Creation of near infrared spectroscopy calibration algorithms for soil water content prediction

HJ Faul

orcid.org 0000-0003-0692-432X

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:	Mr A Kock
Co-Supervisor:	Mr WH Cloete

Graduation May 2024

ACKNOWLEDGEMENTS

I would firstly like to thank my supervisor Prof.George van Zijl for his guidance, insights, and support through the course of this project. I would like to thank my co-supervisors Mr. Kock and Mr. Cloete for their inputs and aid with fieldwork, as well as Ms. Molebaleng Sehlapelo for her assistance with fieldwork.

I would also like to thank all of my friends and family for their love and support, especially my mother Mariëtte Nicholenas and my father Terry Nicholenas for all of their sacrifices, love, support, and kindness along the way.

I would like to thank my girlfriend Mariska Kleyn for her undivided support, her undeserved love, and unwavering kindness that kept me going through the last year.

Lastly, I would like to thank my Lord and Saviour Jesus Christ for firstly creating me, equipping me with the abilities and skills I have, and carrying me through to the very end. I would also like to thank Him for everyone He has sent me along the way.

Matthew 5:3-18 (MSG): “You’re blessed when you’re at the end of your rope. With less of you there is more of God and His rule”.

ABSTRACT

Water is arguably the most important human resource, and fresh water is becoming increasingly scarce. Irrigation in agriculture is one of the most water-demanding sectors, which urges the importance of effective management decisions within this sector. Conducting accurate measurements of soil water content is paramount in making productive decisions in agricultural water management. The problem, however, is those traditional methods of water content measurement in soils, are expensive and immobile. Methods such as the neutron probe is required to calibrate for each soil type and carries a health hazard for users. Additionally, tensiometers and electrical resistance probes are stationary and cannot be relocated to measure different areas. Near infrared spectroscopy (NIRS) can provide a solution to the need for a rapid, accurate, and cost-effective method for measuring soil moisture content. A calibration model is however needed by NIRS in order to make sense of observations. There are unfortunately no such freely available calibration models for South African soils.

This study aimed to create such calibration algorithms for soil water content and dry bulk density (DBD) predictions on approximately 213 soil core samples taken at predetermined locations in five catchments in South Africa. Algorithms were also created to compare to the only freely available calibrations algorithms from the Open Soil Spectroscopy Library (OSSL). The samples were scanned with a portable, handheld NeoSpectra NIR scanner at different moisture contents, which was achieved by saturating the samples and then placing them in pressure chambers to gradually reduce the moisture, at pressures of 33 (Drained upper limit), 100, 500, and 1500 kPa (lower limit). At each scan the soil was weighed to determine the gravimetric water content, which was later converted to volumetric water content using the bulk density. Calibration algorithms were then created on the R coding platform using the spectral data and were compared against the lab-determined volumetric water content. The calibration algorithms were trained on 75% of the dataset, selected by K-means clustering on the spectra. Calibration algorithms were created using Random Forest, Cubist, and Partial Least Squares machine learning algorithms and various pre-processing methods including Savitzky-Golay, Multiplicative Scatter Correction, Standard Normal Variate, and Normalization. Various statistical measures were used to evaluate the accuracy of the different machine learning and pre-processing combinations. Validation was performed on the remaining 25% samples of the dataset and the results were evaluated using a range of statistical methods, which include the mean error that measures the bias of the model, the root mean square error (RMSE) indicating overall model accuracy, the correlation coefficient (R^2) showing the correlation between predicted and actual values, Lin's concordance coefficient (rhoC) which is an adapted correlation coefficient used for model precision and accuracy, the ratio of prediction to deviation (RPD) that indicates predictive performance, and lastly the ratio of

performance to interquartile distance (RPIQ) that measures predictive performance and model robustness.

The best results were obtained by using a combination of Savitzky-Golay pre-processing paired with the Random Forest machine learning technique, with an RMSE of 6.9%, R^2 of 0.62, bias of 0.75, rhoC of 0.75, and RPIQ of 1.8. Algorithms for DBD were created, where a combination of removed outliers, Savitzky-Golay with a Cubist model provided the best results with an RMSE of 0.16 g/cm^3, an R^2 of 0.72, a bias value of 0.01, a rhoC of 0.82, and an RPIQ of 2.31. Calibration algorithms for volumetric water content and DBD were also created for each catchment, where the results showed an overall further improvement over the regional algorithms that used all of the samples. The best catchment calibration algorithm for water content was the upper Olifants catchment's water content, with a low mean error and RMSE of 0.59 and 5.07% respectively, accompanied by a high rhoC, RPD and RPIQ of 0.81, 1.85, and 1.94 respectively. The best catchment algorithm for DBD was the Sabie catchment with a 0 mean error and low RMSE of 0.08 g/cm^3 a high rhoC, RPD, and RPIQ of 0.96, 3.52, and 3.75 respectively.

Algorithms were then created for water retention at 33 kPa, 1500 kPa, and bulk density using the regional dataset to compare with the freely available international OSSL algorithms. The validation set of each created algorithm was uploaded to the OSSL prediction service to obtain predicted values, which were then compared to the created algorithms. Although the created algorithms were poorly predicted, they were still superior to the OSSL calibrations. The OSSL algorithm for the drained upper limit water retention indicate poor results of an RMSE of 17.12%, a mean error of 12.38, and very low RPD and RPIQ of 0.59 and 0.51 respectively. The created algorithm for the drained upper limit water retention performed better, with an RMSE of 8.89%, a mean error of 1.98, and an RPD and RPIQ of 1.13 and 0.99 respectively. The OSSL model for the lower limit water retention performed very poorly with an RMSE, mean error, RPD, and RPIQ of 20.85%, 19.24, 0.39, and 0.5 respectively. The created algorithm performed better with an RMSE, mean error, RPD, and RPIQ of 8.35%, 1.57, 0.97, and 1.24 respectively. The results that were found supported the idea that local calibrations are necessary for accurate and reliable predictions. The eventual compilation of sufficient local calibrations can lead to effective regional calibration algorithms, where sufficient regional algorithms can lead to the effective and accurate use of international algorithms like the OSSL algorithms.

Index words:

Soil moisture, machine learning, near-infrared spectroscopy, lower limit, drained upper limit

Table of Contents

Table of Figures

List of Tables

Table of Equations

1 CHAPTER 1: INTRODUCTION AND CONCEPTUALIZATION

1.1. Background

It has been suggested that the potential cause of the next world war could be a struggle over freshwater resources (Singh *et al*., 2012; Tignino, 2010), given that it is arguably the most precious resource known to humankind. It has also been said that water is the cornerstone of all life (Skalko, 2013). The importance of water on our planet can thus not be overstated, as it forms the foundation of the biochemical function in all living organisms (Chaplin, 2001). Simply put, without water, life on earth would simply cease to exist.

Only about 2.5% of water on earth is freshwater usable by humans, whereas approximately 1.7% of global freshwater reserves are found below the surface as groundwater, while the rest is confined to glaciers and the atmosphere (Kikkas & Kulik, 2018). The availability of freshwater has declined over recent years due to the rapid growth of global population, the unrelenting progression of urbanization, and the surge in demand for goods and services that often involve clean, freshwater to manufacture and deliver (Hanjra & Qureshi, 2010). The situation in South Africa is worse than the global average, as the country is seen as being water scarce (Viljoen & van der Walt, 2018),

Soil is a first order control of the hydrological cycle (Yamanaka *et al*., 2007). Understanding its role in the hydrological cycle is thus of the utmost importance when it comes to the management of water resources. A critical factor to consider then, is the dynamic, interactive relationship between water and soil, also known as hydropedology (Van Tol *et al*., 2013). Hydropedology is an important field of study, especially in countries where agriculture under irrigation plays a key role in the economy, like in the case of South Africa (Rapanyane & Ngoepe, 2019). Understanding how water interacts with soil can aid in effectively managing irrigation schedules and will prevent issues like waterlogging and soil erosion (Yerro & Ceccato, 2023). Because South Africa is a third-world country with a struggling economy (Rapanyane & Ngoepe, 2019), the expenses associated with measuring soil properties is an important issue that needs to be addressed, especially soil moisture (Paterson *et al*., 2015). The quality and success of land management decisions also rely greatly on the accuracy of soil measurements upon which they are based (Packer *et al*., 2019). Soil characteristics tend to vary on a fine scale (Paterson *et al*., 2015), making the measurements of soil characteristics even more difficult, demanding sufficient accuracy when are being taken.

South Africa is also faced with the growing issue of land availability accompanied by a growing population. In a country that is deeply plagued by erosion and improper land-use practices, the availability of usable land for agriculture only decreases, making effective water management extremely crucial (Phinzi *et al*., 2020).

In addition to South Africa's land, soil, and water issues, there is also the lack of existing soil databases to monitor the health and development of soils (Paterson *et al*., 2015). Understanding soil properties and identifying certain soil characteristics, especially in agriculture, can drastically increase productivity by helping people utilising the soil more effectively. Not only can it increase the productivity of agriculture and crop production but may increase the efficiency at which water is managed as well. Conducting soil measurements to ensure effective productivity, especially any water related soil measurements, can prove to be challenging due to the expensive costs, inaccuracy in readings, as well as tedious processes they pertain (Afzali *et al*., 2021; Placidi *et al*., 2021).

Water content in soils have traditionally been measured using various methods, like tensiometers, electrical conductivity using moisture probes, neutron moisture meters, and gravimetric measurements. Methods like moisture probes and tensiometers are stationary, while methods like the neutron moisture meter and remote sensing are sometimes expensive. Gravimetric and volumetric soil moisture measurements are furthermore not always reliable in accuracy (Afzali *et al*., 2021; Placidi *et al*., 2021). This naturally, leaves room to find a better, more cost-effective solution. When observing newer technologies and methods, spectroscopy shows real promise. Near infrared spectroscopy (NIRS) stands out in spectroscopy since it provides portable and seamlessly accurate measurements without disturbing the soil in any way (Knox *et al*., 2015). NIRS can also provide real time data to accommodate fluctuating moisture levels. This method is like remote sensing, other than that it is portable, less prone to noise, and more economical and more rapid than remote sensing solutions (Knadel *et al*., 2017).

For the scanner to identify what it scans however, requires calibration using an algorithm which acts as a lexicon by which the scanner identifies features in the spectra it retrieves. The issue, however, is that there are no freely and readily available NIRS algorithms for predicting soil water content in South Africa. Most of these algorithms either demand membership or payment to access, and algorithms from other countries are not suitable due to the diverse and unique spectral signatures of South African soils.

1.2. Problem statement

Measuring water content in soil can be a difficult and tedious task, and most methods are labor-intensive, which can lead to errors or inaccuracy, especially since water levels in soil tend to fluctuate. This can result in inefficient management choices in water management. Near infrared spectrometry can provide a non-invasive and fast solution to this problem, however, it requires a calibration model to interpret results. There are unfortunately no such freely accessible soil water calibration models specifically designed for South African soils yet.

1.3. Hypothesis

The hypothesis tested in this study is that near-infrared spectroscopy (NIRS) can accurately determine soil water content in a variety of South African soils if efficient calibration algorithms are available.

1.4. Research aim and objectives

The aim of the study is to create near-infrared calibration models for soil water content and bulk density prediction for a wide variety of soils in South Africa.

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

1. To establish regional NIRS soil water content and bulk density calibration algorithms for soils from five catchments in South Africa.

2. To determine at which scale NIRS calibration algorithms perform best by establishing water content and bulk density prediction calibration algorithms for each individual catchment.

3. To compare created algorithms of drained upper limit, lower limit and bulk density against feely available international calibration algorithms for the same properties.

2. CHAPTER 2: LITERATURE STUDY

2.1. Soil water and physical properties

The soil water content can be described as the quantity of water contained within the pore spaces between soil aggregates and is usually measured in mass or volume (Brocca *et al*., 2017). When water permeates the soil, it forces air to displace from the pore spaces. Soil water content is then considered to be indirectly proportional to soil air content (Brady & Weil, 2017; Kome *et al*., 2019). Larger macro pores within soils tend to aerate quicker after being filled with water due to the ease at which the water evaporates and *vice versa*, while smaller or micro pores generally take longer to aerate or fill with water. Soils with a high clay percentage are thus considered to have a greater water holding capacity, while soils with larger macro pores like sandy soils have poor water holding capacity. This can also be attributed to particle size, where clayey soils have smaller particles, thus having a larger contact surface area enabling them to hold onto water molecules more effectively. Where sandy particles are larger and have a smaller contact surface area, resulting in a less effective hold on water within the soil (Brady & Weil, 2017; Kome *et al*., 2019; Sujatha *et al.,* 2016). This means that clayey soils with smaller pore spaces are poorly aerated in comparison to sandy soils with larger pore spaces. Since clay soils have a higher volume of pore spaces due to the tiny size of the clay particles, they have the potential to have a higher average water content than sandy soils, especially when saturated (Amoakwah *et al*., 2017; Bordoloi *et al*., 2018). Water content also affects friability and strength of soil particles, especially fine textured soils. Literature suggests that when water content levels increase in a soil, it can progressively reduce the shear strength of the soil by hindering cohesion forces between soil particles, leading to the disintegration of soil aggregates (Brady & Weil, 2017; Wei *et al*., 2018). Each soil type then, has an optimal water content level that should be considered especially in the field of construction.

2.2. Soil colour and spectral properties

Figure 1: Image showing the effect water has on the soil colour (Brady & Weil, 2017).

Water can also influence a soil's colour. Figure 1 shows how water alters the colour of soil, where the darker colour indicates the wet soil, while the lighter colour indicates the natural colour of the soil without excess water. This is mainly due to the shift of refraction index (*n*) from air which is (*n = 1)* to water which is (*n = 1.33)*. This reduces the contrast between soil particles (*n = 1.5*) and their surroundings, increasing the average degree of forward diffraction by incoming light and consequently the increasing the likelihood of absorption before the light exits from the medium (Lobell & Asner, 2002; Twomey *et al*., 1986).

This alteration affects the spectral reflectance characteristics of soil, particularly in the near-infrared (NIR) region. In the NIR spectrum, soil moisture exhibits distinctive absorption features that NIRS can capture (Oliveira et al., 2013; Slaughter et al., 2001). The presence of water in the soil alters the vibrational modes of water molecules, leading to unique absorption bands in the NIR range. These unique bands are also called spectral signatures. Some of these signatures are directly related to the amount of water present and are crucial for predicting soil moisture levels. The water absorption bands in the NIR region are usually found around 1400 nm to 1900 nm (Oliveira et al., 2013; Slaughter et al., 2001). Furthermore, the reflectance spectrum in the visible and NIR regions is sensitive to the physical and chemical properties of soil, providing a plethora of information for soil characterization. The information it can provide includes mineral

composition, organic matter, soil texture, bulk density, and soil moisture (Ahmadi et al., 2021; (García-Sánchez et al., 2017).

2.3. Soil organic carbon (SOC), biological, and microbial activity

Water has, however, more intricate effects on a soil's colour than one might realize. Temperature, water content, pH, as well as the diversity and quantity of microorganisms, all influence soil microbial processes, including the carbon cycle (Ghezzehei *et al*., 2019).

Soil water content influences microbial activity, which in turn, influences the rate at which organic matter accumulate and decompose. Second, the soil water content present impacts the amount of organic matter extractable by water, which is a vital substrate for microbial metabolism (Hafizah & Khairunniza, 2011; Waldrip *et al*., 2022). The amount of SOC furthermore influences the water holding capacity of a soil, where higher amounts of SOC contribute to better water holding capacity (Manns *et al*., 2016). Agricultural soils typically react differently to the effects of soil water, with drought typically producing more severe changes in respiration rates and microbial activities than flooding (Schnecker *et al*., 2021).

The water content affects the quantity of oxygen in the soil, which will furthermore influence the organic activity in the soil (Moyano *et al*., 2012). Organic activity, for example the presence of earth worms can be very beneficial to the soil, as they contribute to the amount of organic matter in the soil as well as increasing aeration by breaking apart aggregates, which can also aid in healthy plant root growth (Sharma *et al*., 2019).

2.4. Redox reactions

Water furthermore influences the rate at which manganese and iron reduce and oxidize. The reduction of both manganese and iron can produce greyish colours due to the removal of these ions by the movement of soil water, due to the fact that these reduced ions are soluble in water. When these soluble ions oxidize in high concentrations, colouring is increased in the areas where they oxidize. In the case of the oxidation of manganese, black mottles may be increased, while the oxidation of iron can create red or yellow colours depending on the ion (Kalvani *et al*., 2021). Figure 2 shows the effect of iron redox reactions in soil.

Figure 2: The left image shows red/brown oxidized iron mottles, while the right image shows dark manganese mottles (Brady & Weil, 2017).

2.5. Soil genesis and type

The redox reactions in soil can also help in describing why soils in arid, dry areas are more prone to physical weathering where water is mostly absent, while soils in humid, wet areas are more likely to be weathered by chemical processes that flourish with water as a medium for various reactions. Without redox reactions, physical weathering processes such as freeze-thaw cycles and abrasion dominate, leading to the breakdown of soil particles (Zhang & Furman, 2021). These processes of categorising physical and chemical weathering in soils on grounds of location is neatly summarized under the term called the 'Weinert-N value'. The Weinert-N value categorizes regions by the type of weathering soils and geological units will undergo depending on the climate conditions in the region (Faul, 2018; Weinert, 1984). The N-value is a ratio that depicts the annual rainfall (*Pa*) to evaporation during the hottest month (*Ej*), which is displayed in Equation 1.

$$\boldsymbol{N = 12\frac{Ej}{Pa}} \quad [1]$$

This value represents the influence of climate on the type of weathering that will be present. Where wetter, more humid climates will more likely result in chemical weathering, and dryer climates will likely result in physical weathering. Since the one of the primary soil-forming factors in South Africa is the climate (le Roux & du Preez, 2006), the Weinert-N value will thus indirectly affect the type of soil that will form (Futshane *et al.,* 2022).

Figure 3 divides Southern Africa up into its different climatic regions through the use of the Weinert-N value. A Weinert-N value larger than 5 would typically mean a dry climate lacking precipitation, indicating physical weathering and disintegration of soil aggregates. A low Weinert-N value of lower than 2 would indicate a humid/wet climate, indicating chemical weathering and decomposition of soil aggregates. While a moderate value of between 2 and 5 would indicate a mixture of physical and chemical weathering processes (Faul, 2018; Weinert, 1984).

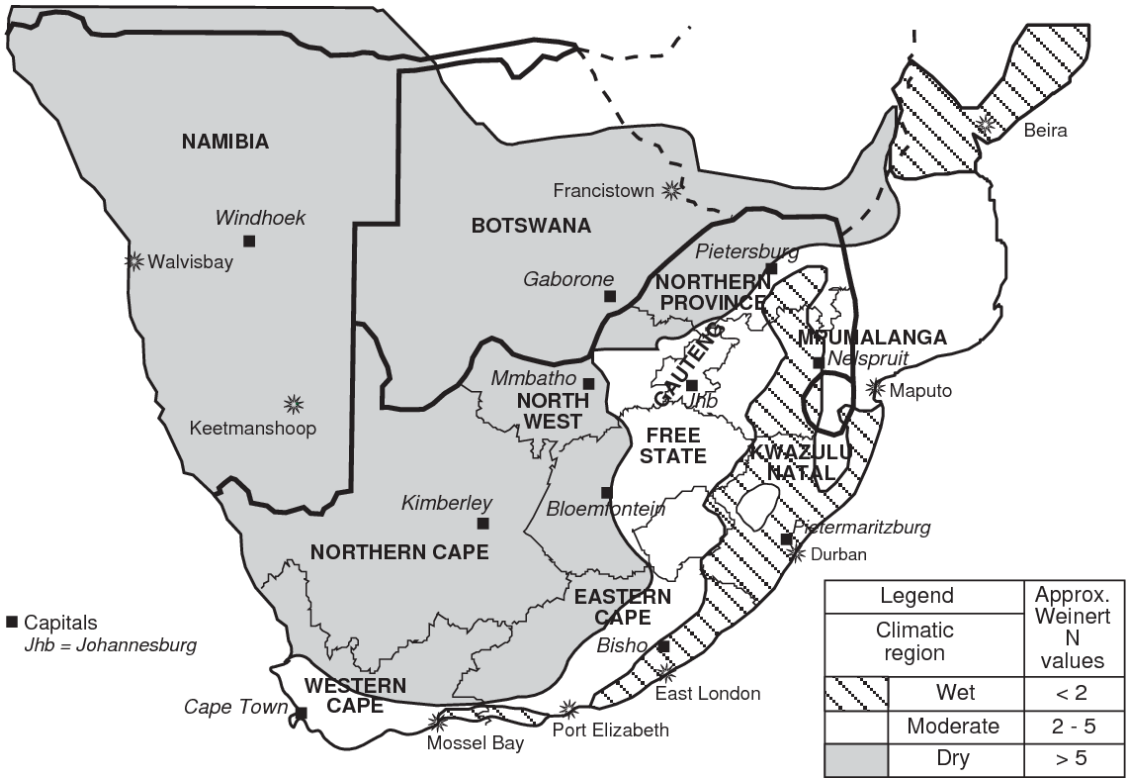

Figure 3: Map that categorizes South Africa into different climatic regions according to the Weinert-N values (Van Zyl, 2004).

