problem in some of these complex machine learning methods.

2.6.3 Cubist

Cubist is a rule-based algorithm which investigates nonlinear associations in observed data according to a series of “if-then” rules represented by multivariate linear models of the predictor (Kuhn et al., 2016). Cubist is not as widely used as PLSR but is of interest due to its simplicity. In predicting Ca, the cubist model have outperformed PLSR and RF, with R^2 values of 0.95, 0.89, and 0.93; respectively (Dangal *et al.*, 2019). Cubist tries to simplify models whilst maintaining

predictive accuracy by building a model containing multiple rules expressed linearly which aims to minimize average absolute error of the predicted values (Anon, 2019). Therefore, the rules may be biased concerning the predicted mean value compared to the measured mean value. Generally, with over-simplified models there is a trade-off with simplicity and prediction accuracy (Anon, 2019).

Cubist as prediction model is used less than PLSR and more than RF. Prediction accuracies vary widely for different properties and is generally classed like PLSR prediction models (Hong *et al.*, 2014; Morellos *et al.*, 2016; Dangal *et al.*, 2019; Haghi *et al.*, 2021). However, it seems that there is more variation between studies. Dangal *et al.* (2019) found more success overall than Hong *et al.* (2014) but may be attributed to using MIR as opposed to VNIR spectroscopy.

2.7 Algorithm validation

Validation is the process of determining the degree of accuracy of a model (Sornette *et al.*, 2007). Industries depend increasingly on predictions by computers models in multiple sectors (Sornette *et al.*, 2007). Vabalas *et al.* (2019) suggest that statistical power increases with sample size. Lucà *et al.* (2017) found that different regression methods, such as PLSR, are more sensitive to calibration sizes. The creation of predictive calibration algorithms requires the data to be split into training and evaluation datasets to be able to determine the model performance (Gates, 2018). Training and evaluation datasets are generally split into 75% and 25% for training and evaluation datasets, respectively (Gates, 2018). These values can be changed to better suit different sample sizes. Sabetizade *et al.* (2021) split the data into 70% and 30% for training and evaluation datasets, respectively, but only had 302 data points.

Different problems can be more adequately solved by different calibration models (Mayer & Butler, 1993). Reviewing multiple machine learning studies, the most prominent validation algorithm is coefficient of determination followed by root mean square error. Also frequently encountered in soil property predictions studies is the ratio of RPD, which accounts for the offset of root mean square error value (Chang *et al.*, 2001; Niederberger *et al.*, 2015).

2.7.1 Coefficient of determination

The coefficient of determination (R^2) is a statistical measure of the proportion of variance for calibration models and are standard across most studies. It describes the degree of correlation between predicted and observed values (Gates, 2018). R^2 is calculated by Equation 1.

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

n represent the sample size, $\overline{obs}$ the total measured soil property values, obs_i represent the vector of the measured- and $pred_i$ the vector of the predicted soil property values, respectively (Wadoux *et al.*, 2021).

An R^2 value greater than 0.9 may be used to replace conventional analysis techniques, a value between 0.7 and 0.9 may be suitable to use for analysis that is too time consuming and costly whereas an R^2 value below 0.7 is considered poor (Janik *et al.*, 1998; Reeves III & Smith, 2009). R^2 is not a good standalone algorithm validation to represent the prediction accuracy of a model, and should be accompanied by RMSE (Davies & Fearn, 2006; Gates, 2018).

2.7.2 Root mean squared error

Model prediction performance is often described by the RMSE statistical parameter (Estienne *et al.*, 2001; Bellon-Maurel *et al.*, 2010; Gates, 2018). The RMSE is calculated by the squares of the variance between predicted and observed values, summed (Gates, 2018). It is calculated by Equation 2.

$$\mathrm{RMSE}_{fo} = [\sum_{i=1}^{N} (z_{fi} - z_{oi})^2 / N]^{1/2} \quad [2]$$

N is the number of samples in the test set, z_{fi} and z_{oi} is the predicted and observed values, respectively. RMSE is an estimate of average for the property of interest and its units are directly related to the unit of the property tested, in this study it's $mg.kg^{-1}$. Ratio of performance to deviation is very much related to the RMSE, however makes the equation relevant to properties with varying RMSE values by making use of the standard deviation.

2.7.3 Ratio of performance to deviation

The ratio of performance to deviation (RPD) is specifically formulated to explain the accuracy of predictive models. Dividing the standard deviation of the response by the root RMSE, results in the RPD. The RPD is an excellent way to minimize confusion caused by the RMSE as the RMSE between different variables can differ significantly in value and not in percentage related to the observed value. It allows one to compare results between different properties with one another. Model quality is classified by Chang *et al.* (2001) and Niederberger *et al.* (2015) and in Table 2-1.

Table 2-1: Model quality explained by RPD and R^2 as explained by (Chang *et al.*, 2001) & (Niederberger *et al.*, 2015)

Model quality	(Chang *et al.*, 2001) RPD	(Niederberger *et al.*, 2015) RPD	(Niederberger *et al.*, 2015) R^2
Excellent	>2	>4	> 0.95
Successful	1.4 - 2	3 – 4	0.9 – 0.95
Moderately successful		2.25 – 3	0.8 – 0.9
Moderately useful		1.75 – 2.25	0.7 – 0.8
Poor	<1.4	<1.75	< 0.7

RPD is calculated by Equation 3.

$$RPD = \frac{\sigma}{RMSE} \quad [3]$$

σ represents standard deviation and RMSE is the abbreviation for root mean squared error.

2.7.4 Ratio of performance to interquartile distance

Ratio of performance to Interquartile distance (RPIQ), is the calculation of the interquartile range of the measured values divided by the Root Mean Square Error, displayed by Equation 4.

$$RPIQ = \frac{IQR}{RMSE} \quad [4]$$

A higher RPIQ value is a representation of a more accurate prediction model.

2.7.5 Bias

Bias is a calculation of the average value of which the measured value is greater than the predicted value and is calculated by using the mean error and the RMSE (Wadoux & McBratney, 2021). The closer to zero the value of the bias, the more unbiased the predictor is (Wadoux & McBratney, 2021). Bias is calculated by Equation 5:

$$Bias = \frac{1}{n}\sum_{i=1}^{n}(obs_i - pred_i) \quad [5]$$

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

2.8 Exchangeable base cations

Exchangeable base cations (EBC) are defined as being able to be replaced by a cation of an added salt solution (Thomas, 1983). Base cations are prevalent, exchangeable, weak acid cations present in the soil. They garner their name because they offset acidifying effects of SO_x and NO_x (Johnson, 1992). The four cations that most generally occur in soils are calcium (Ca^{2+}), magnesium (Mg^{2+}), potassium (K^{+}), sodium (Na^{+}) (Hazelton & Murphy, 2016). In strongly acidic soils, aluminium (Al^{3+}) may be high in concentration (Hazelton & Murphy, 2016). Hydrogen (H^{+}) and Al^{3+} are referred to as acid cations (Gillespie *et al.*, 2021). Manganese (Mn^{2+}), iron (Fe^{2+}), copper (Cu^{2+}) and zinc (Zn^{2+}) are usually in negligible amounts in comparison to the other EBC (Hazelton & Murphy, 2016). The affinity of base cations to occupy an exchange site is also dependent on the soil pH as seen in Figure 2.6.

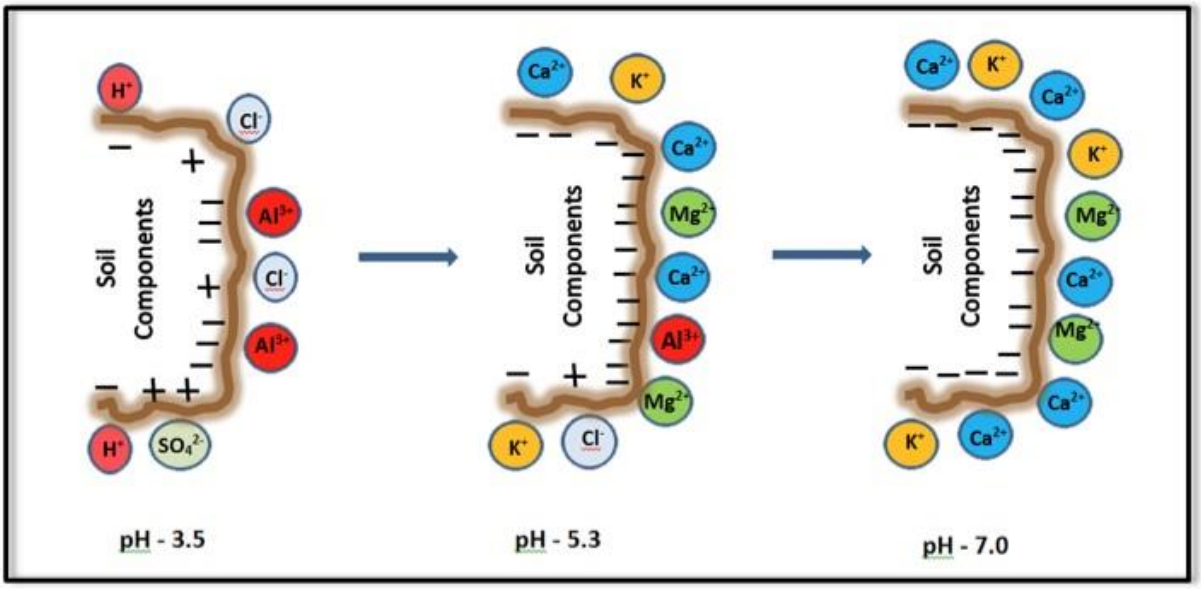


Figure 2.6: The affinity of base cations to occupy an exchange site is dependent on soil pH (Saha, 2014).

The exchangeable cations are held onto the soil particles by the negative charge of the clay minerals as seen in Figure 2.7.

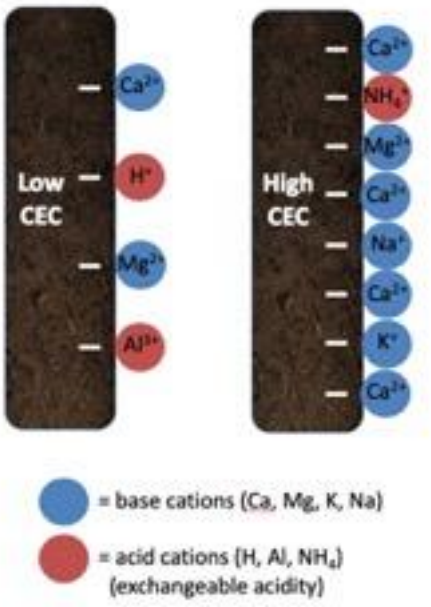


Figure 2.7: The comparison of the attraction of cations to a soil particle with low vs high CEC (Culman *et al.*, 2019).

The degree at which the cations are drawn to the soil particle, is related to the cation exchange capacity CEC. A soil with a higher CEC can thus hold onto more nutrients and exchangeable cations can only be replaced by other cations. The affinity of a cation to occupy an exchange site is related to the size of its charge density as can be seen in Table 2-2 (Moore & Bradley, 2018).

Table 2-2: Soil cations represented with their size in picometer (pm) and charge (mono-, di-, and trivalent)

Monovalent (1+)	Divalent (2+)	Trivalent (3+)
Sodium (116 pm)	Magnesium (86 pm)	Aluminium (41 pm)
Potassium (152 pm)	Calcium (114 pm)	Gallium (68 pm)
Rubidium (166 pm)	Strontium (132 pm)	Indium (94 pm)

The ratio of base- vs acid cations occupying the exchange sites, is called base saturation of a soil, and can be seen in Figure 2.8.

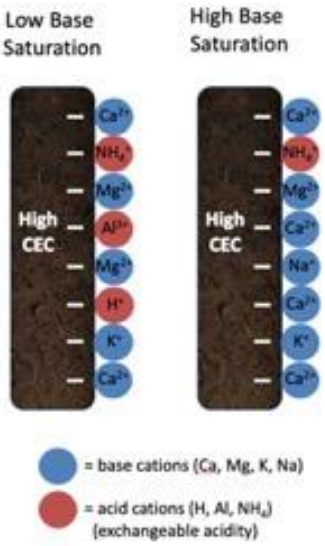


Figure 2.8: Comparison of the occupation by cations on soil particles with high and low base saturation (Culman *et al.*, 2019).

2.8.1 Exchangeable base cation analysis

Exchangeable base cations are generally tested in South Africa by extraction with ammonium acetate and measured with inductively couple plasma (ICP) spectrometry or atomic absorption spectrometry (AAS). Another used method is the Mehlich 3 extraction stock solution (Warncke & Brown, 1998). Warncke and Brown (1998) propose the methods and reagents to be used. In soils above 7.3 pH, Mehlich 3 values are overestimated as it starts to extract non-extractable Mg and Ca form soils (Rutter *et al.*, 2021).

Wavebands where exchangeable basic cations are active do not fall into the MIR absorption bands. However, clay particles contain mineral compounds which do fall within the MIR absorption bands. The correlation of exchangeable basic cations with these mineral compounds may attribute to their prediction accuracy. Gates (2018) measured correlation of clay percentage with the exchangeable basic cation and can be seen in Figure 2.9. Freeman *et al.* (2008) discussed feldspars and Madejová *et al.* (2017) discussed micas, especially the peaks in the region of 650-, 760-, 800-, 1000-, 1100-, and 1200 cm^{-1}. Feldspars contain calcium, potassium and sodium ions which balance the negative charge on the silicate framework (Nash & Marshall, 1956). Micas also contain calcium, potassium, and sodium but contain double the number of sites that can contain magnesium (Allpress & Sanders, 1967). In general MIR is quite good at predicting total elements but is far more variable with solution and exchangeable elements. Ca and Mg having higher charge density and are generally found in higher concentrations than K and Na in soil may be why Ca and Mg do better than K and Na in predictive analysis.

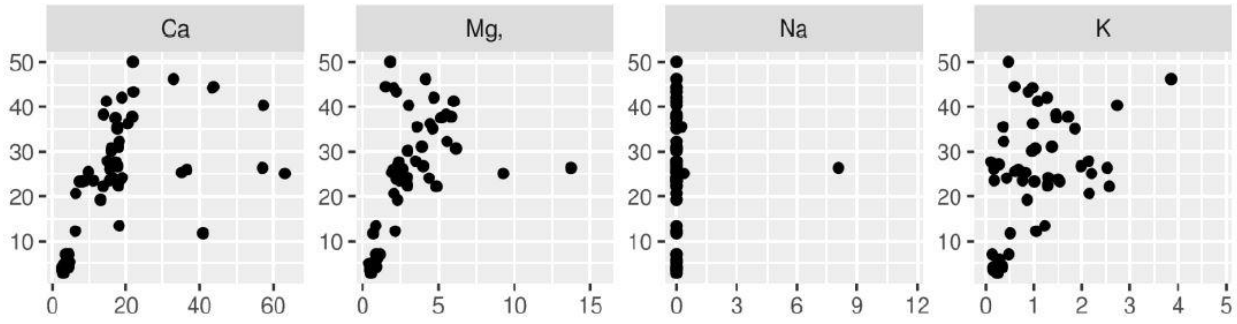


Figure 2.9: Correlation of clay percentage (x-axis) with exchangeable calcium; magnesium; sodium; and potassium; respectively (y-axis), in a study by Gates (2018), measured in cmol.kg^{-1}.

