

https://dspace.org/

Detecting soil properties in agricultural lands using field spectroscopy and regression models

Zahinda, Franck Mugisho

2020

https://hdl.handle.net/10539/31000

Zahinda, Franck Mugisho (2020) Detecting soil properties in agricultural lands using field spectroscopy and regression models, University of the Witwatersrand, Johannesburg, <http://hdl.handle.net/10539/31000>

Downloaded from a DSpace repository. https://dspace.org/

DETECTING SOIL PROPERTIES IN AGRICULTURAL LANDS USING FIELD SPECTROSCOPY AND REGRESSION MODELS

By

Franck Mugisho Zahinda – Student No 2288618

A research report submitted to the Faculty of Science, University of the Witwatersrand, Johannesburg, in partial fulfilment of the requirements for the degree of Master of Science in Geographical Information Systems and Remote Sensing

Supervisor: Dr Elhadi Adam

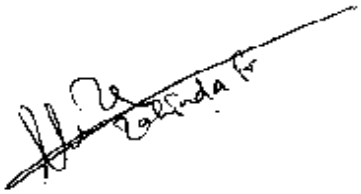
Co-supervisor: Dr Mohamed A. M. Abd Elbasit

Johannesburg, 2020

DECLARATION

I, Franck Mugisho Zahinda, declare that the present research report is my own unaided work. It is being submitted to the Degree of Master of Science in Geographical Information Systems and Remote Sensing to the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other University.

Signature of candidate



14th day of September 2020 in Johannesburg

ABSTRACT

Reflectance spectroscopy can be used to non-destructively characterize materials for a wide range of applications. In this study, visible-near infrared (Vis-NIR) spectroscopy was evaluated for the prediction of diverse soil properties (clay content, SOC, TN, and pH) related to different soil samples from the Eastern Cape Province in South Africa.

Soil samples were scanned by a portable spectrometer at 1 nm wavelength resolution from 350 to 2500 nm. Calibrations between soil properties obtained from digital soil maps and reflectance spectra were then developed using cross-validation under partial least squares regression (PLSR) and support vector machine regression (SVMR). Raw reflectance and Savitzky-Golay first derivative data were used separately for all the samples in the data set. Key wavelengths to predict clay content, SOC, TN, and pH were identified using the variable importance projection (VIP) and Boruta algorithms. Data were additionally divided into two random subsets of 70 and 30% of the full data, which were each used for calibration and validation.

The results indicated that Vis-NIR spectroscopy can be successfully used to predict soil clay content, SOC, TN and pH. For clay content, SOC, and pH, the best results were obtained by SVMR with first derivative data (RPD = 2.05, R_p^2 = 0.83, RMSEP = 1.95% for clay content; RPD = 2.40, R_p^2 = 0.87, RMSEP = 2.48 $g.kg^{-1}$ for SOC; and RPD = 2.87, R_p^2 = 0.89, RMSEP = 0.16 for pH). In contrast, PLSR with raw data outperformed SVMR models for TN prediction (RPD = 2.15, R_p^2 = 0.77, RMSEP = 0.20 $mg.kg^{-1}$). Key wavelengths to predict the four properties were identified mostly around 400-700 nm and 2200-2450 nm. In conclusion, Vis-NIR spectroscopy was variably good in estimating clay content, SOC, TN and pH in laboratory conditions, and showed potential for substituting traditional wet laboratory analyses or providing inexpensive data.

Keywords: Soil properties, Vis-NIR spectroscopy, PLSR, SVMR.

ACKNOWLEDGEMENTS

I would like to express my sincere appreciation and gratitude to my supervisors Dr Elhadi Adam and Dr Mohamed A. M. Abd Elbasit for the support and supervision they offered during the present research. Your consistency and commitment provided me with an opportunity to learn and improve my knowledge.

I appreciate the valuable contribution of the Agricultural Research Council (ARC-ISWC) for providing me with resources and support to process my data.

I am very grateful to all the lecturers and staff members of the School of Geography, Archaeology and Environmental Sciences for providing me with advanced knowledge in Geographic Information Systems and remote sensing.

My sincere thanks also go to Professor Gerhard Bringmann and Professor Karine Ndjoko for instilling in me the sense of excellence, and for granting me the BEBUC scholarship (Bourse d'Excellence Bringmann aux Universités Congolaises) for a Master degree.

CONTENTS

ABSTRACT ..... **ii**

ACKNOWLEDGEMENTS ..... **iii**

LIST OF FIGURES ..... **vii**

LIST OF TABLES ..... **ix**

LIST OF ABBREVIATIONS ..... **x**

CHAPTER ONE **1**

GENERAL INTRODUCTION **1**

1.1 Background 1

1.2 Problem statement 4

1.3 Aim and objectives 5

1.4 Scope of study 6

CHAPTER TWO **7**

LITERATURE REVIEW **7**

2.1 Understanding soil properties for agriculture 7

2.1.1 Soil physical properties 8

2.1.2 Soil chemical properties 8

2.1.3 Importance of clay content, SOC, TN, and pH for agriculture 9

2.2 Remote sensing of soils 10

2.2.1 Satellite remote sensing 10

2.2.2 Proximal remote sensing: laboratory and *in situ* spectroscopy to predict soil properties 11

2.3 Spectral reflectance and soil properties 12

2.3.1 General trends of soil spectra in Vis-NIR spectroscopy 12

2.3.2 Relationship between soil reflectance spectra and soil properties 14

2.4 Use of regression models in soil spectroscopy 15

2.4.1 Multicollinearity issue in spectral data 16

2.4.2 Linear vs non-linear models 17

2.5 Chapter summary 18

CHAPTER THREE 20

MATERIALS AND METHODS 20

3.1 Study area 20

3.1.1 Soil sample selection 21

3.1.2 Determination of soil properties of the selected samples 22

3.1.3 Characteristics of the downloaded digital soil maps 23

3.2 Laboratory spectral measurements 24

3.3 Laboratory spectra pre-processing and transformation 26

3.4 Statistical analysis 27

3.5 Establishment of calibration models and variable selection 29

3.5.1 Partial least squares regression (PLSR) and selection of key wavelengths with the Variable importance projection (VIP) algorithm 29

3.5.2 Variable selection by the Random Forest (RF) Boruta algorithm 30

3.5.3 Support vector machines regression (SVMR) 31

3.5.4 Validation and comparison of PLSR and SVMR models 32

CHAPTER FOUR 34

RESULTS 34

4.1 Descriptive statistics of the soil properties 34

4.2 Spectral characteristics of the soil samples 36

4.3 Determination of key wavelengths 37

4.3.1 PLSR-variable importance projection (VIP) 37

4.3.2 Feature selection with the Boruta algorithm 38

4.3.3 Position of wavelengths and interpretation 39

4.4 Development of PLSR and SVMR models 41

4.5 Prediction accuracy of the multivariate methods 42

4.5.1 PLSR model performance 42

4.5.2 SVMR model performance 45

4.6 Comparison between PLSR and SVMR models 49

CHAPTER FIVE 52

DISCUSSION 52

5.1 Qualitative analysis of collected spectra 52
5.2 Identifying key wavelengths for predicting soil properties 53
5.3 Performance of PLSR and SVMR in predicting clay content, SOC, TN, and pH ... 55
5.4 Limitations of the study 56
CONCLUSIONS 57
REFERENCES 58
APPENDICES 74
Appendix A – PLSR-VIP code 74
Appendix B – RF-Boruta code 75

LIST OF FIGURES

Figure 2.1 Principal types of NIR absorption bands and their locations (Butkuté and Slepetiene, 2004)..... 14

Figure 2.2 The main characteristic spectral signatures and corresponding soil attributes (Zelikman and Carmina, 2013). 15

Figure 3.1 Geographical location of the study area showing the spatial distribution of soil samples. 21

Figure 3.2 Location of sampling points in different geology types. 22

Figure 3.3 Primary digital soil maps used (0-5 cm): (a) clay content (%); (b) SOC (g.kg-1); (c) TN (mg.kg-1); (d) pH. 24

Figure 3.4 Laboratory spectral measurements: (a) selection of soil samples from the store; (b) preparation of soil samples; (c) spectral measurements with the spectrometer. 25

Figure 3.5 Spectral data: (a) collected spectra; (b) spectra with noisy and water absorption regions removed (< 380 nm, 1350-1640 nm, 1790-1960 and > 2450 nm); (c) and first derivative spectra. 27

Figure 3.6 Detection of outliers after principal component analysis (PCA) of soil samples (n = 200) 28

Figure 4.1 Mean laboratory soil reflectance spectra: (a) raw spectra; (b) first derivative spectra. 36

Figure 4.2 Identification of key wavelengths with VIP algorithms: (a) clay raw spectra; (b) clay first derivative spectra; (c) SOC raw spectra; (d) SOC first derivative spectra; (e) TN raw spectra; (f) TN first derivative spectra; (g) pH raw spectra; (h) pH first derivative spectra. .. 38

Figure 4.3 Performance of PLSR in predicting clay content: (a) clay PLSR raw model with all wavelengths; (b) clay first derivative PLSR model; (c) clay PLSR model with key wavelengths; (d) clay derivative PLSR model with key wavelengths. 42

Figure 4.4 Performance of PLSR in predicting SOC: (a) SOC PLSR raw model with all wavelengths; (b) SOC first derivative PLSR model; (c) SOC PLSR model with key wavelengths; (d) SOC derivative PLSR model with key wavelengths. 43

Figure 4.5 Performance of PLSR in predicting TN: (a) TN PLSR raw model with all wavelengths; (b) TN first derivative PLSR model; (c) TN PLSR model with key wavelengths; (d) TN derivative PLSR model with key wavelengths. 44

Figure 4.6 Performance of PLSR in predicting pH: (a) pH PLSR raw model with all wavelengths; (b) pH first derivative PLSR model; (c) pH PLSR model with key wavelengths; (d) pH derivative PLSR model with key wavelengths. 45

Figure 4.7 Performance of SVMR in predicting clay content: (a) clay SVMR raw model with all wavelengths; (b) clay first derivative SVMR model; (c) clay SVMR model with key wavelengths; (d) clay derivative SVMR model with key wavelengths. 46

Figure 4.8 Performance of SVMR in predicting SOC: (a) SOC SVMR raw model with all wavelengths; (b) SOC first derivative SVMR model; (c) SOC SVMR model with key wavelengths; (d) SOC derivative SVMR model with key wavelengths. 47

Figure 4.9 Performance of SVMR in predicting TN: (a) TN SVMR raw model with all wavelengths; (b) TN first derivative SVMR model; (c) TN SVMR model with key wavelengths; (d) TN derivative SVMR model with key wavelengths. 48

Figure 4.10 Performance of SVMR in predicting pH: (a) pH SVMR raw model with all wavelengths; (b) pH first derivative SVMR model; (c) pH SVMR model with key wavelengths; (d) pH derivative SVMR model with key wavelengths. 49

LIST OF TABLES

Table 3.1 Waveband range of Analytical Spectral Devices (Bangelesa, 2017). 25

Table 3.2 Qualitative model performance based on RPD and R^2 (Gates, 2018). 33

Table 4.1 Descriptive statistics of clay content, SOC, TN, and pH within three different datasets. 35

Table 4.2 Pearson's correlation coefficients (r) for clay content, SOC, TN, and pH values (n=187). 36

Table 4.3 Number of key wavelengths selected by the RF-Boruta algorithm, RMSE on the validation dataset and location of important features on the raw and first derivative spectral data. 39

Table 4.4 Comparison of the PLSR-VIP and RF-Boruta algorithms in selecting key wavelengths in the visible and near-infrared ranges. 40

Table 4.5 Functional groups and vibration modes of wavelengths considered for interpretation (Stuart, 2004) 41

Table 4.6 Performance of all SVMR and PLSR models in the calibration and validation datasets. 50

LIST OF ABBREVIATIONS

AfSIS:	Africa Soil Information Service
AIC:	Akaike Information Criterion
ASD:	Analytical Spectral Device
BPNN:	Backpropagation Neural Network
CEC:	Cation Exchange Capacity
DSM:	Digital Soil Map
ET:	Electronic Transfer
MSEOOB:	Mean Square Error of the Out-Of-Bag sample
NIR:	Near-Infrared
OOB:	Out-Of-Bag sample
PA:	Precision Agriculture
PCR:	Principal Component Regression
pH:	Potential of Hydrogen
PLSR:	Partial Least Squares Regression
R^2:	Coefficient of determination
RF:	Random Forest
RMSEP:	Root Mean Square Error of the validation data set
RPD:	Residual Prediction Deviation
RS:	Remote Sensing
SOC:	Soil Organic Carbon
SOM:	Soil Organic Matter

SPA:	Successive Projections Algorithm
SVMR:	Support Vector Machines Regression
SWIR:	Short-Wave Infrared
TN:	Total Nitrogen
UV:	Ultraviolet
VIP:	Variable Importance Projection
Vis:	Visible
Vis-NIR:	Visible and Near-Infrared

CHAPTER ONE

GENERAL INTRODUCTION

This chapter introduces the study with background elements of the research, the problem statement, the aim and objectives, and the scope of the study. Thus, this chapter aims to describe the context of the study, give it meaning, and clarify what it hopes to achieve.

1.1 Background

Healthy soils are essential to agricultural production (FAO and ITPS, 2015). Poor quality of the soil is among the major factors of food insecurity and the situation seems to be alarming by looking at the trend of soil degradation and food demand (Lal, 2009). A variety of estimates suggest that current food production must be doubled by 2050 to respond to the world population increase (Tilman *et al.*, 2011; Ray *et al*., 2013). Therefore, it is essential to manage soil fertility or many regions of the world will have adverse environmental and agricultural consequences (McBratney *et al*., 2014).

Soil fertility, a feature of soil health, is the capacity of soil to provide nutrients required by plants for growth, the foundation of the food system. Fertile soils produce healthy crops that in turn nourish people (FAO, 2005). Soil fertility combines several properties (biological, chemical and physical), which affect directly or indirectly the soil nutrient dynamics and availability (Gates, 2018). In most cases, the term soil fertility describes the present state of the soil, which means that soil fertility is a combination of the inherent soil quality (e.g. mineral composition, soil texture) and achieved qualities resulting from the influence of climate, relief and organisms interacting overtime (e.g. soil structure, soil organic matter content, phosphorus concentration) (Karltun *et al.*, 2011).

Fertility is a “manageable” soil property and its management is of utmost importance for optimizing crop nutrition on both a short-term and a long-term basis to achieve sustainable crop production (FAO, 2015). Hence, understanding soil fertility and the various processes taking place in the soil helps farmers make prudent management decisions (Foster *et al.*, 2013).

Precision agriculture (PA), important for sustainable crop production, has become common practice for growers around the world (Ji *et al.*, 2016).

Nowadays, PA technologies are being developed to optimize farm profit and minimize environmental impacts by adjusting production inputs to the needs of individual areas within fields (Ji *et al.*, 2016). One of the first and most important areas in which PA has been applied is in managing variability in soil fertility (Thomasson *et al.*, 2001). However, the data for this purpose come from laborious soil sample collection efforts and expensive laboratory analyses (Thomasson *et al.*, 2001, Ji *et al.*, 2016).

The study of soil attributes implies the determination of their analytical values as well as their mapping. This is recommended before soil management since soil analyses are essential for the assessment and monitoring of the soil's chemical and physical conditions (FAO, 2008). The most applied methods to determine soil attributes in laboratories are those called traditional wet analyses (FAO, 2008). However, wet analysis usually involves a great number of samples to be analysed (Demattê *et al*., 2019). Furthermore, one analysis may take several days for delivering results, which is not adequate considering the speed required in PA. Also, despite being time-consuming, traditional laboratory analyses use several types of chemical substances, and some are toxic (e.g. sulfuric acid) (Demattê *et al*., 2019).

Under those circumstances, new methods for determining variability in soil characteristics are needed (Thomasson *et al.*, 2001; Babaeian *et al*., 2015). The use of methods that quickly allow the evaluation of soil properties, at low cost and without residues production may facilitate the evaluation of more samples to characterize soils in more detail and for different purposes (Silva *et al*., 2018). Therefore, soil remote sensing has been viewed as an efficient alternative that could significantly increase the affordability of soil measurements (Ben-Dor *et al*., 2009).

Remote sensing (RS) has been widely used in agriculture and one of its earliest applications is the characterization of soil properties. For example, Bushnell (1932) described efforts in the 1920s to use aerial photos to map boundaries of different soil series (Barnes *et al*., 2003). Aerial photographs have been used as a mapping aid in most of the soil surveys since the late 1950s. Satellite RS has also been used in soil science and agriculture since the 1970s (Mulla, 2013).

However, aerial photographs and most of the optical RS means cannot detect the entire soil body ("pedon") that extends from the surface to the parent material. Moreover, the thin, upper layer that is eventually sensed by optical sensors may be affected by many factors such as dust,

rust, ploughing, particle size distribution, vegetation coverage, litter, and physical and biogenic crusts (Ben-Dor *et al*., 2009). Thus, optical RS of soils from far distances becomes a significant challenge.

The spectral resolution of remotely sensed data constitutes another issue. Single or even multi-band RS means is rather limited and problematic when striving for quantitatively accurate information. For that purpose, high spectral resolution data are required since higher resolution predetermines higher information content. Thus, several studies over the past decade used hyperspectral data from the visible and near-infrared spectroscopy to recognize soils qualitatively and quantitatively (Ben-Dor *et al*., 2009).

Visible and near-infrared (Vis-NIR) spectroscopy has garnered a wide interest in soil assessment studies (Kopackova and Ben-Dor, 2016), and its benefits have been documented extensively. Vis-NIR spectroscopy enables the quantification of several important properties of soil samples from their Vis-NIR spectral responses in a cheaper and faster way than by conventional laboratory methods (Ramirez-Lopez *et al*., 2019). It also enables soil surveyors to increase sampling densities without incurring substantial additional costs (Ramirez-Lopez *et al*., 2019) and has the potential to provide both rapid and high-resolution prediction of multiple soil properties for PA, soil health assessment, and other applications related to environmental protection and agronomic sustainability (Veum *et al*., 2018).

Vis-NIR spectroscopy has been widely used to quantify soil physical and chemical properties with excellent accuracy (Schirrmann *et al*., 2013). Properties such as soil organic matter (SOM), soil organic carbon (the C in SOM), moisture, cation exchange capacity (CEC), total nitrogen (TN), total phosphorus (TP), total potassium (TK), pH and texture have been well predicted by many authors using Vis-NIR (Xu *et al*., 2018).