Climate type narrowed down to water content in soils ultimately affects a myriad of processes limited not only to agriculture but includes the ultimate land-use of the country collectively. This map also indicates why South Africa is considered a semi-arid country, since the majority of the country's surface area has a high Weinert-N value (N>5), indicating dry climate conditions accompanied by physical weathering (Faul, 2018; Weinert, 1984). Since the major soil forming factors in South Africa are mainly parent material and climatic conditions (Breytenbach, 2010), the Weinert-N value can be an indicator of the nutrient availability, evaporation rate, soil pH, and soil organic matter (Breytenbach, 2010; Faul, 2018; Nesbitt, 2014).

2.6. The importance of soil water content in agriculture

Along with affecting the physical and chemical aspects of soils, water also has the capacity to influence biological activity within a soil. For this reason, the water content of a soil holds significant value in the field of agriculture (Défossez *et al*., 2003; Roper *et al*., 2013). Water content not only influences the plant-available water, which is described as the amount of water retained in the soil between the permanent wilting point or lower limit (1500 kPa) and the soil's field capacity or drained upper limit (33 kPa), but it also largely dictates how essential nutrient transport in the soil, ultimately affecting how organisms and plants will grow (Brady & Weil, 2017; Chen *et al*., 2019). The water content between the drained upper limit (DUL) and lower limit (LL) is known as the plant available water content (Brady & Weil, 2017). If the water content of the soil exceeds that which is found at the DUL, the plant will drown and wilt, whereas of the water content is below that which is found at the LL, the plant will dry out (Eitzinger *et al*., 2004). Scheduled irrigation makes use of these pressure to ensure that the water content in the soil is always sufficient for a plant to survive (Sutcliffe *et al*., 2021).

Figure 4 displays the relationship between soil water content and matric potential. As the water content increases, the energy needed by plants to extract water from the soil decreases due to the absence of capillary action and abundance of water, whereas plants will need to exert more energy to extract water if the water content is low (Brady & Weil, 2017; Chowdhury *et al*., 2011). As the water content increases the matric potential also increases and vice versa, making the correlation direct. Figure 4 also shows the interesting phenomena known as hysteresis.

Hysteresis in terms of soil water can be defined as the difference seen between soil's water content and the associated water potential produced through the wetting and drying process. When soil dries, the water molecules stick more tightly to the soil particles, making them harder to remove. When soil is wetted, the water molecules stick less tightly to the soil particles, making them easier to remove (Zhang *et al*., 2018). Moreover, the term soil water potential simply describes the difference in energy status between water present in the soil and pure water (Brady & Weil, 2017).

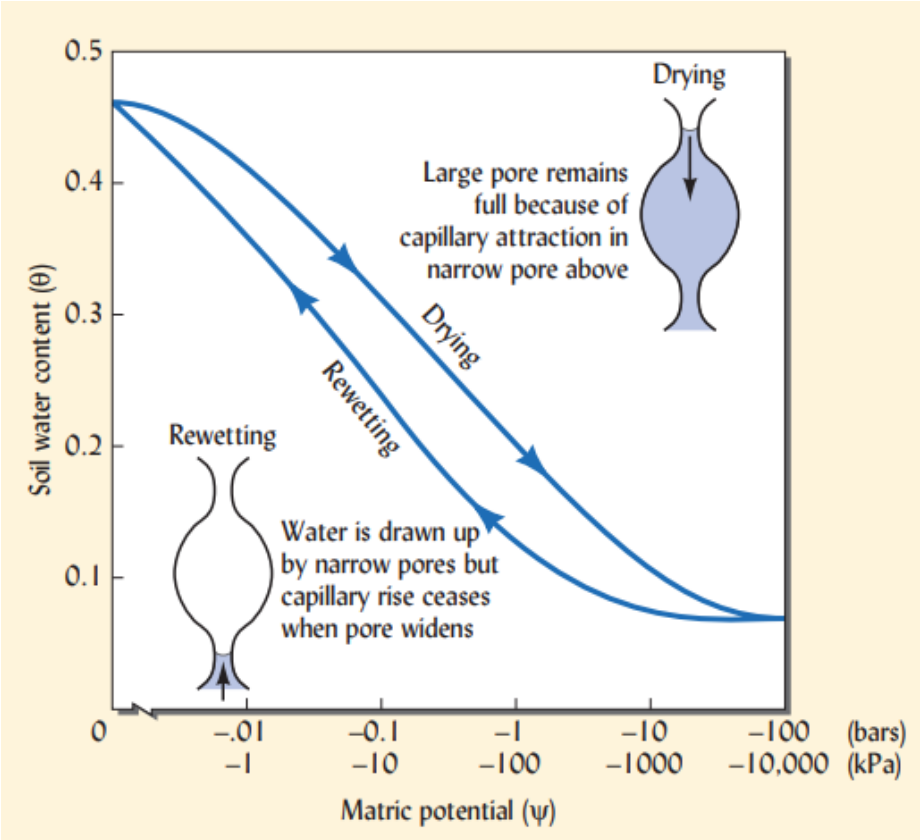

Figure 4: Diagram showing the relationship between soil water content and matric potential where drying and wetting takes place (Brady & Weil, 2017).

In agriculture, it is also imperative for farmers to plough at a certain soil water content (Patil *et al*., 2015). Farmers often plough their fields after rain when the residual water content makes the ploughing easier. The water content should however not be higher than approximately 30%, as it may contribute to excess compaction caused by heavy machinery as excess water content can act as a lubricant for soil to move into a more compact state (Rahmat & Ismail, 2018) This makes measuring the water content in soils necessary for accurately determining when the water content is optimal for ploughing (Ahmadi & Mollazade, 2009).

Knowing the water content level in soil can however be useful for the planting season as well. Planting when the soil is soft due to sufficient water can help reduce soil structure by weakening soil aggregates, making it possible for plant roots to penetrate the soil easier, where dry soils may present stronger structure that can hinder plant growth (Sadeghi *et al*., 2014). Having an optimal soil water content, approximately 60-80% of the soil's water holding capacity when planting can also boost plant growth when water can be readily absorbed by crop seeds (Nguyen *et al*., 2021). Furthermore, water in the soil before planting serves as a medium for essential nutrient transport, as well as dissolving salts. (Ghezzehei *et al*., 2019). Excessive water in soil can nonetheless have adverse effects on plant growth. Nutrient absorption by plants is a metabolic process, meaning

that it requires energy, and can be hindered by factors that limit root metabolism. Excessive soil water in the form of waterlogging can lead to reduced oxygen levels in the soil as well as create favourable conditions for plant root pathogens that may pose a threat to plant health (Ahmad *et al*., 1992; Zhang *et al*., 2019).

2.7. Conventional methods of measuring soil water content

2.7.1. Gravimetric method

The gravimetric method is arguably the most used method, as manifold studies have utilised this method and is often used as a reference to determine the accuracy of other methods (Bittelli *et al*., 2008; Chow *et al*., 2009; Serbin & Or, 2004) proving that this method is reliable when it comes to accuracy (Cosh *et al*., 2005). The method operates by weighing a moist soil sample and then oven drying it for 24 hours to remove all the water, and then weighing it again to determine the mass of the soil particles by subtracting the mass of the core, making it possible to calculate the bulk density using the known volume of the core sample with the mass of the soil particles (Chow *et al*., 2009).

The gravimetric water content is determined by using Equation 2 (Chow *et al*., 2009)

$$\boldsymbol{Gravimetric\ moisture\ content\ (w) = \frac{Mass\ of\ the\ moisture\ (g) - Mass\ of\ the\ dry\ sample(g) * 100}{Mass\ of\ the\ dry\ sample(g)}} \quad [2]$$

Whereas the bulk density can be determined using Equation 3 (Chow *et al*., 2009):

$$\boldsymbol{Bulk\ density(\rho b) = \frac{Mass\ of\ the\ dry\ sample\ (g)}{Volume\ of\ the\ dry\ sample\ (cm^3)}} \quad [3]$$

Even though this method is quite simplistic, it is very time-consuming, labour-intensive, and destructive since soil samples must be taken (Cosh *et al*., 2005) which is why other methods of predicting water content in soils are being explored.

2.7.2. Volumetric method

Similar to the gravimetric method, a soil sample is taken in a core with a volume that is predetermined. Oven drying takes place to remove all water present within the sample, where the mass of the remaining dry soil is used to calculate the mass of the removed water (Chow *et al*., 2009; Serbin & Or, 2004). The mass of the removed water, as well as the density of water is then used to calculate the volumetric water content.

The volumetric water content can then be determined based on the gravimetric water content and the bulk density using Equation 4 (Chow *et al*., 2009):

$$\boldsymbol{Volumetric\ water\ content}\ (\boldsymbol{\theta}) = \boldsymbol{gravimetric\ moisture\ content} * \boldsymbol{bulk\ density} \quad [4]$$

This method makes use of the gravimetric method's process, meaning that it is also time-consuming, invasive, and labour-intensive (Rasheed *et al*., 2022).

2.7.3. Electrical resistance

The electrical conductivity or resistance in soil is dependent on numerous factors, of which include the soil's water content. Resistance blocks are embedded within the soil where the resistance to the path of an electrical current between the two blocks are measured as a product of the water content in the soil. The electrical resistance is measured at different water content levels in order for calibration to be possible to interpret changes in the water content. In reality, though, the contact resistance between the resistance blocks and the soil might significantly surpass the actual resistance of the soil water since it is a fluctuating parameter. (Schwartz *et al*., 2008)

To prevent this error of occurring, the absorbent gauge was developed, where instead of resistance blocks being placed in the soil, they are placed in an absorbent material which is then embedded in the soil. When buried in the soil, the absorbent material of the blocks rapidly collects or release water, allowing the block's water content to remain in balance with the soil's water content. These changes in water content result in variations in electrical resistance, which may be measured using a data logger at the surface. A calibration chart is used to translate the meter's resistance readings to water content figures (Croney *et al*., 1951). This method is non-invasive and non-destructive and is relatively easy and affordable to maintain. It also gives continuous measurements. The main drawback, however, is that it is not portable and thus placed in a fixed position. The accuracy is also limited in dry soils where there is no water for conduction to take place, thus limiting its applicability (Schwartz *et al*., 2008).

2.7.4. Tensiometers

Tensiometers employ a porous clay cup linked to a water-filled tube to detect the matric potential of soil water in situ. Either a vacuum gauge or a mercury manometer can be connected to the water cup and tube (Gaddikeri *et al*., 2021; Shock & Wang, 2010; Shreeja, s.a.). Figure 5 shows the tensiometer layout in a soil.

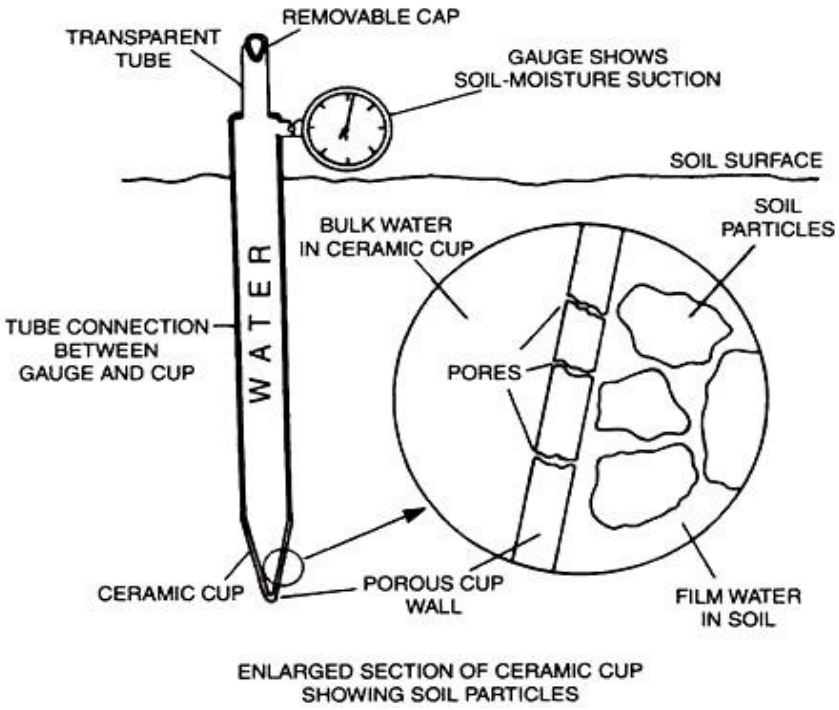

Figure 5: Diagram of a tensiometer, with an enlarged representation of the porous ceramic cup that is in contact with soil water (Shreeja, s.a.)

As the soil dries, water drains through the permeable cup, producing a vacuum on the water column. These vacuum measurements are then calibrated on the gauge to the variety of soil present to determine the percentage of water. (Gaddikeri *et al*., 2021; Shock & Wang, 2010).

The biggest drawback of tensiometers is that they do not measure soil matric potential values as low as average wilting values. The range of accurate measurement is however limited from 0 to 85 kPa. Like the electrical resistance method, tensiometers are also fixed in a certain space, and is not portable. Tensiometers are more useful for measuring soil water content in sandy soils than that of clay soils, because of higher matric potentials found in clay soils (Gaddikeri *et al*., 2021; Shock & Wang, 2010).

2.7.5. Neutron scattering

The neutron scattering method is applied through neutron moisture meters or probes (Hardie, 2020; Sutcliffe *et al*., 2021). Neutron moisture probes function by measuring the amount of neutron scattering in the soil. The probe consists of a source that rapidly emits neutrons into the soil. Neutrons are electrically neutral subatomic particles. When neutrons are emitted into the soil, they interact with the atom nuclei. The quantity of interactions that take place are determined by the type of atom. Soil water, which is prevalently hydrogen, is very good at scattering neutrons.

This means that if a neutron collides with a hydrogen atom, it will most likely be deflected from its initial route. The probe also has a detector that counts the number of neutrons that are reflected to it. The more water in the soil, the more neutrons are deflected back to the detector. This makes the measurement of water content in soil possible (Sutcliffe et al., 2021; Yin *et al*., 2013).

Neutron probes are non-invasive, rapid, portable, and accurate. The drawback of this technique, however, is that it is very expensive, requires calibration for each different soil profile, and can be dangerous to operate due to radiation (Sutcliffe *et al*., 2021; Yin *et al*., 2013).

2.8. Spectroscopy

Spectroscopy in soil science is a somewhat new and growing application that allows for the uninterrupted and accurate acquisition of a variety of soil properties at a relatively inexpensive cost (Poppiel *et al*., 2022; Seybold *et al*., 2019). Spectroscopy operates on the foundation of how distinct objects reflect energy differently, often due to their unique physical or chemical characteristics. NIRS scanners emit energy at known wavelengths, which interact with the soil and the reflected energy at these wavelengths is measured (Aenugu *et al*., 2011). Calibration algorithms are needed to correlate the measured reflected energy values to soil properties (Aenugu *et al*., 2011; Du & Zhou, 2008).

Although there are numerous spectroscopy methods that operate at different wavelengths and for different purposes, this study will use near infrared spectroscopy (NIRS), as it is an instrument that can be used under field conditions that is also portable and highly cost-effective.

2.8.1. Near infrared spectroscopy

Near infrared spectroscopy is a rapid, cost-effective, and non-destructive method of soil analysis that operates using energy waves with wavelengths between 700-2500nm on the electromagnetic spectrum. Figure 6 indicates where the near infrared section is on the electromagnetic spectrum.

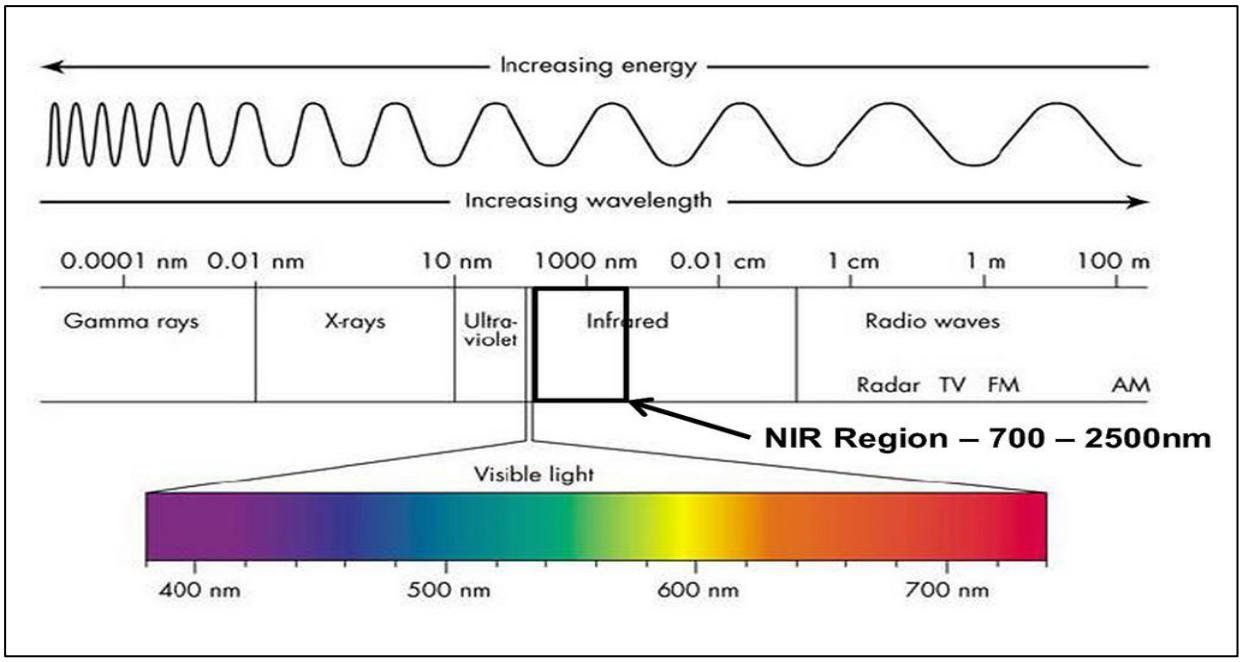

Figure 6: Where the NIR region is on the electromagnetic spectrum (kpm analytics, s.a.)

This method is used by a myriad of studies for soil analysis, providing data for a variety of soil properties including clay content, organic material content, mineral composition, pH, nutrient status, as well as properties like soil texture, structure, and bulk density (Aenugu *et al*., 2011; Cambule *et al*., 2012; Janse Van Vuuren & Groenewald, 2013; Knox *et al*., 2015).

The near infrared waves of energy used in this method cause hydrogen bonds like O-H, N-H as well as weaker covalent bonds like H-C to vibrate in unique ways (Bullock *et al*., 2004). When near infrared spectroscopy is used, two types of vibrations may occur, which are: *bending* which alters the angle between two bonds and *stretching* which alters the length of a bond (Bullock *et al*., 2004; García-Sánchez *et al*., 2017).

2.8.1.1. Advantages and limitations

Near infrared spectroscopy is relatively inexpensive, as well as time-efficient and non-invasive. It is advantageous over most other methods due to its ease of operation, especially in hand-held scanners, the large number of readings it can measure in a short amount of time, as well as the capacity to measure various soil properties in a single scan (García-Sánchez *et al*., 2017). The main drawback of this method, however, is the requirement for calibration algorithms for measurements to be meaningful. The creation of calibration algorithms can be a complex and tedious process and requires multivariate analysis for data to be usable. NIRS also suffers highly from scatter effects, as well as overtones, further demanding the need for calibration (García-Sánchez *et al*., 2017).

2.8.2. Vis-Near infrared spectroscopy (vis-NIRS)

Similar to NIR, vis-NIR is simply the NIR region of the electromagnetic spectrum that extends into the visual light of the spectrum. Referring to Figure 6, vis-NIR consists of wavelengths of approximately 350 nm to 2500 nm. As mentioned in Section 2.1, soil water alters a soil's colour, which naturally means that vis-NIR can pick this change up since it includes the visual part of the spectrum. This allows for more information to be obtained about the soil but will evidently require more processing time and power. But this may however mean that certain soil properties can be more accurately determined.

From Table 1 it is evident that the accuracy of models tends to increase as the scale gets smaller, where regional and international models like the ones in Koirala *et al*. (2022) and Liu *et al*. (2020) are slightly less accurate than site-specific model that were created for specific farm fields or other similarly sized areas like in the studies done by Gou *et al*. (2020) and Liang *et al*. (2012). This supports the idea that local and site-specific calibrations are required for more accurate predictions since it is easier for local calibrations to express the soil spatial variability in an area.

Furthermore, the sample size correlates well with the accuracy of the calibrations, where smaller samples sizes <100 like in the studies of Bullock *et al*. (2004) and Liu *et al*. (2020) have lower RMSE and R^2 values of 6.4 %, 0.95 and 7.67 %, 0.714 respectively, compared to studies that have more samples like Chen *et al*. (2021) and Mouazen & Al-Asadi (2018) with 1.6 %, 0.989, and 1.2 %, 0.95-1 respectively.

Another important aspect to consider is the wavelength selection. Looking at the studies done by Gou *et al.* (2020) and Mouazen & Al-Asadi, (2018), their wavelength selection includes the visual spectrum from 350 nm to 2500 nm, and their results are significantly better than studies that only use the NIR part of the spectrum like Bullock *et al.* (2004) and Minasny *et al.* (2009). But one can also argue that the age of the study also plays a role, where the older studies like the ones done by Bullock *et al.* (2004) and Minasny *et al.* (2009) had less accuracy than the newer ones due to technological improvements over the years. This is also evident in the wavelength selection and spectral resolution, as more recent studies have better spectral resolutions and expanded wavelength selections when compared to older studies. This trend suggests an ongoing refinement in NIRS technology for soil moisture prediction.