2.8.2 Potassium

Potassium can be present in the soil solution, adsorbed onto the soil (mainly clay) particles and in organic matter (Moore & Bradley, 2018). Primary sources of potassium in soil are from the weathering of minerals such as feldspars and micas (Hillel, 2008). Weathering of minerals makes the nutrients available to plants. In the case of ultisols and oxisols, where prolonging or extreme cases of weathering is detected, potassium is easily leached out of the soil profile and not in reach of the plant's root system (Hillel, 2008). Layered-aluminosilicate clay minerals have the capacity to adsorb potassium to the inside of their layered crystal lattices, inducing fixation of potassium ions, making it unavailable to the plant (Hillel, 2008). Additional potassium, in the form of fertilizer, will have to be added to these types of soils when being utilized as agricultural soils (FERTASA, 2016; Moore & Bradley, 2018).

As a primary macronutrient, potassium has numerous important functions to fill as a plant nutrient. It aids essential enzymes, helps plants with drought tolerance by increasing water use efficiency, disease resistance and stalk rigidity as well as helping plants with nitrogen uptake (Hillel, 2008; FERTASA, 2016; Moore & Bradley, 2018). Crops that require large amounts of potassium, such as bananas and potatoes, may show symptoms of deficiency more quickly than other crops. Deficiencies include chlorosis and shedding of older leaves (Hillel, 2008; FERTASA, 2016; Moore & Bradley, 2018). Desirable proportions of potassium for plants is between 1 – 5% of the CEC (Hazelton & Murphy, 2016).

Potassium deficiencies are far less common in Southern Africa than nitrogen or phosphorus; however, thresholds for potassium are much higher and South African staple crops such as maize, also require high potassium inputs (FERTASA, 2016). Past studies utilizing MIR as analysis method for predicting potassium concentrations ranged from poor (0.28 R^2) (Ji *et al.*, 2016) to sufficient (0.59 – 0.7 R^2) accuracy (Rossel *et al.*, 2008; Gates, 2018; Haghi *et al.*, 2021). Prediction accuracies of exchangeable potassium are variable and are displayed in Table 2-3.

Table 2-3: Results produced from previous studies for the MIR prediction of exchangeable potassium

Reference	R^2	RMSE	RPD	Country
(Gates, 2018)	0.83	0.49 cmol/kg	2.48	USA
(Ji *et al.*, 2016)	0.28	39.2 mg/kg	1.16	Quebec, Canada
(Rossel *et al.*, 2008)	0.59	2.3 mmol/kg	1.48	Australia
(Haghi *et al.*, 2021)	0.84	0.23 cmol/kg	2.41	Scotland

2.8.3 Sodium

Although sodium is not found in the essential plant nutrients list, it is important to test its concentration in the soil and irrigation water because of its negative effects on soil and plants. Where Na is in excess relative to Ca and Mg, sodic condition develops. Sodic soils are high in soluble and exchangeable sodium and pose great threat to agricultural land and crop growth (FERTASA, 2016). Excess sodium salt prevents normal metabolism and limits nutrient and water uptake of plants and beneficial soil biota, cause rooting problems and increase pH to unsafe levels, which also inhibits plant uptake of other essential nutrients (Provin & Pitt, 2001; FERTASA, 2016). Furthermore, when a sodic soil dries out, a hard soil crust is the result, which also limits germination (FERTASA, 2016). The high pH caused by excess Na can promote dispersion of clays and result in the dissolution of organic materials leading to so called “black-alkali” soils. Sodium levels below 1% are preferred for most plants’ optimal growth (Hazelton & Murphy, 2016).

Saline soils in South Africa are localised to several soil groups, and occur mainly under arid conditions (Van der Merwe, 1962). A very small percentage of the saline and sodic soils fall in the sodic category (Nell & van Huyssteen, 2018). In South Africa, 3.8% of soils are considered non-alkaline saline-sodic; 6.3% are alkaline saline-sodic and 0.4% can be considered sodic (Nell & van Huyssteen, 2018).

Previous studies found similar accuracies for sodium concentrations tested with MIR spectroscopy, ranging from poor (0.39 R^2) (Rossel *et al.*, 2008) to moderate although insufficient accuracy (0.63 R^2) (Ji *et al.*, 2016; Gates, 2018). Wavebands where sodium is active does not fall into the MIR absorption bands and may contribute to low accuracy of determination as seen in Table 2-4.

Table 2-4: Results produced from previous studies for the MIR prediction of exchangeable sodium

Reference	R^2	RMSE	RPD	Country
(Gates, 2018)	0.5	0.76 cmol/kg	1.42	USA
(Ji *et al.*, 2016)	0.63	3.2 mg/kg	1.58	Canada
(Rossel *et al.*, 2008)	0.39	20.3 mmol/kg	1.23	Australia

2.8.4 Calcium

Animals and plants require Ca in large amounts for healthy and vigorous growth. Apart from plant nutrition, calcium also maintains physical properties in soil and reclaim sodic soils (Norton, 2013). Calcium is the most dominant cation because of its stronger affinity for exchange sites (Table 2-2). Base cation exchange site affinity is also pH dependent as seen in Figure 2.6. In turn improving soil structure and water holding capacity by replacing sodium in these exchange sites (Norton, 2013). Cell walls and membranes are stronger with adequate amounts of calcium, which conversely increase shoot and root growth, whilst supressing a variety of diseases (Norton, 2013; Moore & Bradley, 2018). High calcium concentrations between 65 and 80% of the CEC is optimal for plant growth (Hazelton & Murphy, 2016).

Calcium losses primarily occur when Ca^{2+} leaches as counter ion with NO_3^- from the topsoil or the root zone or when toxic levels of manganese and/or aluminium present, compete with Ca for exchange sites (FERTASA, 2016). Deficiency of calcium usually presents itself as browning or die-back of newly formed leaves or roots (Norton, 2013). Some deficiency markers include browning of leafy vegetables, blossom end rot of tomato, peppers or watermelon, bitter pit of apples and peanuts with empty pods. Calcium toxicity is also possible although rare and may prevent germination or lead to slow growth rates (FERTASA, 2016). An example of calcium toxicity is gold spot in the cell walls of the tomato fruit (De Kreij *et al.*, 1992; Norton, 2013). This problem may also arise when excess Ca antagonises uptake of other nutrients, or in calcareous soils leading to lock-up of certain nutrients rather than a direct Ca toxicity (Wilkinson *et al.*, 1990).

Calcium is the EBC that is the most accurately predicted with R^2 values ranging between 0.73 and 0.94 R^2 when tested with MIR spectroscopy as displayed in Table 2-5 (Rossel *et al.*, 2008; Ji *et al.*, 2016; Gates, 2018; Dangal *et al.*, 2019; Sanderman *et al.*, 2020). Wavebands where calcium is active does not fall into the MIR absorption bands. It is suggested that the accurate predictions of exchangeable calcium by the calibration algorithms are based on soil properties which are active in the MIR waveband region that correlate with exchangeable calcium as discussed in Exchangeable base cation analysis.

Table 2-5: Results produced from previous studies for the MIR prediction of exchangeable calcium

Reference	R^2	RMSE	RPD	Country
(Ji *et al.*, 2016)	0.73	0.50 g/kg	1.88	Canada
(Rossel *et al.*, 2008)	0.84	22.7 mmol/kg	2.51	Australia
(Sanderman *et al.*, 2020)	0.94	6.2 cmol/kg	-	USA
(Dangal *et al.*, 2019)	0.89	6.85 cmol/kg	3	USA
(Gates, 2018)	0.96	1.43 cmol/kg	5.1	USA

2.8.5 Magnesium

In the soil exchange complex, magnesium is generally the second most abundant after calcium, which is dependent on parent material (FERTASA, 2016) and soil pH as seen in Figure 2.6 (Saha, 2014). Magnesium influences multiple metabolic processes, especially with regards to carbon dioxide (CO_2), protein and chlorophyll mechanisms in plants (Cakmak & Yazici, 2010).

Magnesium is located not only on exchange sites on clay surfaces but are also found inside clay minerals. Thus, magnesium can be locked up in chlorite, vermiculite, and montmorillonite clay's internal structure even after intermediate weathering (Cakmak & Yazici, 2010). Even though magnesium plays such an important role in plant health, it has been overlooked by agronomists in the past two decades (Cakmak & Yazici, 2010; Rosanoff *et al.*, 2012; Guo *et al.*, 2016). This was mainly due to the focus on NPK only fertilizers, resulting in potassium inhabiting most of the exchange sites (Cakmak & Yazici, 2010). Magnesium in quantities of 10 – 15% of the total CEC, are recommended for optimal plant growth (Hazelton & Murphy, 2016).

Consequently, growth and yield are severely affected by deficiencies of magnesium leading to abnormal development of important physiological and biochemical processes in plants (Cakmak & Yazici, 2010). In acidic-, sandy- and highly weathered soils, where magnesium is easily leached, an interaction with Al is also of serious concern. Magnesium is needed by plants to release organic acids which can chelate toxic Al ions, deeming it no longer toxic (Yang *et al.*, 2007; Cakmak & Yazici, 2010).

Prediction accuracy when using MIR spectroscopy is sufficient to excellent in testing magnesium levels. Previous studies found R^2 values ranging between 0.66 to 0.89 as seen in Table 2-6 (Rossel *et al.*, 2008; Ji *et al.*, 2016; Gates, 2018; Haghi *et al.*, 2021). Wavebands where magnesium is active does not fall into the MIR absorption bands. It is suggested that the accurate predictions of exchangeable magnesium by the calibration algorithms are based on soil properties which are active in the MIR waveband region that correlate with exchangeable magnesium.

Table 2-6: Results produced from previous studies for the MIR prediction of exchangeable magnesium

Reference	R^2	RMSE	RPD	Country
(Gates, 2018)	0.89	0.45 cmol/kg	3.07	USA
(Ji *et al.*, 2016)	0.66	100 mg/kg	1.7	Quebec, Canada
(Rossel *et al.*, 2008)	0.75	23 mmol/kg	2.03	Australia
(Haghi *et al.*, 2021)	0.83	1.77 cmol/kg	2.42	Scotland

CHAPTER 3 MATERIALS & METHODS

3.1 Introduction

Various steps are required to create an adequate soil spectroscopy algorithm in terms of its representation, size, and generalization ability (Ramirez-Lopez *et al.*, 2013; Ramirez-Lopez *et al.*, 2014). The steps taken to complete the study involves creating a soil spectral database from the soil property database, and its corresponding spectra, and combining of both to create a spectral library which will be used to create calibration algorithms for the prediction of the four major exchangeable base cations (Ca, Mg, K, and Na).

3.2 Study area

Plots of the samples received from NWK are displayed on the map in Figure 3.1. The samples were all collected within the Western 26 Highveld grain producing area, located at 26°27'24.3", S 26°04'58.0"E in the North-West province, South Africa. The area is an arid to semi-arid region, found at an average height of 1,769 m above sea-level on average (USGS, 2022). The area has an annual temperature and rainfall of 19.7°C and 567 mm, respectively (Malherbe *et al.*, 2016). The site map for GWK samples were not made available.

The Kalahari group covers the westerns portions of the study area consisting of calc-conglomerate, mudstone, gritstone, siliceous/calcareous sandstone, silcrete, diatomaceous limestone and calcrete. The Klipriviersberg group, Malmani subgroup and Rietgat formations covers the eastern portions of the study area consisting of dolomite, chert, shale, limestone, quartzite, basalts, andesites, tuff, agglomerate, gneiss, and granites (Council for Geoscience, 2019, (Kock, 2022). Broad land types 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, (Kock, 2022). Samples were all collected on commercial farmland hosting *Zea Mays L.* and *Helianthus annus L* as annual crops.

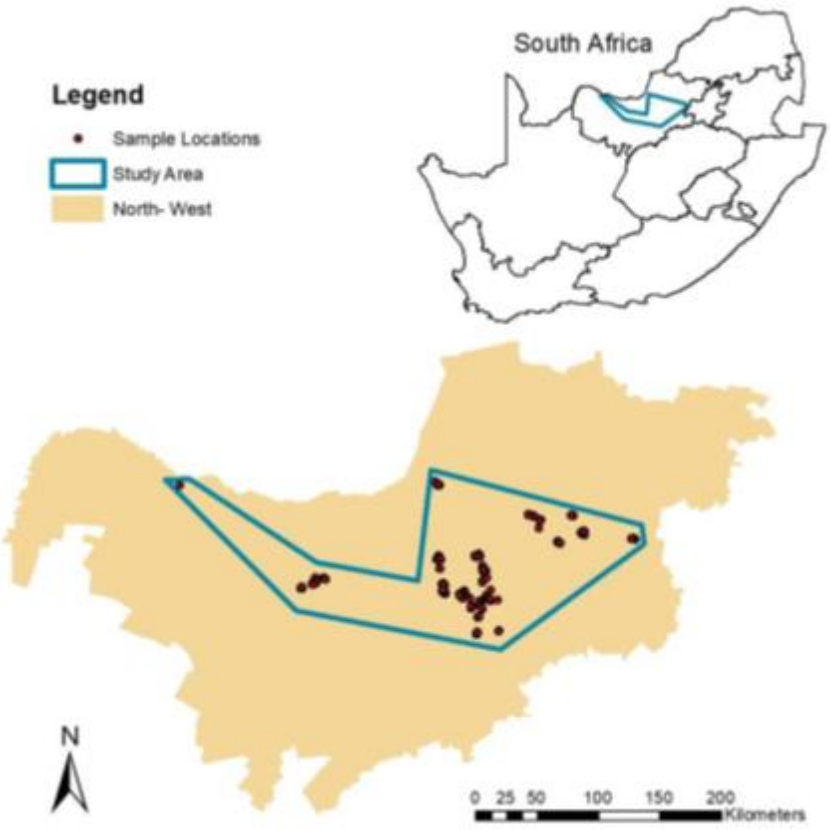


Figure 3.1: Map showing the location of the collected samples from NWK in the North-West province, South Africa (Kock, 2022).

3.3 Soil property database

The soil property database was created with topsoil sample information received from Noordwes Kooperasie (NWK) and Griekwaland Wes Korporatief (GWK), that was collected and underwent chemical testing by the two co-operations respectively. NWK and GWK provided a soil property database of 4,393 and 175 samples, respectively. From the 4,393 samples, 1,500 was selected with conditioned Latin Hypercube Sampling (cLHS) developed by Minasny and McBratney (2006). Kock (2022) ran the selection with cLHS by using pH, phosphorus (P) and effective cation exchange capacity (T-value) as covariates and the same samples were used in this study. The 1,500 NWK soil samples were then made available by NviroTek Central (Hartbeespoort, North-West) for collection. The 175 soil samples from GWK were made available by Dries Bloem (Potchefstroom, North-West) for collection. The collected samples were reduced to 900 using cLHS, also with pH, Ca, and T-value as covariates. GWK made a total of 175 samples available, of which 100 was selected for analysis using the cLHS method with pH, P, and T-value as covariates. Therefore, a total of 1,000 samples were selected for spectroscopic analysis. Of these, 21 samples were removed due to incorrect naming of samples, resulting in 979 total samples. The following soil properties were determined for NWK samples and GWK samples:

exchangeable- calcium, magnesium, potassium, and sodium, with the ammonium acetate standard method.