For example, Vâgen *et al*. (2006) successfully predicted SOC (R^2 = 0.92, RMSE = 8.40 $g.kg^{-1}$) and TN (R^2 = 0.93, RMSE = 0.64 $g.kg^{-1}$) using PLSR in laboratory conditions. Todorova *et al*. (2011) found good model predictions for pH using NIR (R^2 = 0.91; RPD = 2.3). Total P was predicted successfully with Vis-NIR by Bogrekci and Lee (2005) (R^2 = 0.92, RMSE = 273.3 $mg.kg^{-1}$) and Todorova *et al*. (2011) (R^2 = 0.89, RPD = 2.0). Chang *et al*. (2001) using Vis-NIR and Zornoza *et al*. (2008) with NIR were successful in predicting CEC at regional scales (R^2 = 0.81; RPD = 2.28 and R^2 = 0.92; RPD = 3.46 respectively).

The technique used in Vis-NIR spectroscopy is based on the fact that energy is absorbed or reflected by the vibrations of atomic bonds (basically O-H, C-H, N-H and C-O groups) (Xu *et al*., 2018). Over the past decades, Vis-NIR reflectance spectroscopy has also proved to be environmental-friendly, reproducible, and repeatable (Nocita *et al*., 2015). The technique is mainly used in the laboratory, but its application *in situ* (Viscarra Rossel *et al.*, 2009), as well as from air- and spaceborne sensors is growing (Ben Dor *et al*., 2009; Nocita *et al*., 2015).

Though Vis-NIR spectroscopy is a method that has proven useful in quantifying constituents of soil samples, little is known about how it performs in many areas due to the absence of local models since a model that works for one area may not work for another. In this respect, it is important to develop local models of soil properties using laboratory and/or *in situ* spectroscopy in different regions of South Africa. Bangelesa (2017) mentioned that in Southern Africa, few investigations address the modelling of soil properties with either laboratory or field measurements and, to the best of our knowledge, there are not many investigations that address the modelling of multiple physical and chemical soil properties using laboratory spectral measurements.

1.2 Problem statement

Soil fertility and its management play an important role in farm productivity (FAO, 2006). A characteristic of most South African soils is that they are extremely vulnerable to degradation and have low recovery potential. Thus, poor land management at a small scale can be devastating, with little chance of recovery (WWF, 2015). Agricultural soils of South Africa are subject to physical, chemical and biological degradation (Hensley *et al*., 2006). Consequently, there is a crucial need to assess and monitor soil properties.

The Eastern Cape Province of South Africa covers an area of about 17.1 million ha and has a diversity of soils and climatic conditions permitting a variety of different forms of agriculture (Erasmus, 1996; Mnkeni *et al*., 2005). Approximately 30% of the area consists of smallholdings on which farmers mostly practice mixed farming for home consumption purposes. This involves the grazing of cattle, goats, and sheep on communally owned natural rangeland; production of maize, beans, and pumpkins on individual arable holdings of between 1 and 5 ha. The province also produces a wide selection of grains and vegetables, such as maize, potatoes, cabbages, Swiss chard, onions, peas, and carrots in gardens (Mnkeni *et al*., 2005).

On the other hand, the Eastern Cape is among the provinces that are badly affected by soil degradation (Paterson *et al.,* 2015). Hence, PA which implies proper management of soil physical and chemical properties is vital to enhance farm activities by sustaining soil fertility. This can be done by estimating the soil clay content, soil organic carbon (SOC), total N (TN) and the potential of Hydrogen (pH).

To predict different soil properties, numerous statistical techniques have been applied to soil spectroscopy and include multiple linear regression, partial least square regression, generalized linear models and linear mixed models, etc. (Were *et al*., 2015). Recently, many studies applied new methods from the machine learning field, such as artificial neural networks, support vector machines, boosted regression trees, and random forests (Were *et al.*, 2015).

Although these approaches have been previously tested to measure soil properties using spectroscopy, their results vary according to different factors (e.g., type of soil and geological heterogeneity), thus cannot be applied everywhere (Viscarra Rossel and Behrens, 2010). Also, in Southern Africa, few studies focused on detecting different soil properties using either laboratory or *in situ* measurements (Bangelesa, 2017), though it would be ideal to predict soil properties of agricultural fields with a rapid and relatively inexpensive method.

Estimating soil clay content, SOC, TN, and pH using traditional laboratory analyses would be laborious, time-consuming and expensive for the Eastern Cape Province because of the high number of soil samples the study would require. Thus, we assume that Vis-NIR spectroscopy and regression models will be good at rapidly predicting soil properties under laboratory conditions at a lower cost.

1.3 Aim and objectives

This research aims to predict agricultural fields' soil properties using reflectance spectroscopy and multivariate methods. The specific objectives of this study are:

(1) To qualitatively assess the spectra collected on soils;
(2) To determine key wavelengths for estimating clay content, SOC, TN, and pH with satisfying accuracy using partial least squares regression (PLSR) and random forest (RF);
(3) To evaluate the predictive ability of partial least squares regression (PLSR) and support vector machine regression (SVMR) in predicting the selected properties.

1.4 Scope of study

This study is limited at developing spectral models using PLSR and SVMR to quantify clay content, SOC, TN, and pH in soil samples without running chemical analyses. In this research, no mapping was considered.

CHAPTER TWO

LITERATURE REVIEW

This chapter comprises a concise review of the literature related to this research as well as concepts and theories that support the content and context of the research. Amongst these include different topics such as soil physical and chemical properties, spectral reflectance, remote sensing of soil properties and spectroscopy modelling.

2.1 Understanding soil properties for agriculture

Comprehensive, accurate and up-to-date soil information is an essential input into agricultural and ecological decision-making models (Gourlay *et al.*, 2017; Silva *et al.*, 2018). Soil information can help predict scenario dependent crop yields as well as water and nutrient dynamics. It can also help identify areas at risk of soil degradation and support choosing appropriate preventive and rehabilitative soil management interventions (Hengl *et al.*, 2015).

For a given soil, its properties depend on the history of the soil formation and can be substantially modified by human intervention (e.g. through agricultural practices) (FAO and ITPS, 2015). A proper understanding of soil characteristics and adequate interpretation of the magnitudes of its properties, both combined under the broader term of soil quality, is required for proper management of agricultural soils (Delgado and Gomez, 2016).

Soil is a complex material that is extremely variable in its physical and chemical composition. It is formed from exposed masses of partially weathered rocks and minerals composing the earth's crust (Ben-Dor *et al*., 2009). However, discrimination of physical, chemical, biological properties and soil quality is very difficult because many soil properties are interrelated (Warkentin, 1995).

The soil quality indicators like chemical and biological properties have a prominent connection between them. Nitrogen in its mineralizable form for example, is considered by many researchers to be part of both chemical and biological properties (Jamil *et al*., 2016).

2.1.1 Soil physical properties

Soil physical properties determine many key soil processes and thus the agronomical potential of a soil (Delgado and Gomez, 2016). Soil is composed of minerals, soil organic matter (SOM), water, and air. The composition and proportion of these components greatly influence soil physical properties, including texture, structure, and porosity. In turn, these properties affect air and water movement in the soil, and thus the soil's ability to function (McCauley, 2005).

Soil texture, bulk density, and structure affect the management practices required to maintain water potential, oxygen diffusion rate, temperature, and mechanical resistance in a range suitable for good production (Letey *et al.*, 1958). Soil texture, which is a description of the size distribution of the mineral soil particles composing the solid fraction of the soil (from clay < 2 μm to coarse particles > 2000 μm) is perhaps the most important since it determines many other physical properties (such as infiltration rate) and some chemical properties (such as cation exchange capacity, CEC) (Delgado and Gomez, 2016).

2.1.2 Soil chemical properties

According to Jamil *et al*. (2016), the chemistry of clays and humus determines soil chemical properties. Chemical properties of soils include: (i) inorganic matters of soil (soil mineralogy); (ii) organic matters in soil, most of which originates from plant tissues (the dry matter consists of carbon (C), oxygen, hydrogen (H) and small amounts of sulphur (S), nitrogen (N), phosphorus (P), potassium (K), calcium (Ca) and magnesium (Mg); (iii) colloidal properties of soil particles, and (iv) soil reactions and buffering action referring to the change in pH of a system (acidic soils and basic soils) (Balasubramanian, 2017).

In soils, there are about 20 nutrients required for plant health. Three of them, C, H, and O are considered part of the protoplasm, and the remainder is considered to be mineral elements. Hydrogen comes from water, oxygen from water and air (Lohry, 2007). Three main elements, nitrogen, phosphorus and potassium (N, P, K) are required in abundance. They must be readily available through soil medium or fertilizer. The secondary elements are sulphur, calcium, and magnesium (S, Ca, Mg). The quantities required for secondary elements in the soil are much less than the quantities of macroelements but secondary elements are also needed in reasonably large concentrations (Lohry, 2007).

Micronutrients are just as essential as macronutrients but are required by plants in smaller amounts (Penny, 2004). There are eight essential micronutrients (iron, zinc, copper, manganese, boron, chloride, molybdenum and nickel) plus others that are considered to be beneficial (sodium, silicon, and vanadium). Cobalt is also included since it is required for nitrogen fixation by microorganisms (Rhizobia and blue-green algae) (Penny, 2004).

2.1.3 Importance of clay content, SOC, TN, and pH for agriculture

In managing soils for agricultural production, soil texture or particle-size distribution, and the amount of clay present are very important. Soil structure depends very much on clay: soils with little clay have a simple structure, whereas soils with much clay have complex structures and multimodal pore size distributions (Newman, 1984). Clay particles are the smallest particles of the mineral fraction (< 2μm) and are an important component of soil as they have a negative electrical charge. This enables them to hold and exchange nutrients (which also have an electrical charge) (Baxter and Williamson, 2001). Clay particles in the soil provide an exchange site for plant nutrients. Clay soils are normally more fertile than sandy soils. The type and quantity of clay in the soil can affect the amount of nutrients held for plant use and the ease at which these nutrients are released to the plant (Baxter and Williamson, 2001).

Carbon is contained in soils in two different forms, inorganic and organic. Soil organic carbon (SOC) is produced by soil organisms, plant roots, manure branches and litter. On the other hand, soil inorganic carbon comes from carbonic acids and weathering of rocks that precipitate as carbonite minerals (Walcott *et al*., 2009). SOC is one of the main constituents of soil organic matter (SOM) and plays a crucial role in soil chemical and physical properties (Novara *et al*., 2011). From an agricultural point of view, SOC highly impacts bulk density, hydraulic conductivity, water retention, nutrient availability, structural stability, and soil biodiversity (Novara *et al*., 2011).

Nitrogen (N) is by far the most important nutrient in most agricultural systems (Stenberg *et al*., 2010). It is an important component of all protein, so it is integral to the plant structure. Soil total nitrogen (TN) is the total amount of nitrogen in the soil including all organic and inorganic forms (Gates, 2018). TN has a significant impact on plant growth, appearance, yields, and quality, and is also crucial in the formation of organic compounds in new crop cells and tissues (Zhou *et al*., 2019).

Soil pH, an excellent indicator of a soil's suitability for plant growth, is a measure of its acidity or basicity (alkalinity). pH is defined as the negative logarithm (base 10) of the activity of hydronium ions (H^+ or, more precisely, $H_3O^+_{aq}$) in a solution (Slessarev *et al*., 2016). In soils, it is measured in a slurry of soil mixed with water or a salt solution, such as 0.01 M $CaCl_2$. For most crops, a range of 6 to 7.5 is best, with 7 being neutral. Acid soils have a pH below 7 and alkaline soils have a pH above 7. Ultra-acidic soils (pH < 3.5) and strongly alkaline soils (pH > 9) are rare (Slessarev *et al*., 2016).

2.2 Remote sensing of soils

2.2.1 Satellite remote sensing

Traditional soil mapping approaches have mostly relied on ground-based surveys (Mulder *et al*., 2011). Classical field surveys including soil sampling and laboratory analyses are reported to be time-consuming and expensive, especially when mapping is being done at national, regional or global scales (Dobos *et al*., 2001).

Over the past decades, remote sensing (RS) data have been viewed as major secondary data sources for improving digital soil mapping at all scales (Forkuor *et al.*, 2017). Remotely sensed data sources: (1) contain extractable soil information, e.g. spectral reflectance, (2) have large spatial coverage and therefore permit mapping of inaccessible areas, (3) produce consistent and comprehensive data both in time and space, and (4) offer possibilities of supplementing or at least reducing traditional soil sampling in soil surveys. Based on these advantages, numerous studies have explored the use of RS data with varying spatial, temporal and spectral characteristics in digital soil mapping (Mulder *et al*., 2011; Forkuor *et al.*, 2017).

Since the 1970s, satellites (e.g., Landsat, SPOT, IKONOS, QuickBird, RapidEye, GeoEye, WorldView) have been used for RS imagery in agriculture (Mulla, 2013). RS applications in agriculture are based on the interaction of electromagnetic radiation with soil or plant material (Mulla, 2013). Cohen *et al.* (2005) showed that typically, RS involves the measurement of reflected radiation, rather than transmitted or absorbed radiation from agricultural fields. Furthermore, as the spatial and spectral resolution of satellite imagery has improved, the suitability of using reflectance data from these platforms for precision agriculture (PA) applications has increased (Chan *et al*., 2004, Forkuor *et al.*, 2017). However, the most appropriate spatial and spectral resolution for PA applications depends on factors such as crop

management objectives, capacity of farm equipment to vary farm inputs, and farm unit area (Chan *et al*., 2004).

According to Croft *et al*. (2012), there have been considerably fewer studies using airborne and satellite platforms compared to proximal RS. Access to data, cost and training requirements affect the accessibility of airborne/satellite-derived reflectance products. Moreover, there is an increased complexity in deriving reflectance data from a pixel compared to controlled laboratory conditions, due to changes in illumination angles, terrain effects, atmospheric attenuation and low signal-to-noise ratios (Ben-Dor *et al*., 2002; Croft *et al*., 2012). Thus, using satellite RS to study soil properties involves multiple challenges. In this respect, proximal RS is recommended (Ben-Dor *et al*., 2009).

2.2.2 Proximal remote sensing: laboratory and *in situ* spectroscopy to predict soil properties

It has been shown by Pei *et al*. (2018) that the soil property that is most directly correlated to reflectance-based data is soil albedo. Additional soil properties are inferred from reflectance measurements under laboratory conditions such as moisture, SOC, TN, and other chemical properties (Barnes *et al*., 2003; Pinheiro *et al*., 2017; Pei *et al*., 2018).

Many authors used proximal RS (mainly Vis-NIR spectroscopy) to predict soil chemical and physical properties with fair accuracy. The predicted properties are soil organic matter (SOM) (He *et al*., 2007; Wetterlind *et al*., 2008), soil organic carbon (SOC) (Morellos *et al*., 2016), total N (Madari *et al*., 2006; Schirrmann *et al*., 2013), total P (Abdi *et al*., 2012), total K (Schirrmann *et al*., 2013), pH, cation exchange capacity (CEC), exchangeable Ca and Mg, $CaCO_3$ contents (Ji *et al*., 2016), moisture, etc. (Xu *et al*., 2018). In addition to estimating basic soil properties from spectral information, several studies explored the potential of using spectral information to estimate soil-water content at specific pressure head values (e.g. Babaeian *et al*., 2015). In most studies, SOM and SOC are among the well-predicted attributes (Stenberg *et al*., 2010; Zhang *et al*., 2017).

In comparison to laboratory spectroscopy, O'Rourke *et al*. (2016) stated that field spectroscopy is not a mature method for soil analysis, but along with the measured elemental content, spectral data can be extracted and chemometrics can be applied efficiently. Ben-Dor *et al*. (2009) pointed out that most *in situ* RS methods cannot detect the entire soil body that extends from

the surface to the parent material. Furthermore, the sensed upper layer may be affected by many factors such as dust, rust, ploughing, particle size distribution, vegetation coverage and litter.

Regarding the prediction of soil properties, few studies compared the performance of laboratory spectroscopy versus field spectroscopy (Bangelesa, 2017). According to Stevens *et al*. (2010), field spectroscopy is generally less accurate than laboratory spectroscopy due to the surface roughness and moisture contents found *in situ*. This has been shown by Ji *et al*. (2016) who worked in two agricultural fields in Quebec, Canada. They found that Mid-infrared (MIR) soil spectroscopy showed applicability to predict selected properties through various laboratory studies, but the use of MIR instruments in field conditions (*in situ*) was limited. However, all of these results are specific to the characteristics of the study areas.

2.3 Spectral reflectance and soil properties

2.3.1 General trends of soil spectra in Vis-NIR spectroscopy

According to Cieniewski *et al.* (2010), it is mainly the solid phase of the soil, composed of particles of different sizes covered with organic matter and minerals, that decides the reflectance of the soil. These elements, combined with the direction of the incident radiation and the direction in which the reflected radiation is observed by a sensor, are considered to be the main factors influencing the reflectance of a soil sample under laboratory conditions (Cieniewski *et al*., 2010). In the optical range from 300 to 2500 nm, almost all the radiation is either absorbed or reflected and only a small part is transmitted (Dwivedi, 2017).

It has also been shown by Stoner and Baumgardner (1981) that the mineral and organic fractions of soil materials absorb more the shorter wavelengths than the longer ones. Thus, the reflectance spectrum of soils has the form of a rising curve. In the visible range, soils have an increasing reflectance as a function of the wavelength (Xu *et al*., 2018). Then in the mid-infrared zone (1300-2500 nm), the reflectance stabilizes or decreases a little. Between 1450 and 1950 nm, the radiation is strongly absorbed by water molecules, and this is generally shown by two minima at two different wavelengths. It has also been reported that the impact of the presence of water (moisture) is observed in particular in soil spectra collected *in situ* (Stoner and Baumgardner, 1981).

The reflectance of soils across the entire spectral region of the solar illumination (400-2500 nm) carries more information, as was reviewed by Baumgardner *et al*. (1985) and later also by

Ben-Dor *et al*. (2017). In general, a wide range of information can be obtained from reflectance properties related to the nature and chemical composition of the soil material (Ben-Dor *et al*., 2017). This is mainly based on specific absorption of spectrally active groups (known as chromophores), such as Fe, OH in water and minerals, CO_3 in minerals, and many others in organic matter (Viscara-Rossel and Behrens, 2010; Bayer *et al*., 2012). Whereas the visible (Vis; 400-700 nm) information of soils and minerals is characterized by broader spectral features (typical of the electronic process at that range), the near-infrared (NIR; 700-1100 nm) and the short-wave infrared (SWIR; 1100-2500 nm) regions are characterized by intensive and strong absorption features that emerge from a combination mode and overtones of the fundamental processes in the infrared region (> 2.5 μm) (Viscara-Rossel and Behrens, 2010; Qi *et al.*, 2017).