2.8.3. Studies that have used NIRS for soil water content prediction

Table 1 displays studies that predicted or determined soil water using vis-NIR/NIR with the wavelengths they used, resolution of the scanners, samples size, location, and their best results.

Table 1: Studies that have used NIRS for soil water content prediction.

Study	Wavelengths used	Resolution	Sample Size	RMSE %	R^2	Location	Scale	Soil texture
Bullock *et al.*, 2004	1100-2500 nm	2 nm	45	6.4	0.95	Canada	Site specific	Different soil textures
Minasny *et al.*, 2009	1400-1900 nm	2 nm	416	5.1	-	Australia	Site-specific	Different soil textures
Fabre *et al.*, 2015	400-2500 nm (vis-NIR)	10 nm	32 (190 spectral signatures)	2.9	0.96	France	National	Different soil textures
Liang *et al*., 2012	900-2550 nm	6.8 nm	116	1.2	0.9913	China	Site-specific	Similar textures
Chen *et al.*, 2021	400-2500 nm (vis-NIR)	10 nm	251 (3479 spectral signatures)	1.6	0.989	China	International	Different soil textures
Mouazen & Al-Asadi, 2018	350-2500 nm (vis-NIR)	10 nm	300	1.2	0.95-1	UK	Site specific	Different soil textures
Liu *et al*., 2020	400-999 nm (vis-NIR)	-	54	7.67	0.714	China	Regional	Different soil textures
Gou *et al*., 2020	350-2500 nm (vis-NIR)	1.4 nm	72	0.88	0.973	China	Site-specific	Similar textures
Koirala *et al.*, 2022	350-2500 nm (vis-NIR)	1 nm	153 (from different studies)	2-7	-	Multiple locations	International	Different soil textures

Root mean square error = RMSE, Correlation coefficient = R^2 , Visual near infrared = vis-nir

The texture of the soil also plays an important role in calibration predictive capabilities. The studies done by Liang *et al*. (2012) and Gou *et al*. (2020) both conducted studies on soils with similar textures, and both achieved impressive results with RMSE values lower than 1.2% and R^2 values higher than 0.97, indicating that the predictive capabilities of NIRS within soils of similar texture has great potential.

The spectral resolution on the other hand, does not show significant effects on the accuracy of calibrations. This can be seen when comparing the studies of Chen *et al*. (2021) with Koirala *et al*. (2022). Koirala *et al*. (2022) utilised a finer resolution of 1 nm on an international scale and obtained RMSE values of 2-7%, where Mouazen & Al-Asadi (2018) utilised a coarser spectral resolution of 10 nm on a local scale and obtained a much better RMSE of 1.2%, suggesting that the resolution does not play a significant role in the accuracy of the calibration, but that the scale of the study plays a much more important role in the accuracy of calibrations.

Overall, the studies demonstrate good accuracy in predicting soil moisture using NIRS, with some obtaining exceptionally low RMSE and high R^2 values.

2.9. Creating a soil spectral library

Creating a soil spectral library is crucial because it allows for the estimation of soil attributes with high accuracy, efficiency, and low cost (Mohammedzein *et al.*, 2023). This information is imperative for soil assessment, agronomic and environmental management decisions, spectroscopy, precision agriculture, and even digital soil mapping (Mendes *et al*., 2022; Shepherd *et al*., 2022).

A soil water spectral library is created by collecting and analysing soil samples in order to establish a link between spectral data and other lab measured soil properties so that calibration can take place (Bellinaso *et al*., 2010; Kock, 2022). This procedure entails collecting representative soil samples, preparing them for spectroscopic examination, and measuring their spectra with a handheld spectrometer. It is essential to fully quantify the spectral reflectance of soil samples by collecting spectral data at different wavelengths (Kock, 2022). The measured soil properties and the spectra will then be joined to form one big spectral library, which can be used to create calibration algorithms.

2.10. Spectral pre-processing

Research indicates that spectral pre-processing, especially in NIRS, can significantly improve calibration results (Carvalho *et al*., 2022; Mishra & Lohumi, 2021). Pre-processing approaches try to reduce undesired spectral fluctuation such as baseline offsets, instrument drift, and noise

(Kock, 2022; Wadoux *et al*., 2021). Baseline correction, smoothing, standardization, and derivatives are all prominent pre-processing approaches. Smoothing decreases noise and improves peak clarity while baseline correction removes any background noise in the spectra, normalization adjusts the spectra to a common reference value, and derivatives emphasize spectral characteristics (Kock, 2022; Wadoux *et al*., 2021).

2.10.1. Multiplicative Scatter Correction (MSC):

Multiplicative scatter correction (MSC) is a method used to remove the effects of scattering from spectral data. Scattering is a phenomenon that causes light to be scattered in different directions when it interacts with a sample (Mishra & Lohumi, 2021). In the case of soil, this might occur due to plants or roots on the surface, or due to the roughness in texture that can scatter incoming light or radiation. MSC works by dividing the measured spectrum by a reference spectrum. The reference spectrum is a spectrum that represents the scattering effects but does not contain any information about the sample's chemical composition but does however contain information about its physical properties. This results in a corrected spectrum that is more representative of the sample's inherent properties (Silalahi *et al*., 2018). Figure 7 is a representation of multiplicative scatter corrected spectrum (red line) versus the original spectrum in black. It makes it possible to compare multiple spectra from different samples by standardizing the scale of the spectral data to identify similarities and differences (Silalahi *et al*., 2018).

MSC is a widely used method for pre-processing spectral data in a variety of applications, including food science, agriculture, and pharmaceutical analysis. It is a relatively simple and effective method for removing the effects of scattering, and it can be used to improve the accuracy and robustness of spectral calibration models (Maleki *et al*., 2007; Silalahi *et al*., 2018).

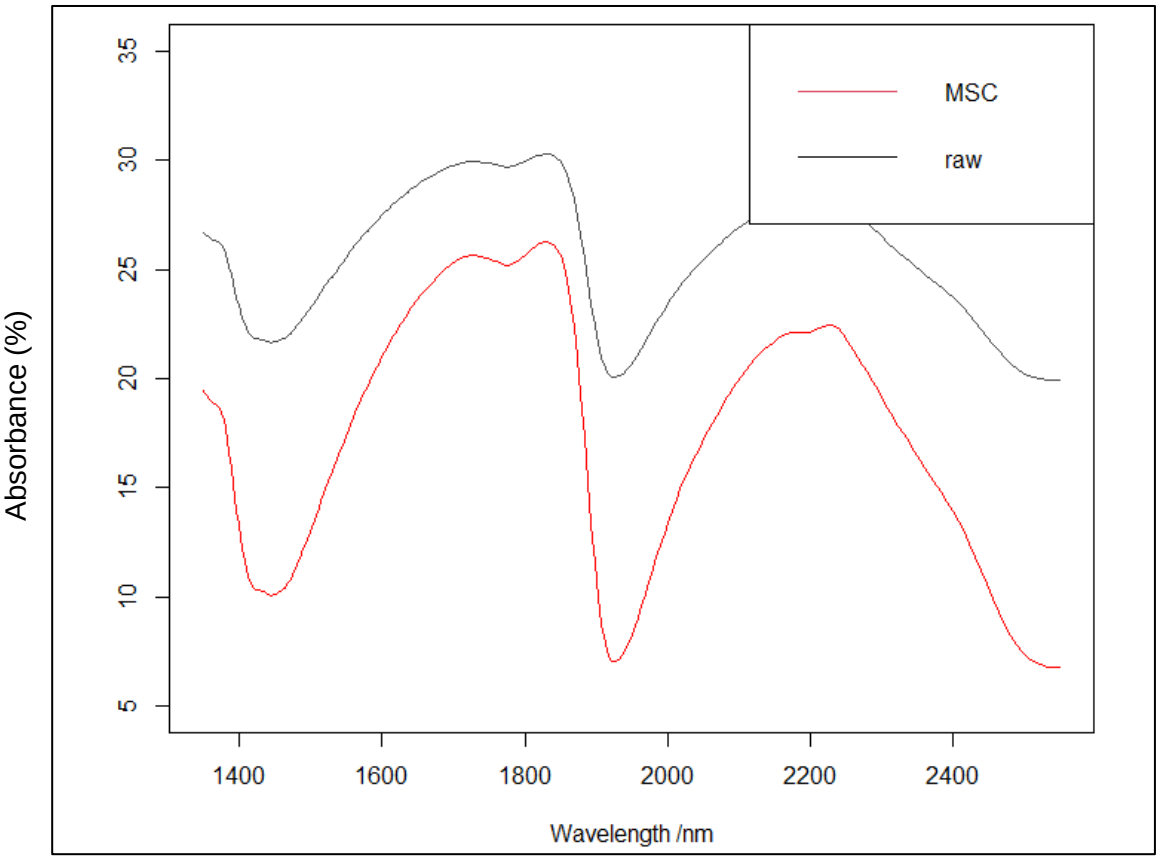

Figure 7: Multiplicative scatter-corrected (red line) and original (black line) spectrum.

2.10.2. Savitzky-Golay (SG):

Savitzky-Golay smoothing is a pre-processing method that is used to remove high-frequency noise in spectra while preserving the overall shape of the spectra, thus smoothing the data. SG works by fitting a polynomial to a portion of the spectra's data points. The smoothing qualities of the filter are determined by the size of the portion and the position of the polynomial. A bigger selection and a higher-order polynomial will provide a smoother spectrum, but the spectral shape may be slightly distorted (Chen *et al*., 2013; Delwiche & Reeves, 2010). SG is also able to calculate the derivatives of spectra (Schmid *et al*., 2022) making it a relatively popular pre-processing method that is applied in various fields including: geoscience, chemistry, and medicine (Schmid *et al*., 2022). Figure 8 shows how SG changes an example spectrum.

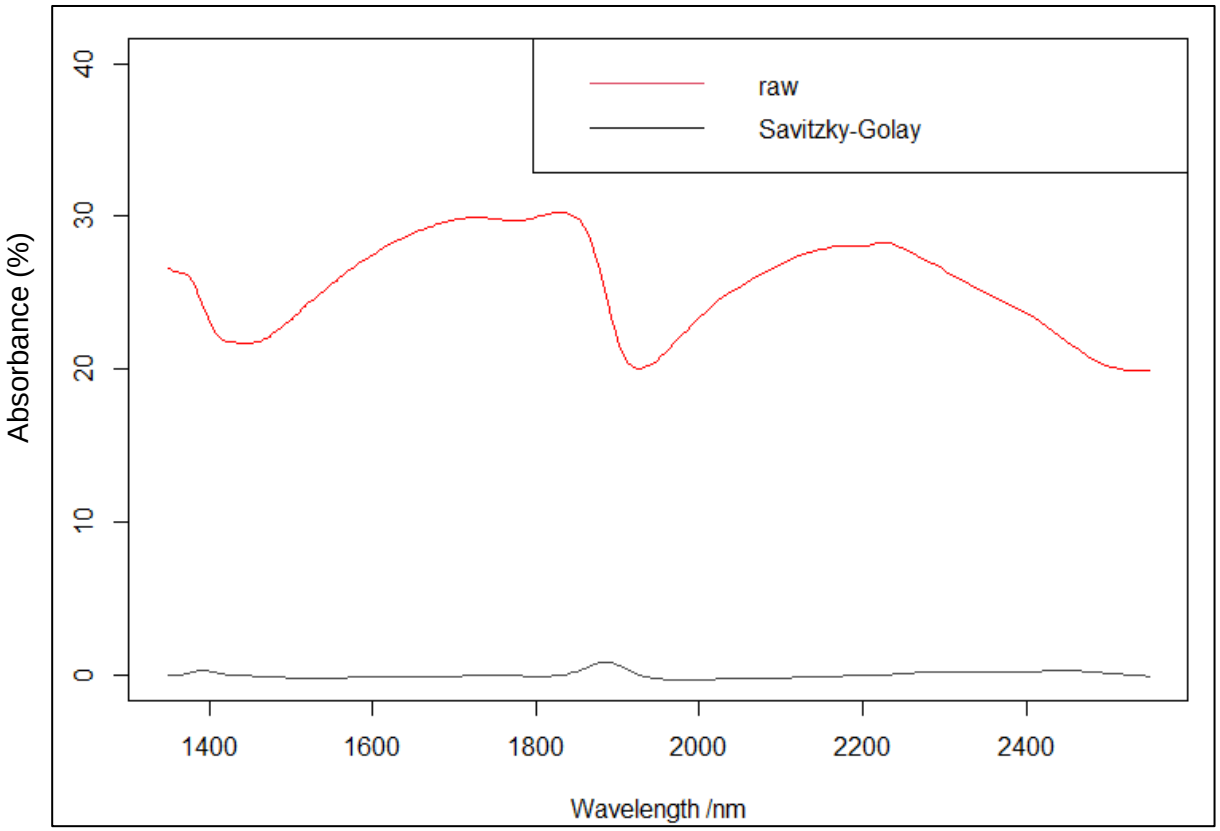

Figure 8: Savitzky-Golay filtering with second order derivative (black line) spectrum against a raw spectrum (red line).

By smoothing noisy data through polynomial fitting within a moving window, lowering noise, and sharpening spectral peaks, Savitzky-Golay filtering improves the interpretability of spectra. Derivatives, particularly the second derivative, amplify spectral features, highlighting changes in slope and assisting in peak identification. This filtering technique enhances the signal-to-noise ratio, making it less difficult to distinguish relevant information in complicated spectra and, as a result, allowing for more reliable and accurate analytical observations in NIRS (Chen *et al*., 2021).

2.10.3. Standard Normal Variate (SNV):

Standard normal variate (SNV) is a pre-processing method that focusses on removing the baseline offsets and reducing spectral intensity variations. This is achieved by subtracting the mean of all the spectra from each individual spectrum and dividing it by the standard deviation of all the spectra. SNV is particularly effective at managing fluctuations in baseline and intensity, which improves the robustness and accuracy of models applied to spectroscopic data (Al Riza *et al*., 2023; Carvalho *et al*., 2022; Engel *et al*., 2013).

SNV is a prominent spectral pre-processing method in NIRS. It is also fairly easy to configure and may be used with a wide array of algorithms (Carvalho *et al*., 2022; Li *et al*., 2021). Figure 9 shows how SNV transforms a spectrum.

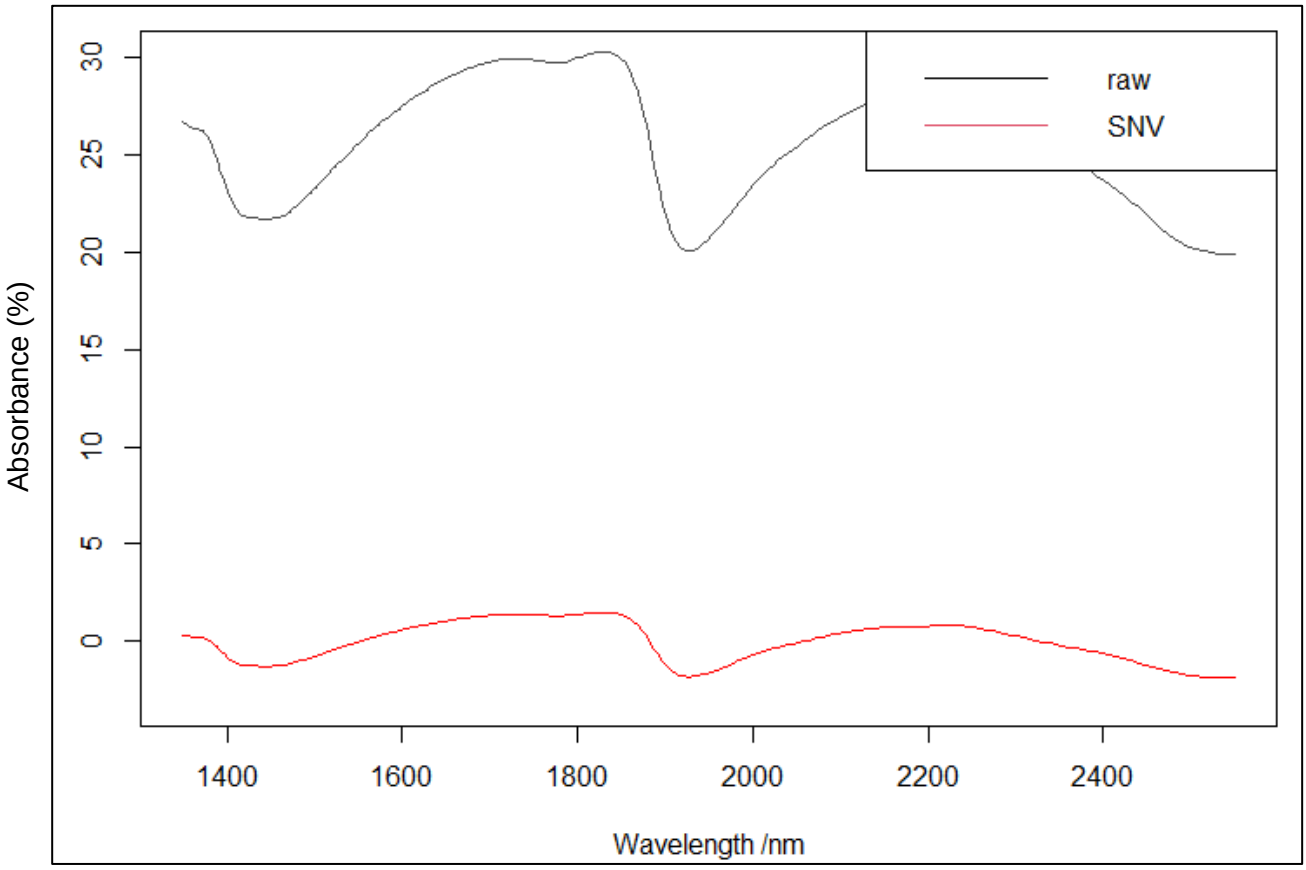

Figure 9: SNV spectra (red line) against raw spectra (black line).

2.10.4. Centring and standardization:

Similar to SNV, standardization also subtracts the total spectral mean from a single spectrum and then dividing it by the total spectral standard deviation (Engel *et al*., 2013; Wadoux *et al*., 2021). The difference, however, is the centring. Centring a wavelength is calculated by subtracting the spectral wavelength value by the average of all spectral values for that specific wavelength. In essence, centring and standardizing is applied to each specific wavelength, whereas SNV is applied to each individual spectrum (Engel *et al*., 2013; Wadoux *et al*., 2021). Figure 10 visualizes the effect of centring and standardization on an example spectrum. Standardization can improve prediction algorithm performance and interpretability by standardizing variable scales, lowering noise, and enhancing data comparability (Watanabe *et al*., 2021).

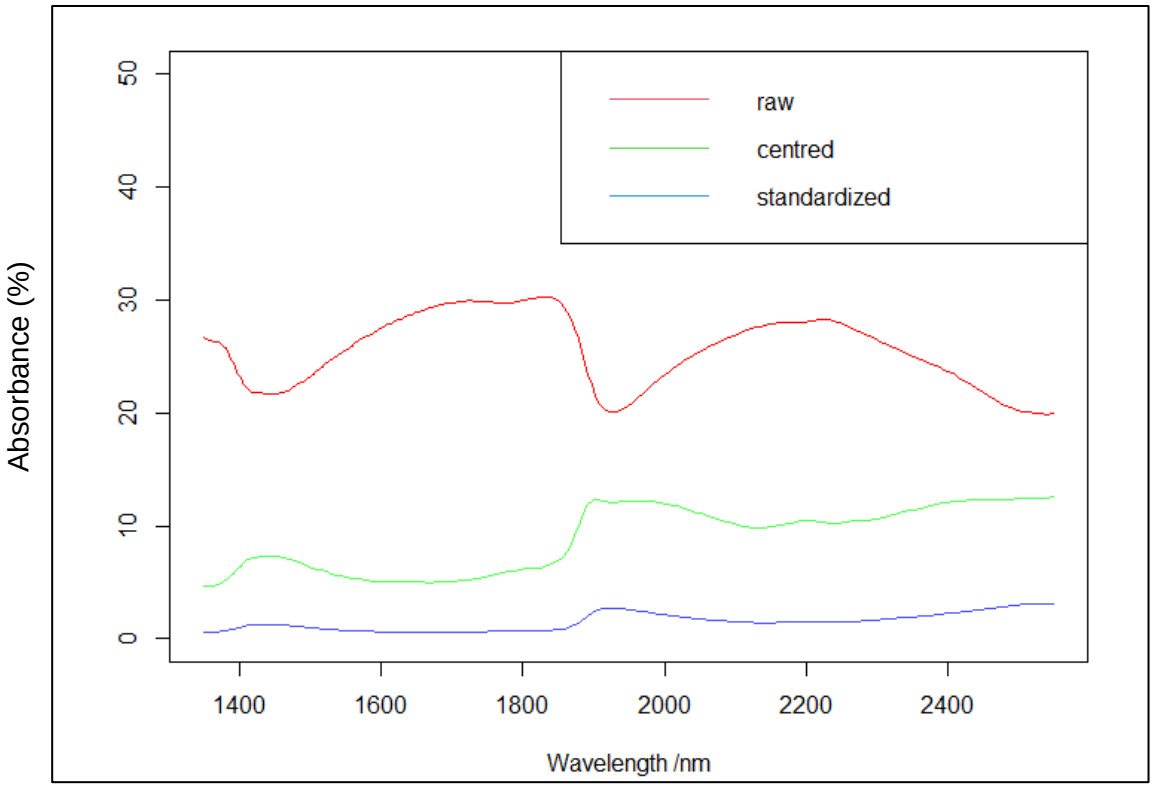

Figure 10: Visualisation of centring and standardization.