3.4 Soil spectral database

The samples from the soil property database were ground with a Retsch Mortar Grinder RM 200 (Retsch GmbH, 2021), sieved with a 53-micron sieve, and scanned to obtain the MIR spectra with the Bruker Alpha II (Figure 3.2) with DRIFT module instrumentation. Baseline scans were required by the software every hour and was done with a gold-plated sample. A spectral range of 4,000 - 600 cm^{-1} and a resolution of 2 cm^{-1} was used. A spectral database was created with the MIR spectra. Roughly 2 g of sample was loaded into the sample holder and levelled to provide consistent results from scanning. The sample holder was then placed in the DRIFT module and the instrument was gently closed to avoid spilling any soil from the sample holder. Bruker (2021) provided OPUS Base package software integrated with the spectrometer. Soil spectral data was created from scanning each sample once with 36 iterations per sample to display the spectra as an .opus file format. The .opus file format was then converted with Spectrograph 1.2 software to a .csv format.

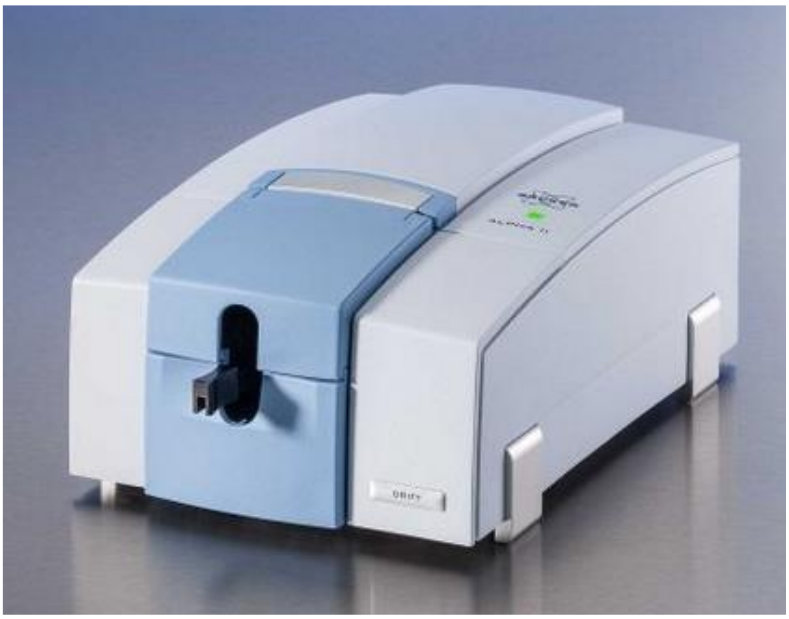


Figure 3.2: Bruker Alpha II with FT-IR DRIFT module attached (Bruker, 2021).

3.5 Soil spectral library

A soil spectral library was created which consist of information from the soil spectral database and spectral data. Specifically, the exchangeable base cation concentrations of each sample as well as the MIR spectra associated with each sample. The data from said databases were merged in Excel and read into R Studio as a .csv file.

3.6 Creating calibration algorithms

The merging of the soil spectral database and the spectral data, and creation of the calibration algorithm and prediction models for the exchangeable base cation concentrations, were done using R Software (Core, 2020) in R Studio user interface software. The cubist-, plsr-, random forest functions included in the Cubist- (Kuhn *et al.*, 2022), pls- (Mevik & Wehrens, 2015), and randomForest (Breiman, 2001) packages, respectively, was used for the soil property predictions. The soil spectral library was divided into a training and evaluation dataset in R at 75% and 25% (or 743 and 236) samples, respectively, using a randomize and split function. The functions used to set up the model trees and predict the exchangeable base cations are displayed in Appendix 2, with potassium used as the example.

3.7 Evaluation

The accuracy of the calibration algorithms was tested on the evaluation dataset in R Studio. Statistical analysis includes the R^2, RMSE, RPD, RPIQ and bias. Information regarding the formulas used for each can be found in the subheading “Algorithm validation”. The scripts used for each the training and prediction of potassium as an example, can be found in Appendix 4.

The accuracy of the predictive models on the evaluation dataset is visually represented by scatterplots of measured vs predicted values. They were calculated using the scripts in Appendix 4. The script for the scatterplot used as an example, is the accuracy of cubist prediction model on potassium.

CHAPTER 4 RESULTS AND DISCUSSION

4.1 Soil property database (SPD)

The soil property database contains valuable information regarding the specified soil properties. The soil property database was created from data obtained from NWK and GWK for four soil properties used for calibration in this study. The four soil properties and their statistical analysis is presented in Table 4.1 and represent the 1,675 samples in the soil property database.

Table 4-1: Descriptive statistics of base cation concentration (mg.kg^{-1}) of 1675 samples used of the creation of the soil property database

Property	Mean	Median	Std. dev	Min	Max	Range
Potassium	169	162	85	17	789	772
Sodium	10.6	7.4	33.1	0.5	1303	1302
Calcium	520	458	293	56	2505	2449
Magnesium	129	99.0	105	14	1128	1114

Std. dev = Standard deviation, Min = Minimum, Max = Maximum, Units for potassium, sodium, calcium, and magnesium = mg.kg^{-1}.

The mean for exchangeable K, Na, Ca, and Mg is 169, 10.6, 520- and 129 mg.kg^{-1}; respectively. The medians are slightly lower than the mean for all the base cations at 162-, 7.4-, 458-, and 99 mg.kg^{-1} for exchangeable K, Na, Ca, and Mg; respectively. The standard deviations for exchangeable K and Ca are lower than the median at 84 and 293 mg.kg^{-1} respectively; compared to the standard deviations of Mg (105 mg.kg^{-1}); which is slightly higher than the median (99 mg.kg^{-1}). The standard deviation of Na (33.1 mg.kg^{-1}) is much higher than the median (7.4 mg.kg^{-1}), which speaks of possible outliers. The Na maximum seems excessive and may be an error or outlier.

4.2 Soil Spectral database (SSD)

The soil spectral database is compiled from the data from the soil property database selected with conditioned Latin Hypercube Sampling (cLHS). The SSD is being compared to the SPD to ensure that they SSD is a good representation of the SPD so that basing predictions on the SSD is sufficient for the calibration models. The mean, median and standard deviation for K is slightly lower in the soil spectral dataset at 147, 134 and 74.4 mg.kg^{-1}, respectively; showing the samples removed from the soil property database were overall larger values with respect to the exchangeable base cations. The minimum is slightly higher (25 mg.kg^{-1}) and the maximum slightly lower (526 mg.kg^{-1}) than for the full SPD which indicates outliers from both extremes were removed. Box-and-whiskers plots of the SSD of K (Figure 4.1) greatly resembles the SPD of K

(Figure 4.1), except for the three uppermost values from the soil property database removed as outliers. This indicates the SSD of K is a satisfactory representation of the SPD of K. Positive skewness indicate that the mean is larger than the median for all EBCs and are right skewed. Kurtosis on all the EBCs indicate a fat-tailed distribution and are considered leptokurtic.

Table 4-2: Descriptive statistics for 979 samples in the soil spectral database for the four exchangeable base cations of interest

Property	Mean	Median	Std. Dev	Min	Max	Range	Skewness	Kurtosis
Potassium	147	134	74	25	526	501	1.01	1.53
Sodium	11.2	9	7.4	1	85	84	4.4	32.3
Calcium	464	408	258	45	2243	2198	1.53	4.59
Magnesium	114	98	73	16	786	770	2.21	9.4

All units are in $mg.kg^{-1}$

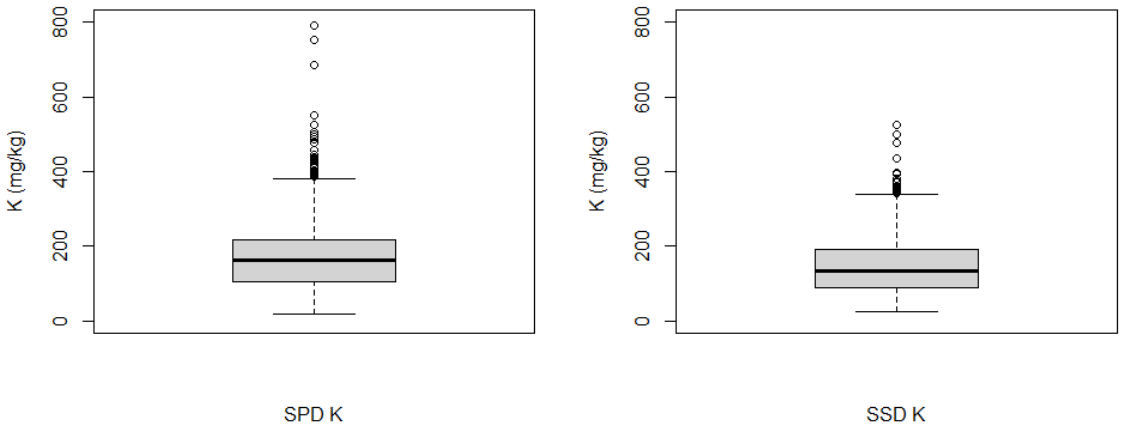


Figure 4.1: Box and whiskers plot of the soil property database of potassium (left) and the soil spectral database of potassium (right).

The mean, median and standard deviation of sodium are 11.2-, 9- and 7.4 $mg.kg^{-1}$ respectively. The minimum value is slightly higher in the SSD with a value of 1 $mg.kg^{-1}$. The maximum value however is extremely lower at 85 $mg.kg^{-1}$ in comparison to 1303 $mg.kg^{-1}$ in the SPD. When reviewing the box and whiskers plot for the SPD (Figure 4.2), it is evident that the maximum value of 1303 $mg.kg^{-1}$ is a definite outlier and may be a faulty value.

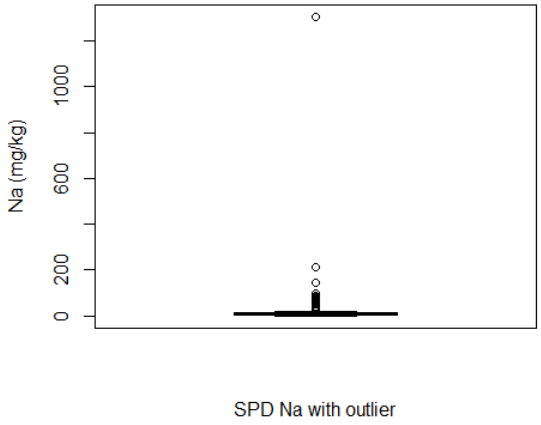


Figure 4.2: Box-and-whiskers plot of sodium showing the faulty maximum value to be removed as outlier.

Excluding the anomalous maximum value from the database and comparing the box-and-whiskers plots of the SSD and the SPD, it shows that the SSD sufficiently represents the larger dataset (Figure 4.3).

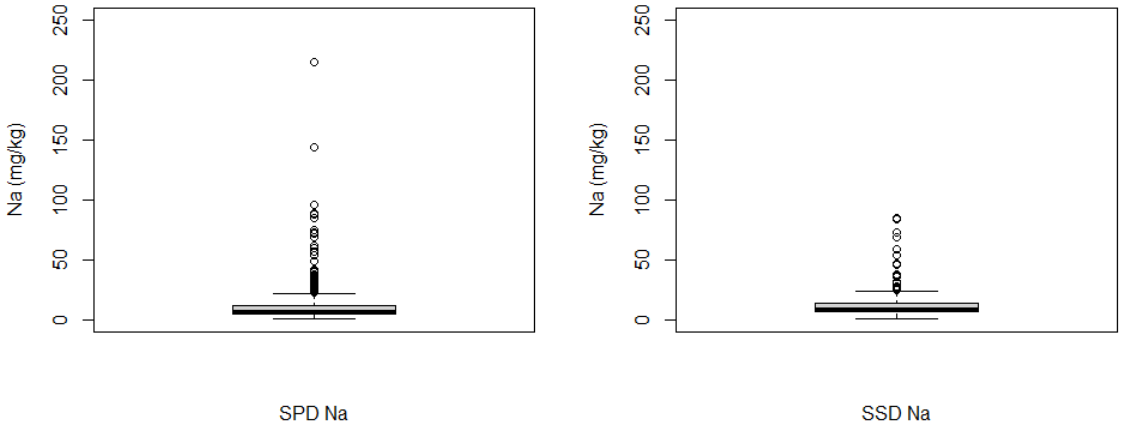


Figure 4.3: Box-and-whiskers plot of the soil property database of sodium (left) and the soil spectral database of sodium (right).

Values for calcium in the SSD was closely related to the SPD with values of 464-, 408- and 257 $mg.kg^{-1}$ for the mean, median and standard deviation respectively. Minimum and maximum values of 45- and 2,243 $mg.kg^{-1}$, respectively, from the SSD is also a good representation of the SSD values. The box-and-whiskers plots for both the SPD and SSD of calcium confirms a good representation of the SSD on the SPD for calcium, as seen in Figure 4.4.

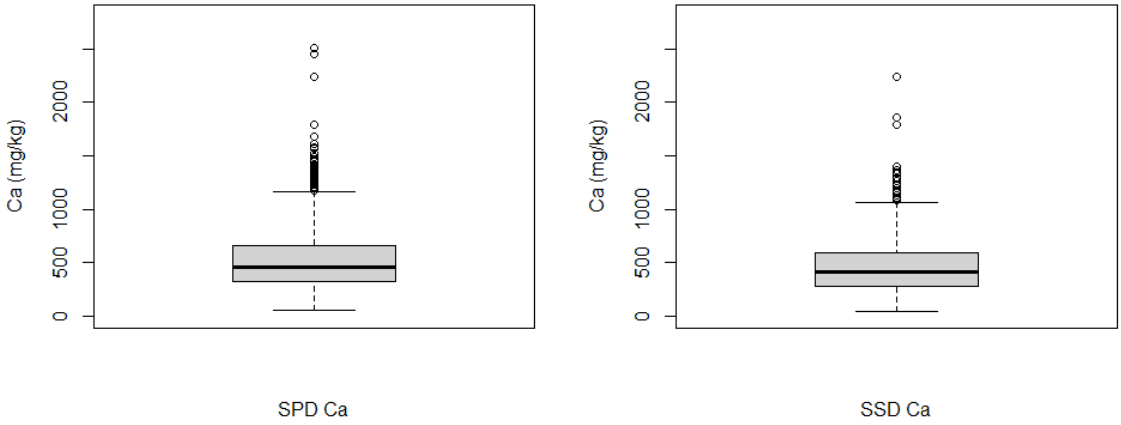


Figure 4.4: Box-and-whiskers plot of the soil property database of calcium (left) and the soil spectral database of calcium (right).

The mean, median, standard deviation, and minimum values of magnesium in the SSD is 114-, 98, 73.5- and 16 mg.kg^{-1}, respectively. This closely resembled the values in the SPD for magnesium. However, the maximum value of magnesium in the SSD (786 mg.kg^{-1}) is moderately lower than the value of 1,128 mg.kg^{-1}. When reviewing the box-and-whiskers plots (Figure 4.5) of both the SSD and SPD of magnesium, it is evident that the larger SSD of magnesium is a good representation of the larger SPD for magnesium.