Soil reflectance across the NIR is characterized by well-defined absorption features associated with overtones of O-H and H-O-H stretch vibrations of free water and overtones and combinations of O-H stretch and metal-OH bends in the clay lattice (Nocita *et al.*, 2015). Vibrations of atoms are mostly observed in the thermal and mid-infrared (MIR) wavebands (2500-25000 nm), with low signals located in the Vis-NIR range (Soriano-Disla *et al*., 2014). Al-Abbas *et al*. (1972) and Bayer *et al*. (2012) indicated that soil absorption features are related to biochemical groups such as carboxyl, hydroxyl, and amine functional groups. According to Butkuté and Slepetiene (2004), absorption bands relating to many chemical bonds, such as C-H, N-H, O-H, S-H, C=O, and C=C, are found in the NIR region (780-2500 nm). The NIR spectrum shows overtones and a combination of these groups (Figure 2.1).

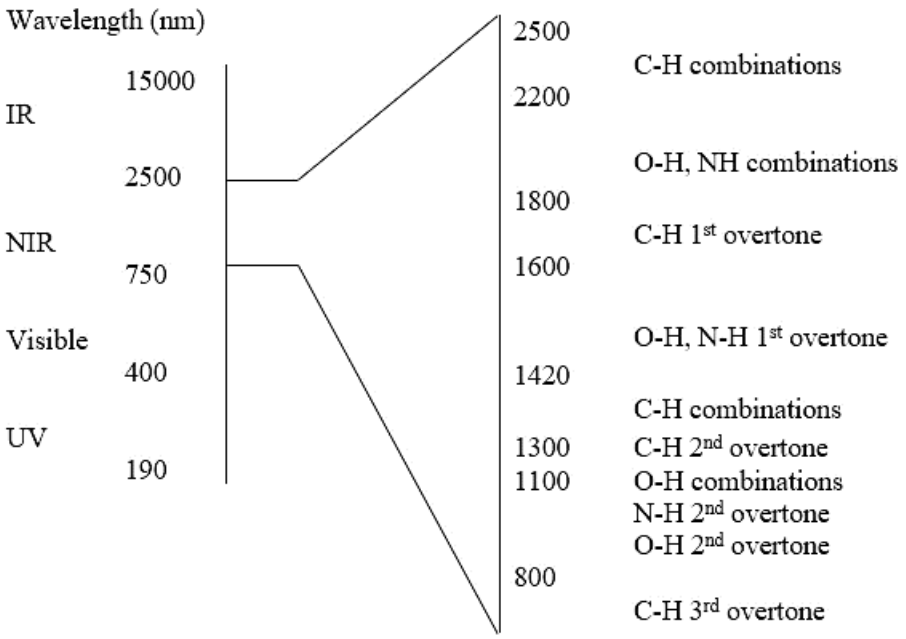


Figure 2.1 Principal types of NIR absorption bands and their locations (Butkuté and Slepetiene, 2004).

2.3.2 Relationship between soil reflectance spectra and soil properties

In the laboratory, the reflectance of soils increases as the size of soil particles decreases (Bowers and Smith, 1972). By testing the reflectance of the materials of the soil texture (from coarse clay to sand), Bowers and Hanks (1965) found that the character of this relationship is exponential. Also, according to Coulson and Reynolds (1971), the decrease in the size of soil aggregates increases the spectral reflectance of the soil. Smaller aggregates have a more spherical shape, but larger ones have an irregular shape with a larger number of spaces and inter-aggregated cracks where incident light is trapped (Cierniewski and Kusnierek, 2010). Hence, when the particle size of the soil decreases from 2 mm to less than 0.06 mm for example, the reflectance of the soil becomes higher (Cierniewski and Kusnierek, 2010).

In addition to the effect of soil particle size, Budak and Gunal (2016) have also shown the impact of soil organic matter content (SOM). They indicated that when the SOM content of the soil is high, the reflectance of the soil is low. For example, for soils that contain less than 1% of SOM, increasing the SOM results in a significant decrease in reflectance. On the other hand, for soils containing 1.5 to 2% of SOM or more, this relationship is less close, because it is weakened by the influence of the variety of mineralogical composition of the soil particles not covered by SOM (Budak and Gunal, 2016). The relationship between SOM content and the

reflectance of the soil in the Vis-NIR, studied by Cierniewski and Kusnierek (2010), indicated that it is the closest in the wavelength range from 600 to 700 nm.

According to Hoffer and Johannsen (1969), the total reflectance is inversely proportional to SOM in the portion of 400-2500 nm. Demattê *et al*. (2003) found, after removing the SOM component using 30% H_2O_2, that the spectral reflectance was higher. So, generally, soil reflectance decreases with organic matter and water content (Nocita *et al*., 2015).

Concerning calcium, under laboratory conditions, the higher the calcium carbonate content of the soil samples, the higher the reflectance. It has been shown by Cierniewski and Kusnierek (2010) that calcium carbonate most strongly absorbs electromagnetic waves in the wavelengths of 2208 nm and 2341 nm (Figure 2.2).

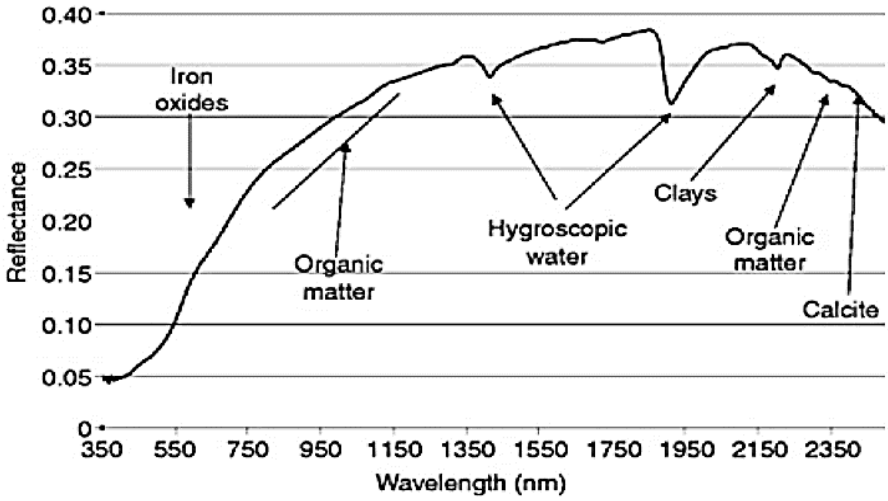


Figure 2.2 The main characteristic spectral signatures and corresponding soil attributes (Zelikman and Carmina, 2013).

Compared to SOM, texture, water and calcium carbonate, spectrally inactive properties like pH, CEC and EC are not directly related to reflectance. However, they can be predicted by the amount of co-variation they have with SOM and the clay mineralogy of the soil (Gates, 2018).

2.4 Use of regression models in soil spectroscopy

Vis-NIR spectroscopy is a suitable method to predict soil chemical and physical properties. However, the often-low concentration of soil constituents and the overlapping absorptions make the spectra broad (Xu *et al.*, 2018). Therefore, the information needs to be mathematically extracted from the spectra so that they may be correlated with soil properties (Viscarra Rossel

and Behrens, 2010). Hence, the analysis of soil diffuse reflectance spectra requires the use of chemometric techniques and multivariate calibration. In these cases, to be useful quantitatively, spectra must be related to a set of known reference samples through a calibration model (Viscarra Rossel and Behrens, 2010).

The set of reference samples used in the models need to be representative of the range of soils in which the models are to be used. PLSR is the most common algorithm used to calibrate Vis-NIR spectra to soil properties (Were *et al.*, 2015). Other approaches have also been used, for example, principal components regression (PCR), stepwise multiple linear regression (SMLR) (Chang *et al*., 2001), artificial neural networks (ANN) (Daniel *et al.*, 2003), multivariate adaptive regression splines (MARS), boosted regression trees, PLSR with bootstrap aggregation (bagging-PLSR), support vector machines (SVM) and penalised spline signal regression (PSSR) (Viscarra Rossel and Behrens, 2010; Ji *et al*., 2016; Gholizadeh *et al*., 2017).

2.4.1 Multicollinearity issue in spectral data

Multi-linear regression was the first method used to predict soil properties with spectral data. However, multicollinearity emerged as a major issue, because it causes uncertainties that decrease the model performance (Martens and Martens, 1986; Bangelesa, 2017). The correlation between independent variables results in large variances in estimating regression coefficients. Thus, serious multicollinearity causes unexplained changes in the dependent variable (Yanli *et al*., 2010).

One way of handling data with a high number of covariates such as Vis-NIR spectra is data reduction. Principal components (PC) and partial least-squares (PLS) are data reduction methods commonly used in chemometrics (Minasny and McBratney, 2008). The principal component regression (PCR) provides a unified way to handle multicollinearity which requires some calculations that are not usually included in standard regression analysis. The principal component analysis follows from the fact that every linear regression model can be restated in terms of a set of orthogonal explanatory variables. These new variables are obtained as linear combinations of the original explanatory variables. They are referred to as the principal components (Alibuhtto and Peiris, 2015).

In PCR, the principal components corresponding to near-zero eigenvalues are removed from the analysis and least squares applied to the remaining components (Alibuhtto and Peiris, 2015). Condit (1970) was the first to use this technique for the analysis of soil spectral

reflectance. However, Adnan *et al*. (2006) indicated that when one is dealing with a great amount of data as in spectroscopy and when data are uncorrelated, PCR performs less well because selecting variables manually becomes a difficult process.

Besides data reduction, other methods of dealing with multicollinearity in Vis-NIR spectroscopy include wavelet analysis, a way of handling large dimensional data by using variable selection techniques. Techniques based on the Bayesian method have been proposed for selecting important variables by discriminating the best predictors between all the variables (Minasny and McBratney, 2008).

2.4.2 Linear vs non-linear models

The commonly used linear regressions such as PCR and PLSR can decompose the original spectral matrix through linear combinations to extract useful components and overcome the problems of collinearity with a high interpretable ability (Wold *et al*., 1984; Qi *et al.*, 2017). PLSR is one of the most popular methods used to predict soil properties since it is handled well with easy-manipulated and accessible software. Another advantage of PLSR is related to the fact that it reduces multi-dimensional data and is not difficult to understand and interpret (Boulesteix and Stimmer, 2007).

In addition to linear algorithms (PCR and PLSR), several nonlinear methods are used in soil spectroscopy. These include machine learning algorithms such as the artificial neural networks (ANNs), least-square support vector machine (LS-SVM), multivariate adaptive regression splines (MARS), random forest regression (RFR) and more, which proved to enhance the prediction performance based on their excellent non-linear learning ability (Xu *et al.*, 2018). Xu *et al*. (2018) indicated that since the relationships between spectral data and soil characteristics are rarely linear, the interest in using non-linear methods has increased.

In the literature, multivariate models are often compared to test their performance in the prediction of different soil properties (Viscarra Rossel *et al*., 2006; Vohland *et al.*, 2011; Were *et al*., 2015). Viscarra Rossel and Behrens (2010) for example, compared linear and non-linear models in predicting various soil properties. They found that SVMR was more powerful than partial least square regression (PLSR), random forest (RF), artificial neural networks (ANN), multiple linear regression (MLR), multiple adaptive regression spline (MARS), and boosted tree (BT).

Xu *et al*. (2018) in their study conducted in Yujiang County in China compared four regression models (PCR, PLSR; backpropagation neural network, BPNN; and SVMR) with the aim of accurately and rapidly predicting soil properties (SOM, total P, TN, and total K). Their results indicated that the SVMR model performed better than PCR, PLSR, and BPNN, for P and N predictions whereas BPNN performed better than all the other models for K.

In many studies, the complexity of SVMR is simplified by combining it with a selection method (e.g., the combination of SVMR and the successive projections algorithm, SPA) (Peng *et al*., 2014). In laboratory conditions, Peng *et al.* (2014) showed that in the presence of outliers and noise, SPA-SVMR performs better than PLSR. In the same way, Li *et al.* (2015), in quantifying SOC, found that the least-squares-SVM (a combination of least-squares and SVM) outperformed PLSR. Other studies have also mentioned the robustness of SVMR and other non-linear methods (e.g., Forkuor *et al.*, 2017; Ludwig *et al*., 2019).

Although studies showed excellent predictions, their results cannot be applied everywhere because contradictory findings for the different soil properties have been reported and can be attributed to a lack of standardised methodology concerning: (i) sample preparation, (ii) spectrum acquisition, (iii) spectrum pre-treatment, (iv) soil texture, (v) geological heterogeneity, (vi) reference method, and (vii) calibration method (Nduwamungu *et al*., 2009). Thus, no method has been universally proven to be better than others.

2.5 Chapter summary

Accurate information on soil physical and chemical properties is required to effectively manage agricultural soils (Silva *et al.*, 2018). To determine soil properties, traditional field and laboratory methods are used the most. However, these methods are laborious and time-consuming hence the usefulness of remote sensing which makes it possible to quickly obtain information on soils at low cost (Forkuor *et al.*, 2017).

RS imagery has been used for several decades to determine soil properties, but the data have limitations in terms of access, cost, training requirements, and spatial and spectral resolution (Croft *et al*., 2012). Vis-NIR spectroscopy is therefore presented as a means to study soil properties with more accuracy since information relating to the nature and composition of the soil can be extracted from the reflectance properties of the soil (Xu *et al.*, 2018).

To predict soil properties from spectral measurements, several regression models have been developed and compared in the literature (e.g. PCR, PLSR, SVMR, RFR, BPNN, etc.), but

none has been universally declared to be more efficient than the others. This is not only due to issues related to the characteristics of the study area but also to issues that are specific to spectral data such as multicollinearity and the non-linear nature of the relationship between spectral measurements and soil properties (Nduwamungu *et al*., 2009). Thus, this study will hope to predict soil properties using reflectance spectroscopy and different regression models and make a contribution by way of results and/or methodology.

CHAPTER THREE

MATERIALS AND METHODS

This chapter outlines the methodological approaches by describing the study area in section 3.1 while the laboratory spectral measurements are described in section 3.2. Furthermore, the laboratory spectra pre-processing and transformation are described in section 3.3, the statistical analysis in section 3.4 and the establishment of calibration and variable selection in section 3.5.

3.1 Study area

The study area of this research was the Eastern Cape Province (Figure 3.1), which covers an area of close to 169 000 km^2 (13.9% of South Africa's land area), making it the second-largest province in South Africa after the Northern Cape (StatsSA, 2003). The province is characterised by high spatial and seasonal rainfall variability, similar to the situation in the entire South Africa. The Eastern Cape exhibits a bimodal rainfall pattern, with a winter rainfall (or all year rainfall) zone in the west, and a summer rainfall zone in the east (Hamann and Tuinder, 2012). The prevailing climate condition varies according to proximity to the ocean as well as west-east direction, becoming progressively wetter towards the east (Hosu *et al*., 2016). The climatic conditions of the Eastern Cape's coastal areas lie between the subtropical conditions prevalent in KwaZulu-Natal and the Mediterranean climate of the Western Cape. The Karoo in the west experiences long hot summers and moderate winters, whereas the high altitudes of the Great Escarpment towards Lesotho and the Free State regularly experience snow in winter (Hamann and Tuinder, 2012).

There is much fertile land in the Eastern Cape, and agriculture remains important (Mnkeni *et al*., 2005). The province has a diversity of soils and climatic conditions permitting a variety of different forms of agriculture. Approximately 30% of the area consists of smallholdings on which farmers mostly practice mixed farming for home consumption purposes (Mnkeni *et al*., 2005). Generally, in the province, nutrient supply is moderate to low in-home gardens and very low to non-existent in field crop production, suggesting that soil fertility depletion may be a

major cause for the decline of productivity in the smallholder cropping system (Bembridge 1984; Andersson and Galt 1998).

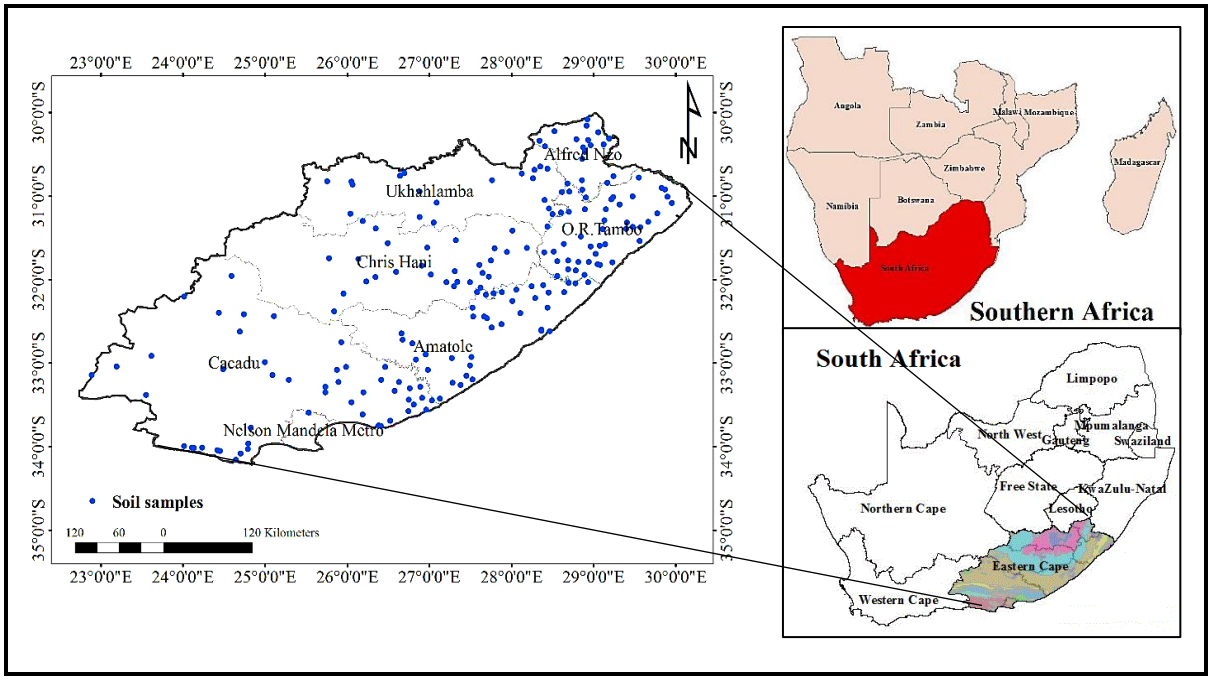


Figure 3.1 Geographical location of the study area showing the spatial distribution of soil samples.