2.10.5. Outlier removal

Similarity or distance metrics between soil spectra are required in digital soil spectroscopy for a variety of applications, including assessing the accuracy of a spectrometer over continued scans, searching for a similar soil sample based on spectra from an extensive database, classifying spectra into groups with similar characteristics, and detecting outlier spectra (Engel *et al*., 2013; Wadoux *et al*., 2021). Usually, when spectral similarity is determined, the distance between the spectra is calculated. A single similarity measure is calculated by averaging the distance between the spectra. It is considered that the closer said spectra are to the other, the more similar the soil attributes are that they express (Engel *et al*., 2013).

2.11. Machine learning

Machine learning is a form of artificial intelligence (AI) which entails the development of algorithms that can learn from previously used data to recognize patterns and make intelligent decisions with minimal human involvement, apart from the creation thereof. It automates the process of creating analytical models that can utilise various types of data ranging from numerical data to categorical data or image pixels.

Machine learning techniques can broadly be classified into two categories, which are supervised and unsupervised techniques, as depicted in Figure 11. Firstly, supervised learning involves the categorization and interpretation of data based only on input data, while unsupervised learning develops a predictive model based on both the input data and output data (Tangirala, 2020). Supervised and unsupervised machine learning techniques can further be divided under subcategories, where supervised techniques are either *classification* which aims to obtain a categorical variable, or *regression* which aims to obtain a numerical variable. Unsupervised is further categorised under *clustering* which identifies a pattern or groups of similar objects, or *dimension reduction* which simply reduces the number of variables that are being considered in a search for information (Tangirala, 2020).

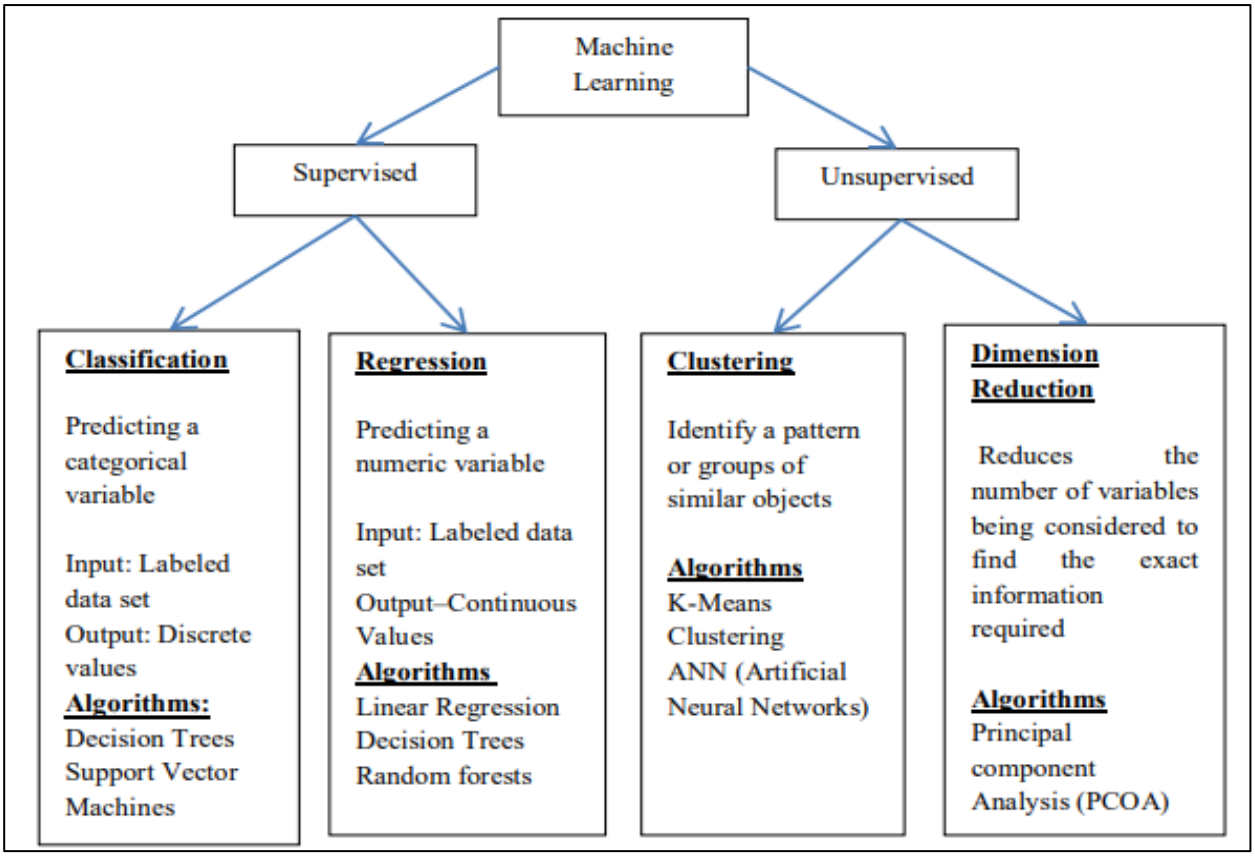

Figure 11: Diagram visualising the different categories of machine learning techniques (Tangirala, 2020).

There are a wide variety of regression algorithms to choose from. In the following section a few multivariate regression models will be discussed, as well as why they are chosen in this study.

2.11.1. Partial least squares regression (PLSR)

Partial least squares regression (PLSR) is a predictive, multivariate statistical model that studies the correlation between continuous, target predicted variables (soil water content), and a larger

number of observed variables (NIRS spectra). PLSR works by constructing a linear model that considers the correlations between observed variables and target variables and uses these correlations to make predictions. Unlike ordinary linear regression, PLSR works exceptionally well in cases where there are a large number of observed variables, as is the case for NIRS. This is because PLSR reduces the dimensionality of large amounts of spectra and enables modelling the correlations thereof (Ge *et al*., 2014). This is desirable, as it can reduce the amount of time it takes to create the model as well as the hardware power required for its creation. PLSR is arguably one of the most common multivariate regression models as it is used by a myriad of studies for calibration purposes and works exceptionally well with capturing complex and non-linear relationships in the data, making it one of the most preferred algorithms for accurately predicting soil parameters (Chen *et al*., 2018; Dangal *et al*., 2019; Ge *et al*., 2014; Khorshidi & Niazi, 2018).

2.11.2. Random Forest (RF)

Random forest (RF) is an ensemble, supervised machine learning approach that was designed to improve the prediction accuracy of the algorithm as an extension of classification and regression trees (CART). The algorithm creation process is similar to that of CART, wherein the dataset is continuously fragmented to examine the relationship between the response and observation variables. The decision tree is trained using a subset of the features in the training data. This is called feature sampling (Schonlau & Zou, 2020). Figure 12 visualizes how bootstrap samples and samples of predictors are selected under the RF model (Montesinos López *et al*., 2022).

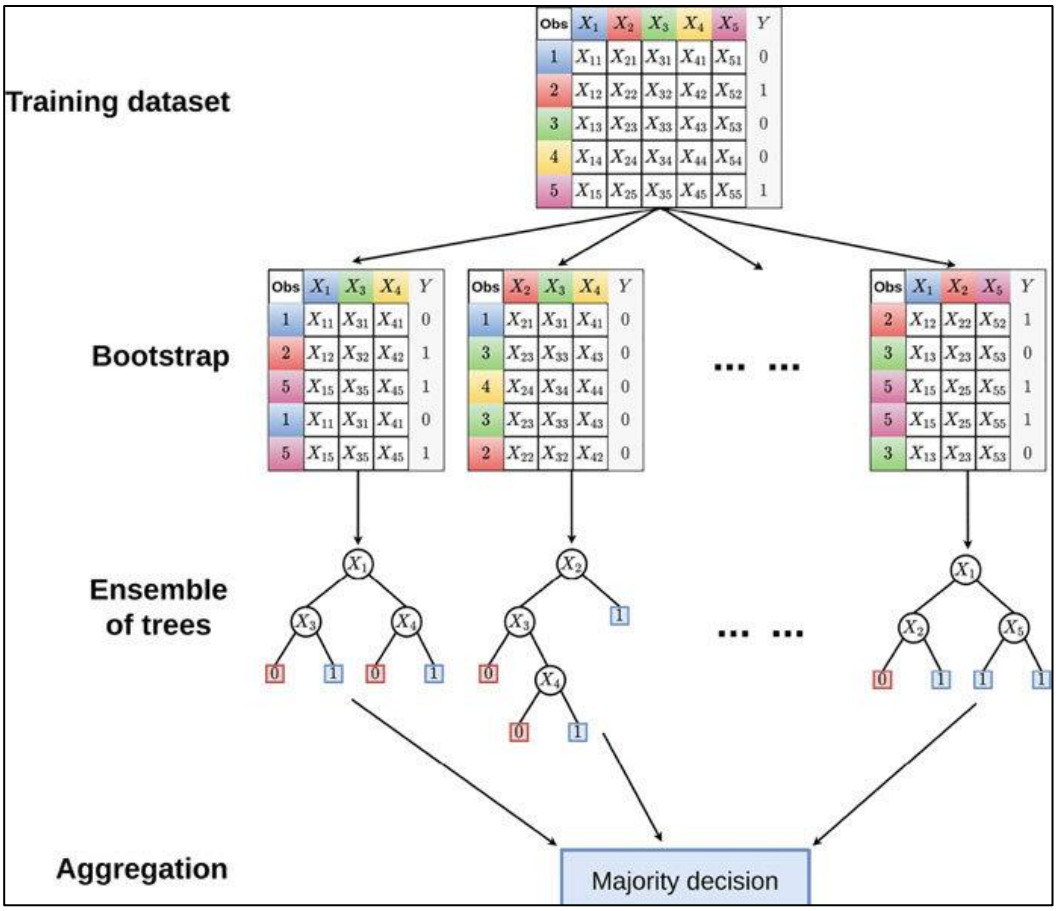

Figure 12: Illustration of how samples are selected under the RF model.

Once all decision trees have been trained, they are used to make predictions. Each decision tree makes a prediction for a given data point. The final prediction is the average of the predictions from all of the decision trees. While creating each tree, a bootstrap sample of the source observation data is selected, and the efficiency of each tree is evaluated using a third of the samples that were not used for creating the tree (Dega *et al*., 2023; Montesinos López *et al.,* 2022). This makes the predictions of RF more stable and reliable, making it one of the most commonly used machine learning algorithms in a myriad of studies in a wide range of fields (Acharya *et al*., 2022; Dega *et al*., 2023; Schonlau & Zou, 2020).

2.11.3. Cubist

Cubist is a prediction-oriented machine learning algorithm that constructs decision trees using the CART approach, with added instance-based corrections. It is based on the work of Quinlan (1992) and is described as the M5 model with additional corrections based on closest neighbours in the training set. Unlike the random forest technique that retrieves the final model on the basis of

discrete values, at each terminal node of the Cubist algorithm, a set of multivariate models and a set of rules are applied. The linear model that meets the set of conditions associated with the predictor variables is used to make the final prediction. Cubist primarily utilises standard deviation as a measure of error in the splitting criteria. By developing rule collections using a multivariate linear model for each rule, it maximizes accuracy. The linked multivariate linear model creates predictions when an observation satisfies a rule's requirement. Literature suggests that cubist is a candidate machine learning technique for calibration for a variety of purposes, since it operates efficiently even when working with large amounts of different data, while simultaneously having relatively low system requirements to operate (Helfer *et al.*, 2021; John *et al*., 2021; Shahbazi *et al*., 2019).

2.11.4. Comparison of machine learning algorithms

Table 2 shows a short comparison of the different machine learning algorithms. It is evident that these models are all viable methods for spectroscopy for various reasons. PLSR firstly for its low computational requirements and its ability to handle small to medium datasets well (Chen *et al*., 2018). It is however limited by its inability to handle the absence of variables in a dataset (Dangal *et al*., 2019; Ge *et al*., 2014). Cubist, in turn, also has relatively low computational requirements (John *et al*., 2021), and handles small to medium datasets well (Helfer *et al.*, 2021). Unlike PLSR, cubist is rather robust when it comes to missing variables in a dataset, making it more favoured since it will handle interference such as noise better (John *et al*., 2021; Shahbazi *et al*., 2019). RF, in turn, would also handle missing data and noise well. It is however limited by its ability to handle smaller datasets well and requires significant computational resources (Dega *et al*., 2023; Schonlau & Zou, 2020). Random forest is furthermore classified as an ensemble technique, which makes robust against overfitting, however, it may be prone to overfitting when aggregation takes place (Acharya *et al*., 2022).

Table 2: Comparison of the different machine learning techniques.

Algorithm	Type	Handling Missing Data	Ensemble Method	Computational Complexity	Scalability
PLSR	Regression	Limited	No	Low to moderate	Small to medium datasets
Cubist	Regression/Classification	Robust	No	Low to moderate	Small to medium datasets
RF	Classification/Regression	Robust	Ensemble of decision trees	Moderate to high	Large datasets

Partial Least Squares Regression = PLSR, Random Forest = RF

2.12. Model validation

As previously stated, the statistical analysis form part of the evaluation of the regression algorithm's accuracy. Different statistical parameters are used to determine the model's accuracy and will be discussed in detail in the following section.

2.12.1. Root mean square error (RMSE)

The Root Mean Square Error (RMSE, Equation 5) is a statistical parameter used to evaluate how good a calibration algorithm can predict a certain parameter. It demonstrates the overall magnitude of the algorithm's errors when it predicts the values of a certain parameter. The RMSE is essentially a measure of an algorithm's deviation between predicted values and actual or observed values of a specific parameter (Wadoux *et al*., 2021).

It is arguably one of the most important parameters to consider when evaluating an algorithm's performance due to its wide range of applicability in model evaluation. It is also relatively simple to interpret, and numerous studies utilise this parameter in their model performance evaluations.

$$\boldsymbol{RMSE} = \sqrt{\frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}(\boldsymbol{obs_i} - \boldsymbol{pred_i})^{\mathbf{2}}} \quad [5]$$

Where n is the size of the validation or test sample size, and obs and $pred$ are vector values of specific soil properties.

The RMSE is arguably one of the most used statistical parameters used in model evaluation. Literature suggests that acceptable RMSE values for calibration algorithms, specifically for soil water prediction, range from anything less than ~5% (Bullock *et al*., 2004; Fabre *et al.,* 2015; Gou *et al*., 2020; Koirala *et al.,* 2022; Liu *et al*., 2020)

2.12.2. Coefficient of determination (R^2)

The coefficient of determination (Equation 6), known as R^2, serves as a metric to assess the level of agreement between an algorithm and a specific dataset. It provides insights into the extent to which the algorithm elucidates the variability present in the dataset. R^2 values range from 0 to 1, with higher values indicating a stronger statistical fit. A greater R^2 value indicates that the algorithm more accurately captures the data. When the R^2 value equals 1, the algorithm perfectly represents the variance within the dataset; conversely, an R^2 value of 0 suggests that the algorithm failed to account for any of the dataset's variation (Mittal *et al*., 2021; Wadoux *et al*., 2021).

$$\mathbf{R2} = \mathbf{1} - \frac{\sum_{\mathbf{i=1}}^{\mathbf{n}}(\mathbf{obs_i} - \mathbf{pred_i})^{\mathbf{2}}}{\sum_{\mathbf{i=1}}^{\mathbf{n}}(\mathbf{obs_i} - \mathbf{obs})^{\mathbf{2}}} \quad [6]$$

Where $\sum_{i=1}^{n}(obs_i - pred_i)^2$ represents the sum of squared errors, and $\sum_{i=1}^{n}(obs_i - obs)^2$ represents the sum of total squares. A vast majority of studies use R^2 as one of the main statistical parameters in their studies and is widely used in studies focused on machine learning (Chen *et al.,* 2021; Koirala *et al.,* 2022). It is however imperative to not base the performance of a model entirely on this statistical parameter, and to make use of multiple parameters to summarize a model's performance (Mouazen & Al-Asadi, 2018).

2.12.3. Bias/Mean error (ME)

Calibration algorithms are constructed based on training data, which is collected by humans that are always in some shape or form prone to bias. Humans demonstrate cognitive bias, which is transferred to the collected data and can unfortunately be mimicked by machine learning algorithms (Shoukri *et al*., 2016; Zhang & Knoll, 2016). Biased data can impact the performance of machine learning algorithms and lead to biased results that aren't necessarily accurate. It is

thus imperative to measure the bias of a created algorithm to ensure the results are unbiased and as accurate as possible. The lower the ME the better, however anything under 10% is acceptable in typical environmental studies (Berardi & Zhang, 2003).

The ME or bias can be calculated using Equation 7:

$$\boldsymbol{ME} = \frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}(\boldsymbol{obs_i} - \boldsymbol{pred_i}) \quad [7]$$

2.12.4. Ratio of performance to deviation (RPD)

The ratio of performance to deviate (Equation 8), otherwise referred to as the calibration ratio or simply RPD is a metric utilised to assess a calibration algorithm. It is obtained by dividing the observed error rate of the model by the expected error rate. (Esbensen *et al*., 2014; Wadoux *et al*., 2021). A calibration algorithm is considered unreliable with an RPD value of less than 1.4, acceptable if the value is between 1.4 and 2, and excellent when the value is above 2 (Minasny *et al*., 2009), this is however seldomly the case, since the majority of data that is collected in the real world, tend to have some degree of inaccuracies (Esbensen *et al*., 2014).

It can be visualized by the following formula:

$$\boldsymbol{RPD} = \frac{\sqrt{\frac{\mathbf{1}}{\boldsymbol{n-1}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}(\boldsymbol{obs_i} - \overline{\boldsymbol{obs}})^{\mathbf{2}}}}{\sqrt{\frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}(\boldsymbol{obs_i} - \boldsymbol{pred_i})^{\mathbf{2}}}} \quad [8]$$

The RPD can be utilised to evaluate the performance of a model and identify any potential biases or inaccuracies in its predictions. It is often employed in conjunction with the other evaluation metrics that have been discussed to gain an in-depth, comprehensive understanding of the model's performance (Esbensen *et al*., 2014; Kamruzzaman *et al*., 2022). The equation of RPD can be simplified to the following (Equation 9):

$$\boldsymbol{RPD} = \frac{\boldsymbol{SD}}{\boldsymbol{SEP}} \quad [9]$$

Where SD represents the standard deviation, and SEP represents the standard error prediction.

2.12.5. Concordance correlation coefficient (rhoC)

Similar to the coefficient of determination (R^2), the concordance correlation coefficient, sometimes referred to as Lin's concordance coefficient, ρc, CCC or rhoC, evaluates the agreement between two sets of data (Barnhart *et al*., 2002; Lin, 1989; Wadoux *et al*., 2021). The rhoC can be between -1 and 1, where -1 shows perfect disagreement between two datasets, 0 shows no agreement, and 1 shows perfect agreement. Even though this parameter is seldomly used in studies, it provides valuable insights in agreements between datasets, as well as reproducibility in algorithms (Kwame *et al.,* 2020; Lin, 1989). It can be calculated using Equation 10.

$$\boldsymbol{\rho c} = \frac{2r\sigma_{pred}\sigma_{obs}}{\sigma_{obs}^2 + \sigma_{pred}^2 + (\mu_{obs} - \mu_{pred})^2} \quad [10]$$

where σ represents the standard deviation, μ represents the mean, *r* represents Pearson's correlation coefficient, and $r\sigma_{pred}\sigma_{obs}$ represents the covariance between predicted and observed data (Lin, 1989).

2.12.6. Ratio of performance to interquartile distance (RPIQ)

Proposed by Bellon-Maurel *et al* (2010) this statistical parameter was created in order to better represent the population spread of a dataset regardless of how it is distributed. It operates on the same basis as RPD but replaces the standard deviation with the interquartile distance (Q3-Q1) and accounts for a better visualization of data spread. The following Equation 11 shows how the RPIQ can be calculated:

$$\boldsymbol{RPIQ} = \frac{(\boldsymbol{Q3}(\boldsymbol{obs}) - \boldsymbol{Q1}(\boldsymbol{obs}))}{\sqrt{\frac{\mathbf{1}}{\boldsymbol{n}}\sum_{i=1}^{n}(\boldsymbol{obs}_i - \boldsymbol{pred}_i)^2}} \quad [11]$$

The equation can thus be simplified as Equation 12:

$$\boldsymbol{RPIQ} = \frac{\boldsymbol{Q3-Q1}}{\boldsymbol{SEP}} \quad [12]$$

2.13. Freely available models

Although most created algorithms in general remain privately owned or not freely available, there are projects or communities that aim to create free algorithms that can be used by the public. In the case of soil water content prediction, and various other soil properties, the Open Soil Spectral

Library (OSSL) created by Soil Spec For Greater Good provides free soil parameter algorithms provided they can use uploaded spectra in their databases to improve algorithm calibrations.

They are described as a Food and Agriculture Cyberinformatics Tools Coordinated Innovation Network NIFA Award #2020-67021-32467 project financed by the USDA. They aim to bring together soil scientists, spectroscopists, informaticians, data scientists, and software developers to address some of the present barriers to the large-scale and more efficient use of soil spectroscopy. A number of working groups are organized to address issues such as model selection, community outreach and demonstration, and the use of spectroscopy to inform global carbon cycle modelling. (SoilSpec4GG, s.a.). The prediction service makes use of uploaded spectral data in different format, ranging from NIRS to MIR. Initial calibration algorithms were created with samples that were dried.