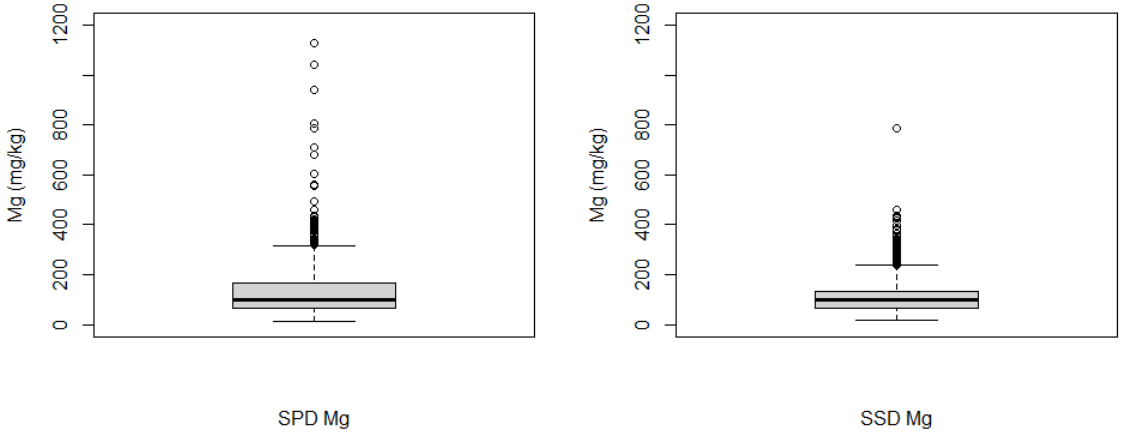


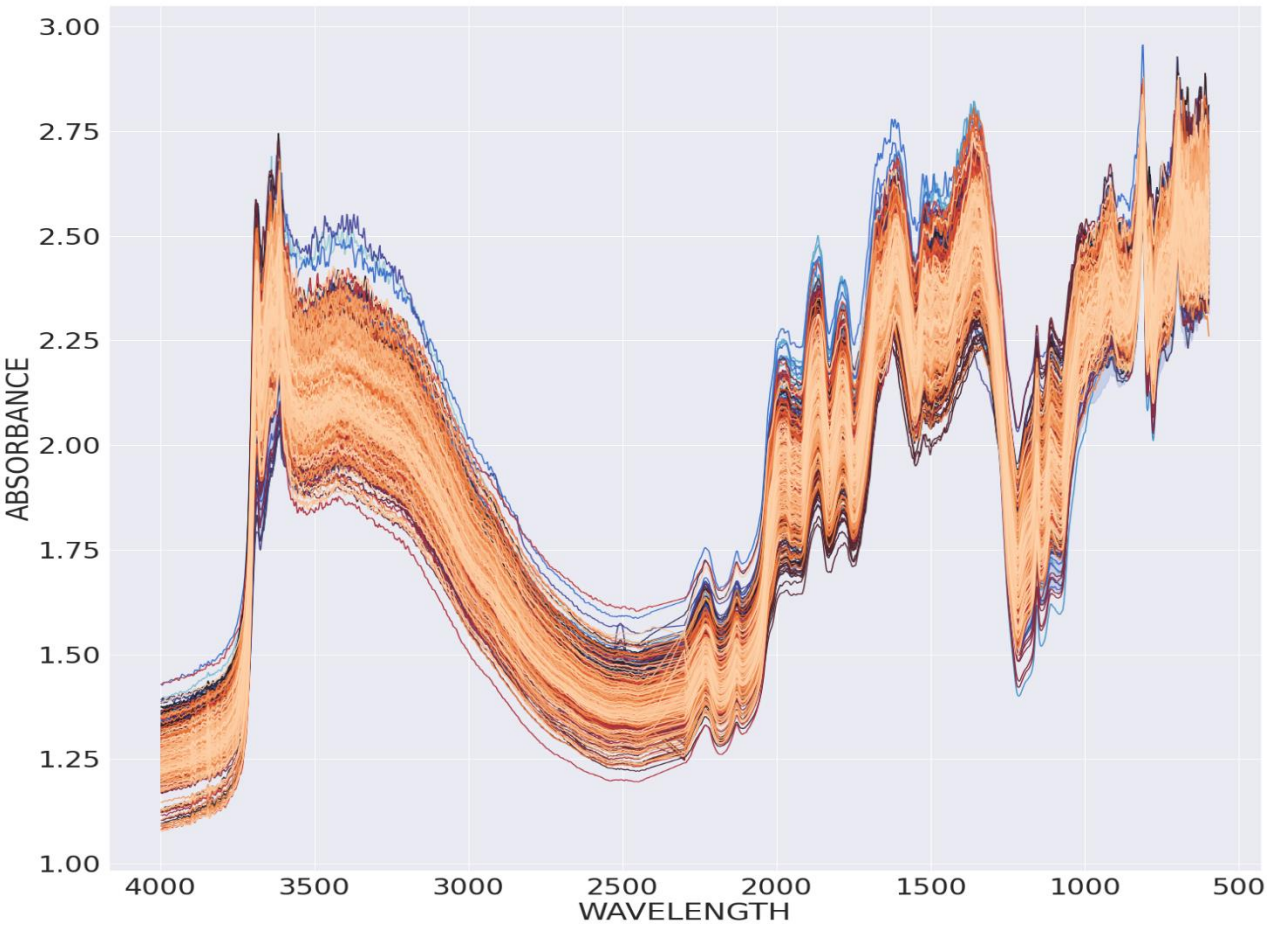
Figure 4.5: Box-and-whiskers plot of the soil property database of magnesium (left) and the soil spectral database of magnesium (right).

4.3 Soil Spectral Library

The mid-infrared spectra obtained from scanning the samples in the SSD, shows similar features as seen in Figure 4.6. A cluster of strong peaks are visible in the 3,700 – 3,600 cm^{-1} range which represent OH stretch bonds. There are two peaks in the range of 2,300 – 2,100 cm^{-1} which

represent either triple CN or triple CC stretch bonds. Another cluster of peaks between 2,000 and 1,300 cm $^{-1}$ represent double CO bonds as esters, ketones, or amides, or double CC alkene and aromatic bonds. Two peaks are present in the region of 1,200 – 1,100cm $^{-1}$ are C-O-C stretch bonds and C-OH stretch bonds. The cluster is evident in the range of 950 – 500 cm $^{-1}$ are possible C-Cl, C-Br, and CI, bonds.

Figure 4.6: All spectra obtained from scanning the 979 soil samples in the soil spectral database using the



Bruker Alpha II with MIR DRIFT Module showing similar features across the board.

Data in the range of 2,400 – 2,300 cm^{-1} was removed due to noise that was related to the instrument. The noise was due to moisture build-up in the detection chamber of the Bruker Alpha II MIR instrument as specified by Bruker. The noise for unscanned samples were rectified by drying the silicone moisture traps at 105 °C for 4 hours. Due to limited time with the instrument, the samples were not rescanned.

4.4 Algorithm validation

4.4.1 Potassium

All the MIR models were insufficient in predicting the exchangeable potassium values as seen in Table 4-4. The Coefficient of determination values for K were the lowest for RF at 0.36, whereas Cubist and PLSR showed slightly better correlation at 0.41 and 0.42 respectively.

RMSE was calculated at 59, 61 and 59 $mg.kg^{-1}$ for Cubist, RF and PLS models respectively, making PLS a slightly better predictive model for K, although none were sufficient. RPD also showed better performance for the PLS model at 1.29 followed closely by Cubist (1.28) and RF (1.25). RPIQ was calculated at 1.61, 1.57, and 1.61 for Cubist, RF and PLS models respectively. Bias was calculated at -0.108, -0.202, -0.132 for Cubist, RF and PLS models respectively.

Gates (2018); Haghi *et al.* (2021); and Rossel *et al.* (2008) achieved better results with the prediction of exchangeable potassium with PLSR as prediction model; with an R^2 value of 0.83, 0.7, and 0.59 respectively; and an RPD value of 2.48, 1.83, and 1.48, respectively. Our model performed better than that of Ji *et al.* (2016) with an R^2 and RPD of 0.28, and 1.16, respectively. Haghi *et al.* (2021) had slightly better results with their cubist model with values of 0.84, and 2.41, for R^2 and RPD; respectively, whilst our cubist model did slightly worse in comparison to the PLSR model in the same study. No studies on the prediction of exchangeable potassium with RF as prediction model and MIR spectroscopy as scanning methods could be found. Our PLSR model outperformed Johnson *et al.* (2019) which showed results of 0.54 R^2 and 0.97 RPIQ.

The scatterplots for all the prediction models of exchangeable potassium are displayed in Figure 4.7, and confirms the poor prediction value of our model. The poor performance is probably not due to the collection of smaller sample size with cLHS as the dataset showed good resemblance in Figure 4.1. It may be possible to achieve better results with spectral pre-processing. However, Kock (2022) did not achieve any significant improvement in predicting pH, T-value, and CEC. A better explanation of the poor results is K not being spectrally active in the MIR wavelength region. The calibration model then uses spectral peaks of properties correlated with the property of interest to predict the property value. Exchangeable potassium has a moderate to high correlation with clay content (Kundu *et al.*, 2014). Gates (2018) shows the correlation of clay percentage with the EBC found in their study. Ng *et al.* (2022) argues that it is the clay particles or soil organic matter that are spectrally active in the MIR region as discussed in Exchangeable base cation analysis.

Table 4-3: The calibration results for exchangeable base cations' spectral data from Cubist, PLSR and RF prediction models.

	R^2			RMSE ($mg.kg^{-1}$)			RPD			RPIQ			Bias		
Property	PLS	RF	Cubist	PLS	RF	Cubist	PLS	RF	Cubist	PLS	RF	Cubist	PLS	RF	Cubist
K	0.60	0.94	0.63	46	24	45	1.58	3.04	1.62	2.16	4.17	2.22	-0.111	-0.090	-0.096
Na	0.45	0.91	0.45	5.5	3.1	3.2	1.35	2.43	1.32	1.28	2.30	1.25	-0.155	-0.117	-0.087
Ca	0.85	0.96	0.83	97	61	103	2.62	4.13	2.46	3.19	5.03	2.99	-0.021	-0.044	-0.030
Mg	0.81	0.94	0.77	31	20	35	2.27	3.48	2.05	2.10	3.21	1.89	-0.050	-0.059	-0.061

Table 4-4: The validation results for exchangeable base cations' spectral data from Cubist, PLSR and RF prediction models.

	R^2			RMSE ($mg.kg^{-1}$)			RPD			RPIQ			Bias		
Property	PLS	RF	Cubist	PLS	RF	Cubist	PLS	RF	Cubist	PLS	RF	Cubist	PLS	RF	Cubist
K	0.42	0.36	0.41	59	61	59	1.29	1.25	1.28	1.61	1.57	1.61	-0.108	-0.202	-0.132
Na	0.11	0.15	0.29	7.44	6.77	6.45	0.99	1.09	1.14	0.84	0.92	0.97	-0.225	-0.249	-0.113
Ca	0.75	0.68	0.77	135	155	128	1.99	1.73	2.09	2.48	2.17	2.6	0.0147	-0.108	0.023
Mg	0.70	0.66	0.75	41	46	40	1.82	1.64	1.89	1.76	1.6	1.83	-0.048	-0.131	-0.035

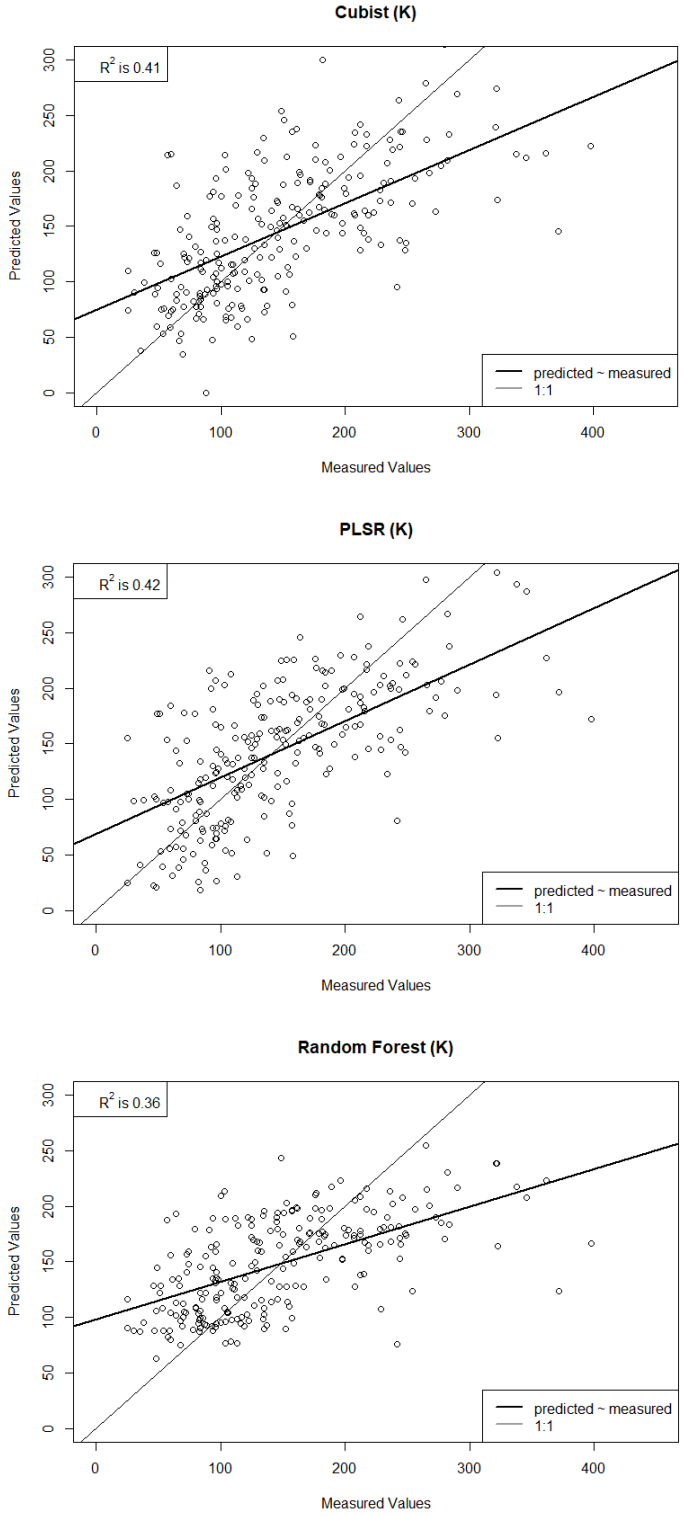


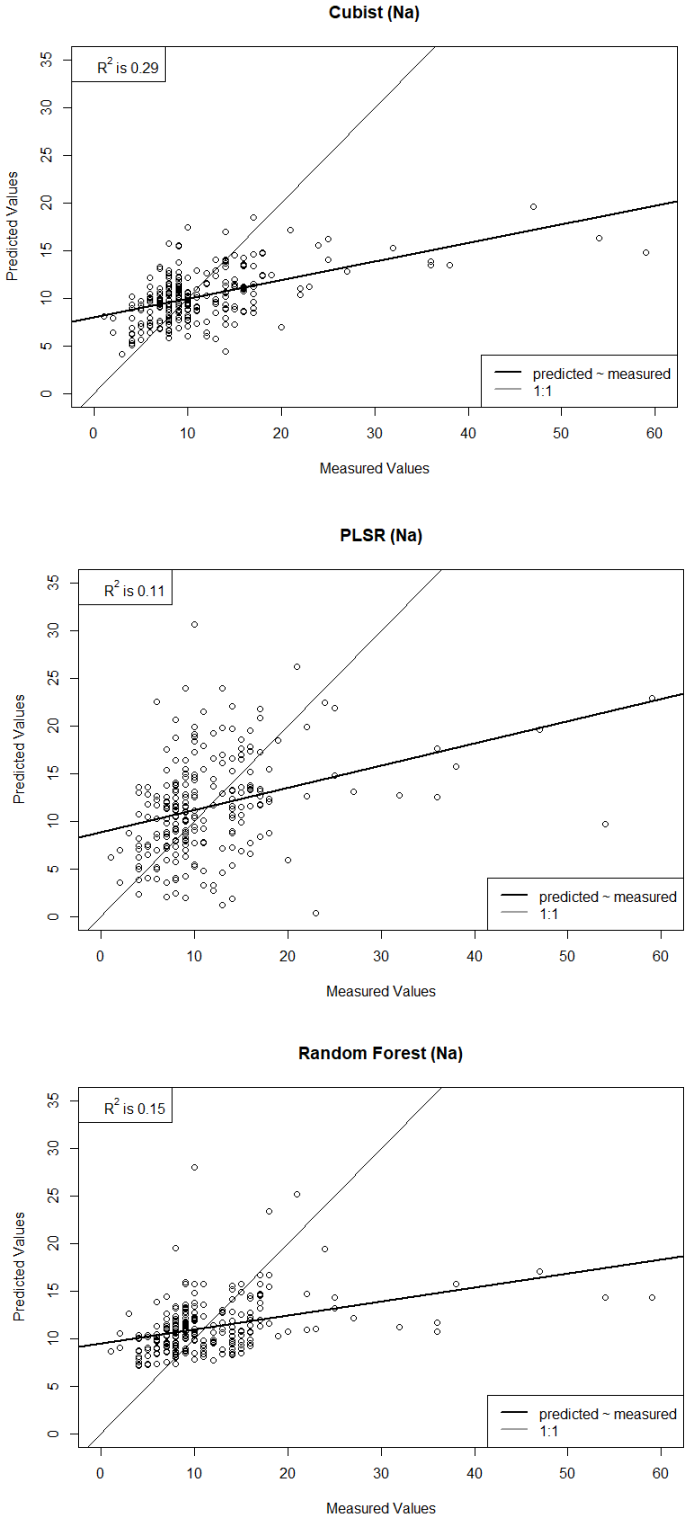
Figure 4.7: Scatter plots and 1:1 line of the validation dataset predictions for exchangeable potassium for the Cubist, PLSR and RF models, respectively.