3.1.1 Soil sample selection

The geographical locations of sampling sites were selected from the database of soil profiles obtained from the National Land Type Survey of South Africa of the Agricultural Research Council (ARC). From the Land Type Survey, a supporting database of around 2500 modal soil profiles of different management systems including croplands, as well as a further 10 000 series identification samples (designed to confirm field soil diagnosis) was created. The survey provided quantitative data about a range of soil properties across the greater part of South Africa (Paterson *et al.*, 2015).

The Eastern Cape Province was selected due to its agricultural potential and its geological heterogeneity which should lead to variability in both physical and chemical properties of soils, depending on the dominant parent material. Geology was a key factor to consider in the sample selection process since it is an important determinant for both inherent (e.g. type of clay) and dynamic (e.g. SOM) soil properties.

In terms of differences in the geology of various areas of the Eastern Cape, 200 soil profiles were selected to be representative of the province from the geology map developed by the Council for Geoscience of South Africa (2008) (Figure 3.2). A total of 13 geology types were represented in the area (Figure 4): (1) Adelaide; (2) Bookeveld; (3) Clarens, Elliot, and Molteno; (4) Dwyka; (5) Ecca; (6) Kalahari; (7) Malmesburg, Kango and Gariep; (8) Natal; (9) Suurberg, Drakensberg and Lebombo; (10) Table mountain; (11) Tarkastad; (12) Uitenhage and (13) Witteberg.

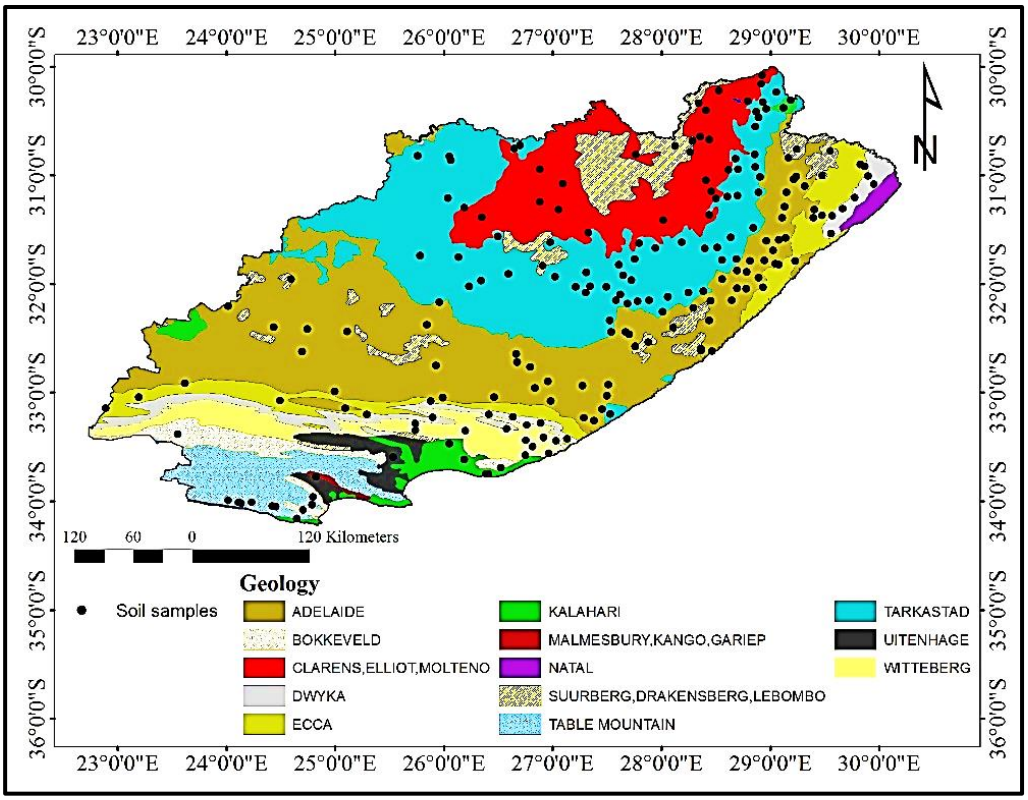


Figure 3.2 Location of sampling points in different geology types.

3.1.2 Determination of soil properties of the selected samples

Digital soil maps (DSMs) were used in this study to obtain reference values of the selected soil properties (clay content, SOC, TN, and pH). Digital soil mapping is the creation of a geographically referenced soil database generated at a given resolution by using field and laboratory observation methods coupled with environmental data (Gourlay *et al.*, 2017). National or regional DSMs offer a potential solution for integrating soil characteristics with agricultural household survey data, particularly when agricultural plots are georeferenced. Improvements in technology have increased both the quantity and quality of geospatial soil data available to the public (for free or for purchase) (Gourlay *et al.*, 2017).

The DSMs data sets used in this study were developed by the African Soil Information Service (AfSIS) project, a collaborative project led by the Tropical Soil Biology and Fertility Institute (TSBF) of the International Center for Tropical Agriculture (CIAT). The AfSIS project used existing soil databases, remote sensing technology, and conventional wet chemistry methods to produce grid-based, national-level DSMs. The project has already shown great success in identifying nutrient deficiencies so that farmers can adjust fertilizing blends accordingly (Gourlay *et al*., 2017).

3.1.3 Characteristics of the downloaded digital soil maps

The DSMs of clay content, SOC, TN, and pH in H_2O at different depths (0-5 cm, 0-15 cm, and 15-30 cm) were obtained from the AfSIS project website (http://africasoils.net/). Since this study focused on the top-soil surface, maps of the three different depths were combined and the average value of each soil property was used for modelling.

Heng *et al*. (2015) showed that none of the AfSIS project's developed DSMs obtained excellent predictions compared with laboratory reference results. However, some maps provided fair accuracy and are suitable for research. In the whole sub-Saharan Africa, soil clay content was mapped with 52.4% of variance, with a root mean square error (RMSE) of 13.7%. SOC, TN, and pH-H_2O were mapped respectively with 61.3, 61.0 and 66.9% of variance explained, with a RMSE of 10.6 $g.kg^{-1}$ for SOC; 0.69 $mg.kg^{-1}$ for TN; and 0.67 for pH-H_2O. They were among the best-predicted properties.

The primary data sets used in this study had a spatial resolution of 250 m (Figure 3.3). Hengl *et al.* (2015) showed that random forests modelling algorithms significantly improved predictions of the AfSIS project's DSMs after using a large compilation of soil profile and soil point observations, in conjunction with a large repository of RS-based images of explanatory environmental variables as input data.

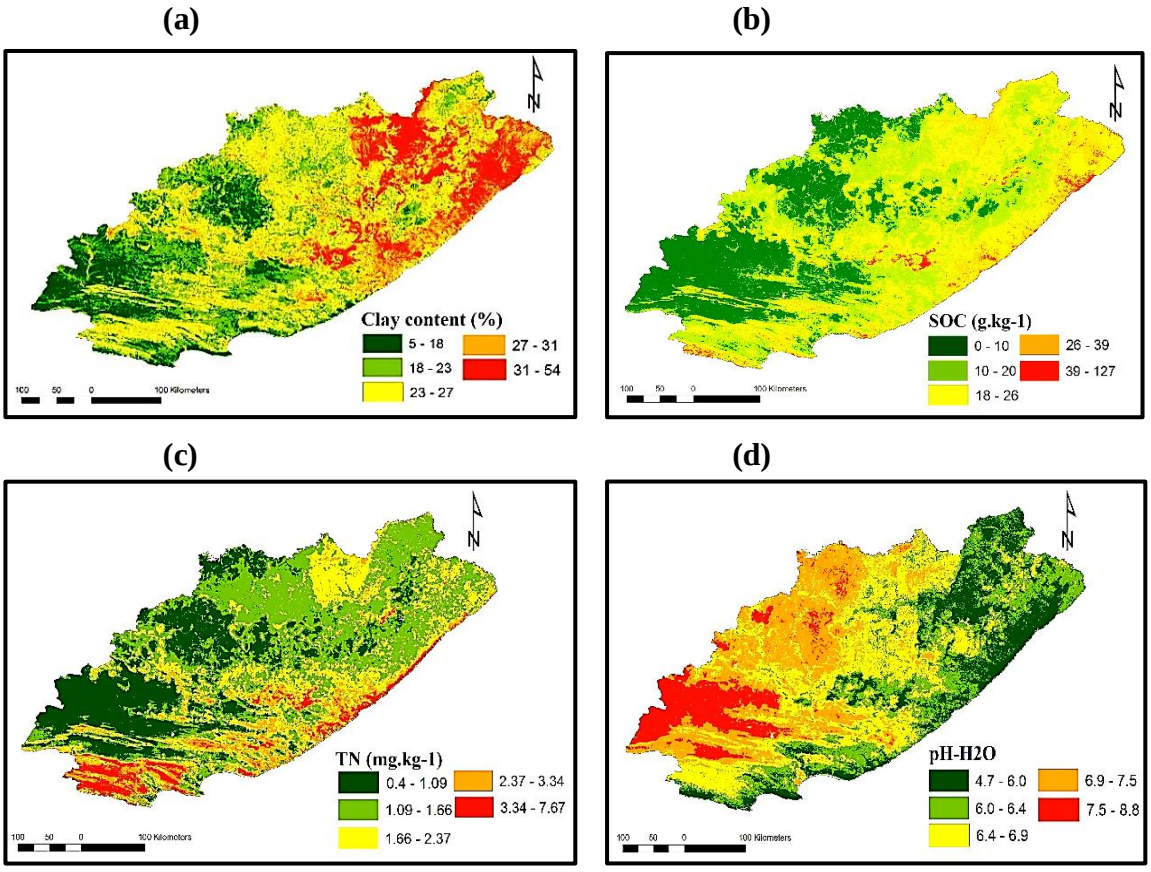


Figure 3.3 Primary digital soil maps used: (a) clay content (%); (b) SOC ($g.kg^{-1}$); (c) TN ($mg.kg^{-1}$); (d) pH.

3.2 Laboratory spectral measurements

Several studies showed that spectral measurements of soil on dried and sieved samples is a standard and routine procedure widely used at the laboratory scale (e.g., Stenberg *et al.*, 2010; Viscarra Rossel and Behrens, 2010; Babaeian *et al.*, 2015; Nocita *et al.*, 2015). After selecting a total of 200 soil profiles to analyse, approximately 5g from the top-soil (0-25 cm) for each profile have been collected in 10 mL tubes from the store for spectral measurements (Figure 3.4a and 3.4b).

Spectra were measured using the ASD FieldSpec Pro FR spectrometer (Analytical Spectral Devices Inc., Boulder CO, USA) with a spectral range of 350 to 2500 nm, and spectral resolution of 1 nm (Table 3.1). Each sample (~5g) was scanned using the contact probe of the spectrometer (Figure 3.4c). Five replicate scans of each sample were conducted and spectra were recorded. The five readings were then averaged to produce a representative spectral

signature for each soil sample. After every 15 outputs, the spectrometer was calibrated with a Spectralon® white tile to maintain consistent and reliable readings of the instrument.

Table 3.1 Waveband range of Analytical Spectral Devices (ASD, 2005; Bangelesa, 2017).

Region names in optical electromagnetic radiation		**Wavelength (nm)**
Ultra Violet (UV)		350-400
Visible (Vis)	Blue light	400-425
	Green-blue	525-605
	Yellow light	605-655
	Red light	655-725
	Far-red	725-750
Near IR	Short wave NIR Infrared (SW-NIR)	750-1100
	Typical 1st NIR region detector (NIR1) or (SWIR1)	1000-1800
	Typical 2nd NIR region detector (NIR2) or (SWIR2)	1800-2500
	Conventional Near Infrared (NIR)	1000-2500

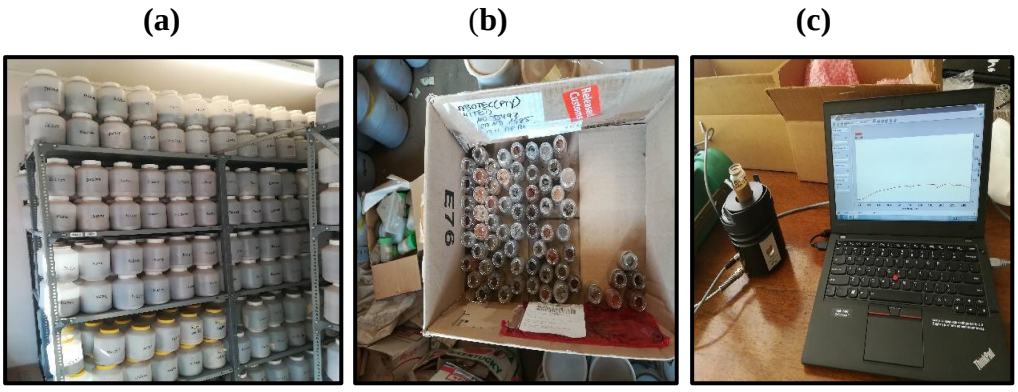


Figure 3.4 Laboratory spectral measurements: (a) selection of soil samples from the store; (b) preparation of soil samples; (c) spectral measurements with the spectrometer.

3.3 Laboratory spectra pre-processing and transformation

Measured spectra are easily influenced by individual differences (the particle size of samples, the intensity of light, the condition of measurement, etc.), baseline variations and substantial noises (Wang *et al.*, 2015). Therefore, pre-treatment was applied to minimize the irrelevant and useless information of the spectra and increase the correlation between spectra and values of soil properties. Pre-processing of spectra was performed to remove artefacts associated with the spectral device or sample geometry (O'Rourke *et al.*, 2016). To correct the radiation of low-intensity that appeared at the edge of spectra, the noisy ends were removed. The moisture absorption features (1350-1460 and 1790-1960 nm) which could affect models were also removed because the impact of soil moisture on reflectance could be greater than the differences in reflectance due to the soil properties (Ben-Dor *et al*., 2009; Qi *et al.*, 2017).

The output resolution of the spectral data was 1 nm and the raw spectra were reduced to between 380 and 2450 nm to eliminate the noise at the edges of each spectrum (Viscarra Rossel *et al*., 2009). To enhance the signal-to-noise ratio, the Savitzky-Golay smoothing and first derivative algorithms were performed (Xu *et al*., 2018), which reduced the baseline variation and enhanced the spectral features (Viscarra Rossel *et al*., 2009). Overall, the first derivative results presented in this study were obtained from the Savitzky-Golay first derivative transformation. The reflectance *R* and the first derivatives *R*';

$$R' = (R_{\lambda} - R_{\lambda-1})/\Delta\lambda, \quad \textbf{(1)}$$

where *R* is the reflectance at wavelength λ and $\Delta\lambda$ is the spectral interval between two closed spectral bands (λ and λ-1) of each wavelength for all samples, were used against the value of the given soil property (clay content, SOC, TN, and pH) to develop regression models. The pre-processing of spectral data was performed in Unscrambler software version 10.5.1 (Camo Analytics Inc., Oslo, Norway).

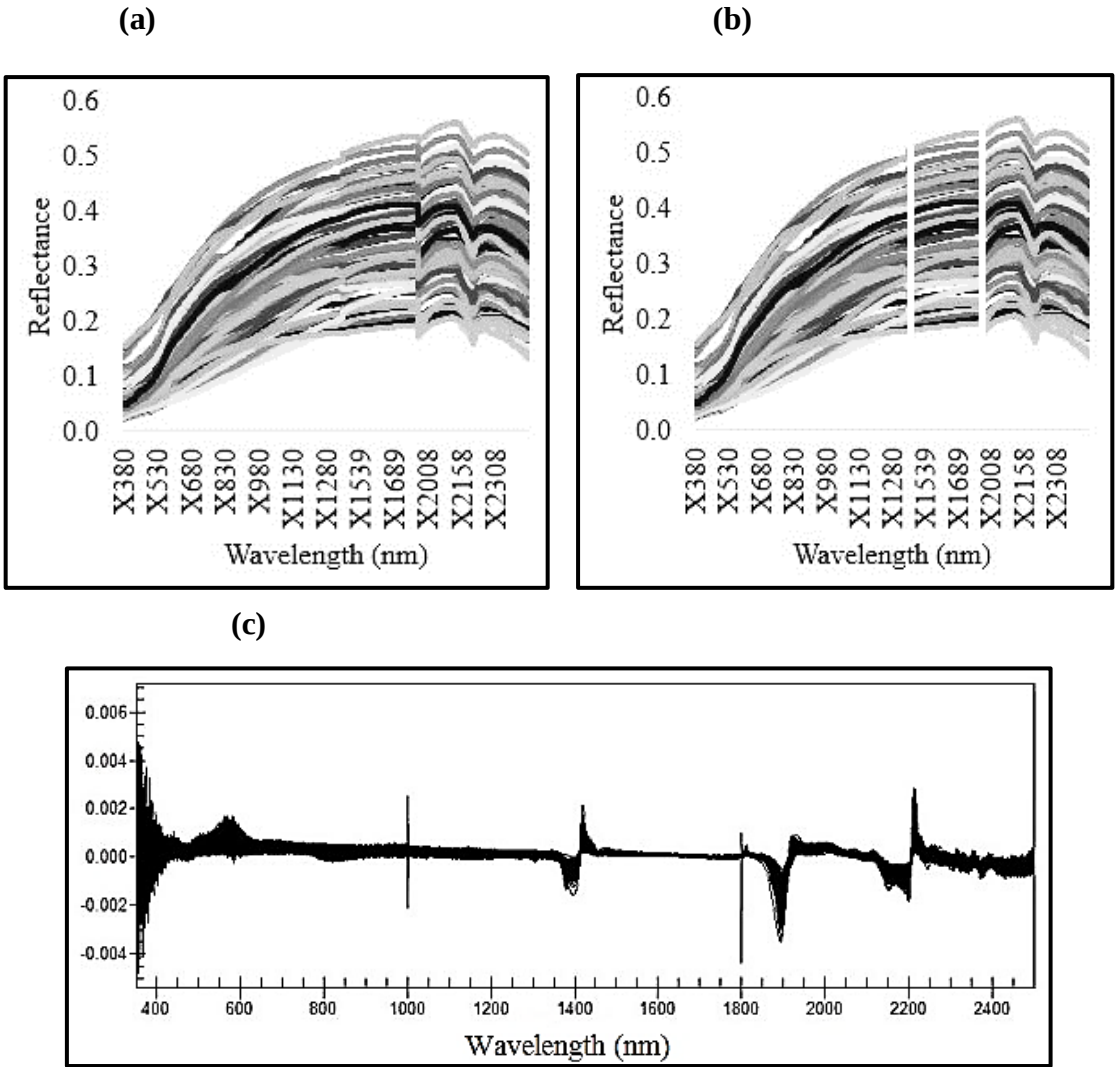


Figure 3.5 Spectral data: (a) collected spectra; (b) spectra with noisy and water absorption regions removed (< 380 nm, 1350-1640 nm, 1790-1960 and > 2450 nm); (c) and first derivative spectra.