3. CHAPTER 3: MATERIALS AND METHODS

3.1. Study Area

Undisturbed soil samples were collected from five diverse catchments spread through South Africa that can be seen in Figure 13. These catchments include the Goukou catchment in the Western province, the Umngeni catchment in KwaZulu-Natal, the Tsitsa catchment in the Eastern Cape, the Sabie catchment in Mpumalanga and Limpopo, and the Olifants catchment in Gauteng and Mpumalanga.

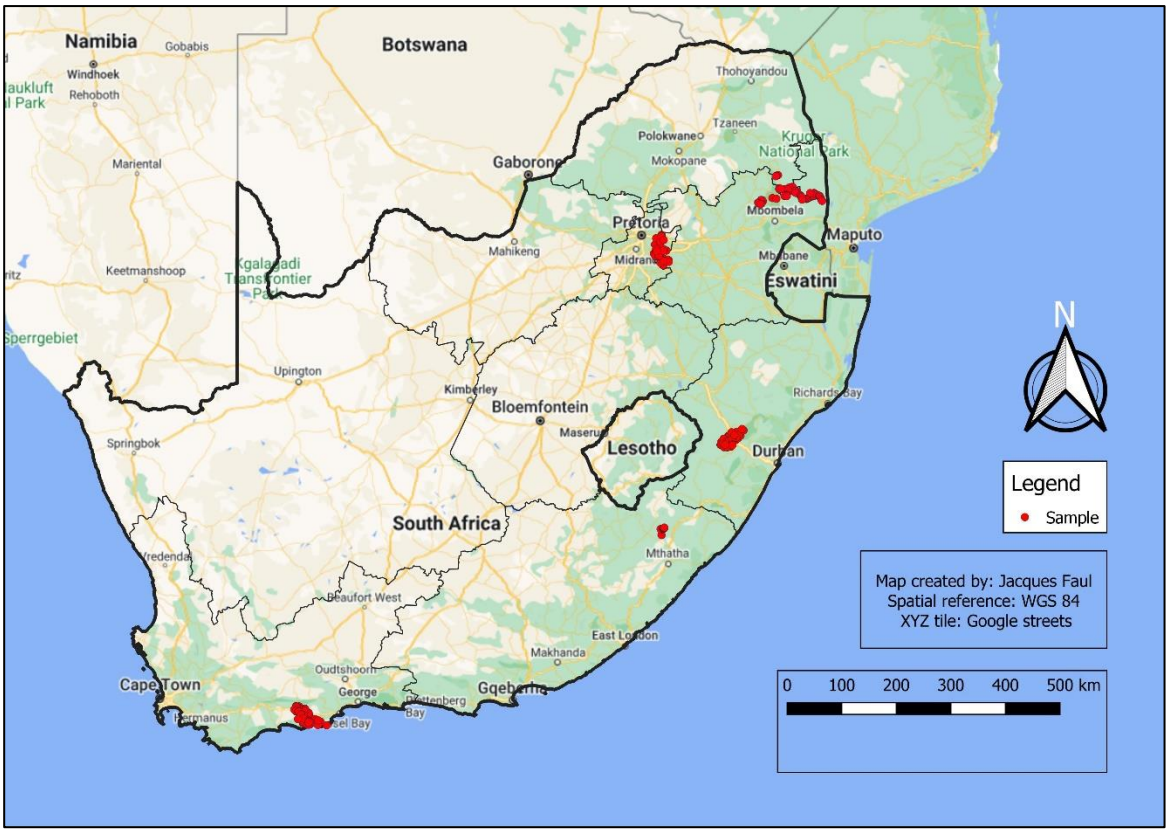

Figure 13: Locality map of all the samples taken.

3.2. Methodology

A total of 213 undisturbed soil core samples with known volume were collected from the surface layer of the soil (depth of 0-20 cm) at various locations within the five catchments. The cores are created from cutting PVC pipes into sections that have an average length of 8 cm, while the diameter of the core and PVC pipe is 10.4 cm.

Sample locations for each catchment were determined using conditioned Latin Hypercube sampling (Minasny & McBratney, 2006), using a mixture of covariates such as Brightness Index

(BI), Colouration Index (CI), Redness Index (RI), Saturation Index (SI), Normalized Difference Vegetation Index (NDVI), as well as various topographic covariates like slope and aspect for each catchment, giving a number of sample points that are all representative of the catchment's spatial variability.

After the samples were collected, they were moved to a laboratory where they were wetted until saturated, weighed and scanned with the handheld Neospectra NIRS by placing the scanner on top of the sample and scanning for a total time of 14 seconds. The spectrometer had a spectral resolution of 16 nm, and the spectral range was from 1250 – 2500 nm.

The samples were then placed in a pressure pot that subjected the samples to different pressures to force the water from the core sample. The different pressures used were from 33 kPa (drained upper limit), 100 kPa, 500 kPa, and 1500 kPa (lower limit). Once the soil water reached a constant level at each pressure, the sample was weighed and then scanned again with the Neospectra spectrometer.

Each sample will thus have a spectral measurement at five different water levels, giving a total of 1065 individual scans. Afterwards all of the samples were oven dried and weighed again to determine the dry bulk density (*pb*). The bulk density can be calculated using Equation 13:

$$\boldsymbol{Bulk\ density(\rho b)} = \frac{\boldsymbol{Mass\ of\ the\ dry\ sample\ (g)}}{\boldsymbol{Volume\ of\ the\ dry\ sample\ (cm^3)}} \quad [13]$$

The saturated and completely dry weights of the samples were used to calculate the gravimetric water content (GWC) at each measurement from which the volumetric water content (θv) was calculated using the pb using Equation 14.

$$\boldsymbol{Volumetric\ water\ content\ (\theta v)} = \boldsymbol{gravimetric\ moisture\ content} * \boldsymbol{bulk\ density} \quad [14]$$

For the calibration process, approximately 75% of the 1065 spectra were used in the creation of the calibration model, where the remaining 25% were used as training data in the creation of the algorithm. The fuzzy k-means clustering sample selection algorithm was used to split the data into a training and validation set at a 75:25 ratio on the spectra. Pre-processing methods applied included Standard normal variate (SNV), multivariate scatter correction (MSC), Standardization, Savitzky-Golay (SG), and outlier removal was used (Table 3).

Table 3: Summary and comparison of the pre-processing methods used.

Pre-processing	Description
No pre-processing	Uses the spectra without any modifications
Savitzky-Golay	Applies a smoothing filter to the spectra to reduce noise (Mouazen & Al-Asadi, 2018).
Savitzky-Golay + removal of outliers	Applies a smoothing filter to the spectra and then removes any remaining multivariate outliers (Mouazen & Al-Asadi, 2018; Wadoux *et al*., 2021).
Standard Normal Variate	Standardizes the spectra by mean centering and dividing by the standard deviation (Zhang *et al.,* 2019).
Multiplicative Scatter Correction	Mean centers the spectra and then scales it to have a uniform standard deviation (Zhang *et al.,* 2019).
Standardization	Centers the spectra by subtracting the mean and then scales it to have a unit standard deviation (Wadoux *et al.,* 2021).

The calibration algorithms were created using the training dataset with different algorithms including partial least squares regression (PLSR), Random Forest (RF), and Cubist for both soil water content. Table 4 shows the different algorithms used, and the pre-processing methods used for each algorithm for the water content prediction. Giving a total of 18 calibrations.

Table 4: Combinations of models and pre-processing methods used for the volumetric water content prediction on the regional dataset.

Model	Pre-processing
PLSR Cubist & Random Forest	No pre-processing
	Savitzky-Golay
	Savitzky-Golay + Removed outliers
	Standard Normal Variate
	Multiplicative Scatter Correction
	Standardization

Partial Least Squares Regression = PLSR, Random Forest = RF

For the bulk density algorithms, the entire spectral dataset of 1065 spectra were again split into a 75:25 calibration and validation ratio using fuzzy k-means clustering. All three models (PLSR, Radom Forest, and Cubist) were used firstly with no pre-processing, and then with Savitzky-Golay

and standardization or outlier removal, depending on which pre-processing method performed the best, giving a total of nine calibrations.

For the catchment specific calibrations, only the best performing model and pre-processing method were used, and this was done for each catchment for volumetric water content, and dry bulk density. Similar to the previous calibrations, fuzzy k-means cluster was used on the spectra to split the data into a 75:25 training and validation sets.

Table 5 shows the number of samples for each catchment, resulting in a total of 213 samples scanned at five different water contents.

Table 5: Number of samples for each catchment at five different water contents.

Catchment	Number of samples	Number of spectra (samples * 5)
Goukou	35	175
Olifants	39	195
Sabie	77	385
Tsitsa	28	140
Umngeni	34	170

The entire dataset for each catchment was used and split into a 75:25 training and validation datasets using fuzzy k-means clustering. This resulted in five calibrations for volumetric water content and five calibrations for dry bulk density.

To compare the created algorithms with the OSSL models, algorithms were created for the DUL, LL, and bulk density using all of the data at the lower limit since the OSSL estimation service does not have a prediction option for volumetric water content, but only for DUL, LL, and bulk density.

It is important to note that the OSSL models were created using samples that were dried and sieved. This study, however, did not scan the samples at completely dried moisture level, since water content below the lower limit is not necessarily important for agricultural purposes. For the comparison then, we used the driest moisture content available (which is the lower limit) data to

compare with the OSSL models, since the moisture content at the lower limit is comparably low to completely dried.

The OSSL models were created using a combination of Cubist and the standard normal variate pre-processing method. For this reason, cubist was used to create models for the DUL, LL, and bulk density to ensure that the comparison is valid. The data was again split into a 75:25 training and validation dataset using fuzzy k-means clustering, and cubist with no pre-processing were used since standard normal variate did not provide any noticeable improvement on calibration results with cubist algorithms. The validation dataset of each created algorithm is exported as a CSV file, which is then uploaded to the OSSL estimation service in order for them to make predictions for the same parameters, which was then loaded into R to compare the calibrations.

Validation of the calibration algorithms were conducted on the independent validation dataset. R^2 displaying the correlation between the model's results and the validation samples, the mean error (ME) indicating any bias, the root mean square error (RMSE) which indicates the magnitude of error uncertainty, the concordance correlation coefficient (rhoC) which measures the agreement between the spectra and the water content and bulk density, while the RPD and RPIQ will measure the ratio of the range of the reference values to the standard deviation of the prediction errors and the interquartile distance of the prediction errors respectively.

Additionally, the texture of the samples was determined using the hydrometer method for particle size analysis, and the total percentage carbon was determined using the Leco combustion method.

4. CHAPTER 4: RESULTS AND DISCUSSION

4.1. Soil water content database

The box and whiskers plots (Figure 14 A-E) of some measured soil properties from the samples indicate that the samples are very diverse and represent a wide range of soils. shows the box and whiskers plot for the total volumetric water content percentage (VWC%), the dry bulk density (DBD) in g/cm3, the carbon percentage of all samples, and the clay percentage of all samples respectively.

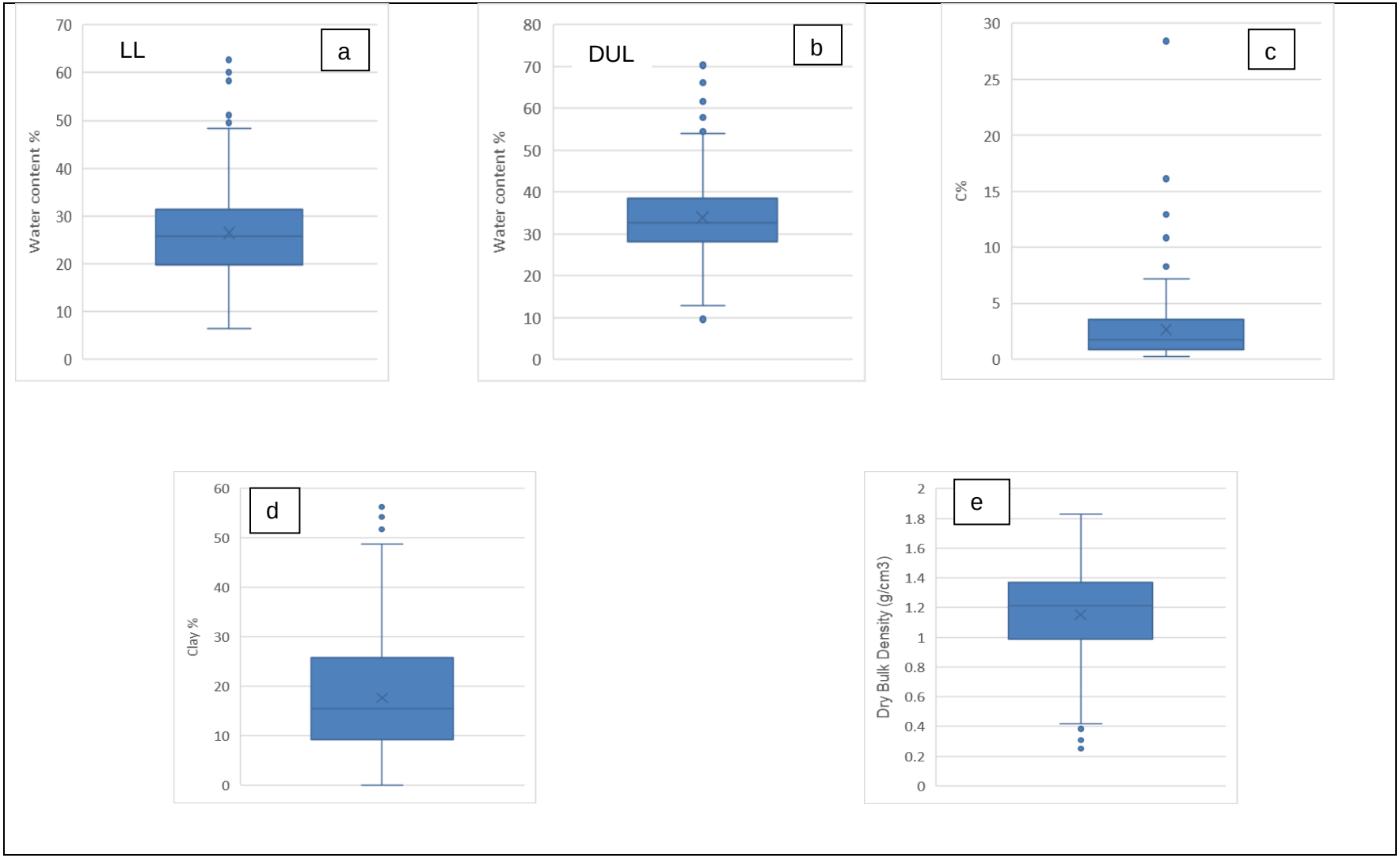

Figure 14: Box and whiskers plots showing the spatial variability of all the samples. a. Lower Limit, b. Drained Upper Limit, c. Organic Carbon, d. Clay% and e.Dry Bulk Density.

Looking at Figure 14 A-E, it is evident that all of the samples cover a diverse range of properties, with the water contents of both the DUL (B) and LL (A) ranging significantly. Looking at the bulk density (E) the samples cover almost the entire range of all bulk density classes according to Hazelton & Murphy (2007), from very low (<1.0) to high (1.6-1.8), lacking only the very high class of bulk density, which is higher than 1.9, which is considered to be soils that are very compacted (Hazelton & Murphy, 2007). According to Hazelton & Murphy (2007), the clay content of the samples (D) also covers the entire range of all texture grades: from sand to heavy clays. The range of organic carbon in the samples (C) also cover the entire range of organic carbon levels according to Hazelton & Murphy (2007), which ranges from extremely low (<0.4% carbon) to organic soil material (>8.7%). The results from the box and whiskers plots indicate that the samples represent a very good range of diverse soil conditions.

Figures 15 to 17 shows the box and whiskers plots indicating the soil variability within each catchment. Where Figure 15 shows the volumetric water content at the DUL for each catchment, Figure 16 shows the volumetric water content at the lower limit, and Figure 17 shows the dry bulk density. From these three figures it is evident that most catchments represent a significant amount of variability, apart from the Tsitsa catchment. The Umngeni catchment has the most variability with both the DUL and the LL ranging significantly, as well as the bulk density ranging from almost 0.3 to 1.3 g/cm3. The extremely low bulk density found within the Umngeni catchment can be attributed to peat and organic soils found in the area, however, it may also be possible that the upper litter layer consisting of predominantly organic matter was sampled by accident instead of the actual A-horizon.

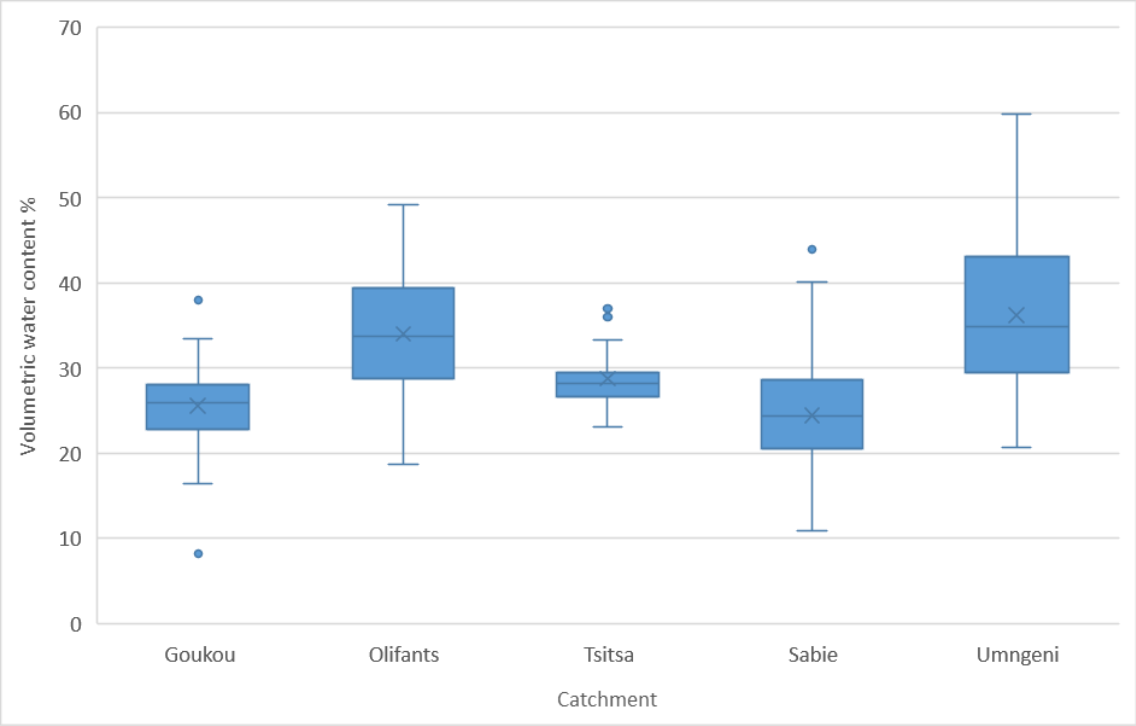

Figure 15: Box and whiskers plot of the volumetric water content % for each catchment at the Drained Upper Limit (DUL).

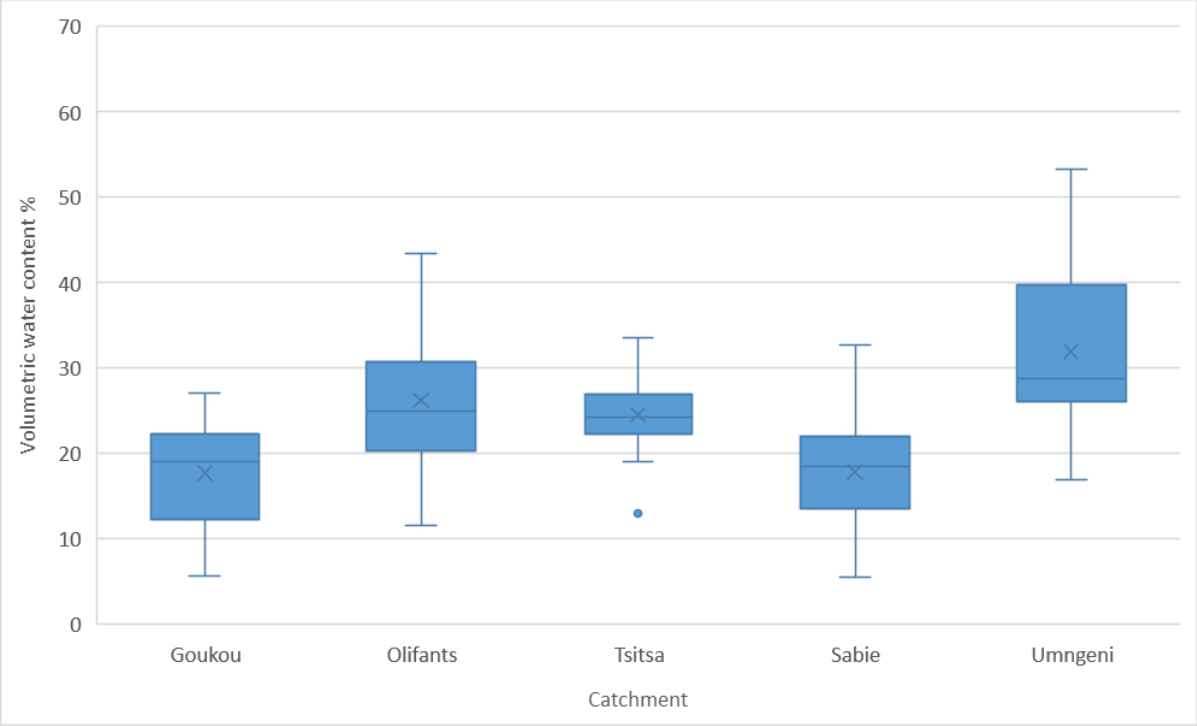

Figure 16: Box and whiskers plot of the volumetric water content % for each catchment at the Lower Limit (LL).