4.4.2 Sodium

Sodium ranged from 1 to 85 mg.kg^{-1} and had a mean of 11 mg.kg^{-1}. Cubist performed better than PLSR and RF, but from a predictive capability perspective it was still of little meaningful use. The R^2 values for Na was 0.29, 0.15 and 0.11 for Cubist, RF and PLS respectively. RMSE was recorded at 6.45-; 6.77-, and 7.44 mg.kg^{-1} for Cubist, RF and PLS models respectively. These values seem low, but with a mean of 11 mg.kg^{-1}, the error associated with the predicted values was high relative to the mean value. RPD also showed better performance by the Cubist model at 1.14 although this is still poor. RF and PLS showed very poor predictive capability with an RPD of 1.09 and 0.99 respectively. RPIQ was calculated at 0.84, 0.92, and 0.97 for Cubist, RF and PLS models respectively. Bias was calculated at -0.225, -0.249, -0.113 for Cubist, RF and PLS models respectively.

Gates (2018) mention that better prediction accuracy can be seen with properties that have large responses to infrared waves, in contrast to ones that have smaller ranges such as exchangeable sodium (Na^+). Previous studies also indicated worse predictive capabilities of exchangeable sodium, compared to the other EBCs. Gates (2018), Ji *et al.* (2016), and Rossel *et al.* (2008), found R^2 values of 0.5, 0.63, and 0.39, respectively; compared to 0.11 in this study; with PLSR as calibration model and MIR as spectroscopy method. However, cubist did have a much better R^2 at 0.29, but is still very poor in relation to the other studies. Our PLSR model had a slightly lower R^2 at 0.29 compared to 0.32 R^2 but an RPIQ of 0.84 compared to 0.60 (Johnson *et al.*, 2019).

The scatterplots for all the prediction models of exchangeable sodium are displayed in Figure 4.8, shows some correlation, however, there are a considerable number of outliers, and they are also dispersed far from the 1:1 line. Exchangeable sodium not being spectral in the MIR waveband region is the probable cause for the low prediction accuracy. Being the base cation with the worst predictive capability by MIR spectroscopy argues that other soil properties correlated with exchangeable sodium, also have low activity in the MIR waveband region. Sodium is not correlated with clay percentage in a study by Gates (2018), as seen in Figure 2.9. Exchangeable sodium, being in low quantities, may also have an effect with regards to its low predictive accuracy with MIR spectroscopy.



7

Figure 4.8: Scatter plots and 1:1 line of the validation dataset predictions for exchangeable sodium for the Cubist, PLSR and RF models, respectively.

4.4.3 Calcium

Calcium is the EBC with the highest predictive potential with MIR spectroscopy. Cubist was the best predictive model while PLS also showed potential. A minimum amount of 45 $mg.kg^{-1}$ and a maximum amount of 2,243 $mg.kg^{-1}$ was recorded, with a mean of 464 $mg.kg^{-1}$. R^2 values of 0.77 and 0.75 for Cubist and PLS respectively, showed that these models are acceptable for the prediction of Ca, whereas RF showed adequate predictive value with an R^2 value of 0.68. RMSE was recorded at 128-, 155- and 135 $mg.kg^{-1}$ for Cubist, RF and PLS models respectively. RPD values are also good for Cubist and PLS at 2.09 and 1.99 respectively, whereas RF had an adequate value of 1.73. RPIQ was calculated at 2.48, 2.17, and 2.6 for Cubist, RF and PLS models respectively. Bias was calculated at 0.0147, -0.108, 0.023 for Cubist, RF and PLS models respectively.

Comparing these results to the studies in Table 2-5, similar predictive capabilities as the study by Ji *et al.* (2016) were obtained, whereas the studies by Gates (2018), Rossel *et al.* (2008), Sanderman *et al.* (2020), and Dangal *et al.* (2019) reported better prediction of exchangeable calcium concentrations in soil with MIR spectroscopy with R^2 values of 0.96, 0.84, 0.94, and 0.89, respectively. Sanderman *et al.* (2020) did not calculate RPD values, but Gates (2018), Rossel *et al.* (2008), and Dangal *et al.* (2019) observed values of 5.1, 2.51, and 3; respectively, by using PLSR as prediction model. Dangal *et al.* (2019) also used cubist and RF as prediction models and in his study, and the results were R^2 values of 0.95 and 0.93, and RPD values of 4.7 and 3.8, respectively, making cubist the most accurate prediction model in their study. Our PLSR model outperformed (Johnson *et al.*, 2019) which showed results of 0.73 R^2 and 1.52 RPIQ.

The scatterplots for all the prediction models of exchangeable calcium are displayed in Figure 4.9. The scatterplots confirm the predictability of exchangeable calcium percentage with MIR spectroscopy. Exchangeable calcium not being spectrally active in the MIR waveband regions, suggest that it is highly correlated with a soil property that is spectrally active in the waveband region. It may be a mineral in the clay particles, that is spectrally active in the MIR waveband region (Ng *et al.*, 2022). According to Gates (2018), exchangeable calcium shows good correlation with clay percentage as seen in Figure 2.9. It is possible that the clay content contains mineral matter that is correlated with exchangeable calcium.

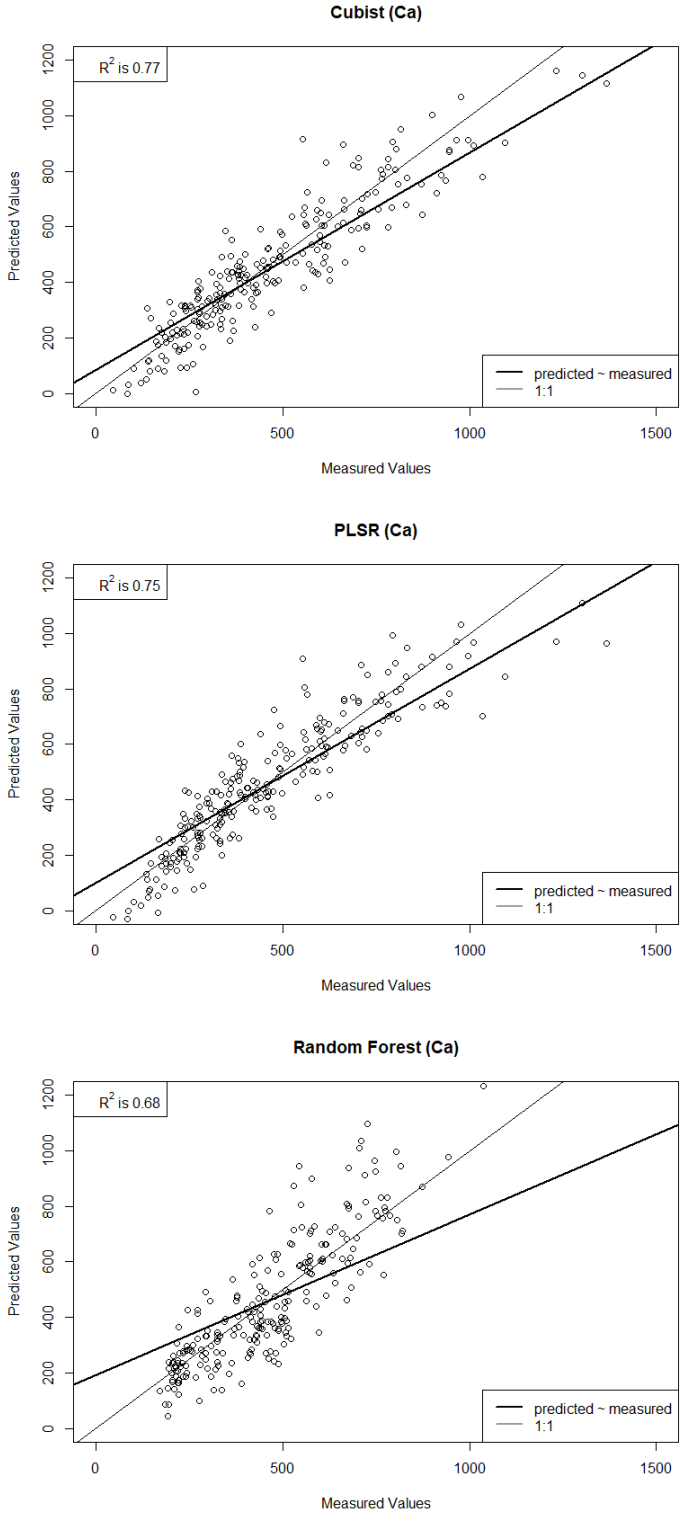


Figure 4.9: Scatter plots and 1:1 line of the validation dataset predictions for exchangeable calcium for the Cubist, PLSR and RF models, respectively.

4.4.4 Magnesium

Magnesium concentrations, like calcium, were also best predicted by the Cubist model, followed by PLS and RF only showing adequate vales. Magnesium ranged between 16 and 786 $mg.kg^{-1}$ with a mean of 114 $mg.kg^{-1}$. R^2 values of 0.75, 0.70 and 0.66 for Cubist, PLS and RF respectively. RMSE was recorded at 40-, 46-, and 41 $mg.kg^{-1}$ for Cubist, RF and PLS models respectively. RPD values where adequate for RF at 1.64 whilst Cubist and PLS showed substantial predictive value at 1.89 and 1.82 respectively. RPIQ was calculated at 1.76, 1.6, and 1.83 for Cubist, RF and PLS models respectively. Bias was calculated at -0.048, -0.131, 0.035 for Cubist, RF and PLS models respectively.

No studies using MIR spectroscopy as scanning method and RF as calibration model could be found. From the studies that used PLSR, Gates (2018) had the best results with R^2 and RPD values of 0.89, and 3.07; respectively, and was the only study that did substantially better than this study regarding PLSR as calibration algorithm. Haghi *et al.* (2021), Ji *et al.* (2016), and Rossel *et al.* (2008) achieved R^2 values of 0.73, 0.66, and 0.75, and RPD values of 1.96, and 1.7, and 2.03, respectively. Haghi *et al.* (2021) achieved R^2, and RPD values of 0.83, and 2.42, respectively. Our PLSR model outperformed Johnson *et al.* (2019), which showed results of 0.77 R^2 and 1.36 RPIQ.

The scatterplots for all the prediction models of exchangeable magnesium are displayed in Figure 4.10. The scatterplots show good correlation with higher values being less predictable as the outliers skew to the right. Exchangeable magnesium is not active in the MIR waveband region and the moderate predictable accuracy suggest that a soil property with moderate correlation reside in the MIR waveband region (Ng *et al.*, 2022). Gates (2018) found good correlation of clay percentage with exchangeable magnesium as seen in Figure 2.9. It may be possible that the clay content contains mineral matter that correlates with exchangeable magnesium.

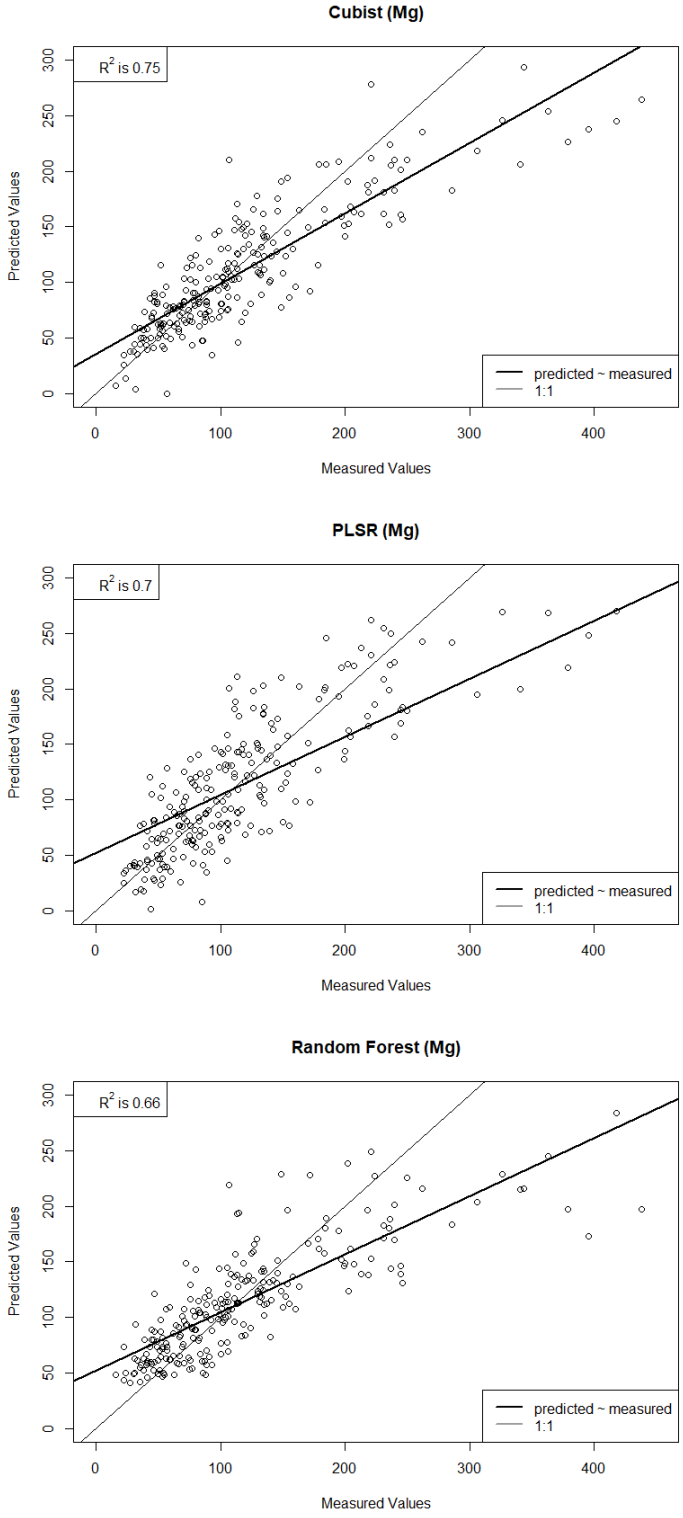


Figure 4.10: Scatter plots and 1:1 line of the validation dataset predictions for exchangeable magnesium for the Cubist, PLSR and RF models, respectively.