3.4 Statistical analysis

The pre-treated spectra and soil properties' values were used to develop calibration models for clay content, SOC, TN, and pH. The test-set validation was used to verify the stability of the prediction models by dividing the dataset into calibration (n = 130) and an independent validation dataset (n = 57). Before sample subsets were created, principal component analysis (PCA) was used to identify spectral outliers (Martens and Naes, 1989). The entire pre-processed spectra were then re-expressed to identify the dimensions that accounted for most of the variation contained in the reflectance spectra (Xu *et al*., 2018). The data points that were outside the 95% confidence ellipse (Hotelling T^2) were strong outliers and were eliminated from the matrix (Morellos *et al*., 2016).

As shown in the score plot (Figure 3.6) of the first two components (representing 90% and 8% of the total variance), thirteen samples (open circles) were removed due to their large deviation in the spectra compared with that of most samples. The remaining spectral dataset (n = 187) was divided into a calibration subset (n = 130) to develop models and an independent validation subset (n = 57) important in the assessment of model performance. Outliers of the four soil properties (clay content, SOC, TN and pH) were also removed after plotting the boxplots of distributions. We assumed that the removal of outliers would improve the performance of regression models. According to Miller and Miller (2010), outliers must only be removed with valid and proven reasons. Hence, a background check from the reference data was done and a decision was taken to remove them to obtain better results.

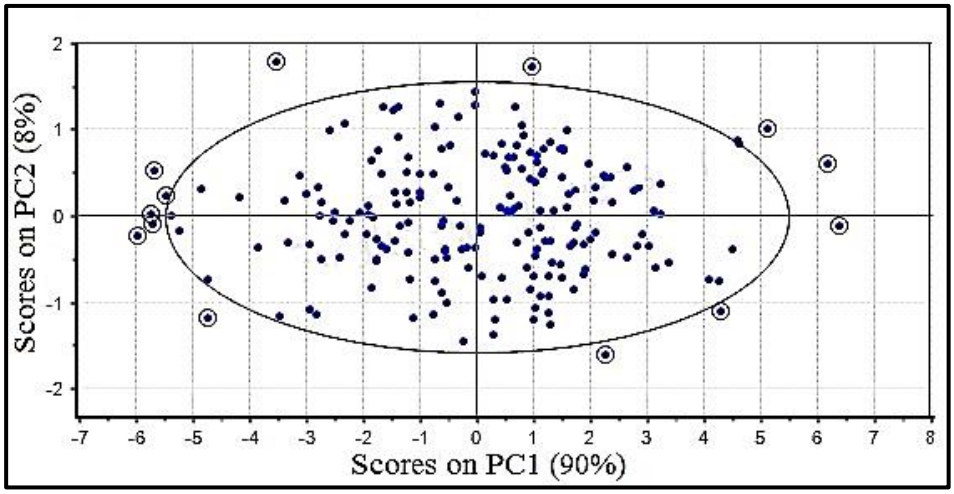


Figure 3.6 Detection of outliers after principal component analysis (PCA) of soil samples (n = 200)

The Krustal-Wallis test was performed to make sure that there is no significant difference between the calibration and the validation datasets. The Kruskal-Wallis test assesses if the difference between at least two samples that do not come from a normal distribution is statistically significant. The null hypothesis states that there is no difference between the datasets. As shown by Bangelesa (2017), the normality of the calibration and validation datasets was checked before the implementation of the Kruskal-Wallis test using the Kolmogorov-Smirnov goodness of fit test, which assumes that the histograms of two different samples should be very similar if those samples are identical (Shorack and Wellner, 2009).

According to Reeves (2010), because soil pH is not a spectrally active soil property, it must be predicted when correlated with other soil properties (soil organic acids, carbonates and soil

minerals). Gates (2018) also stated that similar predictions can be found when there is a strong correlation between pH and SOC for example. In this study, the Pearson's correlation coefficients (*r*) were calculated to assess the correlation among properties.

3.5 Establishment of calibration models and variable selection

Spectral data were calibrated against the selected soil properties obtained from the AfSIS DSMs using PLSR and SVMR. An outline of each of these techniques is provided below, and key references are cited. The regression models were built upon the same dataset.

3.5.1 Partial least squares regression (PLSR) and selection of key wavelengths with the Variable importance projection (VIP) algorithm

Among the available calibration algorithms, PLSR (Wold *et al.*, 1983) is the most popular for spectral calibration and prediction. It is closely related to PCR and yet, it is different. Though both algorithms compress the data before prediction, unlike PCR, PLSR avoids the dilemma of choosing components for regression (Ji *et al.*, 2016).

Briefly, the spectral data matrix X, where X = [x_1, x_2…, x_i] was used as independent variables and each soil property, y as a dependent variable in PLSR. A few linear combinations (called components or factors) T, of the original spectral matrix X were extracted (Ji *et al.*, 2016):

$$T = \omega^T X \quad \textbf{(2)}$$

where ω were the scaled weights and were calculated as the eigenvectors of the matrix $X'yy'X$. Then both X and y were regressed onto T as follows:

$$X = TP^T + E \ \textbf{(3)} \text{ and } y = Tq + f \ \textbf{(4),}$$

where P was the spectral loadings and q was the loadings of soil properties, describing how the variables in T were related to X and y. E and f were residuals and represented noise or irrelevant variability in X and y. Estimated model parameters were then combined into the final prediction model as:

$$\hat{y} = \hat{b}_i x_i + b_0 \ \textbf{(5)}$$

where b_0 was the intercept and $\hat{b}_i$ the regression vectors.

PLSR is particularly useful for predicting a set of dependent variables from a large set of independent variables (Wold *et al*., 1983). To overcome the problem of collinearity between predictors, the PLSR decomposed independent variables and dependent variables by linear combinations to extract latent variables (LVs, or components) and built the regression model based on the LVs instead of the original training variables (Wold *et al*., 1984). To avoid overfitting or underfitting, a leave-one-out cross-validation was used to determine the number of LVs with the smallest mean squared error in calibration (Qi *et al*., 2017).

To assess the influence of each VIS-NIR reflectance band (explanatory variable) on the model results, the variable importance projection (VIP) metric for each band was calculated, as described by Chong and Jun (2005) and implemented by Mevik (2016) (Equations (6) and (7)). For each variable y, the variable importance projection score (VIP) was calculated by:

$$VIP_j = \sqrt{p \sum_{k=1}^{h} (SS(b_k t_k) \left(\frac{\omega_{jk}}{\|\omega_k\|}\right)^2) / \sum_{k=1}^{h} SS(b_k t_k)} \quad \textbf{(6)},$$

$$SS(b_k t_k) = {b_k}^2 {t'}_k t_k \quad \textbf{(7)}$$

where *j* is the index of the explanatory variables, *p* is the number of explanatory variables, *h* is the number of latent variables, SS is the sum of squares, b_k is the y-scores for the *k*-th latent variable, *t* is the loading scores for the *k*-th latent variable, ω_{jk} is the *k*-th value for the *j*-th explanatory variable from the weight matrix, and ω_k is the weights for the *k*-th latent variable (Pinhero *et al*., 2017).

Essentially, the numerator contains the explained sum of squares of y by the PLSR model, and the denominator contains the total sum of squares of y. A spectral band is then considered important in the model if its variable importance projection (VIP) score is considerably large (Pinhero *et al*., 2017). In this study, we used the VIP threshold of 1 put forth by Chong and Jun (2005), Pinhero *et al*. (2017) and Qi *et al*. (2017). PLSR and the VIP algorithm were both implemented in R statistical package version 3.5.3 (R Development Core Team, 2019).

3.5.2 Variable selection by the Random Forest (RF) Boruta algorithm

RF is a machine-learning algorithm that ranks the importance of each predictor included in a model by constructing a multitude of decision trees (Gregorutti *et al*., 2017). Each node of a tree considers a different subset of randomly selected predictors, of which the best predictor is selected and split on. The criterion used to determine the best predictor was decreased in node

impurity, measured with the estimated response variance, which is the default method used for regression trees in the ranger implementation of RF (Wright *et al*., 2017).

Each tree was built using a different random bootstrap sample, which consisted of approximately two-thirds of the total observations and was used as a training set to predict the data in the remaining out-of-bag (OOB) sample, or testing set. Predictions for each variable were aggregated across all trees and the mean square error (MSE) of the OOB estimates was calculated. The MSEOOB and percentage of variance explained were used to evaluate the performance of each RF (Darst *et al*., 2017).

Feature selection is often an important step in applications of machine learning methods and there are good reasons for this. Modern data sets are often described with far too many variables for practical model building. Usually, most of these variables are irrelevant to the classification, and obviously, their relevance is not known in advance (Stuvi and John, 1997).

The Boruta algorithm was implemented in R (R Development Core Team 2019) package "Boruta" and used a wrapper approach built around a random forest (Breiman, 2001) classifier (Boruta is a god of the forest in the Slavic mythology). The algorithm is an extension of the idea introduced by Stoppiglia *et al*. (2003) to determine relevance by comparing the relevance of the real features to that of the random probes (Kursa and Rudnicki, 2010).

3.5.3 Support vector machines regression (SVMR)

The concept of SVMR follows a different approach of supervised learning. Its algorithm is based on the statistical learning theory (Vohland *et al*., 2011). It has been known to strike the right balance between accuracy attained on a given finite amount of training patterns, and an ability to generalize to unseen data. The most valuable properties of SVMs are their ability to handle large input spaces efficiently, to deal with noisy patterns and multi-modal class distributions, and their restriction on only a subset of training data to fit a nonlinear function (Kremer *et al*., 2014; Nalepa and Kawulok, 2018).

SVMR derives a model hyperplane that characterizes the data as correctly as possible while minimizing the distances from the hyperplane to the training data (Vapnik, 2000). An important property of SVMR is that its solution depends only on a subset of training examples called support vectors (Xu *et al*., 2018).

Training SVMR means solving (Smola and Schölkopf, 2004):

$$\text{Minimise } \frac{1}{2} \parallel \omega \parallel^2$$

$$\text{subject to } |y_i - (\omega, x_i) - b| \leq \varepsilon \textbf{ (8)} \text{ and } (\omega, x_i) + b - y_i \leq \varepsilon \textbf{ (9)}$$

where x_i is a training sample with the target value y_i. The inner product plus intercept $(\omega, x_i) + b$ is the prediction for that sample, and ε is a free parameter that serves as a threshold: all predictions have to be within an ε range of the true predictions.

In this study, SVMR was applied to the datasets to compare its performance with PLSR and to inspect the importance of pre-processing and selection of wavenumber ranges. Both PLSR and SVMR were used for the original spectra without data treatment (variant 1), the transformed data (variant 2) and the important wavenumber region selection approach (variant 3). For predictions, SVMR was used as implemented in Unscrambler software version 10.5.1 (Camo Analytics Inc., Oslo, Norway).

3.5.4 Validation and comparison of PLSR and SVMR models

In this study, simple regression was used to compare the soil properties' observed values (from DSMs) and those predicted by the validation dataset. R^2 (coefficient of determination), RMSE (root mean square error), AIC (Aike Information Criterion), and RPD (ratio of performance to deviation) were used to assess model performances. For each property, the coefficient of determination (proportion of total variation, R^2) of the developed model was calculated separately for the calibration and validation datasets (R_c^2 for calibration, and R_p^2 for validation). To compare predictions, the model with the highest R_p^2 was considered the best.

The RMSE of different models were also calculated for the validation (RMSEP) and calibration datasets (RMSEC).

$$RMSEC = \sqrt{\frac{\sum(y_m - y_p)^2}{N}} \textbf{ (10)} \text{ and } RMSEP = \sqrt{\frac{\sum(y_m - y_v)^2}{N}} \textbf{ (11)}$$

where y_m are the observed values of clay content (in %), SOC (in $g.kg^{-1}$), TN (in $mg.kg^{-1}$), and pH obtained from the AfSIS DSMs, y_p are predicted values obtained with the calibration spectral data, y_v are predicted values estimated using the validation set, and N refers to the number of samples. For comparison, the model with the lowest RMSEP was considered the best.

The AIC (Aike Information Criterion) values of the linear models were calculated by:

$$AIC = n\, lnRMSE + 2p \quad \textbf{(12)}$$

where *n* is the number of samples and *p* the number of variables used in the model. The predictive model with the smallest AIC was considered the best.

The prediction accuracy of each regression model was validated using the ratio of performance to deviation (RPD) of the validation set that was calculated as:

$$RPD = SD/RMSEP = \sqrt{m \sum_{i=1}^{m}(y_i - \bar{y})^2 / \sum_{i=1}^{m}(f(X_i) - y_i)^2}/(m-1) \quad \textbf{(13)}$$

where *m* is the number of testing samples in the validation set, y*i* is the observed value of sample *i*, *f* (X_i) is the predicted value of sample *i*, and $\bar{y}$ is the average value of y. *SD* refers to the standard deviation of the property in the calibration dataset and *RMSEP* is the root mean square error of the property in the validation dataset.

In this study, RPD was taken as one of the most important indicators to compare predictive models because it computes the accuracy by integrating the training and testing datasets. The overall assessment of model performances was qualitatively defined by combining threshold values of RPD and R^2 values (Table 3.2). RPD was given more weight than R^2 (Viscarra Rossel *et al.,* 2006). The selection of the best models for each soil property within each calibration sampling method was first determined by RPD then by R^2 value if RPD values were the same.

Table 3.2 Qualitative model performance based on RPD and R^2 (Gates, 2018).

R^2	RPD < 1.4	1.4 < RPD < 2	RPD > 2
< 0.7	Very poor	Poor	-
0.7-0.8	Poor	Fair	Good
0.8-0.9	-	Good	Very good
> 0.9	-	-	Excellent

According to Viscarra Rossel *et al*. (2006), in soil spectroscopy, the use of very poor models (RPD < 1.0) and poor models (1.0 < RPD < 1.4) is not recommended. Fair models (1.4 < RPD < 1.8) may be used for assessment and correlation. Good (1.8 < RPD < 2.0) and very good models (2.0 < RPD < 2.5) can possibly be used for quantitative predictions, and excellent quantitative models (RPD > 2.5) can be used to replace laboratory analysis.

CHAPTER FOUR

RESULTS

This chapter presents the different findings of the research based on the aim and objectives. It is split into the descriptive statistics of the soil properties (section 4.1), the spectral characteristics of soil samples (section 4.2) and the determination of key wavelengths (section 4.3). In addition, the development of PLSR and SVMR models is presented in section 4.4, the prediction accuracy of the multivariate methods in section 4.5, and the comparison between PLSR and SVMR models in section 4.6.

4.1 Descriptive statistics of the soil properties

After spectra pre-processing and removing outliers, 187 soil samples remained. From the 187 samples analysed, 70% (130) were randomly assigned to the calibration dataset and 30% (57) to the validation dataset. Table 4.1 shows the summary statistics for the entire, calibration and validation datasets for the four soil properties (i.e. clay content, SOC, TN, and pH).

The variation of SOC values was larger (values ranging from 3.0 to 36.0 $g.kg^{-1}$) compared to clay content (values ranging from 13.0 to 35.0%), TN (values between 0.56 and 3.21 $mg.kg^{-1}$) and pH (values between 5.2 and 8.0). All the variables presented positive skewness except the clay content, which indicates that distributions were concentrated at low values with relatively few high values. The Kolmogorov-Smirnov test for normality revealed that the clay content, SOC and TN whole datasets do not differ significantly from those which are normally distributed at 5% significance level with *p*-values of 0.91, 0.30 and 0.29 respectively for clay content, SOC and TN. On the other hand, the distribution of pH was skewed ($p = 0.01$).

The Krustal-Wallis test for independent measures indicated that there is no significant difference between the three datasets (whole, validation and calibration datasets) for each property at 5% significance level. The test obtained a *p*-value of 0.63 among the three datasets of clay content, 0.64 among the three datasets of SOC, 0.26 among the three datasets of TN and 0.24 among the three datasets of pH. Hence, for each property, both calibration and validation datasets are statistically representative of the total dataset. Also, descriptive statistics

(e.g. mean and standard deviation) of the calibration and validation datasets were similar. This indicates that calibration models would be well trained to predict soil properties in the validation dataset.

Table 4.1 Descriptive statistics of clay content, SOC, TN, and pH within three different datasets.

Property	**Range**	**Mean**	**Median**	**SD***	**Variance**	**Skewness**	**Kurtosis**
Whole dataset (n = 187)							
Clay (%)	13.0-35.0	24.94	25.0	4.17	17.4	-0.05	0.21
SOC ($g.kg^{-1}$)	3.0-36.0	17.35	17.0	5.75	33.08	0.22	0.43
TN ($mg.kg^{-1}$)	0.56-3.21	1.51	1.47	0.43	0.18	0.61	0.98
pH	5.2-8.0	6.32	6.30	0.44	0.19	1.04	1.62
Calibration dataset (n = 130)							
Clay (%)	14.0-34.0	25.11	25.0	4.0	16.07	-0.07	-0.12
SOC ($g.kg^{-1}$)	3.0-36.0	17.23	17.0	5.97	35.68	0.23	0.27
TN ($mg.kg^{-1}$)	0.56-3.21	1.49	1.42	0.43	0.18	0.72	1.51
pH	5.2-7.70	6.35	6.30	0.46	0.21	0.82	0.75
Validation dataset (n = 57)							
Clay (%)	13.0-35.0	24.56	24.0	4.54	20.6	0.01	0.75
SOC ($g.kg^{-1}$)	4.0-33.0	17.61	17.0	5.25	27.59	0.21	1.09
TN ($mg.kg^{-1}$)	0.75-2.90	1.56	1.51	0.44	0.19	0.37	0.19
pH	5.6-8.0	6.26	6.20	0.39	0.15	1.71	5.85

* Standard deviation

The Pearson correlation coefficients (Table 4.2) indicate that all the soil properties were significantly correlated at 1% significance level. SOC was strongly correlated to TN and pH (r = 0.735 for SOC-TN and r = -0.724 for SOC-pH). Clay content was moderately correlated to SOC but weakly correlated to pH (r = 0.518 for clay-SOC and r = -0.448 for clay-pH). TN was moderately correlated to pH (r = -0.632) but weakly correlated to clay (r = 0.254). The negative correlation between pH and other properties is also shown in the primary maps (Figure 3.3), where visual comparison indicates that areas with high clay content, SOC, and TN have low pH values, and those with low clay content, SOC, and TN have high pH values.