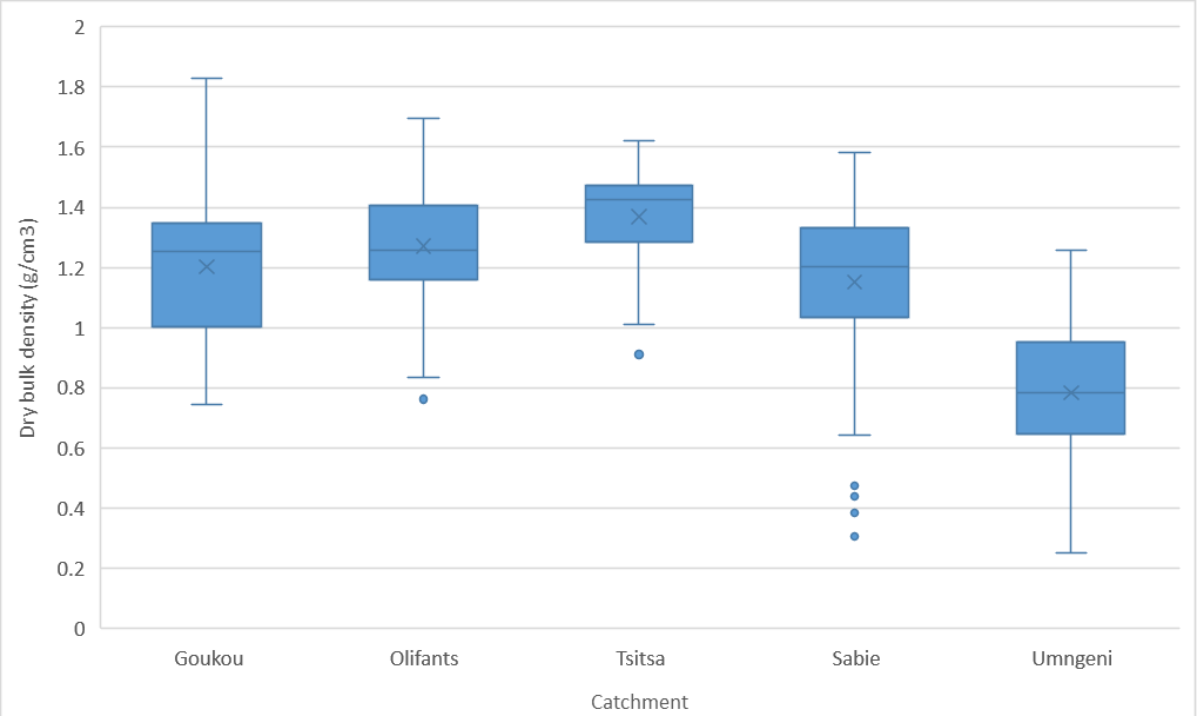

Figure 17: Box and whiskers plot of the dry bulk density (g/cm3) for each catchment.

4.2. Regional calibration model creation

4.2.1. Volumetric water content (VWC)

The results of the all the volumetric water content % algorithms can be seen in Table 6. For the calibration results, most of the models calibrated well with rhoC values above at least 0.6, indicating that the models performed satisfactory in terms of the calibration data. This also means that the models are capturing both the systematic bias and random error in the predictions, resulting in a reliable and consistent relationship between the predicted and observed values. The RPIQ values of the PLSR calibrations did not perform as good, and are all below 2, while the Cubist and RF RPIQ values are all above 2 which is satisfactory. Furthermore, The RMSE values of all the calibrations are all below 8.6%, which is moderately accurate, but leaves desire for improvement.

For the validation, Cubist and Random Forest (RF) performed well with Savitzky-Golay and had relatively similar results, with RF having a slightly lower RMSE of 6.96% in comparison to Cubist with an RMSE of 7.01%. Moreover, RF had a slightly higher R^2, rhoC, and RPD of 0.61, 0.75, and 1.61 respectively, whereas Cubist with 0.6, 0.72, and 1.57 respectively. Cubist and RF also has similar RPIQ values of 1.81 and 1.8 respectively. RF did, however, have a higher bias of 0.75, while Cubist only has 0.38. While the performance of cubist and RF are largely similar, RF shows

signs of overfitting due to the calibration performing well, but the validation performing poor, which may be due to the relatively small dataset size. Random forest typically performs better with larger datasets, due to its usage of decision trees that are prone to overfitting when there is not enough data to train the algorithm (Cosenza *et al*., 2020). The combination of cubist and Savitzky-Golay shows promise, since cubist performs better on smaller datasets (Katuwal *et al*., 2020), while Savitzky-Golay's derivative and smoothing capabilities work well with cubist to ensure a robust calibration algorithm (Zimmermann & Kohler, 2013). Similarly, a study done by Clingensmith & Grunwald (2022) concluded that partial least squares regression (PLSR) underperformed in predicting soil properties in vis-NIR when compared to RF and cubist due to data complexity and non-linear relationships.

The results are promising but can be improved. Studies done by Liang *et al*. (2012) obtained substantially better results in predicting soil water content using NIRS, with an RMSE of 1.2% and an R^2 of 0.99. The study however, created calibration algorithms on a much smaller area with much less spatial variation, accompanied by a spectrometer with a higher spectral resolution of only 6.8 nm, which may significantly increase the accuracy of results. Another study done by Bullock *et al*. (2004) produced desirable results in predicting the water content in five soil horizons with an RMSE of 6.4% and a R^2 of 0.95. Similar to the study done by Liang *et al*. (2012), the study done by Bullock *et al*. (2004) utilised a spectrometer with a very high resolution of only 2 nm with a limited spatial variation of five soil horizons. Considering the results and information from these studies, the results are quite satisfactory, considering the spatial variation of the samples are very high, and the spectral resolution of the scanner used is only 16 nm. Calibrations on a smaller area with more samples will improve results.

Table 6: Results of the volumetric water content algorithms.

Algorithm and pre-processing method	Calibration						Validation					
	ME	RMSE %	R^2	rhoC	RPD	RPIQ	ME	RMSE %	R^2	rhoC	RPD	RPIQ
PLSR (without pre-processing)	0	8.54	0.49	0.66	1.4	1.7	0.7	8.12	0.38	0.65	1.27	1.65
PLSR (Savitzky-Golay)	0	8.21	0.53	0.69	1.46	1.78	1.78	8.6	0.31	0.59	1.21	1.45
PLSR (Savitzky-Golay + outliers)	0	8.22	0.52	0.68	1.44	1.74	0.68	8.94	0.34	0.6	1.23	1.46
PLSR (Standard Normal Variate)	0	8.71	0.46	0.63	1.36	1.66	1.52	8.99	0.32	0.56	1.21	1.43
PLSR (Multiplicative Scatter Correction)	0	8.66	0.45	0.62	1.35	1.66	1.03	9.13	0.37	0.56	1.26	1.41
PLSR (Standardization)	0	8.09	0.54	0.7	1.47	1.79	0.2	8.1	0.42	0.67	1.32	1.63
Cubist (without pre-processing)	0.85	7.23	0.62	0.74	1.63	1.93	1.13	7.96	0.49	0.65	1.41	1.81
Cubist (Savitzky-Golay)	0.96	6.70	0.68	0.77	1.76	2.18	0.38	7.01	0.6	0.72	1.57	1.81
Cubist (Savitzky-Golay + outliers)	0.82	6.31	0.71	0.8	1.87	2.21	-0.51	7.7	0.53	0.65	1.46	1.78
Cubist (Standard Normal Variate)	0.55	6.87	0.66	0.75	1.71	2.12	0.32	7.85	0.5	0.62	1.42	1.5
Cubist (Multiplicative Scatter Correction)	0.57	6.71	0.68	0.77	1.77	2.17	0.25	7.77	0.48	0.61	1.39	1.63
Cubist (Standardization)	0.77	7.05	0.63	0.75	1.65	1.98	0.75	8.18	0.49	0.64	1.41	1.74
RF (without pre-processing)	0.08	3.67	0.9	0.94	3.17	3.84	0.46	8.86	0.41	0.57	1.31	1.57
RF (Savitzky-Golay)	0.02	3.14	0.93	0.96	3.75	4.68	0.75	6.96	0.61	0.75	1.61	1.8
RF (Savitzky-Golay + outliers)	0.01	3.11	0.93	0.96	3.75	4.42	0.5	7.07	0.63	0.75	1.64	2.04
RF (Standard Normal Variate)	0.03	3.86	0.9	0.93	3.1	3.91	1.62	8.68	0.3	0.47	1.2	1.4
RF (Multiplicative Scatter Correction)	0.03	3.82	0.89	0.93	3.06	3.83	1.38	9.08	0.35	0.48	1.24	1.33
RF (Standardization)	0.04	3.76	0.9	0.94	3.17	3.94	0.25	8.28	0.39	0.57	1.29	1.53

Partial Least Squares Regression = PLSR, Random Forest = RF, Mean Error = ME, Root Mean Square Error = RMSE, Correlation Coefficient = R^2, Lin's Concordance Coefficient = rhoC, Ratio of Performance Deviation = RPD, Ratio of Performance to Interquartile Distance = RPIQ

4.2.2. Dry Bulk Density

The results of the *pb* calibrations are displayed in Table 7. For the calibration results it is evident that the models calibrated rather well, with rhoC values ranging from 0.66 to 0.95 highlighting that there is a relatively good agreement between predicted and actual values. The RMSE values are all below 0.2 g/cm^3 which is acceptable but requires further refinement for increased accuracy. Apart from the PLSR and cubist models without pre-processing, the models all have RPIQ values above 2, indicating good predictive accuracy and robustness.

For the validation the cubist and RF models performed the best, yielding similar results with both algorithms having an RMSE of 0.16 g/cm^3. The best Cubist model combination (cubist and Savitzky-Golay with removed outliers) yielded an R^2 value of 0.71, a rhoC of 0.82, an RPD of 1.86, and an RPIQ of 2.31 against the best RF model's R^2 of 0.7, rhoC 0.81, RPD of 1.84, and RPIQ 2.29. All three of the models had a bias value of 0.01. Cubist paired with Savitzky-Golay again performed well, in this case the removal of outliers further improved results.

A study done by (Katuwal *et al*., 2020) predicted soil bulk density using vis-NIRS using a spectrometer with a resolution of 0.5 nm and provided promising results with an RMSE of 0.04 g/cm^3 and an R^2 of 0.94. The study used samples from different datasets, giving a total of 2462 samples from across Denmark. Considering that the number of samples are much higher than this study (1065 data points), and the spectral resolution is much higher, the results can be improved. Davari *et al*. (2021) achieved poorer results for bulk density calibration with an RMSE of 0.15 g/cm^3 and an R^2 of 0.26. The study utilised vis-NIRS on 220 soil samples within a 2000 m^2 area. The study concluded that a larger number of samples could possibly improve calibration results.

Table 7: Results of the dry bulk density algorithms.

Algorithm and pre-processing method	Calibration						Validation					
	ME	RMSE %	R^2	rhoC	RPD	RPIQ	ME	RMSE %	R^2	rhoC	RPD	RPIQ
PLSR (without pre-processing)	0	0.2	0.54	0.71	1.48	1.97	-0.01	0.2	0.54	0.69	1.47	1.8
PLSR (Savitzky-Golay)	0	0.19	0.6	0.75	1.57	2.09	0	0.21	0.44	0.66	1.34	1.69
PLSR (Standardization)	0	0.19	0.59	0.74	1.56	2.06	-0.01	0.19	0.58	0.74	1.54	2
Cubist (without pre-processing)	0.02	0.17	0.7	0.66	0.76	1.71	0.01	0.19	0.64	0.59	0.7	1.57
Cubist (Savitzky-Golay)	0.01	0.11	0.88	0.85	0.91	2.62	0.01	0.16	0.68	0.67	0.79	1.74
Cubist (Savitzky-Golay + outliers)	0.01	0.11	0.88	0.85	0.91	2.6	0.01	0.16	0.72	0.71	0.82	1.86
RF (without pre-processing)	0	0.1	0.94	0.89	0.93	3.01	0	0.23	0.39	0.37	0.5	1.27
RF (Savitzky-Golay)	0	0.07	0.96	0.95	0.97	4.48	0.01	0.16	0.66	0.66	0.78	1.71
RF (Savitzky-Golay + outliers)	0	0.07	0.96	0.95	0.97	4.4	0.01	0.16	0.71	0.7	0.81	1.84

Partial Least Squares Regression = PLSR, Random Forest = RF, Mean Error = ME, Root Mean Square Error = RMSE, Correlation Coefficient = R^2, Lin's Concordance Coefficient = rhoC, Ratio of Performance Deviation = RPD, Ratio of Performance to Interquartile Distance = RPIQ

4.3. Catchment calibrations

Site specific calibration algorithms using were created for each catchment to compare with the regional algorithms that uses all of the samples. Cubist paired with Savitzky-Golay were used, since it showed consistent and robust results previously. It is important to note that this section only focused on the validation statistics for model evaluation. Table 8 shows the results of the catchment calibrations. It is evident that the results are noticeably better for the catchment calibrations than for the regional scale calibrations, with the exception of the Umgeni catchment. The results are satisfactory, with an RMSE of the Tsitsa catchment being 3.83%, the bias being 0.22, RPD 1.23, and RPIQ 1.4. A rhoC value of 0.58 indicates that there is a relatively fair agreement between the predicted values and the observed values but is however not as good as the other catchments' rhoC values. Apart from the Tsitsa algorithm, all of the models' rhoC values are above 0.69, which indicates a good agreement between the predicted and measured values.

Of all the catchments, the Tsitsa catchment had the lowest number of data points of 140, which might explain the poor R^2 value of 0.32 and the lowest rhoC value of 0.58. The most robust catchment algorithms were the Sabie and the Olifants catchments. The Sabie calibration achieved a rhoC value of 0.74 and a satisfactory RPIQ value of 2.06, while the Olifants calibration achieved a high rhoC value of 0.81 and a fair RPIQ value of 1.94 indicating that they may be used for certain practical applications that don't require as precise measurements (Bellon-Maurel *et al*., 2010; Ludwig *et al*., 2018). Catchments with significant spatial variability like the Umngeni performed rather poor as expected.

This is due to the fact that the Sabie catchment had the most data points of 385, and the Olifants the second most of 195. The spatial variability seen in Figure 15-17 can also aid in explaining that the catchments with the least spatial variability should perform better, which is the case for the Tsitsa, Sabie and Olifants catchments.

Studies done by Canal Filho *et al*. (2023) and Koirala *et al*. (2022) conclude that local calibrations prove superior to regional and international calibrations due to less spatial variation and samples being more concentrated to better capture the spatial variation that may be present.

Table 8: Results of the catchment specific calibrations for VWC% using Cubist.

Catchment	ME	RMSE (%)	R^2	rhoC	RPD	RPIQ
Goukou	1.06	5.35	0.58	0.75	1.55	1.33
Olifants	0.59	5.07	0.7	0.81	1.85	1.94
Sabie	1.3	5.52	0.58	0.74	1.56	2.06
Tsitsa	0.22	3.83	0.32	0.58	1.23	1.4
Umgeni	0.6	8.63	0.54	0.69	1.49	1.86

Mean Error = ME, Root Mean Square Error = RMSE, Correlation Coefficient = R^2, Lin's Concordance Coefficient = rhoC, Ratio of Performance Deviation = RPD, Ratio of Performance to Interquartile Distance = RPIQ

Table 9 shows the results of the catchment specific dry bulk density calibrations again using cubist paired with Savitzky-Golay pre-processing. The Sabie catchment yielded the best results with the lowest RMSE 0.08 g/cm3, and R^2 of 0.92, a rhoC of 0.96, an RPD of 3.52, and an RPIQ of 3.75. The predictions for the Goukou, Olifants, and Umgeni catchment are rather poor, where the predictions for the Sabi and the Tsitsa predictions are significantly better with RMSE values that are below 0.1 g/cm3 (Katuwal *et al*., 2020), and even though the R^2 value is low at 0.46. the rhoC which is a better indicator of the correlation in models is at 0.62 which is fair. The results of the catchment calibrations naturally leave the desire for improvement, which can be achieved by increasing the number of samples in the area to better represent the spatial variation and increase algorithm performance. Although these models might not be suitable for precision measurements, the Sabie catchment algorithm can be useful for agricultural water management, irrigation, as well as creating real-time moisture maps to monitor moisture changes in the soil since the RPIQ value is above 3 and the RPD value is above 3, which shows that the model is very robust and reliable (Bellon-Maurel *et al*., 2010; Ludwig *et al*., 2018). Despite of the high RMSE value of the Umngeni catchment, the good RPIQ value of 2.46 also makes it practically usable since it proves that the model is reliable and robust (Ludwig *et al*., 2018).

Table 9: Results of the catchment specific calibrations for dry bulk density (g/cm3) using Cubist.

Algorithm	ME	RMSE (g/cm3)	R^2	rhoC	RPD	RPIQ
Goukou	0.04	0.14	0.64	0.76	1.68	1.72
Olifants	0.02	0.14	0.64	0.74	1.69	1.78
Sabie	0	0.08	0.92	0.96	3.52	3.75
Tsitsa	0	0.1	0.46	0.62	1.39	1.2
Umngeni	0.02	0.17	0.45	0.54	1.37	2.46

Mean Error = ME, Root Mean Square Error = RMSE, Correlation Coefficient = R^2, Lin's Concordance Coefficient = rhoC, Ratio of Performance Deviation = RPD, Ratio of Performance to Interquartile Distance = RPIQ

4.4. Comparing created algorithms against freely available algorithms

For the comparison of the created algorithms against freely available algorithms, the OSSL algorithms from the Soil spectroscopy for global good project were used. The available OSSL models included: Water retention at 33 kPa, water retention at 1500 kPa, and the bulk density using SNV pre-processing and Cubist models. Algorithms for these parameters were created using Cubist and no pre-processing and were then compared with each other. Figure 18 and Table 10 show the VWC% models for 33 kPa (DUL) and Figure 19 and Table 11 the 1500 kPa (LL) compared to the OSSL models, while Figure 20 shows the created bulk density model compared to the OSSL algorithm. The lines represent a one-to-one line, thus, the closer the data point is to the line the more accurate the prediction is to actual values.

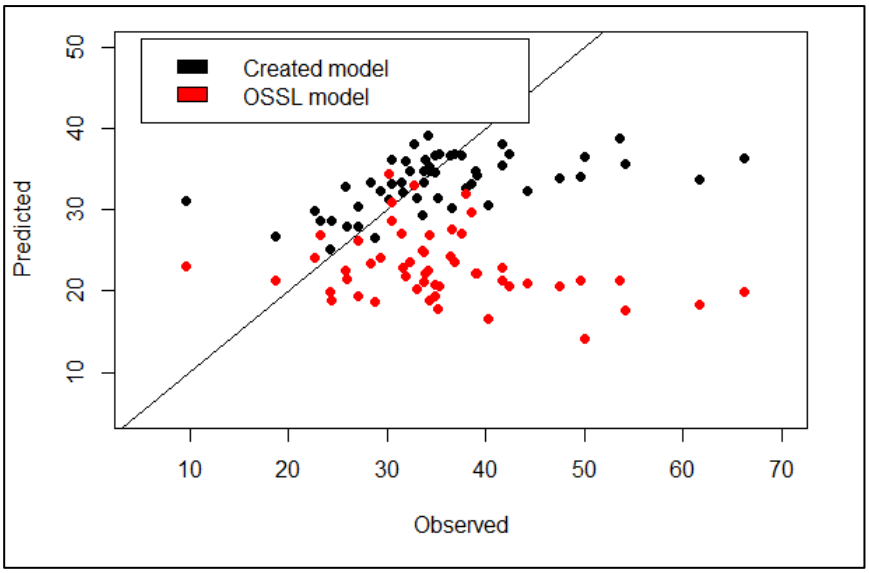

Figure 18: Comparison of the created cubist model of the drained upper limit water content % against the OSSL model.

Table 10: Validation dataset results of the created drained upper limit water content % model against the OSSL model.