4.5 Conclusion

Exchangeable calcium- and magnesium showed promising results with all prediction methods, whereas the prediction of exchangeable K and exchangeable Na was both poor. The same conclusion was made by all the other studies mentioned. Cubist showed the best results.

The results from this study suggest that the higher prediction accuracy for exchangeable calcium and magnesium, is attributed to them correlated to soil properties active in the MIR region. These soil properties are most likely a correlation with minerals within the clay particles or soil organic matter. Furthermore, exchangeable calcium seems to show even higher prediction accuracy in comparison to exchangeable magnesium.

CONCLUSION AND RECOMMENDATIONS

This study aimed to create models from calibration algorithms for the prediction of the exchangeable base cations; calcium, magnesium, potassium, and sodium; with mid-infrared spectroscopy as scanning method, for soils from the North-West province, South Africa. The objectives of the study were met with varying results for the different soil properties as well as the different calibration algorithms.

The first and second objectives were met when creating a soil property database that included 1,675 samples, and then reducing the samples to 979 by using cLHS, followed by the scanning of each sample with mid-infrared spectroscopy to gather spectral information, of each sample. The third objective was met by combining the soil spectral database with the spectral data for each sample, and exporting the data in the correct format, which represented the soil spectral library. The fourth objective was met by importing the soil spectral library into R and creating prediction models from calibration algorithms (partial least square regression, cubist, and random forest) in R Studio, which was used to predict the values of the evaluation dataset for each exchangeable base cation; calcium, magnesium, potassium, and sodium; totalling twelve calibration algorithms. The fifth and final objective was met when evaluating the accuracy of the predictions with coefficient of determination, root mean square error, and ratio of performance to deviation.

The hypothesis tested stated that MIR can be used to aid in the creation of acceptable calibration algorithms to predict EBC concentrations with machine learning techniques. This hypothesis was only partially accepted, as models for exchangeable calcium and magnesium was sufficiently calibrated, while exchangeable potassium and sodium were not. The models created for Ca and Mg are not suitable for single soil sample prediction but may be useful for variability mapping for precision agriculture applications. Cubist was the best overall calibration model, but not by a considerable margin, for exchangeable sodium-, calcium-, and magnesium concentrations. Exchangeable potassium concentrations were slightly better predicted by partial least square regression, also not by considerable margin. Exchangeable calcium was the most accurately predicted soil property followed closely by exchangeable magnesium, which both show great potential to use MIR spectroscopy as a viable soil analysis method. Exchangeable potassium prediction was poor and cannot be used sufficiently for exchangeable potassium analysis for soils from the North-West province, South Africa. Exchangeable sodium showed very poor predictability by the models created, which correlated with other studies, and cannot be used as a viable analysis method.

Base cations are not spectrally active but rely on properties that are spectrally active that are correlated with base cations, to make predictions. The sample size used in this study is small compared to other studies, which in turn produce better accuracy such as Dangal *et al.* (2019) and Gates (2018).

The following recommendations can be made to increase prediction accuracies with the models to make spectroscopy a viable soil analysis technique in the future, especially if large scale commercial use of the analysis methods is of interest.

- The increase in sample numbers per area to ensure spectral variation of soils.
- Use additional machine learning algorithms to create calibration models such as artificial neural network and deep learning.
- The incorporation of different scanning methods to include more waveband regions, which will increase prediction accuracies of different soil properties, such as near-infrared, far-infrared, and ultraviolet.
- Spectral pre-processing, either basic, or waveband specific, can increase model accuracy.

BIBLIOGRAPHY

Adamchuk, V.I., Hummel, J.W., Morgan, M. & Upadhyaya, S. 2004. On-the-go soil sensors for precision agriculture. *Computers and electronics in agriculture*, 44(1):71-91.

Adeleke, O.A., Latiff, A.A.A., Saphira, M.R., Daud, Z., Ismail, N., Ahsan, A., ... Hijab, M. 2019. 2 - locally derived activated carbon from domestic, agricultural and industrial wastes for the treatment of palm oil mill effluent. In: Ahsan, A. & Ismail, A.F., eds*. Nanotechnology in water and wastewater treatment*: Elsevier. pp. 35-62.

Agelet, L.E. & Hurburgh Jr, C.R. 2010. A tutorial on near infrared spectroscopy and its calibration. *Critical Reviews in Analytical Chemistry*, 40(4):246-260.

AL-Zabee, M.R. & AL-Maliki, S.M. 2019. Effect of biofertilizers and chemical fertilizers on soil biological properties and potato yield. *Euphrates Journal of Agriculture Science*, 11(1):1-13.

Allpress, J.G. and Sanders, J.V., 1967. The structure and orientation of crystals in deposits of metals on mica. *Surface Science*, *7*(1), pp.1-25.

Anon, 2019. An overview of Cubist. https://www.rulequest.com/cubist-win.html. Accessed: 2023/02/09.

Barbosa-Cánovas, G.V., Pastore, G.M., Candoğan, K., Meza, I.G.M., da Silva Lannes, S.C., Buckle, K., Yada, R.Y. and Rosenthal, A. eds., 2017. *Global food security and wellness*. New York, NY, USA: Springer.

Barnston, A.G., 1992. Correspondence among the correlation, RMSE, and Heidke forecast verification measures; refinement of the Heidke score. *Weather and Forecasting*, *7*(4), pp.699-709.

Beenen, G.J. 1981. Spectroscopic investigation of atomic and molecular species formed in a laser microprobe plasma using a wavelenth calibrated tunable dye laser.

Bei, D., Ai-Rui, C. & Meng, Z. 2021. The role of machine learning in solving overfitting. *Journal of Psychological Science*, (2):274.

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

Bellon-Maurel, V., Fernandez-Ahumada, E., Palagos, B., Roger, J.M. and McBratney, A., 2010. Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy. *TrAC Trends in Analytical Chemistry*, *29*(9), pp.1073-1081.

Bonilla-Cedrez, C., Chamberlin, J. and Hijmans, R.J., 2021. Fertilizer and grain prices constrain food production in sub-Saharan Africa. *Nature Food*, *2*(10), pp.766-772.

Boufous, S., Hudson, D. and Carpio, C., 2023. Farmers' willingness to adopt sustainable agricultural practices: A meta-analysis. *PLOS Sustainability and Transformation*, *2*(1), p.e0000037.

Biau, G. & Scornet, E. 2016. A random forest guided tour. *Test*, 25(2):197-227.

Bongiovanni, R. & Lowenberg-DeBoer, J. 2004. Precision agriculture and sustainability. *Precision agriculture*, 5(4):359-387.

Bramley, R. & Janik, L. 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.

Branco de Freitas Maia, C.M., Novotny, E.H., Rittl, T.F. & Bermingham Hayes, M.H. 2013. Soil organic matter: Chemical and physical characteristics and analytical methods. A review. *Current Organic Chemistry*, 17(24):2985-2990.

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

Breure, T.S., Prout, J.M., Haefele, S.M., Milne, A.E., Hannam, J.A., Moreno-Rojas, S. & Corstanje, R. 2022. Comparing the effect of different sample conditions and spectral libraries on the prediction accuracy of soil properties from near- and mid-infrared spectra at the field-scale. *Soil and Tillage Research*, 215:105196.

Brown, D.J., Shepherd, K.D., Walsh, M.G., Mays, M.D. & Reinsch, T.G. 2006. Global soil characterization with vnir diffuse reflectance spectroscopy. *Geoderma*, 132(3-4):273-290.

Bruker. 2021.

Compact ft-ir spectrometer alpha ii. **https://www.bruker.com/en/products-and-solutions/infrared-and-raman/ft-ir-routine-spectrometer/alpha-ii-compact-ft-ir-spectrometer.html** Date of access: 2023/02/05

Cai, J., Ma, E., Lin, J., Liao, L. & Han, Y. 2020. Exploring global food security pattern from the perspective of spatio-temporal evolution. *Journal of Geographical Sciences*, 30(2):179-196.
Cakmak, I. & Yazici, A.M. 2010. Magnesium: A forgotten element in crop production. *Better crops*, 94(2):23-25.

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

Chen, S., Peng, J., Ji, W., Zhou, Y., He, J. & Shi, Z. 2016. Study on the characterization of VNIR-MIR spectra and prediction of soil organic matter in paddy soil. *Guang pu xue yu Guang pu fen xi= Guang pu*, 36(6):1712-1716.

Chen, T., Men, J., Zhao, M., Zhang, T. & Li, H. 2021. The spectral fusion of laser-induced breakdown spectroscopy (LIBS) and mid-infrared spectroscopy (MIR) coupled with random forest (RF) for the quantitative analysis of soil ph. *Journal of Analytical Atomic Spectrometry*, 36(5):1084-1092.

Chen, Y., Hu, S., Guo, Z., Cui, T., Zhang, L., Lu, C., Yu, Y., Luo, Z., Fu, H. and Jin, Y., 2021. Effect of balanced nutrient fertilizer: A case study in Pinggu District, Beijing, China. *Science of The Total Environment*, *754*, p.142069.

Ciceri, D. & Allanore, A. 2019. Local fertilizers to achieve food self-sufficiency in africa. *Science of The Total Environment*, 648:669-680. http://www.sciencedirect.com/science/article/pii/S0048969718331188 https://doi.org/10.1016/j.scitotenv.2018.08.154

Core, R. 2020. R: A language and environment for statistical computing (3.6. 3). R foundation for statistical computing. [Software]

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

Cramer III, R.D., 1993. Partial least squares (PLS): its strengths and limitations. *Perspectives in Drug Discovery and Design*, *1*(2), pp.269-278.

Culman, S., Mann, M. and Brown, C., 2019. Calculating cation exchange capacity, base saturation, and calcium saturation. *Agric Nat Resour*, *81*(6).

Dangal, S.R., Sanderman, J., Wills, S. & Ramirez-Lopez, L. 2019. Accurate and precise prediction of soil properties from a large mid-infrared spectral library. *Soil Systems*, 3(1):11.

Davies, A. & Fearn, T. 2006. Back to basics: Calibration statistics. *Spectroscopy Europe*, 18(2):31-32.

De Kreij, C., Janse, J., Van Goor, B. & Van Doesburg, J. 1992. The incidence of calcium oxalate crystals in fruit walls of tomato (lycopersicon esculentum mill.) as affected by humidity, phosphate and calcium supply. *Journal of Horticultural Science*, 67(1):45-50.

Deiss, L., Margenot, A.J., Culman, S.W. & Demyan, M.S. 2020. Tuning support vector machines regression models improves prediction accuracy of soil properties in MIR spectroscopy. *Geoderma*, 365:114227.

Dharumarajan, S., Hegde, R. & Singh, S.K. 2017. Spatial prediction of major soil properties using random forest techniques - a case study in semi-arid tropics of South India. *Geoderma Regional*, 10:154-162. https://www.sciencedirect.com/science/article/pii/S2352009417300731 https://doi.org/10.1016/j.geodrs.2017.07.005

Du, C., Linker, R. & Shaviv, A. 2008. Identification of agricultural mediterranean soils using mid-infrared photoacoustic spectroscopy. *Geoderma*, 143(1-2):85-90.

Elkateb, T., Chalaturnyk, R. and Robertson, P.K., 2003. An overview of soil heterogeneity: quantification and implications on geotechnical field problems. *Canadian Geotechnical Journal*, *40*(1), pp.1-15.

Estienne, F., Pasti, L., Centner, V., Walczak, B., Despagne, F., Rimbaud, D.J., De Noord, O.E. and Massart, D.L., 2001. A comparison of multivariate calibration techniques applied to

experimental NIR data sets: Part II. Predictive ability under extrapolation conditions. *Chemometrics and Intelligent Laboratory Systems*, *58*(2), pp.195-211.

Fang, Q., Hong, H., Zhao, L., Kukolich, S., Yin, K. and Wang, C., 2018. Visible and near-infrared reflectance spectroscopy for investigating soil mineralogy: A review. *Journal of Spectroscopy*, *2018*.

FAO. 2005. The special program for food security, food and agriculture organisation of the united nations. Rome: FAO

Fathizad, H., Ardakani, M.A.H., Sodaiezadeh, H., Kerry, R. & Taghizadeh-Mehrjardi, R. 2020. Investigation of the spatial and temporal variation of soil salinity using random forests in the central desert of iran. *Geoderma*, 365:114233.

FERTASA. 2016. *Fssa fertilizer handbook*. 7. Lynnwood Ridge [South Africa]: Fertilizer Society of South Africa.

Forrester, S.T., Janik, L.J., Soriano-Disla, J.M., Mason, S., Burkitt, L., Moody, P., ... McLaughlin, M.J. 2015. Use of handheld mid-infrared spectroscopy and partial least-squares regression for the prediction of the phosphorus buffering index in australian soils. *Soil Research*, 53(1):67-80.

Fowler, A. 1922. *Report on series in line spectra*. Fleetway Press, Limited. https://archive.org/details/reportonseriesin00fowluoft/mode/2up. Date of access: 2023/02/09.

Freeman, J.J., Wang, A., Kuebler, K.E., Jolliff, B.L. and Haskin, L.A., 2008. Characterization of natural feldspars by Raman spectroscopy for future planetary exploration. *The Canadian Mineralogist*, *46*(6), pp.1477-1500.

García-Jaramillo, M., Cox, L., Cornejo, J. and Hermosín, M.C., 2014. Effect of soil organic amendments on the behavior of bentazone and tricyclazole. *Science of the total environment*, *466*, pp.906-913.

Gates, J.R. 2018. A comparison of vnir and mir spectroscopy for predicting various soil properties. Dissertation and Thesis in Natural Resources. University of Nebraska. https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1269&context=natresdiss. Date of access: 2023/02/09.

Geypens, M., Vanongeval, L., Vogels, N. and Meykens, J., 1999. Spatial variability of agricultural soil fertility parameters in a gleyic podzol of Belgium. *Precision Agriculture*, *1*(3), pp.319-326.

Gholizadeh, A., Borůvka, L., Saberioon, M. & Vašát, R. 2013. Visible, near-infrared, and mid-infrared spectroscopy applications for soil assessment with emphasis on soil organic matter content and quality: State-of-the-art and key issues. *Applied spectroscopy*, 67(12):1349-1362.

Gholizadeh, A., Rossel, R.A.V., Saberioon, M., Borůvka, L., Kratina, J. & Pavlů, L. 2021. National-scale spectroscopic assessment of soil organic carbon in forests of the Czech Republic. *Geoderma*, 385:114832.

Gillespie, C.J., Antonangelo, J.A. & Zhang, H. 2021. The response of soil pH and exchangeable Al to alum and lime amendments. *Agriculture*, 11(6):547.

Gobrecht, A., Roger, J.-M. & Bellon-Maurel, V. 2014. Chapter four - major issues of diffuse reflectance nir spectroscopy in the specific context of soil carbon content estimation: A review. In: Sparks, D.L., ed. *Advances in agronomy*. 123: Academic Press. pp. 145-175.