Table 4.2 Pearson's correlation coefficients (*r*) for clay content, SOC, TN, and pH values (n=187).

Property	Clay	SOC	TN	pH
Clay (%)	1.00			
SOC ($g.kg^{-1}$)	0.518**	1.00		
TN ($mg.kg^{-1}$)	0.254**	0.735**	1.00	
pH	-0.448**	-0.724**	-0.632**	1.00

****** Correlation significant at the 0.01 level (two-tailed).

4.2 Spectral characteristics of the soil samples

All the soil samples had similar reflectance shapes. In the visible region (400-700 nm), the collected spectra presented a higher increasing slope compared to other regions. In this range, the first derivative of the reflectance also showed an increasing trend with a peak at 560 nm. In the NIR region (700-2450 nm), all the samples showed important water absorption features at approximately 1400 nm and 1800 nm. The spectra also showed a remarkable absorption peak at approximately 2200 nm. The first derivative of the reflectance showed a reduction in the baseline shift, and peaks were also observed at 1400, 1800 and 2200 nm. It is shown in Figure 4.1b that the first derivative transformation enhanced the spectral features.

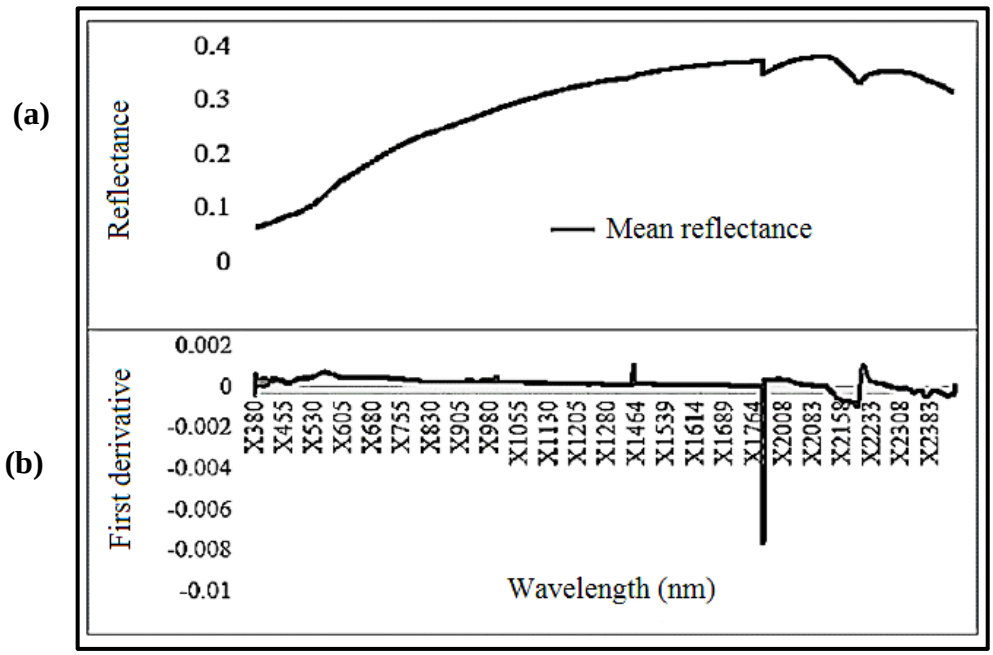
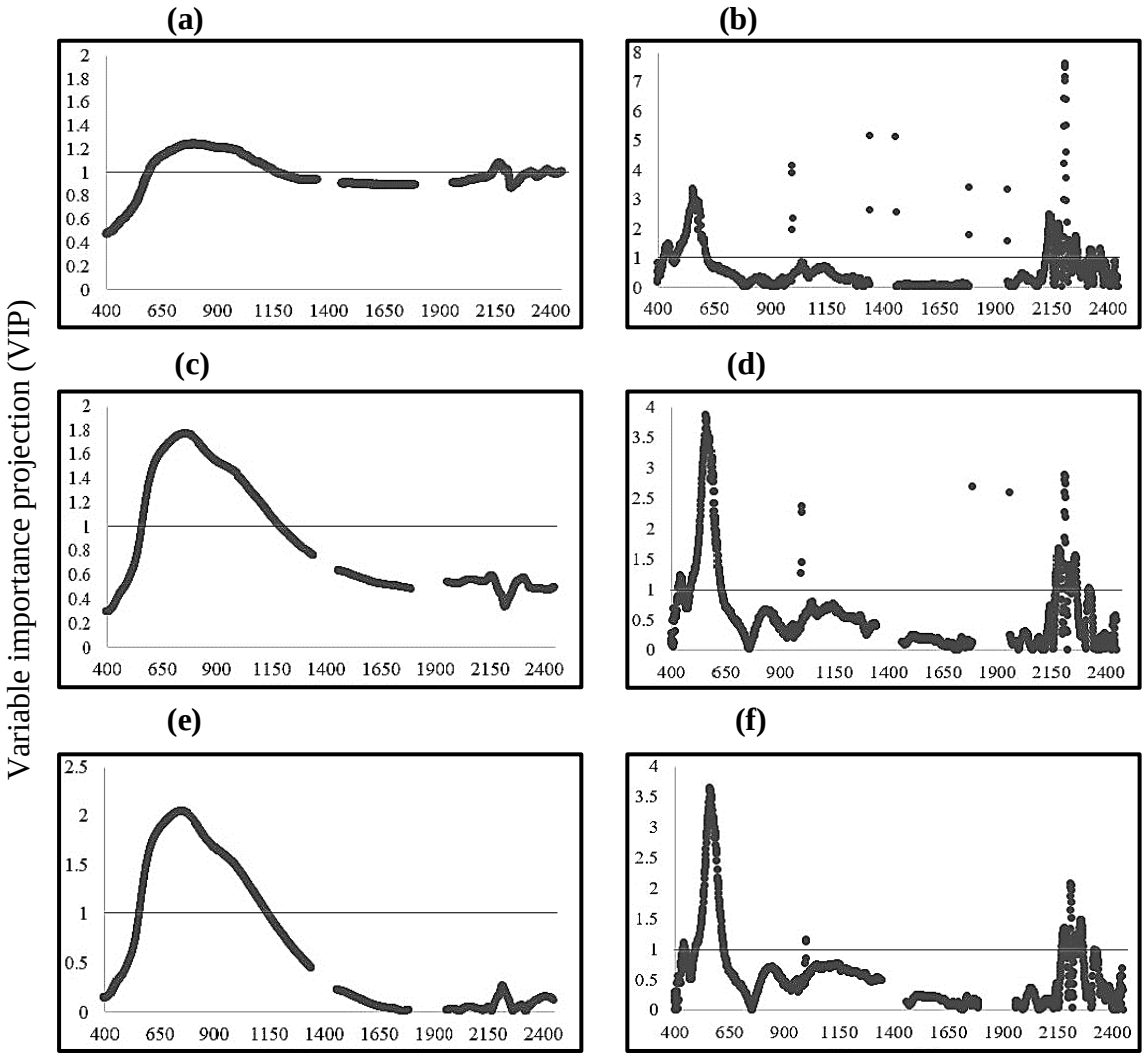


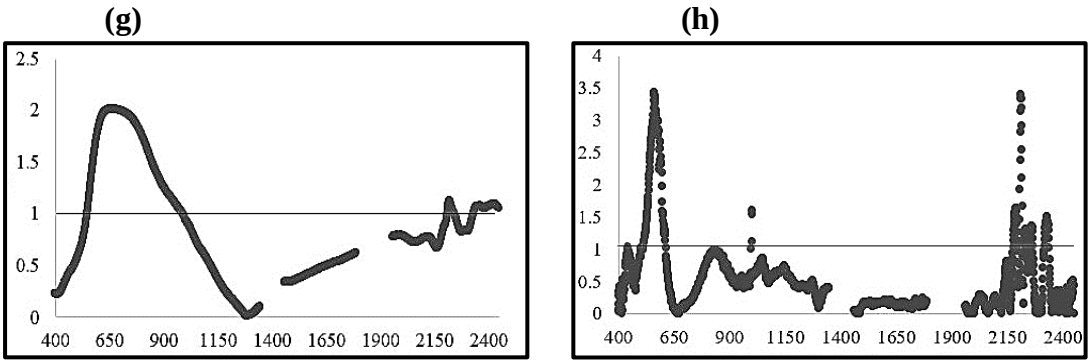
Figure 4.1 Mean laboratory soil reflectance spectra: (a) raw spectra; (b) first derivative spectra.

4.3 Determination of key wavelengths

4.3.1 PLSR-variable importance projection (VIP)

To increase the interpretability and the generalisation of the regression models, VIP algorithms were computed with PLSR for the soil spectra. Values with a peak maximum above 1 were considered to be appropriate wavelengths to predict the soil properties (Figure 4.2). For the raw data, the algorithm broadly selected key wavelengths from approximately 600 to 1150 nm for the four properties, with additional peaks at ~2200 nm and 2300 to 2450 nm wavebands for clay content and pH. The first derivative spectra distinctively selected most key wavelengths with peaks around 600 nm, 1000 nm and 2200 nm for the four properties (i.e. clay content, SOC, TN, and pH).





Wavelength (nm)

Figure 4.2 Identification of key wavelengths with VIP algorithms: (a) clay raw spectra; (b) clay first derivative spectra; (c) SOC raw spectra; (d) SOC first derivative spectra; (e) TN raw spectra; (f) TN first derivative spectra; (g) pH raw spectra; (h) pH first derivative spectra.

4.3.2 Feature selection with the Boruta algorithm

The Boruta algorithm implemented under the random forest machine learning method classified the predictor variables (wavelengths) in three different groups (important features, unimportant features, and tentative features). The algorithm selected 48, 22, 33, and 27 important wavelengths consecutively for clay content, SOC, TN, and pH on raw spectral data. On the first derivative spectral data, 81, 73, 67 and 76 important wavelengths were identified consecutively for clay content, SOC, TN, and pH (Table 4.3).

The lowest number of key wavelengths (22) was obtained with SOC raw spectral data (RMSE = 3.80 $g.kg^{-1}$) during the RF-Boruta classifier run. For the four soil properties, the first derivative data obtained lower RMSE compared to the raw data. Important wavelengths were mostly selected in the Vis range (400-700 nm) for clay raw spectra, SOC raw spectra, TN raw spectra, pH raw spectra, and the SOC first derivative spectra. On the other hand, the first derivative spectra of clay content, TN and pH selected important wavelengths mostly in the NIR range (700-2450 nm).

Table 4.3 Number of key wavelengths selected by the RF-Boruta algorithm, RMSE on the validation dataset and location of important features on the raw and first derivative spectral data.

Property	Number of key wavelengths	RMSE*	Important wavelengths (nm)
Raw spectral data			
Clay	48	3.23	~560 to 690, 946, 2412, 2444, 2450
SOC	22	3.80	~450 to 650, 962, 2215
TN	33	0.30	~600 to 760, ~1980, 2230
pH	27	0.32	~500 to 700, ~2230-2300
First derivative data			
Clay	81	2.90	~450 to 475, 546 to 600, ~712 to 751, 930, 1669, 2058, 2139 to 2148, 2207 to 2371
SOC	73	3.21	~450 to 650, ~1000, ~1450, ~2200
TN	67	0.19	~550 to 600, ~1200-1460, ~2200-2300
pH	76	0.28	~550, ~1050 to 1100, ~1350, ~1600, ~2200 to 2300

*% for clay content, $g.kg^{-1}$ for SOC and $mg.kg^{-1}$ for TN.

4.3.3 Position of wavelengths and interpretation

The performance of PLSR-variable importance projection (VIP) compared to RF-Boruta algorithm in selecting key wavelengths is presented in Table 4.4. For interpretation purposes, the functional groups and vibration modes of wavelengths as suggested by Stuart (2004) and used by Bangelesa (2017) are also presented (Table 4.5). The Boruta algorithm selected most of the key wavelengths in all datasets in the range of 400-700 nm, and 2200-2450 nm for the four properties. The VIP algorithm implemented on the first derivative spectral data also selected the most key wavelengths in the same range. Only the VIP algorithm implemented on the raw datasets for SOC and pH did not select key wavelengths above 1150 nm.

Table 4.4 Comparison of the PLSR-VIP and RF-Boruta algorithms in selecting key wavelengths in the visible and near-infrared ranges.

	PLSR-VIP algorithm								**RF-Boruta algorithm**							
	Clay		**SOC**		**TN**		**pH**		**Clay**		**SOC**		**TN**		**pH**	
Wave length (nm)	**R**	**D**	**R**	**D**	**R**	**D**	**R**	**D**	**R**	**D**	**R**	**D**	**R**	**D**	**R**	**D**
2200-2450	++	+++	-	+++	-	+++	++	+++	+	+++	+	+	+	+++	++	+++
2000-2200	-	+	-	+	-	-	-	-	-	++	-	-	-	+	-	-
1790-1960	-	+	-	+	-	-	-	-	-	-	-	-	+	-	-	-
1650-1780	-	-	-	+	-	-	-	-	-	+	-	-	-	-	-	+
1400-1500	-	+	-	-	-	-	-	-	-	-	-	+	-	-	-	-
1300-1420	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	+
1350-1460	-	-	-	-	-	-	-	-	-	-	-	+	-	+	-	+
1100-1225	+	-	+	-	-	-	-	-	-	-	-	-	-	+	-	-
950-1100	+	+	+	+	+	+	-	+	-	-	+	+	-	+	-	+
850-950	+	-	+	-	+	-	+	-	+	+	-	-	-	-	-	-
775-850	+	-	+	-	+	-	+	-	-	-	-	-	-	-	-	-
400-700	+	++	+	+++	+	+++	+	+++	+++	++	+++	+++	+++	++	+++	+

R = raw data (non-transformed), D = first-order derivative data (transformed). The relative importance of wavelength regions is indicated by "+", "++" and "+++", where "-" indicates and the absence of key wavelengths, "+" region importance is < 10%, "++" region importance is between 20 and 40%, and "+++" indicates that the wavelength region importance is > 40%.

Table 4.5 Functional groups and vibration modes of wavelengths considered for interpretation (Stuart, 2004)

Wavelength (nm)	2200-2450	2000-2200	1790-1960	1650-1780	1400-1500	
Possible assignment	Comb C-H str	Comb N-H str, comb O-str	Water	1st overt C-H str	1st overt N-H str and O-H str	
Wavelength (nm)	1300-1420	1350-1460	1100-1225	950-1100	850-950	400-700
Possible assignment	Comb C-H str	Water	2nd overt C-H str	2nd overt N-H stretch and O-H str	3rd overt C-H str	Mineral (Fe oxides)

∗ str = stretching vibration mode; comb = combination vibration mode; overt = overtone.

4.4 Development of PLSR and SVMR models

As suggested by Bangelesa (2017), eight models were developed for each of the four soil properties using different spectral pre-processed data (raw, Savitzky-Golay derivative, key wavelengths, and the combination of Savitzky-Golay and key wavelengths), four with PLSR and four with SVMR.

- Models developed with PLSR are (1) PLSR-None: PLSR model on raw (non-transformed) data with all wavelengths; (2) FD-PLSR: PLSR model on the first derivative (transformed) data with all wavelengths; (3) K-PLSR: PLSR model on raw data with key wavelengths only; (4) FD-K-PLSR: PLSR model on first derivative data with key wavelengths only.

- Models developed with SVMR are (1) SVMR-None: SVMR model on raw (non-transformed) data with all wavelengths; (2) FD-SVMR: SVMR model on the first derivative (transformed) data with all wavelengths; (3) K-SVMR: SVMR model on raw data with key wavelengths only; (4) FD-K-SVMR: SVMR model on first derivative data with key wavelengths only.

The best prediction models were selected as shown in Table 3.2. The RPD and R_p^2 combined were used for model comparison (with RPD considered the most important criteria).

4.5 Prediction accuracy of the multivariate methods

4.5.1 PLSR model performance

4.5.1.1 PLSR prediction of clay content

Figure 4.3 shows PLSR models developed for clay content. All the models provided poor predictions (1.4 < RPD < 2 and R_p^2 < 0.7). However, the two key wavelengths models, K-PLSR (RPD = 1.50, R_p^2 = 0.64, RMSEP = 2.66%) and FD-K-PLSR (RPD = 1.46, R_p^2 = 0.63, RMSEP = 2.73%) obtained the best predictions. The worst predictions were obtained by the PLSR model on raw data and the first derivative data (RPD < 1.4). RPD = 1.38, R_p^2 = 0.58, RMSEP = 2.89% for FD-PLSR; and RPD = 1.38, R_p^2 = 0.58, RMSEP = 2.90% for PLSR-None.

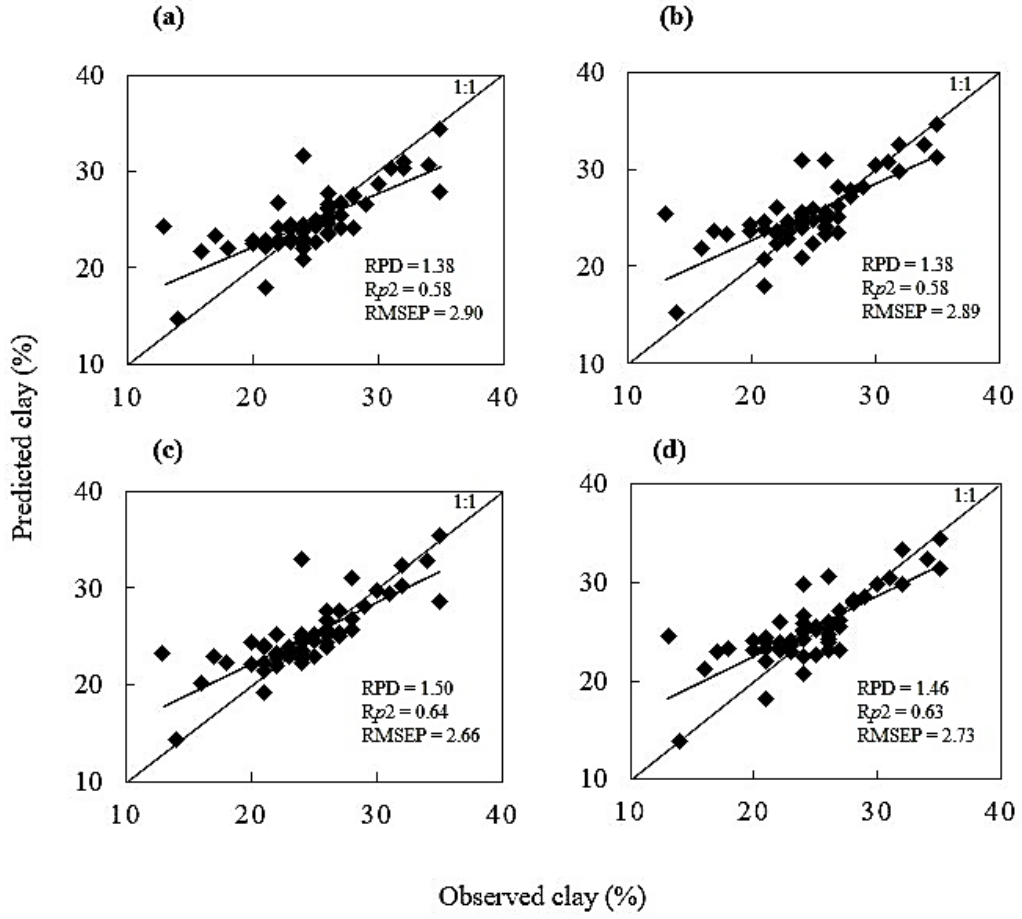


Figure 4.3 Performance of PLSR in predicting clay content: (a) clay PLSR raw model with all wavelengths; (b) clay first derivative PLSR model; (c) clay PLSR model with key wavelengths; (d) clay derivative PLSR model with key wavelengths.