Model	ME	RMSE	R^2	rhoC	RPD	RPIQ
OSSL model	-12.38	17.12%	-1.95	-0.09	0.59	0.51
Created model (without pre-processing)	-1.98	8.89%	0.2	0.31	1.13	0.99

Open Soil Spectral Library = OSSL, Mean Error = ME, Root Mean Square Error = RMSE, Correlation Coefficient = R^2, Lin's Concordance Coefficient = rhoC, Ratio of Performance Deviation = RPD, Ratio of Performance to Interquartile Distance = RPIQ

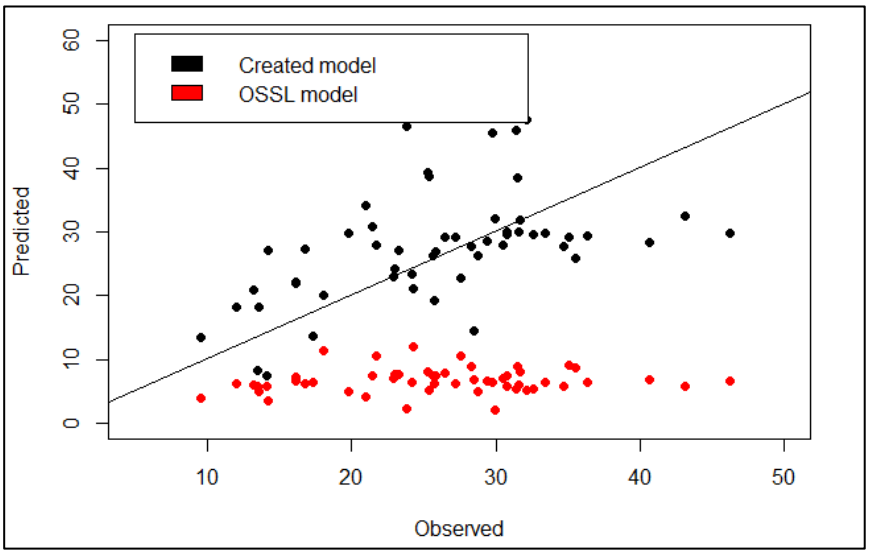

Figure 19: Comparison of the created cubist model of the lower limit water content % against the OSSL model.

Table 11: Validation dataset results of the created lower limit water content % model against the OSSL model.

Model	ME	RMSE	R^2	rhoC	RPD	RPIQ
OSSL model	-19.24	20.85%	-5.72	0.01	039	0.5
Created model (without pre-processing)	-1.57	8.35%	-0.08	0.5	0.97	1.24

Open Soil Spectral Library = OSSL, Mean Error = ME, Root Mean Square Error = RMSE, Correlation Coefficient = R^2, Lin's Concordance Coefficient = rhoC, Ratio of Performance Deviation = RPD, Ratio of Performance to Interquartile Distance = RPIQ

Looking at Figure 18-19 and Table 10-11, it is evident that the created algorithms are noticeably more accurate. It is important to note that the OSSL models were calibrated with dry samples, while the created algorithms were created with the corresponding water data, which might explain why the created algorithms are more accurate. The OSSL algorithm for the drained upper limit water retention indicate poor results of an RMSE of 17.12%, a mean error of 12.38, and very low RPD and RPIQ of 0.59 and 0.51 respectively. The created algorithm for the drained upper limit water retention performed better, with an RMSE of 8.89%, a mean error of 1.98, and an RPD and RPIQ of 1.13 and 0.99 respectively. Although the created algorithms results are better, they are still too poor to be practically applied in the field, which may be due to the spatial variation of all the samples used in the calibration.

With the lower limit calibrations, a similar outcome was found. The OSSL model for the lower limit water retention performed very poorly with an RMSE, mean error, RPD, and RPIQ of 20.85%, 19.24, 0.39, and 0.5 respectively. The created algorithm performed better with an RMSE, mean error, RPD, and RPIQ of 8.35%, 1.57, 0.97, and 1.24 respectively. Even though the results from the created algorithm are better, they still lack the accuracy and reliability that is desired.

These results were expected since the OSSL models utilise samples from across the globe that are scanned under different measuring protocols, which may significantly hinder large scale soil prediction models (Zhou *et al*., 2022). As for the created algorithms, the data points used still vary greatly due to spatial variation, where local calibrations would probably perform better. Figure 20 and Table 12 indicate that the created algorithms are more accurate in determining the DBD of a soil, where the RMSE of the created model is noticeably better at 0.22 g/cm^3 than the RMSE of the OSSL model at 0.38 g/cm^3. These results are unfortunately poor in comparison to other studies (Katuwal *et al*., 2020) that requires an RMSE of <0.1 g/cm^3 to be used effectively in the field.

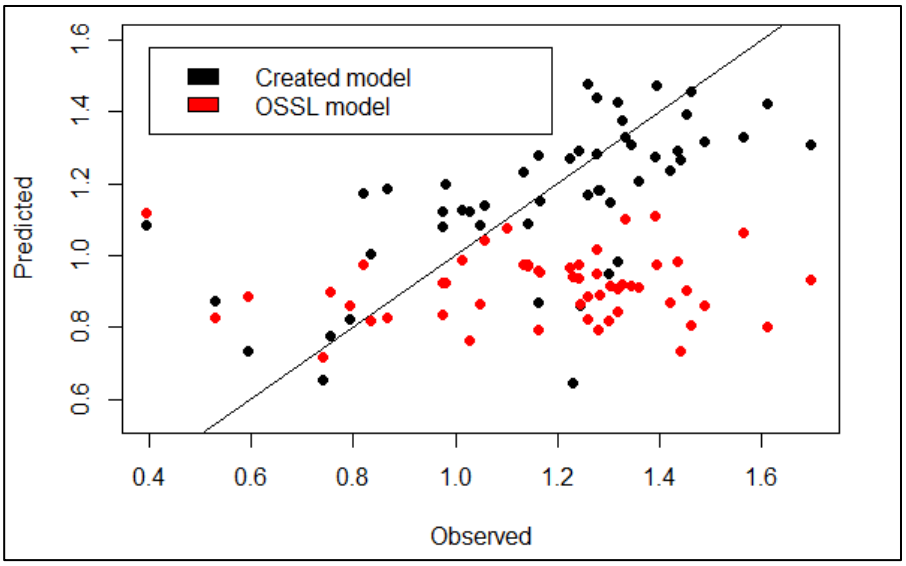

Figure 20: Comparison of the created cubist model of the dry bulk density (g/cm3) against the OSSL model.

Table 12: Validation dataset results of the created dry bulk density (g/cm^3) model against the OSSL model.

Model	ME	RMSE (g/cm^3)	R^2	rhoC	RPD	RPIQ
OSSL model	-0.26	0.38	-1.02	0.01	0.71	0.8
Created model (Without pre-processing)	-0.01	0.22	0.33	0.58	1.23	1.38

Open Soil Spectral Library = OSSL, Mean Error = ME, Root Mean Square Error = RMSE, Correlation Coefficient = R^2, Lin's Concordance Coefficient = rhoC, Ratio of Performance Deviation = RPD, Ratio of Performance to Interquartile Distance = RPIQ

Various studies (Chen *et al*., 2021; Koirala *et al*., 2022; Mouazen & Al-Asadi, 2018) also conclude that increasing the number of samples for calibration can improve results, and that the development of local calibrations can feed regional and national calibrations, and that regional calibrations can furthermore aid in ultimately improving international calibrations. Referring back to Table 1, the models from this section are not as accurate as those from literature, suggesting that they are not practically usable. But for the comparison of international models, they served the purpose of highlighting that local calibrations are, for now at least, the way to go.

5. CHAPTER 5: CONCLUSIONS AND RECOMMENDATIONS

The first objective of this study, the creation of regional water content calibration algorithms was met by the creation of 18 calibration algorithms for volumetric water content prediction. The study found that the most accurate algorithm for the volumetric water content prediction was Cubist with Savitzky-Golay pre-processing, since Cubist is a robust model that performs well with smaller datasets like the one used in this study.

The second objective, which was to determine at which scale NIRS calibration algorithms perform best was achieved by creating calibration algorithms for each catchment using cubist and Savitzky-Golay pre-processing. The achieved results showed that the catchment calibrations performed significantly better than the regional algorithms.

For the third objective, this process was repeated for bulk density, where 9 calibration algorithms for regional dry bulk density prediction were created. The best performing algorithm was cubist with Savitzky-Golay pre-processing. Catchment calibrations for dry bulk density prediction were also created using cubist and Savitzky-Golay pre-processing, and the results were again noticeably better than the regional calibration results, again supporting the theory that local calibrations are necessary for enhanced accuracy and reliability.

For the fourth and final objective, which was to compare created algorithms to freely available international algorithms, calibration algorithms were created for the water content at the 33 kPa (drained upper limit), 1500 kPa (lower limit), and for bulk density predictions, as the OSSL prediction service did not yet have water content algorithms available. Data from 1500 kPa was used since the OSSL algorithms were created using dried samples, and the data from 1500 kPa was the driest water content with spectral data we had available. Even though this is not the same water content, relative conclusions could be drawn from the comparison since the water content at 1500 kPa is comparably low to dry. The results we found indicated that the created algorithms were significantly more accurate than the OSSL algorithms, which was expected since the OSSL algorithms utilise data from across the world that differs from South African soils, and that were collected using different methods. In this study, the calibrations were not done for oven dried soils since portable NIRS scanners are meant to be used in the field.

The hypothesis that NIRS can effectively predict soil water content in a variety of soils is thus accepted, with the condition that local calibrations are to be made for more accurate predictions. Because soil characteristics vary on such a fine scale, especially South African soils that are so diverse, local calibrations for NIRS are highly recommended. Although this might require more

calibrations in the long term, the accuracy and reliability of algorithms will certainly gain from it, which will result in better water management decisions that will promote a more sustainable future where water scarcity is mitigated for future generations. It is also recommended to improve handheld NIRS devices to have a better spectral resolution, as this may show an improvement in results. Furthermore, it would be advisable to increase the number of samples in areas to better capture spatial variation when creating calibration algorithms as this may also add significant value to calibrations. The build and creation of local calibrations may aid in the compilation of regional calibrations, where the continued compilation of data may add to accuracy and reliability international algorithms which would be the ultimate goal.

As for the application of these models, specifically the models that were created to compare with the OSSL models, perform too poorly for high-accuracy measurements. Some of the catchment algorithms, specifically both Sabie catchment algorithms and, are accurate enough to be practically implemented for the use of irrigation management in agriculture and should provide sufficient accuracy due to their RPD and RPIQ values that are satisfactory.

NIRS certainly shows promise as a non-invasive, rapid, and cost-effective method of soil moisture determination, but for its effective and accurate application, the collection of sufficient data on local scale is imperative.

6. CHAPTER 6: REFERENCE LIST

Acharya, U., Daigh, A.L. & Oduor, P.G. 2022. Soil moisture mapping with moisture-related indices, OPTRAM, and an integrated random forest-OPTRAM algorithm from Landsat 8 Images. *Remote sensing*. 14(15):1–20.

Aenugu, H., Kumar, D., Srisudharson, Parthiban, N., Ghosh, S. & Banji, D. 2011. Near infrared spectroscopy- An overview. *International Journal of ChemTech Research*. 3(2):825-836.

Afzali, H., Tasumi, M. & Nishiwaki, A. 2021. Use of hand-held NIR sensor to estimate water status of leaves and soils. *Journal of Rainwater Catchment Systems*. 26(2):1-6.

Ahmad, N., Kanwar, R., Kaspar, T. & Bailey, T. 1992. Effect of soil surface submergence and a water table on vegetative growth and nutrient uptake of corn. *Transactions of the ASAE*. 35(4):1173-1177.

Ahmadi, A., Emami, M., Daccache, A. & He, L. 2021. Soil properties prediction for precision agriculture using visible and near-infrared spectroscopy: A systematic review and meta-analysis. *Agronomy*. 11(3):1–14.

Ahmadi, H. & Mollazade, K. 2009. Effect of plowing depth and soil moisture content on reduced secondary tillage. *Agricultural Engineering International: The CIGR Journal of Scientific Research and Development*. 11:1-9.

Ahmed, M., Seraj, R. & Islam, S. 2020. The K-Means Algorithm: A comprehensive survey and performance evaluation. *Electronics*. 9(8):1–12.

Al Riza, D., Yolanda, J., Tulsi, A., Ikarini, I., Hanif, Z., Nasution, A. & Widodo, S. 2023. Mandarin orange (citrus reticulata Blanco CV. Batu 55) ripeness level prediction using combination reflectance-fluorescence spectroscopy. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy.* 302:1–8.

Amoakwah, E., Frimpong, K.A., Okae-Anti, D., & Arthur, E. 2017. Soil water retention, air flow and pore structure characteristics after corn cob biochar application to a tropical sandy loam. *Geoderma*. 307:189-197.

Barnhart, H.X., Haber, M., & Song, J. 2002. Overall concordance correlation coefficient for evaluating agreement among multiple observers. *Biometrics*. 58(4):1020–1027.

Bellinaso, H., Demattê, J. & Romeiro, S. 2010. Soil spectral library and its use in soil classification. *Revista Brasileira de Ciência do Solo.* 34(3):861–870.

Bellon-Maurel, V., Fernandez-Ahumada, E., Palagos, B., Roger, J. & 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):1073–1081.

Berardi, V. & Zhang, G. 2003. An empirical investigation of bias and variance in time series forecasting: Modeling Considerations and Error Evaluation. *IEEE Transactions on Neural Networks*. 14(3):668–679.

Bittelli, M., Salvatorelli, F. & Pisa, P. 2008. Correction of TDR-based soil water content measurements in conductive soils. *Geoderma*. 143(1-2):133-142.

Bordoloi, R., Das, B., Yam, G., Pandey, P. & Tripathi, O. 2018. Modeling of water holding capacity using readily available soil characteristics. *Agricultural Research*. 8(3):347–355.

Brady, N. & Weil, R. 2017. *The nature and properties of soils*. 15th ed. Harlow: Pearson Education Limited.

Breytenbach, I., Paige-Green, P. & Van Rooy, J. 2010. The relationship between index testing and California Bearing Ratio values for natural road construction materials in South Africa. *Journal of the South African Institution of Civil Engineering*, *52*(2):65-69

Brocca, L., Ciabatta, L., Massari, C., Camici, S. & Tarpanelli, A. 2017. Soil moisture for hydrological applications: Open questions and new opportunities. *Water*. 9(2):1-20

Bullock, P., Li, X. & Leonardi, L. 2004. Near-infrared spectroscopy for soil water determination in small soil volumes. *Canadian Journal of Soil Science*. 84(3):333–338.

Cambule, A., Rossiter, D., Stoorvogel, J. & Smaling, E. 2012. Building a near infrared spectral library for soil organic carbon estimation in the Limpopo National Park, Mozambique. *Geoderma*. 183-184:41-48.

Canal Filho, R., Molin, J.P., Wei, M.C., & Silva, E.R. 2023. Soil attributes mapping with online near-infrared spectroscopy requires spatio-temporal local calibrations. *AgriEngineering*. 5(3):1163–1177.

Carvalho, J., Moura-Bueno, J., Ramon, R., Almeida, T., Naibo, G., Martins, A., Santos, L., Gianello, C. & Tiecher, T. 2022. Combining different pre-processing and multivariate methods for prediction of soil organic matter by near infrared spectroscopy (NIRS) in southern Brazil. *Geoderma Regional.* 29:1–13.

Chaplin, M. 2001. Water: its importance to life. *Biochemistry and Molecular Biology Education*. 29(2):54-59.

Chen, C., Zhou, H., Shang, J., Hu, K. & Ren, T. 2019. Estimation of soil water content at permanent wilting point using hygroscopic water content. *European Journal of Soil Science*. 71(3):392-398.

Chen, H., Song, Q., Tang, G., Feng, Q. & Lin, L. 2013. The combined optimization of Savitzky-Golay smoothing and multiplicative scatter correction for FT-nir pls models. *ISRN Spectroscopy.* 2013:1–9.

Chen, H., Tan, C., Lin, Z. & Wu, T. 2018. Classification and quantitation of milk powder by near-infrared spectroscopy and mutual information-based variable selection and partial least squares. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*. 189:183-189.

Chen, T., Wang, C., Liang, J. & Li, J. 2021. Application of derivative transform spectroscopy in gas detection. *E3S Web of Conferences*. 299:1–8.

Chen, Y., Li, L., Whiting, M., Chen, F., Sun, Z., Song, K. & Wang, Q. 2021. Convolutional neural network model for soil moisture prediction and its transferability analysis based on laboratory vis-nir spectral data. *International Journal of Applied Earth Observation and Geoinformation*. 104:1–9.

Clingensmith, C. & Grunwald, S. 2022. Predicting soil properties and interpreting vis-NIR models from across Continental United States. *Sensors*. 22(9):1–17.

Cosenza, D., Korhonen, L., Maltamo, M., Packalen, P., Strunk, J., Næsset, E., Gobakken, T., Soares, P. & Tomé, M. 2020. Comparison of linear regression, K-nearest neighbour and random forest methods in airborne laser-scanning-based prediction of growing stock. *Forestry: An International Journal of Forest Research.* 94(2):311–323.

Cosh, M., Jackson, T., Bindlish, R., Famiglietti, J. & Ryu, D. 2005. Calibration of an impedance probe for estimation of surface soil water content over large regions. *Journal of Hydrology*. 311(1-4):49-58.

Chow, L., Xing, Z., Rees, H., Meng, F., Monteith, J. & Stevens, L. 2009. Field performance of nine soil water content sensors on a sandy loam soil in New Brunswick, Maritime region, Canada. *Sensors*. 9(11):9398-9413.

Chowdhury, N., Marschner, P. & Burns, R. 2011. Response of microbial activity and community structure to decreasing soil osmotic and matric potential. *Plant and Soil*. 344(1-2):241-254.

Croney, D., Coleman, J. & Currer, E. 1951. The electrical resistance method of measuring soil moisture. *British Journal of Applied Physics*. 2(4):85-91.

Dangal, S., 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):1-23.

Davari, M., Karimi, S., Bahrami, H., Taher Hossaini, S. & Fahmideh, S. 2021. Simultaneous prediction of several soil properties related to engineering uses based on laboratory vis-nir reflectance spectroscopy. *CATENA*. 197:1–12.

Défossez, P., Richard, G., Boizard, H. & O'Sullivan, M. 2003. Modeling change in soil compaction due to agricultural traffic as function of soil water content. *Geoderma*. 116(1-2):89-105.

Dega, S., Dietrich, P., Schrön, M., & Paasche, H. 2023. Probabilistic prediction by means of the propagation of response variable uncertainty through a Monte Carlo approach in regression random forest: Application to soil moisture regionalization. *Frontiers in environmental science*. 14:1-15.

Delwiche, S. & Reeves, J. 2010. A graphical method to evaluate spectral preprocessing in multivariate regression calibrations: Example with Savitzky-Golay filters and partial least squares regression. *Applied Spectroscopy*. 64(1):73–82.

Du, C. & Zhou, J. 2008. Evaluation of soil fertility using infrared spectroscopy: A review. *Environmental Chemistry Letters*. 7(2):97-113.

Eitzinger, J., Trnka, M., Hösch, J., Žalud, Z. & Dubrovský, M. 2004. Comparison of ceres, WOFOST and swap models in simulating soil water content during growing season under different soil conditions. *Ecological Modelling*. 171(3):223–246.

Engel, J., Gerretzen, J., Szymańska, E., Jansen, J., Downey, G., Blanchet, L. & Buydens, L. 2013. Breaking with trends in pre-processing? *TrAC Trends in Analytical Chemistry.* 50:96–106.

Esbensen, K.H., Geladi, P., & Larsen, A. 2014. The RPD myth…. *NIR news*. 25(5):24–28.

Fabre, S., Briottet, X. & Lesaignoux, A. 2015. Estimation of soil moisture content from the spectral reflectance of bare soils in the 0.4–2.5 µm domain. *Sensors*. 15(2):3262–3281.

Faul, C. 2018. *Correlations between vegetation, soil and geology in the semi-arid Bushmanland region of South Africa*. Potchefstroom: North-West University. (Dissertation – MSc).

Futshane, A., Paige-Green, P. & Anochie-Boateng, K. 2022. Understanding durability problems with dolerite in roads in South Africa. *World Journal of Engineering and Technology*. 10(3):574–592.

Gaddikeri, V., Hasan, M., Kumar, D., Sarangi, A., & Alam, W. 2021. Performance analysis and measurement of soil moisture content by piezoresistive sensor. *MAPAN – Journal of metrology society of India.* 37(1):149–160.

Garg, V. & Stogner, R. 2017. Hierarchical latin hypercube sampling. *Journal of the American Statistical Association*. 112(518):673–682.

Ge, Y., Morgan, C.L.S., & Ackerson, J.P. 2014. Visnir spectra of dried ground soils predict properties of soils scanned moist and intact. *Geoderma*. 221-222:61–69.

Ghezzehei, T., Sulman, B., Arnold, C., Bogie, N. & Berhe, A. 2019. On the role of soil water retention characteristic on aerobic microbial respiration. *Biogeosciences*. 16(6):1187–1209.

Gou, Y., Wei, J., Li, J., Han, C., Tu, Q. & Liu, C. 2020. Estimating purple-soil moisture content using vis-nir spectroscopy. *Journal of Mountain Science*. 17(9):2214–2223.

Hafizah, S.N. & Khairunniza, B.S. 2011. Colour spaces for paddy soil moisture content determination. *Journal of tropical agriculture and food science.* 39(1):1-13.

Hardie, M. 2020. Review of novel and emerging proximal soil moisture sensors for use in agriculture. *Sensors*. 20(23):1–23.

Hanjra, M. & Qureshi, M. 2010. Global water crisis and future food security in an era of climate change. *Food Policy*. 35(5):365-377.

Hazelton, P. & Murphy, B. 2007. *Interpreting soil test results: what do all the numbers mean?* Melbourne:CSIRO Publishing.

Helfer, G.A., Barbosa, J.L., Alves, D., da Costa, A.B., Beko, M. & Leithardt, V.R. 2021. Multispectral cameras and machine learning integrated into portable devices as clay prediction technology. *Journal of Sensor and Actuator Networks*. 10(3):1-16.

Jacobs, A. 2020. The importance of natural science collections in South Africa. *South African Journal of Science*. 116(11/12).

Janse Van Vuuren, J. & Groenewald, C. 2013. Use of scanning near-infrared spectroscopy as a quality control indicator for bulk blended inorganic fertilizers. *Communications in Soil Science and Plant Analysis*. 44(1-4):120-135.