Griffin, T.W., Miller, N.J., Bergtold, J., Shanoyan, A., Sharda, A. & Ciampitti, I.A. 2017. Farm's sequence of adoption of information-intensive precision agricultural technology. *Applied Engineering in Agriculture*, 33(4):521.

Guo, W., Nazim, H., Liang, Z. & Yang, D. 2016. Magnesium deficiency in plants: An urgent problem. *The Crop Journal*, 4(2):83-91.

Haas, J. and Mizaikoff, B., 2016. Advances in mid-infrared spectroscopy for chemical analysis. *Annual Review of Analytical Chemistry*, *9*, pp.45-68.

Haghi, R., Pérez-Fernández, E. & Robertson, A. 2021. Prediction of various soil properties for a national spatial dataset of scottish soils based on four different chemometric approaches: A comparison of near infrared and mid-infrared spectroscopy. *Geoderma*, 396:115071.

Hawkes, C. & Fanzo, J. 2017. Nourishing the sdgs: Global nutrition report 2017.

Hazelton, P. & Murphy, B. 2016. *Interpreting soil test results: What do all the numbers mean?* CSIRO publishing.

Herrero, M., Thornton, P.K., Notenbaert, A.M., Wood, S., Msangi, S., Freeman, H.A., Bossio, D., Dixon, J., Peters, M., van de Steeg, J. and Lynam, J., 2010. Smart investments in sustainable food production: revisiting mixed crop-livestock systems. *Science*, *327*(5967), pp.822-825.

Hillel, D. 2008. Soil fertility and plant nutrition. *Soil in the Environment*, 18:151-162.

Hong, S.Y., Lee, K., Minasny, B., Kim, Y. & Hyun, B.K. 2014. Predicting soil chemical properties with regression rules from visible-near infrared reflectance spectroscopy. *Korean Journal of Soil Science and Fertilizer*, 47(5):319-323.

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

Janik, L.J. & Skjemstad, J. 1995. Characterization and analysis of soils using mid-infrared partial least-squares. 2. Correlations with some laboratory data. *Soil Research*, 33(4):637-650.

Janik, L.J., Merry, R.H. & Skjemstad, J. 1998. Can mid infrared diffuse reflectance analysis replace soil extractions? *Australian Journal of Experimental Agriculture*, 38(7):681-696.

Janik, L.J., Soriano-Disla, J.M. & Forrester, S.T. 2020. Feasibility of handheld mid-infrared spectroscopy to predict particle size distribution: Influence of soil field condition and utilisation of existing spectral libraries. *Soil Research*, 58(6):528-539.

Janik, L.J., Merry, R.H., Forrester, S., Lanyon, D. & Rawson, A. 2007. Rapid prediction of soil water retention using mid infrared spectroscopy. *Soil Science Society of America Journal*, 71(2):507-514.

Jean-Philippe, S.R., Labbé, N., Franklin, J.A. & Johnson, A. 2012. Detection of mercury and other metals in mercury contaminated soils using mid-infrared spectroscopy. *Proceedings of the International Academy of Ecology and Environmental Sciences*, 2(3):139.

Ji, W., Adamchuk, V.I., Biswas, A., Dhawale, N.M., Sudarsan, B., Zhang, Y., Rossel, R.A.V. and Shi, Z., 2016. Assessment of soil properties in situ using a prototype portable MIR spectrometer in two agricultural fields. *Biosystems engineering*, *152*, pp.14-27.

Johns, T.J., Angove, M.J. & Wilkens, S. 2015. Measuring soil organic carbon: Which technique and where to from here? *Soil Research*, 53(7):717-736.

Johnson, D. 1992. Base cations. In. *Atmospheric deposition and forest nutrient cycling*: Springer. pp. 233-340.

Johnson, J.-M., Vandamme, E., Senthilkumar, K., Sila, A., Shepherd, K.D. & Saito, K. 2019. Near-infrared, mid-infrared or combined diffuse reflectance spectroscopy for assessing soil fertility in rice fields in sub-saharan africa. *Geoderma*, 354:113840.

Kamnev, A.A., Dyatlova, Y.A., Kenzhegulov, O.A., Vladimirova, A.A., Mamchenkova, P.V. & Tugarova, A.V. 2021. Fourier transform infrared (ftir) spectroscopic analyses of microbiological samples and biogenic selenium nanoparticles of microbial origin: Sample preparation effects. *Molecules*, 26(4):1146.

Kasprzhitskii, A., Lazorenko, G., Khater, A. & Yavna, V. 2018. Mid-infrared spectroscopic assessment of plasticity characteristics of clay soils. *Minerals*, 8(5):184.

Kendall, H., Clark, B., Li, W., Jin, S., Jones, G., Chen, J., ... Frewer, L. 2021. Precision agriculture technology adoption: A qualitative study of small-scale commercial “family farms” located in the north china plain. *Precision Agriculture*:1-33.

Kock, A.-L. 2022. *Creation of mid-infrared spectroscopy calibration algorithms for soil property predictions.* North-West University (South Africa).

Kuhn, M., Weston, S., Keefer, C. & Kuhn, M.M. 2022. Package ‘cubist’. *Rule-and instance-based regression modeling: version 0.2*, 3,

Kundu, M., Hazra, G., Biswas, P., Mondal, S. & Ghosh, G. 2014. Forms and distribution of potassium in some soils of hooghly district of west bengal. *Journal of Crop and Weed*, 10:31-37.

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.

Li, F., Xu, L., You, T. & Lu, A. 2021. Measurement of potentially toxic elements in the soil through nir, mir, and xrf spectral data fusion. *Computers and Electronics in Agriculture*, 187:106257.

Li, J., Yao, J., Li, Y. & Shao, Y. 2012. Controlled release and retarded leaching of pesticides by encapsulating in carboxymethyl chitosan/bentonite composite gel. *Journal of Environmental Science and Health, Part B*, 47(8):795-803.

Li, M., Feng, Y., Yu, Y., Zhang, T., Yan, C., Tang, H., ... Li, H. 2021. Quantitative analysis of polycyclic aromatic hydrocarbons in soil by infrared spectroscopy combined with hybrid variable selection strategy and partial least squares. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 257:119771.

Linker, R., 2008. Soil classification via mid-infrared spectroscopy. In *Computer And Computing Technologies In Agriculture, Volume II: First IFIP TC 12 International Conference on Computer and Computing Technologies in Agriculture (CCTA 2007), Wuyishan, China, August 18-20, 2007 1* (pp. 1137-1146). Springer US.

Linker, R., Shmulevich, I., Kenny, A. & Shaviv, A. 2005. Soil identification and chemometrics for direct determination of nitrate in soils using ftir-atr mid-infrared spectroscopy. *Chemosphere*, 61(5):652-658.

Liu, N., Xu, L., Zhou, S., Zhang, L. & Li, J. 2020. Simultaneous detection of multiple atmospheric components using an NIR and MIR laser hybrid gas sensing system. *ACS sensors*, 5(11):3607-3616.

Liu, N., Xu, L., Zhou, S., Zhang, L. & Li, J. 2021. Soil respiration analysis using a mid-infrared quantum cascade laser and calibration-free wms-based dual-gas sensor. *Analyst*, 146(12):3841-3851.

Louppe, G., 2014. Understanding random forests: From theory to practice. *arXiv preprint arXiv:1407.7502*.

Lucà, F., Conforti, M., Castrignanò, A., Matteucci, G. & Buttafuoco, G. 2017. Effect of calibration set size on prediction at local scale of soil carbon by vis-nir spectroscopy. *Geoderma*, 288:175-183.

Ma, Y., Minasny, B., McBratney, A., Poggio, L. & Fajardo, M. 2021. Predicting soil properties in 3d: Should depth be a covariate? *Geoderma*, 383:114794.

https://www.sciencedirect.com/science/article/pii/S0016706120325490
https://doi.org/10.1016/j.geoderma.2020.114794

Madeira, M., Auxtero, E. & Sousa, E. 2003. Cation and anion exchange properties of andisols from the azores, portugal, as determined by the compulsive exchange and the ammonium acetate methods. *Geoderma*, 117(3):225-241. https://www.sciencedirect.com/science/article/pii/S0016706103001253
https://doi.org/10.1016/S0016-7061(03)00125-3

Maine, N., Lowenberg-DeBoer, J., Nell, W.T. and Alemu, Z.G., 2010. Impact of variable-rate application of nitrogen on yield and profit: a case study from South Africa. *Precision Agriculture*, *11*, pp.448-463.

Malherbe, J., Dieppois, B., Maluleke, P., Van Staden, M. & Pillay, D. 2016. South African droughts and decadal variability. *Natural Hazards*, 80(1):657-681.

Mayer, D.G. & Butler, D.G. 1993. Statistical validation. *Ecological Modelling*, 68(1):21-32. https://www.sciencedirect.com/science/article/pii/0304380093901052
https://doi.org/10.1016/0304-3800(93)90105-2

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

McCarty, G.W. & 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.

McClure, W.F. 2003. 204 years of near infrared technology: 1800–2003. *Journal of Near Infrared Spectroscopy*, 11(6):487-518.

Merry, R. & Sabljic, A. 2009. Acidity and alkalinity of soils. *Environmental and ecological chemistry*, 2:115-131.

Merry, R., Janik, L., Spouncer, L. & Weissman, D. 1997. New methodology for lime requirements and use in decision support. *RIRDC Project CSO-7A, Final Report*,

Metzger, K., Zhang, C. & Daly, K. 2021. From benchtop to handheld MIR for soil analysis: Predicting lime requirement and organic matter in agricultural soils. *Biosystems Engineering*, 204:257-269.

Metzger, K., Zhang, C., Ward, M. & Daly, K. 2020. Mid-infrared spectroscopy as an alternative to laboratory extraction for the determination of lime requirement in tillage soils. *Geoderma*, 364:114171.

Mevik, B.-H. & Wehrens, R. 2015. Introduction to the pls package. *Help Section of The "Pls" Package of R Studio Software*:1-23.

Meyer, N., Meyer, H., Welp, G. & Amelung, W. 2018. Soil respiration and its temperature sensitivity (q10): Rapid acquisition using mid-infrared spectroscopy. *Geoderma*, 323:31-40.

Minasny, B. & McBratney, A.B. 2006. A conditioned latin hypercube method for sampling in the presence of ancillary information. *Computers & Geosciences*, 32(9):1378-1388. https://www.sciencedirect.com/science/article/pii/S009830040500292X https://doi.org/10.1016/j.cageo.2005.12.009

Mitchell, S., Weersink, A. & Erickson, B. 2018. Adoption of precision agriculture technologies in ontario crop production. *Canadian Journal of Plant Science*, 98(6):1384-1388.

Moore, K. and Bradley, L.K. eds., 2018. *North Carolina Extension gardener handbook*. NC State Extension, College of Agriculture and Life Sciences, NC State University.

Morellos, A., Pantazi, X.E., Moshou, D., Alexandridis, T., Whetton, R., Tziotzios, G., Wiebensohn, J., Bill, R. and Mouazen, A.M., 2016. Machine learning based prediction of soil total nitrogen, organic carbon and moisture content by using VIS-NIR spectroscopy. *Biosystems Engineering*, *152*, pp.104-116.

Naik, K., Mishra, S., Srichandan, H., Singh, P.K. & Sarangi, P.K. 2019. Plant growth promoting microbes: Potential link to sustainable agriculture and environment. *Biocatalysis and Agricultural Biotechnology*:101326.

Nash, V.E. and Marshall, C.E., 1956. *The surface reactions of silicate minerals. Part II, Reactions of feldspar surfaces with salt solutions*. University of Missouri, College of Agriculture, Agricultural Experiment Station.

Nath, D., Laik, R., Meena, V.S., Pramanick, B. & Singh, S.K. 2021. Can mid-infrared (MIR) spectroscopy evaluate soil conditions by predicting soil biological properties? *Soil Security*:100008.

National Academies of Sciences, E. & Medicine. 2019. *Reproducibility and replicability in science*. National Academies Press.

Nell, J.P. and Van Huyssteen, C.W., 2018. Prediction of primary salinity, sodicity and alkalinity in South African soils. *South African Journal of Plant and Soil*, *35*(3), pp.173-178.

Ng, L.M. & Simmons, R. 1999. Infrared spectroscopy. *Analytical chemistry*, 71(12):343-350.

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

Ng, W., Minasny, B., Jeon, S.H. and McBratney, A., 2022. Mid-infrared spectroscopy for accurate measurement of an extensive set of soil properties for assessing soil functions. *Soil Security*, *6*, p.100043.

Nocita, M., Kooistra, L., Bachmann, M., Müller, A., Powell, M. & 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*, 167:295-302.

Nocita, M., Stevens, A., van Wesemael, B., Aitkenhead, M., Bachmann, M., Barthès, B., Dor, E.B., Brown, D.J., Clairotte, M., Csorba, A. and Dardenne, P., 2015. Soil spectroscopy: an alternative to wet chemistry for soil monitoring. *Advances in agronomy*, *132*, pp.139-159.

Norton, R., 2013. Focus on Calcium: Its role in crop production. *GRDC Updates Pap. Available online: https://grdc.com.au/resources-and-publications/grdc-update-papers/tab-content/grdc-updatepapers/2013/02/focus-on-calcium-its-role-in-crop-production. Date of access: 2023/02/05*

Nyam, Y., Kotir, J., Jordaan, A., Ogundeji, A. & Turton, A. 2020. Drivers of change in sustainable water management and agricultural development in south africa: A participatory approach. *Sustainable Water Resources Management*, 6(4):1-20.

Palm, C., Sanchez, P., Ahamed, S. & Awiti, A. 2007. Soils: A contemporary perspective. *Annu. Rev. Environ. Resour.*, 32:99-129.

Parent, E.J., Parent, S.-É. & Parent, L.E. 2021. Determining soil particle-size distribution from infrared spectra using machine learning predictions: Methodology and modeling. *PloS one*, 16(7):e0233242.

Pasquini, C. 2018. Near infrared spectroscopy: A mature analytical technique with new perspectives – a review. *Analytica Chimica Acta*, 1026:8-36. https://www.sciencedirect.com/science/article/pii/S0003267018304793 https://doi.org/10.1016/j.aca.2018.04.004

Paterson, G., Turner, D., Wiese, L., Van Zijl, G., Clarke, C. & Van Tol, J. 2015. Spatial soil information in south africa: Situational analysis, limitations and challenges. *South African Journal of Science*, 111(5-6):1-7.

Paz, A.M., Castanheira, N., Farzamian, M., Paz, M.C., Gonçalves, M.C., Monteiro Santos, F.A. & Triantafilis, J. 2020. Prediction of soil salinity and sodicity using electromagnetic conductivity imaging. *Geoderma*, 361:114086. https://www.sciencedirect.com/science/article/pii/S0016706119307347 https://doi.org/10.1016/j.geoderma.2019.114086

Peng, J., Ji, W., Ma, Z., Li, S., Chen, S., Zhou, L. & 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.

Pirie, A., Singh, B. & Islam, K. 2005. Ultra-violet, visible, near-infrared, and mid-infrared diffuse reflectance spectroscopic techniques to predict several soil properties. *Soil Research*, 43(6):713-721.

Pouladi, N., Møller, A.B., Tabatabai, S. & Greve, M.H. 2019. Mapping soil organic matter contents at field level with cubist, random forest and kriging. *Geoderma*, 342:85-92.

Provin, T. and Pitt, J.L., 2001. Managing soil salinity. *Texas FARMER Collection*.