4.5.1.2 PLSR prediction of SOC

Figure 4.4 shows the performance of PLSR for SOC predictions. All the PLSR models developed for SOC indicated poor predictions (1.4 < RPD < 2 and R_p^2 < 0.7). However, the PLSR-None (RPD = 1.93, R_p^2 = 0.65, RMSEP = 3.09 g.kg^{-1}) and FD-PLSR (RPD = 1.91, R_p^2 = 0.63, RMSEP = 3.13 g.kg^{-1}) outperformed the K-PLSR model (RPD = 1.82, R_p^2 = 0.60, RMSEP = 3.27 g.kg^{-1}) and the FD-K-PLSR model (RPD = 1.86, R_p^2 = 0.62, RMSEP = 3.20 g.kg^{-1}).

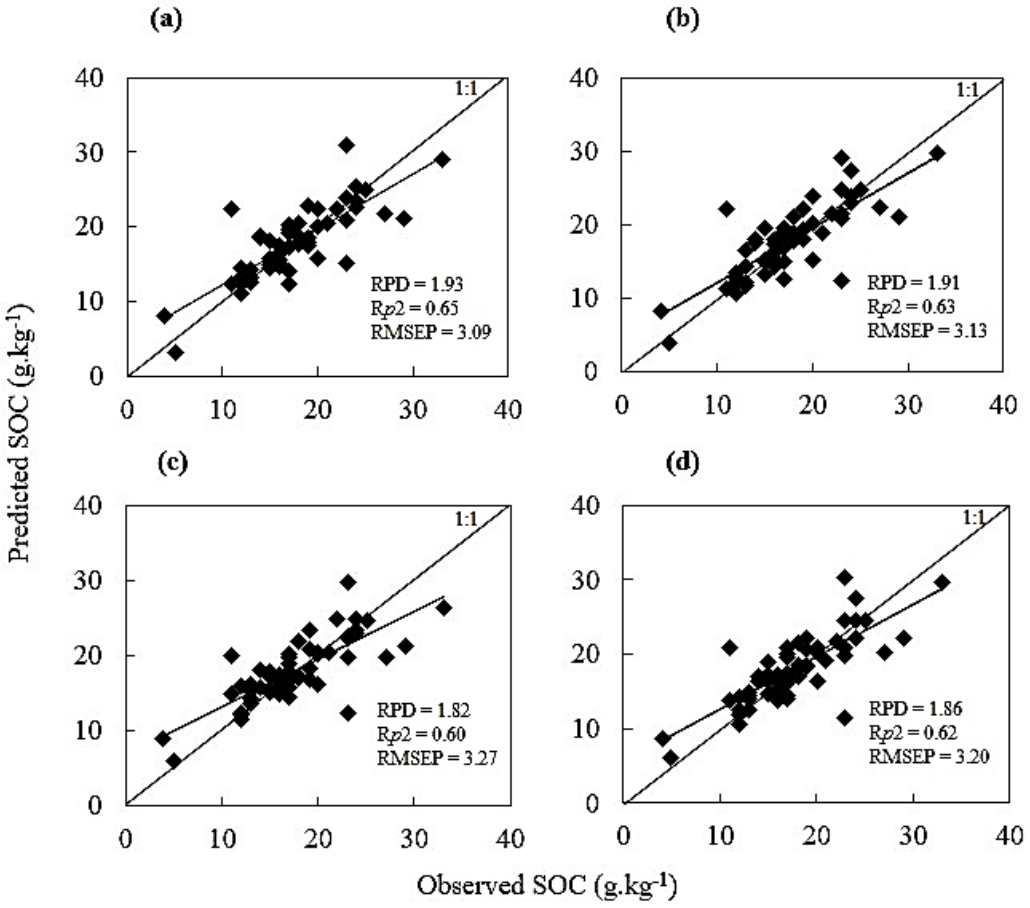


Figure 4.4 Performance of PLSR in predicting SOC: (a) SOC PLSR raw model with all wavelengths; (b) SOC first derivative PLSR model; (c) SOC PLSR model with key wavelengths; (d) SOC derivative PLSR model with key wavelengths.

4.5.1.3 PLSR prediction of TN

Figure 4.5 shows the PLSR models developed for TN predictions. The PLSR-None model (RPD = 2.15, R_p^2 = 0.77, RMSEP = 0.20 mg.kg^{-1}) showed good predictions (RPD > 2 and 0.7 < R_p^2 < 0.8). The FD-PLSR and FD-K-PLSR models indicated fair predictions (1.4 < RPD < 2

and 0.7 < R_p^2 < 0.8). The first derivative model K-PLSR (RPD = 1.72, R_p^2 = 0.65, RMSEP = 0.25 mg.kg^{-1}) indicated poor predictions (1.4 < RPD < 2 and R_p^2 < 0.7).

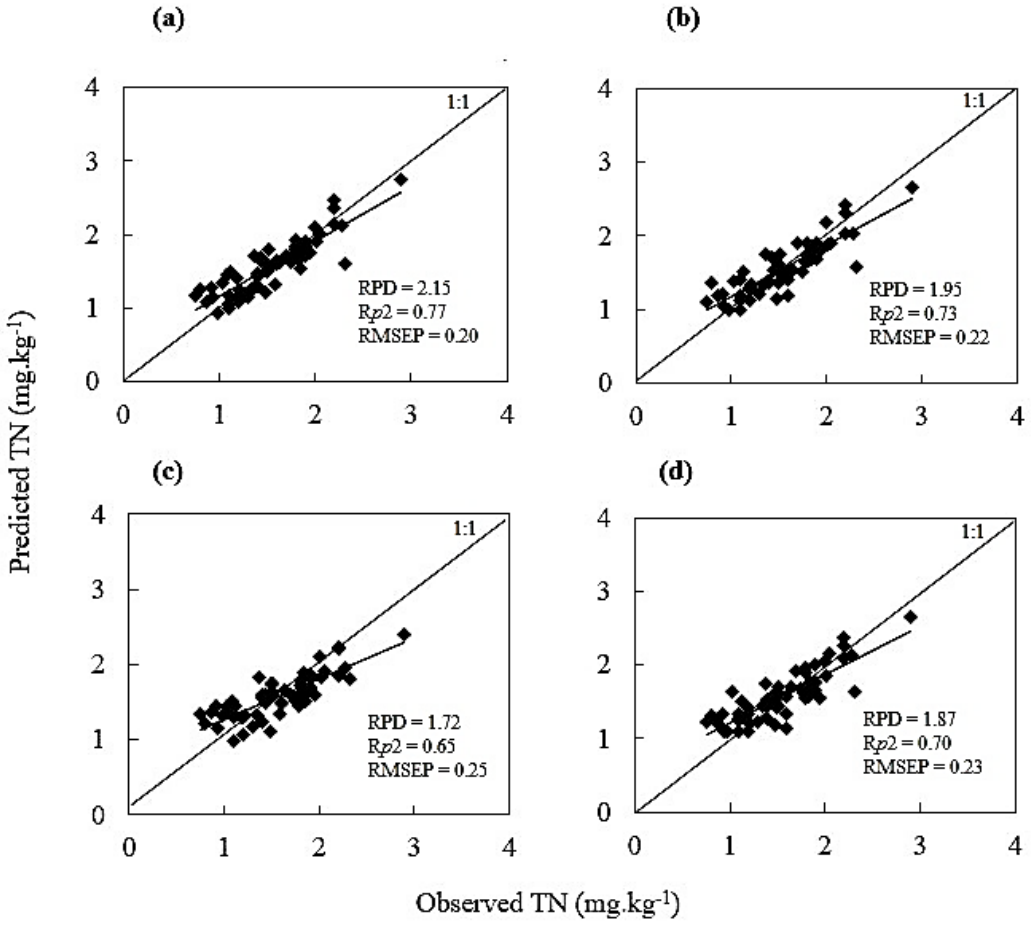


Figure 4.5 Performance of PLSR in predicting TN: (a) TN PLSR raw model with all wavelengths; (b) TN first derivative PLSR model; (c) TN PLSR model with key wavelengths; (d) TN derivative PLSR model with key wavelengths.

4.5.1.4 PLSR prediction of pH

The performance of PLSR in predicting pH is presented in Figure 4.6. The best predictive model was obtained by the PLSR model on raw spectra (RPD = 2.55, R_p^2 = 0.77, RMSEP = 0.18). The other predictive models were good (RPD > 2), with FD-PLSR, K-PLSR, and FD-K-PLSR having consecutively RPD values of 2.30, 2.19 and 2.42, and R_p^2 values of 0.71, 0.68 and 0.75.

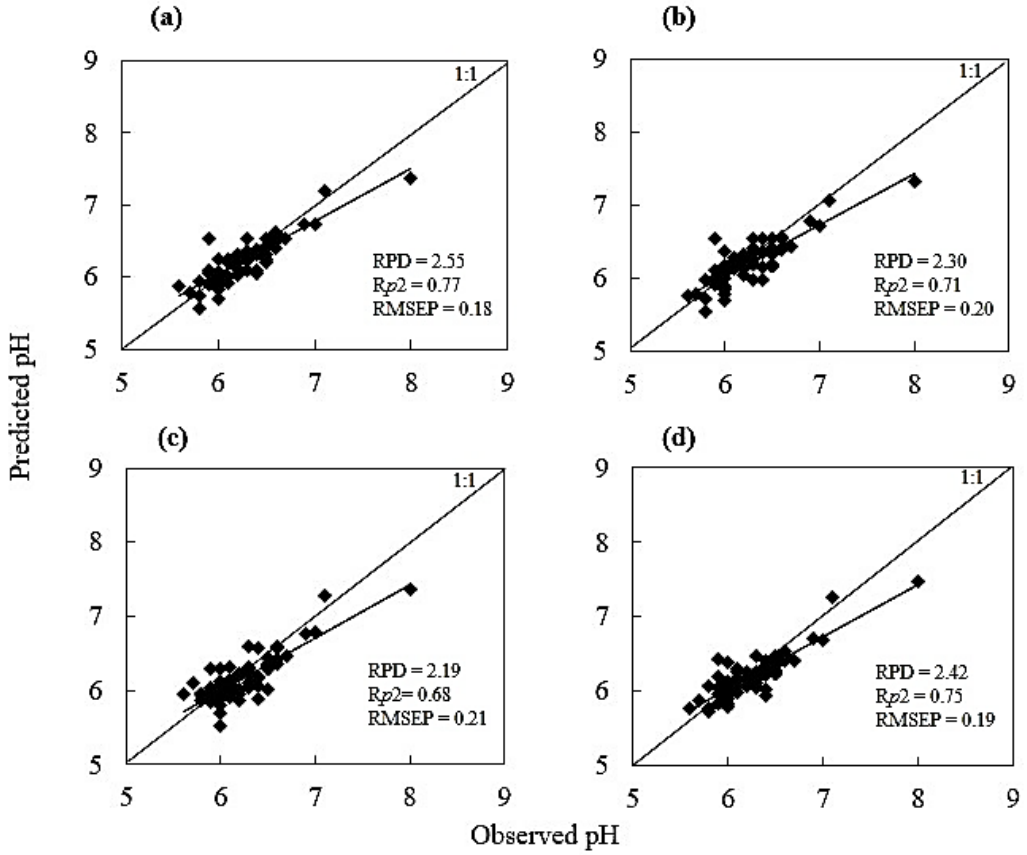


Figure 4.6 Performance of PLSR in predicting pH: (a) pH PLSR raw model with all wavelengths; (b) pH first derivative PLSR model; (c) pH PLSR model with key wavelengths; (d) pH derivative PLSR model with key wavelengths.

4.5.2 SVMR model performance

4.5.2.1 SVMR prediction of clay content

The performance of SVMR in predicting clay content is presented in Figure 4.7. The best predictive model was obtained by the first derivative models FD-SVMR (RPD = 2.05, R_p^2 = 0.83, RMSEP = 1.95) and FD-K-SVMR (RPD = 1.85, R_p^2 = 0.76, RMSEP = 2.16). The SVMR-None and K-SVMR models poorly performed; RPD = 1.30, R_p^2 = 0.48, RMSEP = 3.07 for SVMR-None; and RPD = 1.43, R_p^2 = 0.62, RMSEP = 2.80 for K-SVMR.

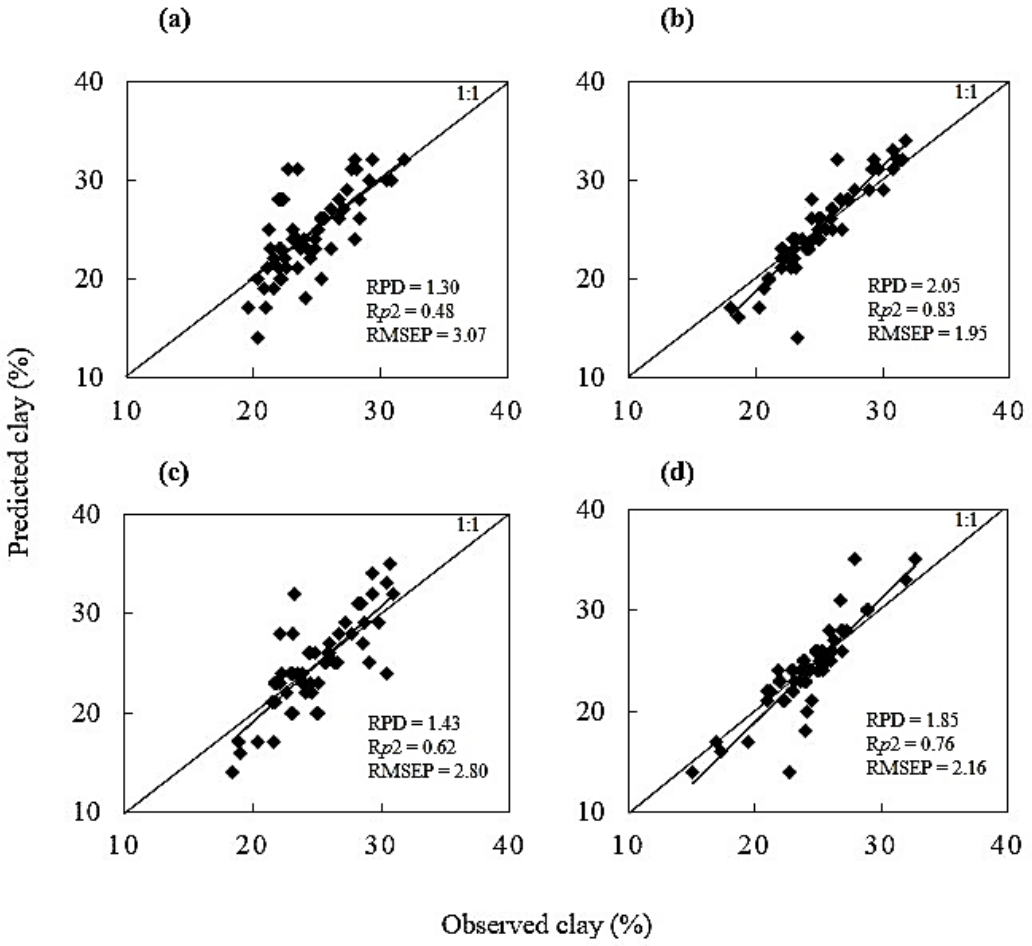


Figure 4.7 Performance of SVMR in predicting clay content: (a) clay SVMR raw model with all wavelengths; (b) clay first derivative SVMR model; (c) clay SVMR model with key wavelengths; (d) clay derivative SVMR model with key wavelengths.

4.5.2.2 SVMR prediction of SOC

Figure 4.8 shows the performance of SVMR models in predicting SOC contents. The first derivative models showed very good predictions (RPD > 2 and 0.8 < R_p^2 < 0.9). RPD = 2.45, R_p^2 = 0.86, RMSEP = 2.43 g.kg^{-1} for FD-K-SVMR; and RPD = 2.40, R_p^2 = 0.87, RMSE = 2.48 g.kg^{-1} for FD-SVMR. The SVMR-None and K-SVMR obtained poor predictions (RPD = 1.61, R_p^2 = 0.68, RMSEP = 3.69 g.kg^{-1} for SVMR-None; and RPD = 1.60, R_p^2 = 0.67, RMSEP = 3.73 g.kg^{-1} for K-SVMR).