John, K., Kebonye, N.M., Agyeman, P.C. & Ahado, S.K. 2021. Comparison of cubist models for soil organic carbon prediction via portable XRF measured data. *Environmental Monitoring and Assessment*. 193(4):1-15.

Kalvani, N., Mesdaghinia, A., Yaghmaeian, K., Abolli, S., Saadi, S., Alimohammadi, M., & Rashidi Mehrabadi, A. 2021. Evaluation of iron and manganese removal effectiveness by treatment plant modules based on water pollution index; a comprehensive approach. *Journal of Environmental Health Science and Engineering*. 19(1):1005–1013.

Kamruzzaman, M., Kalita, D., Ahmed, M.T., ElMasry, G., & Makino, Y. 2022. Effect of variable selection algorithms on model performance for predicting moisture content in biological materials using spectral data. *Analytica Chimica Acta*. 1202:1-9.

Katuwal, S., Knadel, M., Norgaard, T., Moldrup, P., Greve, M. & de Jonge, L. 2020. Predicting the dry bulk density of soils across Denmark: Comparison of single-parameter, multi-parameter, and vis–nir based models. *Geoderma*. 361:1–10.

Khorshidi, N. & Niazi, A. 2018. Moving window partial least squares after orthogonal signal correction as a coupling method for determination of uranium and thorium by ultrasound-assisted emulsification microextraction. *Journal of Chemometrics.* 33(1):1-15.

Kikkas, K. & Kulik, S. 2018. Modelling the effect of human activity on freshwater extraction from the earth's reserves. *IOP Conference Series: Earth and Environmental Science*. 180(2018):12-17.

Knadel, M., Gislum, R., Hermansen, C., Peng, Y., Moldrup, P., de Jonge, L. & Greve, M. 2017. Comparing predictive ability of laser-induced breakdown spectroscopy to visible near-infrared spectroscopy for soil property determination. *Biosystems Engineering*. 156:157-172.

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

Kock, A. 2022. *Creation of mid-infrared spectroscopy calibration algorithms for soil property predictions.* Potchefstroom: North-West University. (Dissertation – MSc).

Koirala, B., Zahiri, Z., & Scheunders, P. 2022. A robust supervised method for estimating soil moisture content from spectral reflectance. *IEEE Transactions on geoscience and remote sensing*. 60:1–13.

Kome, G.K., Enang, R.K., Tabi, F.O., & Yerima, B.P. 2019. Influence of clay minerals on some soil fertility attributes: A Review. *Open Journal of Soil Science*. 9(9):155-188.

Kpm analytics. S.a. *Near infrared measurements – how do they work?* https://www.kpmanalytics.com/articles-insights/near-infrared-measurements-how-do-they-work Date of access: 10 Oct. 2023.

Kwame, A.E., Zhiguang, Q., Ahudey, E. & Clement, A.P. 2020. Kendall's coefficient of concordance ranking of the effectiveness of single machine learning models in predicting stock price movement. *International journal of engineering research & technology (IJERT).* 9(9):1098-1107.

Kyprianidis, K.G. & Skvaril, J. 2017. *Developments in near-infrared spectroscopy*. Rijeka:InTech.

Le Roux, P. & du Preez, C. 2006. Nature and distribution of South African plinthic soils: Conditions for occurrence of soft and hard plinthic soils. *South African Journal of Plant and Soil.* 23(2):120–125.

Li, X., Li, Z., Yang, X. & He, Y. 2021. Boosting the generalization ability of vis-nir-spectroscopy-based regression models through Dimension Reduction and transfer learning. *Computers and Electronics in Agriculture.* 186:1–12.

Liang, X., Li, X. & Lei, T. 2012. *A new NIR technique for rapid determination of soil moisture content.* Paper delivered at the International conference on Systems and Informatics (ICSAI 2012), Yantai. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6223659 Date of access: 6 Oct. 2023.

Lin, L.I.K. 1989. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*. 45(1):255-268.

Liu, Z., Xia, Z., Chen, F., Hu, Y., Wen, Y., Liu, J., Liu, H. & Liu, L. 2020. Soil moisture index model for retrieving soil moisture in semiarid regions of China. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing.* 13:5929–5937.

Lobell, D.B. & Asner, G.P. 2002. Moisture effects on soil reflectance. *Soil science society of america journal*. 66(3):722-727.

Logunova, I. 2023. *K-means clustering algorithm in machine learning*. https://serokell.io/blog/k-means-clustering-in-machine-learning Date of access: 30 Oct. 2023.

Lu, B., Wang, X., Liu, N., He, K., Wu, K., Li, H. & Tang, X. 2020. Feasibility of NIR spectroscopy detection of moisture content in Coco-peat substrate based on the optimization characteristic variables. *Spectrochimica acta part A: Molecular and biomolecular spectroscopy.* 239:1–10.

Ludwig, B., Murugan, R., Parama, V.R. & Vohland, M. 2018. Use of different chemometric approaches for an estimation of soil properties at field scale with near infrared spectroscopy. *Journal of Plant Nutrition and Soil Science*. 181(5):704–713.

Ma, T., Brus, D., Zhu, A.-X., Zhang, L. & Scholten, T. 2020. Comparison of conditioned Latin hypercube and feature space coverage sampling for predicting soil classes using simulation from soil maps. *Geoderma*. 370:1–11.

Maleki, M., Mouazen, A., Ramon, H. & De Baerdemaeker, J. 2007. Multiplicative scatter correction during on-line measurement with near infrared spectroscopy. *Biosystems Engineering.* 96(3):427–433.

Manns, H., Parkin, G. & Martin, R. 2016. Evidence of a union between organic carbon and water content in soil. *Canadian Journal of Soil Science*. 96(3):305–316.

Mendes, W., Demattê, J., Rosin, N., Terra, F., Poppiel, R., Urbina-Salazar, D., Boechat, C., Silva, E., Curi, N., da Silva, S., José dos Santos, U. & Souza Valladares, G. 2022. The Brazilian soil mid-infrared Spectral Library: The power of the fundamental range. *Geoderma*. 415:57–76.

Minasny, B., McBratney, A., Pichon, L., Sun, W. & Short, M. 2009. Evaluating near infrared spectroscopy for field prediction of soil properties. Soil Research. 47(7):664–673.

Mishra, P. & Lohumi, S. 2021. Improved prediction of protein content in wheat kernels with a fusion of scatter correction methods in NIR Data Modelling. *Biosystems Engineering*. 203:93–97.

Mittal, M., Satapathy, S., Pal, V., Agarwal, B., Goyal, L. & Parwekar, P. 2021. Prediction of coefficient of consolidation in soil using machine learning techniques. *Microprocessors and Microsystems*. 82:1-15.

Mohammedzein, M., Csorba, A., Rotich, B., Justin, P., Melenya, C., Andrei, Y. & Micheli, E. 2023. Development of Hungarian Spectral Library: Prediction of soil properties and applications. Eurasian Journal of Soil Science (EJSS). 12(3):244–256.

Montesinos López, O.A., Montesinos López, A. & Crossa, J. 2022. Random Forest for genomic prediction. In: Montesinos López, O.A., Montesinos López, A., & Crossa, J. Eds. *Multivariate statistical machine learning methods for genomic prediction*. Pp. 633–681. Available from springer book collection: https://link.springer.com/content/pdf/10.1007/978-3-030-89010-0.pdf?pdf=button Date of access: 10 Oct. 2023.

Mouazen, A. & Al-Asadi, R. 2018. Influence of soil moisture content on assessment of bulk density with combined frequency domain reflectometry and visible and near infrared spectroscopy under semi field conditions*. Soil and Tillage Research*. 176:95–103.

Moyano, F., Vasilyeva, N., Bouckaert, L., Cook, F., Craine, J., Curiel Yuste, J., Don, A., Epron, D., Formanek, P., Franzluebbers, A., Ilstedt, U., Kätterer, T., Orchard, V., Reichstein, M., Rey,

A., Ruamps, L., Subke, J.-A., Thomsen, I. & Chenu, C. 2012. The moisture response of soil heterotrophic respiration: Interaction with soil properties. *Biogeosciences*. 9(3):1173–1182.

Nesbitt, K. 2014. An investigation into pan hydrology and ecology in the Makuleke concession, Northern Kruger, South Africa. Johannesburg: University of Witwatersrand. (Dissertation – MSc).

Ng, W., Minasny, B., Malone, B. & Filippi, P. 2018. In search of an optimum sampling algorithm for prediction of soil properties from Infrared Spectra. *PeerJ*. 6:1–27.

Nguyen, H., Nguyen, D. & Taguchi, K. 2021. A portable soil microbial fuel cell for sensing soil water content. *Measurement: Sensors*. 18:1–4.

Oliveira, J., Brossard, M., Vendrame, P., Mayi III, S., Corazza, E., Marchão, R. & Guimarães, M. 2013. Soil discrimination using diffuse reflectance vis–NIR spectroscopy in a local toposequence. *Comptes Rendus Geoscience*. 345(11–12):446–453.

Packer, I., Chapman, G. & Lawrie, J. 2019. On-ground extension of soil information to improve land management. *Soil Use and Management*. 35(1):75–84.

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(56):1-7.

Patil, B., Zope, P. & Patil, K. 2015. Development of low-cost TDR system for soil moisture measurement. *International Journal of Advanced Research in Education Technology (IJARET)*. 2(3):17-26.

Phinzi, K., Ngetar, N. & Ebhuoma, O. 2020. Soil erosion risk assessment in the Umzintlava catchment (T32E), Eastern Cape, South Africa, using RUSLE and random forest algorithm. *South African Geographical Journal*. 103(2):139-162.

Placidi, P., Morbidelli, R., Fortunati, D., Papini, N., Gobbi, F. & Scorzoni, A. 2021. Monitoring soil and ambient parameters in the IoT precision agriculture scenario: An original modeling approach dedicated to low-cost soil water content sensors. *Sensors*. 21:1-28.

Poppiel, R., Paiva, A. & Demattê, J. 2022. Bridging the gap between soil spectroscopy and traditional laboratory: Insights for routine implementation. *Geoderma*. 425:1–11.

Quinlan, J. 1992. *Learning with continuous classes.* Paper delivered at the 5th Australian Conference on Artificial Intelligence, Singapore. https://sci2s.ugr.es/keel/pdf/algorithm/congreso/1992-Quinlan-AI.pdf Date of access: 15 Nov. 2022.

Rahmat, M. & Ismail, N. 2018. Effect of optimum compaction moisture content formulations on the strength and durability of sustainable stabilised materials. *Applied Clay Science*. 157:257–266.

Rapanyane, M. & Ngoepe, C. 2019. The impact of illicit financial flows on the South African political economy under Jacob Zuma, 2009–2018. *Journal of Public Affairs*. 20(2):1-7.

Rasheed, M., Tang, J., Sarwar, A., Shah, S., Saddique, N., Khan, M., Imran Khan, M., Nawaz, S., Shamshiri, R., Aziz, M. & Sultan, M. 2022. Soil moisture measuring techniques and factors affecting the moisture dynamics: A comprehensive review. *Sustainability*. 14(18):1–23.

Roper, M., Ward, P., Keulen, A. & Hill, J. 2013. Under no-tillage and stubble retention, soil water content and crop growth are poorly related to soil water repellency. *Soil and Tillage Research*. 126:143-150.

Sadeghi, A., Tonazzini, A., Popova, L. & Mazzolai, B. 2014. A novel growing device inspired by plant root soil penetration behaviors. *PLoS ONE*. 9(2):1-10.

Schmid, M., Rath, D. & Diebold, U. 2022. Why and how Savitzky–Golay filters should be replaced. *ACS Measurement Science Au*. 2(2):185–196.

Schnecker, J., Meeden, D., Calderon, F., Cavigelli, M., Lehman, R., Tiemann, L. & Grandy, A. 2021. Microbial activity responses to water stress in agricultural soils from simple and complex crop rotations. *Soil*. 7(2):547–561.

Schonlau, M. & Zou, Y. 2020. The Random Forest Algorithm for Statistical Learning. *The Stata Journal: Promoting communications on statistics and Stata*. 20(1):3–29.

Schwartz, B., Schreiber, M. & Yan, T. 2008. Quantifying field-scale soil moisture using electrical resistivity imaging. *Journal of Hydrology*. 362(3–4):234–246.

Serbin, G. & Or, D. 2004. Ground-penetrating radar measurement of soil water content dynamics using a suspended horn antenna. *IEEE Transactions on Geoscience and Remote Sensing*. 42(8):1695-1705.

Seybold, C., Ferguson, R., Wysocki, D., Bailey, S., Anderson, J., Nester, B., Schoeneberger, P., Wills, S., Libohova, Z., Hoover, D. & Thomas, P. 2019. Application of mid-infrared spectroscopy in soil survey. *Soil Science Society of America Journal*. 83(6):1746-1759.

Shahbazi, F., McBratney, A., Malone, B., Oustan, S. & Minasny, B. 2019. Retrospective monitoring of the spatial variability of crystalline iron in soils of the east shore of Urmia Lake, Iran using remotely sensed data and digital maps. *Geoderma.* 337:1196–1207.

Sharma, B., Vaish, B., Monika, Singh, U., Singh, P. & Singh, R. 2019. Recycling of organic wastes in agriculture: An environmental perspective. *International Journal of Environmental Research*. 13(2):409-429.

Shepherd, K., Ferguson, R., Hoover, D., van Egmond, F., Sanderman, J. & Ge, Y. 2022. A Global soil spectral calibration library and estimation service. *Soil Security*. 7:1–13.

Shock, C., Wang, F. 2010. Soil water tension, a powerful tool for productivity and stewardship. *Hortscience*. 46(2):178-185.

Shoukri, M., Al-Hassan, T., DeNiro, M., El Dali, A., & Al-Mohanna, F. 2016. Bias and mean square error of reliability estimators under the one and two random effects models: The effect of Non-Normality. *Open Journal of Statistics*. 6(2):254–273.

Shreeja, D. S.a. *Measurement of soil moisture: 4 methods.* https://www.soilmanagementindia.com/soil-water/measurement-of-soil-moisture-4-methods/1783 Date of access: 16 Nov. 2022.

Silalahi, D., Midi, H., Arasan, J., Mustafa, M. & Caliman, J. 2018. Robust generalized Multiplicative Scatter Correction algorithm on Pretreatment of Near Infrared Spectral Data. *Vibrational Spectroscopy.* 97:55–65.

Singh, C., Shashtri, S., Rina, K. & Mukherjee, S. 2012. Chemometric analysis to infer hydro-geochemical processes in a semi-arid region of India. *Arabian Journal of Geosciences*. 6(8):2915-2932.

Skalko, J. 2013. If food and water are proportionate means, why not oxygen?. *The National Catholic Bioethics Quarterly*. 13(3):453-467.

Slaughter, D., Pelletier, M. & Upadhyaya, S. 2001. Sensing soil moisture using NIR spectroscopy. *Applied Engineering in Agriculture*. 17(2):241-247.

SoilSpec4GG. S.a. OSSL Manual. https://soilspectroscopy.github.io/ossl-manual/ Date of access: 31 Oct. 2023.

Sujatha, K.N., Kavya, G., Manasa, P., & Divya, K. 2016. Assessment of soil properties to improve water holding capacity in soils. *International research journal of engineering and technology (IRJET).* 3(3):1777–1783.

Sutcliffe, C., Knox, J. & Hess, T. 2021. Managing irrigation under pressure: How supply chain demands and environmental objectives drive imbalance in agricultural resilience to water shortages. *Agricultural Water Management*. 243:1–10.

Tangirala, S. 2020. Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm. *International Journal of Advanced Computer Science and Applications*. 11(2):612-619.

Tignino, M. 2010. Water, international peace, and security. *International Review of the Red Cross.* 92(879):647–674.

Twomey, S.A., Bohren, C.F., & Mergenthaler, J.L. 1986. Reflectance and albedo differences between wet and dry surfaces. *Applied Optics*. 25(3):431–437.

Van Tol, J., Le Roux, P., Lorentz, S. & Hensley, M. 2013. Hydropedological classification of South African hillslopes. *Vadose Zone Journal.* 12(4):1-10.

Van Zyl, G. 2004. *Surface enrichment sprays as a cost effective solution for preventative maintenance.* Paper delivered at the 8th Conference on asphalt pavements for Southern Africa, Sun City. https://www.researchgate.net/profile/Gerrie-Van-Zyl/publication/242231035_SURFACE_ENRICHMENT_SPRAYS_AS_A_COST_EFFECTIVE_SOLUTION_FOR_PREVENTIVE_MAINTENANCE/links/555c89b608ae91e75e7847ff/SURFACE-ENRICHMENT-SPRAYS-AS-A-COST-EFFECTIVE-SOLUTION-FOR-PREVENTIVE-MAINTENANCE.pdf?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6Il9kaXJlY3QiLCJwYWdlIjoicHVibGljYXRpb24iLCJwcmV2aW91c1BhZ2UiOiJfZGlyZWN0In19 Date of access: 11 Oct. 2023

Viljoen, G. & van der Walt, K. 2018. South Africa's water crisis - an interdisciplinary approach. *Tydskrif vir Geesteswetenskappe*. 58(3):483-500.

Viscarra Rossel, R., Behrens, T., Ben-Dor, E., Brown, D., Demattê, J., Shepherd, K., Shi, Z., Stenberg, B., Stevens, A., Adamchuk, V., Aïchi, H., Barthès, B., Bartholomeus, H., Bayer, A.,

Bernoux, M., Böttcher, K., Brodský, L., Du, C., Chappell, A., Fouad, Y., Genot, V., Gomez, C., Grunwald, S., Gubler, A., Guerrero, C., Hedley, C., Knadel, M., Morrás, H., Nocita, M., Ramirez-Lopez, L., Roudier, P., Campos, E., Sanborn, P., Sellitto, V., Sudduth, K., Rawlins, B., Walter, C., Winowiecki, L., Hong, S., & Ji, W. 2016. A global spectral library to characterize the world's soil. *Earth-Science Reviews*. 155:198-230.

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

Waldrip, H., Schwartz, R., He, Z., Todd, R., Baumhardt, R., Zhang, M., Parker, D., Brauer, D. & Min, B. 2022. Soil water extractable organic matter under long-term dryland cropping systems on the Texas high plains. *Soil Science Society of America Journal*. 86(5):1249–1263.

Watanabe, A., Kuramata, M., Majima, K., Kiyohara, H., Kensho, K. & Nakata, K. 2021. Constrained Generalized Additive 2 model with consideration of high-order interactions. *2021 International Conference on Electrical, Computer and Energy Technologies (ICECET)*. 1–7.

Wei, H., Xu, W., Wei, C. & Meng, Q. 2018. Influence of water content and shear rate on the mechanical behavior of soil-rock mixtures. *Science China Technological Sciences*. 61(8):1127-1136.

Weinert, H. 1984. Climate and the durability of South African road aggregates. *Bulletin of the International Association of Engineering Geology.* 29(1):463-466.

Xu, Q., Yang, Y., Liu, Y. & Wang, X. 2017. An improved Latin hypercube sampling method to enhance numerical stability considering the correlation of input variables. *IEEE Access.* 5:15197–15205.

Yamanaka, T., Kaihotsu, I., Oyunbaatar, D. & Ganbold, T. 2007. Summertime soil hydrological cycle and surface energy balance on the Mongolian steppe. *Journal of Arid Environments*. 69(1):65-79.

Yang, L., Li, X., Shi, J., Shen, F., Qi, F., Gao, B., Chen, Z., Zhu, A., & Zhou, C. 2020. Evaluation of conditioned Latin hypercube sampling for soil mapping based on a machine learning method. *Geoderma*. 369:1–15.

Yerro, A. & Ceccato, F. 2023. Soil–water–structure interactions. *Geotechnics*. 3(2):301–305.

Yin, Z., Lei, T., Yan, Q., Chen, Z. & Dong, Y. 2013. A near-infrared reflectance sensor for soil surface moisture measurement. *Computers and Electronics in Agriculture.* 99:101–107.

Zhang, F. & Knoll, A. 2016. Systematic error modeling and bias estimation. *Sensors*. 16(5):1-10

Zhang, Q., Phillips, R., Manzoni, S., Scott, R., Oishi, A., Finzi, A., Daly, E., Vargas, R. & Novick, K. 2018. Changes in photosynthesis and soil moisture drive the seasonal soil respiration-temperature hysteresis relationship. *Agricultural and Forest Meteorology*. 259:184–195.

Zhang, R., Zhou, Y., Yue, Z., Chen, X., Cao, X., Xu, X., Xing, Y., Jiang, B., Ai, X. & Huang, R. 2019. Changes in photosynthesis, chloroplast ultrastructure, and antioxidant metabolism in leaves of sorghum under waterlogging stress. *Photosynthetica*. 57(4):1076–1083.

Zhang, Z. & Furman, A. 2021. Soil redox dynamics under dynamic hydrologic regimes - A Review. *Science of The Total Environment.* 763:1–17.

Zhou, Y., Chen, S., Hu, B., Ji, W., Li, S., Hong, Y., Xu, H., Wang, N., Xue, J., Zhang, X., Xiao, Y. & Shi, Z. 2022. Global soil salinity prediction by open soil vis-nir spectral library. Remote Sensing. 14(21):1–13.

Zimmermann, B. & Kohler, A. 2013. Optimizing Savitzky–golay parameters for improving spectral resolution and quantification in infrared spectroscopy. *Applied Spectroscopy.* 67(8):892–902.