Qiao, C., Penton, C.R., Xiong, W., Liu, C., Wang, R., Liu, Z., ... Shen, Q. 2019. Reshaping the rhizosphere microbiome by bio-organic amendment to enhance crop yield in a maize-cabbage rotation system. *Applied Soil Ecology*, 142:136-146.

Rahman, K. & Zhang, D. 2018a. Effects of fertilizer broadcasting on the excessive use of inorganic fertilizers and environmental sustainability. *Sustainability*, 10(3):759.

Rahman, K. & Zhang, D. 2018b. Effects of fertilizer broadcasting on the excessive use of inorganic fertilizers and environmental sustainability. *Sustainability (Switzerland)*, 10, 10.3390/su10030759
Raidimi, E. & Kabiti, H. 2019. A review of the role of agricultural extension and training in achieving sustainable food security: A case of south africa. *South African Journal of Agricultural Extension*, 47(3):120-130.

Ramirez-Lopez, L., Behrens, T., Schmidt, K., Stevens, A., Demattê, J.A.M. & Scholten, T. 2013. The spectrum-based learner: A new local approach for modeling soil VNIR spectra of complex datasets. *Geoderma*, 195:268-279.

Ramirez-Lopez, L., Schmidt, K., Behrens, T., van Wesemael, B., Demattê, J.A.M. & Scholten, T. 2014. Sampling optimal calibration sets in soil infrared spectroscopy. *Geoderma*, 226-227:140-150.

Ray, S.K., Barman, K., Chowdhury, P. & Deka, B.C. Soil testing for optimization crop production.
Reeves III, J.B. 2010. Near-versus mid-infrared diffuse reflectance spectroscopy for soil analysis emphasizing carbon and laboratory versus on-site analysis: Where are we and what needs to be done? *Geoderma*, 158(1-2):3-14.

Reeves III, J.B. 2012. Mid-infrared spectral interpretation of soils: Is it practical or accurate? *Geoderma*, 189:508-513.

Reeves III, J.B. & Smith, D.B. 2009. The potential of mid-and near-infrared diffuse reflectance spectroscopy for determining major-and trace-element concentrations in soils from a geochemical survey of north america. *Applied Geochemistry*, 24(8):1472-1481.

Robertson, M., Carberry, P. and Brennan, L., 2007. The economic benefits of precision agriculture: case studies from Australian grain farms. *Crop Pasture Sci*, *60*, p.2012.

Rosanoff, A., Weaver, C.M. & Rude, R.K. 2012. Suboptimal magnesium status in the united states: Are the health consequences underestimated? *Nutrition reviews*, 70(3):153-164.

Rossel, R.V., Jeon, Y., Odeh, I. & McBratney, A. 2008. Using a legacy soil sample to develop a mid-ir spectral library. *Soil Research*, 46(1):1-16.

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

Rutter, E.B., Diaz, D.R. and Hargrave, L., 2021. Comparison of Mehlich-3 and Ammonium Acetate Extractable Calcium and Magnesium in Kansas Soils. *Kansas Agricultural Experiment Station Research Report s*, p.11.

Ryan, J., Rashid, A., Torrent, J., Yau, S.K., Ibrikci, H., Sommer, R. and Erenoglu, E.B., 2013. Micronutrient constraints to crop production in the Middle East–West Asia region: significance, research, and management. *Advances in Agronomy*, *122*, pp.1-84.

Sabetizade, M., Gorji, M., Roudier, P., Zolfaghari, A.A. & Keshavarzi, A. 2021. Combination of MIR spectroscopy and environmental covariates to predict soil organic carbon in a semi-arid region. *Catena*, 196:104844.

Sanderman, J., Savage, K. & Dangal, S.R. 2020. Mid-infrared spectroscopy for prediction of soil health indicators in the united states. *Soil Science Society of America Journal*, 84(1):251-261.
Savci, S. 2012. An agricultural pollutant: Chemical fertilizer. *International Journal of Environmental Science and Development*, 3(1):73.

Say, S.M., Keskin, M., Sehri, M. and Sekerli, Y.E., 2018. Adoption of precision agriculture technologies in developed and developing countries. *The Online Journal of Science and Technology-January*, *8*(1), pp.7-15.

Schimmelpfennig, D. and Ebel, R., 2016. Sequential adoption and cost savings from precision agriculture. *Journal of Agricultural and Resource Economics*, pp.97-115.

Shah, F. & Wu, W. 2019. Soil and crop management strategies to ensure higher crop productivity within sustainable environments. *Sustainability*, 11(5):1485.

Shepherd, K.D. & 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. & Walsh, M.G. 2007. 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.

Shi, T., Chen, Y., Liu, Y. & Wu, G. 2014. Visible and near-infrared reflectance spectroscopy—an alternative for monitoring soil contamination by heavy metals. *Journal of Hazardous Materials*, 265:166-176. https://www.sciencedirect.com/science/article/pii/S030438941300914X https://doi.org/10.1016/j.jhazmat.2013.11.059

Sirsat, M.S., Cernadas, E., Fernández-Delgado, M. & Barro, S. 2018. Automatic prediction of village-wise soil fertility for several nutrients in india using a wide range of regression methods. *Computers and Electronics in Agriculture*, 154:120-133. https://www.sciencedirect.com/science/article/pii/S0168169917311006 https://doi.org/10.1016/j.compag.2018.08.003

Sordo, M. and Zeng, Q., 2005, November. On sample size and classification accuracy: A performance comparison. In *International Symposium on Biological and Medical Data Analysis* (pp. 193-201). Springer, Berlin, Heidelberg.

Sornette, D., Davis, A., Ide, K., Vixie, K., Pisarenko, V. & Kamm, J. 2007. Algorithm for model validation: Theory and applications. *Proceedings of the National Academy of Sciences*, 104(16):6562-6567.

Stamatiadis, S., Schepers, J.S., Evangelou, E., Tsadilas, C., Glampedakis, A., Glampedakis, M., Dercas, N., Spyropoulos, N., Dalezios, N.R. and Eskridge, K., 2018. Variable-rate nitrogen fertilization of winter wheat under high spatial resolution. *Precision Agriculture*, *19*, pp.570-587.

Stenberg, B., Rossel, R.A.V., Mouazen, A.M. & Wetterlind, J. 2010. Visible and near infrared spectroscopy in soil science. *Advances in agronomy*, 107:163-215.

Sutton, M.A., Bleeker, A., Howard, C.M., Erisman, J.W., Abrol, Y.P., Bekunda, M., Datta, A., Davidson, E., De Vries, W., Oenema, O. and Zhang, F.S., 2013. *Our nutrient world. The challenge to produce more food & energy with less pollution*. Centre for Ecology & Hydrology.

Takele, C. & Iticha, B. 2020. Use of infrared spectroscopy and geospatial techniques for measurement and spatial prediction of soil properties. *Heliyon*, 6(10):e05269.

Tey, Y.S., Li, E., Bruwer, J., Abdullah, A.M., Brindal, M., Radam, A., Ismail, M.M. and Darham, S., 2017. Factors influencing the adoption of sustainable agricultural practices in developing countries: a review. *Environmental Engineering & Management Journal (EEMJ)*, *16*(2).

Thomas, G.W. 1983. Exchangeable cations. *Methods of soil analysis: Part 2 chemical and microbiological properties*, 9:159-165.

Thompson, N.M., Bir, C., Widmar, D.A. & Mintert, J.R. 2019. Farmer perceptions of precision agriculture technology benefits. *Journal of Agricultural and Applied Economics*, 51(1):142-163.

Tian, Z., Wang, J.W., Li, J. & Han, B. 2021. Designing future crops: Challenges and strategies for sustainable agriculture. *The Plant Journal*, 105(5):1165-1178.

Triki Fourati, H., Bouaziz, M., Benzina, M. & Bouaziz, S. 2015. Modeling of soil salinity within a semi-arid region using spectral analysis. *Arabian Journal of Geosciences*, 8(12):11175-11182.

Tümsavaş, Z., Tekin, Y., Ulusoy, Y. and Mouazen, A.M., 2019. Prediction and mapping of soil clay and sand contents using visible and near-infrared spectroscopy. *Biosystems Engineering*, *177*, pp.90-100.

UN. 2015. Resolution adopted by the general assembly on 25 september 2015.

USGS. 2022. *United States Geological Survey*. https://earthexplorer.usgs.gov/ Date of access: 20 Jun. 2021.

Vabalas, A., Gowen, E., Poliakoff, E. & Casson, A.J. 2019. Machine learning algorithm validation with a limited sample size. *PloS one*, 14(11):e0224365.

Van der Berg, S., 2006. Public spending and the poor since the transition to democracy[1]. *Poverty and policy in post-apartheid South Africa*, p.201-231.

Van der Merwe, C.R., 1962. *Soil groups and subgroups of South Africa* (Doctoral dissertation, Stellenbosch: Stellenbosch University).

Von Grebmer, K., Bernstein, J., Hossain, N., Brown, T., Prasai, N., Yohannes, Y., Patterson, F., Sonntag, A., Zimmerman, S.M., Towey, O. and Foley, C., 2017. *2017 Global Hunger Index: the inequalities of hunger*. Intl Food Policy Res Inst.

Wadoux, A.M.C. and McBratney, A.B., 2021. Hypotheses, machine learning and soil mapping. *Geoderma*, *383*, p.114725.

Wall, D. & Plunkett, M. 2021. *Major and micro nutrient advice for productive agricultural crops*.

Wang, C., Li, W., Guo, M. & Ji, J. 2017. Ecological risk assessment on heavy metals in soils: Use of soil diffuse reflectance mid-infrared fourier-transform spectroscopy. *Scientific reports*, 7(1):1-11.

Warncke, D. & Brown, J. 1998. Potassium and other basic cations. *Recommended chemical soil test procedures for the North Central Region*, 1001:31.

Wei, W., Yang, M., Liu, Y., Huang, H., Ye, C., Zheng, J., ... Zhu, S. 2018. Fertilizer n application rate impacts plant-soil feedback in a sanqi production system. *Science of The Total Environment*, 633:796-807. https://www.sciencedirect.com/science/article/pii/S0048969718309781 https://doi.org/10.1016/j.scitotenv.2018.03.219

Whelan, B. and Taylor, J., 2013. *Precision agriculture for grain production systems*. Csiro publishing.

Wilkinson, S.R., Welch, R.M., Mayland, H.F. and Grunes, D.L., 1990. Magnesium in plants: uptake, distribution, function, and utilization by man and animals. Compendium on Magnesium and Its Role in Biology, Nutrition, and Physiology. Volume 6.

Wold, S., Sjöström, M. & Eriksson, L. 2001. Pls-regression: A basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, 58(2):109-130.

Workman Jr, J. & Shenk, J. 2004. Understanding and using the near-infrared spectrum as an analytical method. *Near-infrared spectroscopy in agriculture*, 44:1-10.

Xiao, Q., Wang, J., Liang, D., Xia, H., Wang, T. & Lyu, X. 2019. Effects of formulated fertilization on soil physical and chemical characteristics of early ripe peach orchard. In. IOP Conference Series: Earth and Environmental Science. IOP Publishing. p 012086.

Yang, J.L., You, J.F., Li, Y.Y., Wu, P. and Zheng, S.J., 2007. Magnesium enhances aluminum-induced citrate secretion in rice bean roots (Vigna umbellata) by restoring plasma membrane H+-ATPase activity. *Plant and cell physiology*, *48*(1), pp.66-73.

Yun-peng, H., Jia-gui, X. & Cai-xia, Y. 2010. Effects of soil testing and formula fertilization on maize yield and fertilizer utilization efficiency [j]. *Journal of Anhui Agricultural Sciences*, 18,

Zorin, I., Gattinger, P., Ebner, A. and Brandstetter, M., 2022. Advances in mid-infrared spectroscopy enabled by supercontinuum laser sources. *Optics Express*, *30*(4), pp.5222-5254.

APPENDIX

Appendix 1: Table of soil properties tested by Janik et al. (1998)

Soil property	R^2	Range	n	Soil property	R^2	Range	n
Air-dry moisture (%)	0.70	0–16	303	Total organic carbon (%)	0.93	1.1–8.0	188
Moisture content (%)				Aromatic NMR OC (%)	0.86	0–1.3	61
@ 10 kPa	0.83	21–34	23	GC-FAME ECL16 peak area	0.88	22 000–82 000	16
@ 30 kPa	0.90	11–21	23	Biomass (g/kg)	0.69	20–70	23
Particle size composition (%)				Lime requirement (t/ha.pH unit)	0.86	0.7–5.0	188
Sand	0.94	21–96	88	P capacity (mg/kg)	0.87	50–900	47
Silt	0.84	0–44	88	Atrazine absorption (L/kg)	0.69	0–4.6	31
Clay	0.79	2–49	88	Fe DTPA (mg/kg)	0.55	17–540	183
Field texture (arbitrary unit)	0.54	1–18	88	Mn DTPA (mg/kg)	0.57	0–181	183
XRF analysis (%)				pH $CaCl_2$ (1 : 5 w/v)	0.67	3.8–6.0	183
Al_2O_3	0.92	0–30	298	pH water (1 : 5 w/v)	0.56	4.9–6.6	183
CaO	0.70	0–4.5	298	Cation exchange capacity (cmol/kg)	0.88	26–224	183
Fe_2O_3	0.93	0–23	298	Sum of cations (cmol/kg)	0.84	0–96	298
K_2O	0.63	0–3	298	$CaCl_2$-extractable Al (cmol/kg)	0.53	0–7.6	183
MgO	0.80	0–6	298	$CaCl_2$-extractable Mn (cmol/kg)	0.60	0.2–35	183
P_2O_5	0.60	0–0.8	298	Exchangeable Ca (cmol/kg)	0.89	11–160	183
SiO_2	0.97	10–100	298	Exchangeable Mg (cmol/kg)	0.76	1.5–70	183
TiO	0.78	0–7	298	KCl-exchangeable Al (cmol/kg)	0.64	0–17	183
Total organic nitrogen (%)	0.86	0.0–0.7	188	KCl-exchangeable Mn (cmol/kg)	0.66	0–4	183

Appendix 2: Specific scripts used for the training and predicting of potassium as an example, by each of the three functions used, cubist, partial least squared regression, and random forest

Function	Script
Cubist: training	cubist (train [], train$K, control = cubistControl (rules=9))".
Cubist: predict	predict (Model_Cubist_K, test []).
PLSR: training	plsr (K ~., data = trainK, ncomp = maxc, validation = "CV")
PLSR: predict	predict (Model_PLS_K, ncomp = nc, newdata = testK).
RF: training	randomForest (K~., data = trainK))
RF: predict	pred_RF_Na = predict (Model_RF_Na,testNa) testNa$pred_RF_Na = pred_RF_Na

Appendix 3: Specific scripts used for calculating the values of the statistical markers; R2, RMSE, and RPD, representing the accuracy of the prediction models for potassium as an example

Statistical marker	Script

R^2	(cor (pred_Cubist_K, test$K)) ^2
RMSE	sqrt (mean ((pred_Cubist_K - test$K) ^2))
RPD	RPD (pred_Cubist_K, test$K, na.rm = FALSE)

Appendix 4: Specific script used for the creation of the scatterplots which give an indication of the accuracy of the prediction model, in this case cubist with potassium as an example

Figure	Script
Scatterplot	plot (y=pred_Cubist_K, x=test$K, ylab='Predicted Values', xlab='Measured Values', main='Cubist (K)', abline (a=0, b=1))