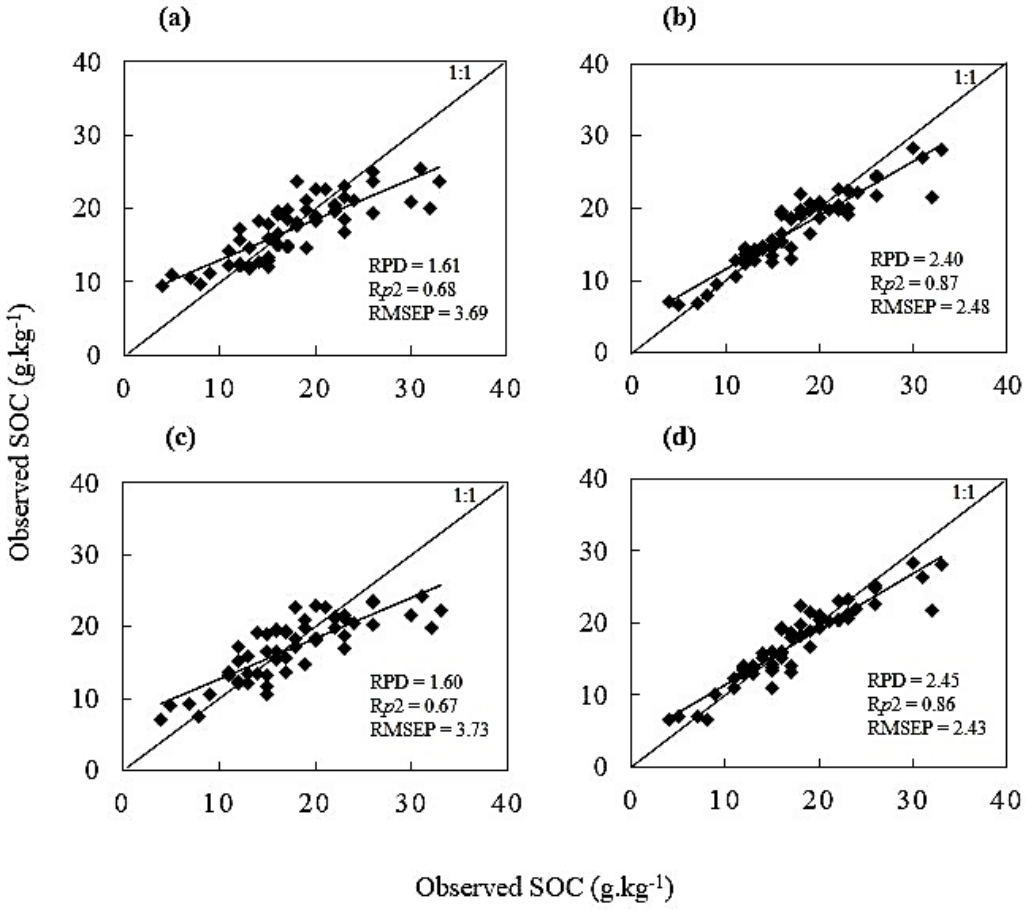


Figure 4.8 Performance of SVMR in predicting SOC: (a) SOC SVMR raw model with all wavelengths; (b) SOC first derivative SVMR model; (c) SOC SVMR model with key wavelengths; (d) SOC derivative SVMR model with key wavelengths.

4.5.2.3 SVMR prediction of TN

Figure 4.9 shows the SVMR models developed for TN predictions. The FD-SVMR and FD-K-SVMR models obtained fair predictions (RPD = 1.59, R_p^2 = 0.73, RMSEP = 0.27 mg.kg^{-1} for the FD-SVMR model and RPD = 1.59, R_p^2 = 0.72, RMSEP = 0.27 mg.kg^{-1} for the FD-K-SVMR model). The SVMR-None and K-SVMR models indicated very poor performance (RPD < 1.4).

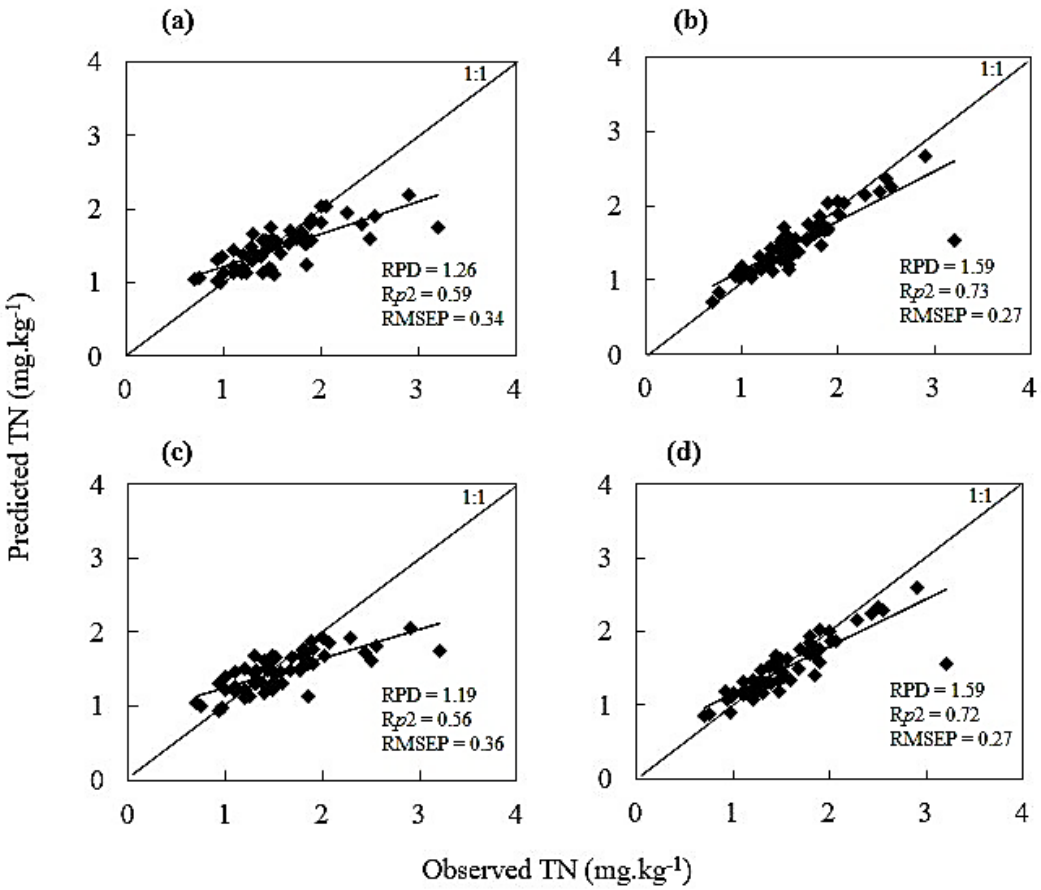


Figure 4.9 Performance of SVMR in predicting TN: (a) TN SVMR raw model with all wavelengths; (b) TN first derivative SVMR model; (c) TN SVMR model with key wavelengths; (d) TN derivative SVMR model with key wavelengths.

4.5.2.4 SVMR prediction of pH

The performance of SVMR in predicting pH is presented in Figure 4.10. The best predictions were obtained by the first derivative models FD-SVMR (RPD = 2.87, R_p^2 = 0.89, RMSEP = 0.16) and FD-K-SVMR (RPD = 2.87, R_p^2 = 0.88, RMSEP = 0.16). The SVMR-None and K-SVMR models obtained poor predictions (1.4 < RPD < 2 and R_p^2 < 0.7), with the SVMR-None model and the key wavelengths models on raw data having consecutively RPD values of 1.53 and 1.58 with R_p^2 values of 0.63 and 0.64.

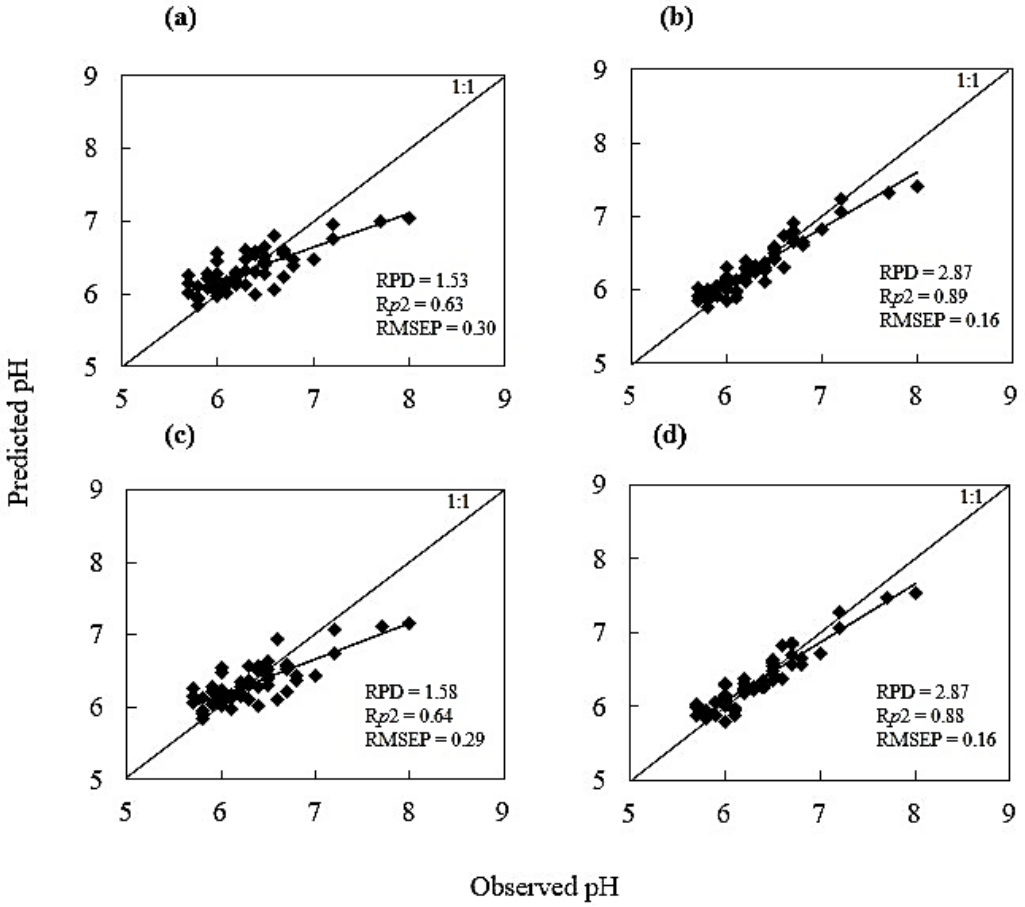


Figure 4.10 Performance of SVMR in predicting pH: (a) pH SVMR raw model with all wavelengths; (b) pH first derivative SVMR model; (c) pH SVMR model with key wavelengths; (d) pH derivative SVMR model with key wavelengths.

4.6 Comparison between PLSR and SVMR models

Table 4.6 summarises the performance of all SVMR and PLSR models in predicting clay content, SOC, TN, and pH. The model performance was assessed on both calibration and validation datasets. Prediction results of clay content indicate that PLSR outperformed SVMR when raw and key wavelengths data are used but overall, SVMR obtained the best results with the first derivative (RPD = 2.05, R_p^2 = 0.83, RMSEP = 1.95%) and the first derivative data with key wavelengths (RPD = 1.85, R_p^2 = 0.76, RMSEP = 2.16%).

In the same way, for the prediction of SOC, PLSR outperformed SVMR on raw and key wavelengths data, but the best results were obtained with SVMR models, FD-K-SVMR (RPD = 2.45, R_p^2 = 0.86, RMSEP = 2.43 g.kg^{-1}) and FD-SVMR (RPD = 2.40, R_p^2 = 0.87, RMSEP = 2.48 g.kg^{-1}). For the prediction of TN, results show that PLSR outperformed SVMR in all the four models (raw data, FD, K, and FD-K), with the best predictions provided by PLSR on raw

data (RPD = 2.15, R_p^2 = 0.77, RMSEP = 0.20 mg.kg^{-1}). For pH predictions, like clay content and SOC, PLSR outperformed SVMR on raw and key wavelengths data but overall, the best models to predict pH were provided by the first derivative SVMR models; FD-K-SVMR (RPD = 2.87, R_p^2 = 0.88, RMSEP = 0.16) and FD-SVMR (RPD = 2.87, R_p^2 = 0.89, RMSEP = 0.16).

The worst predictions were provided by SVMR-None for clay content (RPD = 1.30; R_p^2 = 0.48, RMSEP = 3.07%); K-SVMR for SOC (RPD = 1.60, R_p^2 = 0.67, RMSEP = 3.73 g.kg^{-1}); K-SVMR for TN (RPD = 1.19, R_p^2 = 0.56, RMSEP = 0.36 g.kg^{-1}); and SVMR-None for pH (RPD =1.53, R_p^2 = 0.63, RMSEP = 0.30).

Table 4.6 Performance of all SVMR and PLSR models in the calibration and validation datasets.

Property	Model	Pre-treatment	Calibration set (n = 130)			Validation set (n = 57)		
			R_c^2	RMSEC	AIC	R_p^2	RMSEP	RPD
Clay content	PLSR	None	0.99	0.17	-73.77	0.58	2.90	1.38
		FD	0.82	1.67	508.64	0.58	2.89	1.38
		K	0.90	1.22	427	0.64	2.66	1.50
		FD-K	0.79	1.82	531	0.63	2.73	1.46
	SVMR	None	0.65	2.60	-	0.48	3.07	1.30
		FD	0.79	2.40	-	0.83	1.95	2.05
		K	0.56	2.65	-	0.62	2.80	1.43
		FD-K	0.81	1.83	-	0.76	2.16	1.85
SOC	PLSR	None	0.88	2.05	561.58	0.65	3.09	1.93
		FD	0.89	1.97	551.99	0.63	3.13	1.91
		K	0.88	1.99	554.87	0.60	3.27	1.82
		FD-K	0.88	2.05	561.84	0.62	3.20	1.86
	SVMR	None	0.55	3.66	-	0.68	3.69	1.61
		FD	0.83	2.29	-	0.87	2.48	2.40
		K	0.54	3.71	-	0.67	3.73	1.60
		FD-K	0.84	2.21	-	0.86	2.43	2.45

TN	PLSR	None	0.76	0.20	-34.13	0.77	0.20	2.15
		FD	0.81	0.18	-69.86	0.73	0.22	1.95
		K	0.79	0.19	-50.35	0.65	0.25	1.72
		FD-K	0.75	0.19	-45.43	0.70	0.23	1.87
	SVMR	None	0.51	0.28	-	0.59	0.34	1.26
		FD	0.78	0.17	-	0.73	0.27	1.59
		K	0.44	0.30	-	0.56	0.36	1.19
		FD-K	0.74	0.20	-	0.72	0.27	1.59
pH	PLSR	None	0.84	0.18	-66.15	0.77	0.18	2.55
		FD	0.84	0.17	-70.96	0.71	0.20	2.30
		K	0.99	0.01	-910.0	0.68	0.21	2.19
		FD-K	0.83	0.18	-61.69	0.75	0.19	2.42
	SVMR	None	0.48	0.31	-	0.63	0.30	1.53
		FD	0.84	0.18	-	0.89	0.16	2.87
		K	0.48	0.31	-	0.64	0.29	1.58
		FD-K	0.85	0.17	-	0.88	0.16	2.87

*SVMR = support vector machine regression; PLSR = partial least square regression; none = non-transformed (raw) data with all wavelengths; FD = first derivative (transformed) data; K = raw data with key wavelengths only; FD-K = first derivative data with key wavelengths only; R_c^2 = coefficient of determination in the calibration (training) dataset; RMSEC = root mean square error of calibration in % for clay content, $g.kg^{-1}$ for SOC and $mg.kg^{-1}$ for TN; AIC = Akaike Information Criterion; R_p^2 = coefficient of determination in the validation (testing) dataset; RMSEP = root mean square error of validation in % for clay content, $g.kg^{-1}$ for SOC and $mg.kg^{-1}$ for TN ; RPD = Ration of prediction to deviation.

CHAPTER FIVE

DISCUSSION

This chapter discusses the results, concludes the study and outlines suggestions for future research. Section 5.1 discusses the qualitative analysis of collected spectra while section 5.2 discusses the identification of key wavelengths in the prediction of soil properties. The performance of PLSR and SVMR in predicting the selected soil properties and the limitations of the study are discussed successively in sections 5.3 and 5.4.

5.1 Qualitative analysis of collected spectra

The collected spectra showed similar reflectance shapes because the same spectrally active elements were present in the soil samples. In the visible range (400-700 nm), the samples presented an increasing slope explained by the presence of iron oxides (Zelikman and Carmina, 2013). The samples also showed absorption features caused by the O-H functional group related to water, at approximately 1400 and 1800 nm in the NIR range (Shepherd and Walsh, 2002; Xu *et al*., 2018). According to Clark *et al*. (1990), the absorption region at about 1400 nm is the first overtone of O-H stretching (moisture adsorbed to the clay surface), and the region at approximately 1900 nm is the combination of O-H stretching and H-O-H bending in water molecules trapped in the crystal lattice.

Besides water absorption features, the raw reflectance and the first derivative spectra showed an absorption peak at ~2200 nm. According to Clark *et al*. (1990), this peak is related to the clay lattice Al-OH absorption band. It was also shown that spectral features were enhanced after the first derivative transformation. This has also been reported by various authors (Stoner and Baumgardner, 1981; Henderson *et al*., 1992; Li *et al*., 2015; Xu *et al*., 2018).

The collected Vis-NIR spectra contained useful information to derive estimates of soil properties. For example, absorption features in the 400-1000 nm range are characteristics of the presence of soil carbon and iron oxides (Gee and Bauder, 1986; Jensen *et al*., 2007; Gholizadeh *et al*., 2017), and those in the 1000-2500 nm range are from water, clay minerals and organic matter (Araujo *et al*., 2014). Mnkeni *et al.* (2005) who studied the mineralogical

and chemical composition of top-soils from different croplands of the Eastern Cape Province, found that iron oxides as hematite and a clay fraction that is dominated by quartz, mica, and/or kaolinite are present in soils of the province.

5.2 Identifying key wavelengths for predicting soil properties

All Vis-NIR wavebands were used to evaluate the feature importance in the prediction of each soil property. Two feature-blocks regions with high importance existing in the entire Vis-NIR region have been identified. The RF-Boruta algorithm selected most of the key wavelengths in the range of 400-700 nm and 2200-2450 nm. Most of the key wavelengths were also selected in the same region by the VIP algorithm implemented on the first derivative data.

The first region ranging from 400 to 700 nm, is mostly related to the Fe oxides (Qi *et al.*, 2017). In the 400-700 nm range, 490 nm is assigned to the electronic transition (ET) band of Fe^{3+} (Hunt and Salisbury, 1970; Viscara-Rossel and Behrens, 2010), ~ 503 nm to goethite (Grove *et al*., 1992), ~510 nm to the ET band of Fe^{2+} (Hunt and Salisbury 1970), 529 nm to the ET band of hematite (Viscara-Rossel and Behrens, 2010), 535 to hematite, and ~550 nm to the ET band of Fe^{2+} and hematite (Hunt and Salisbury 1970; Bayer *et al*., 2012). Approximately 650 nm is assigned to the ET band of hematite and goethite, 665 nm to the ET band of goethite, and ~700 nm to the ET band of Fe^{3+} (Viscara-Rossel and Behrens, 2010; Bayer *et al*., 2012).

The second region, ranging from 2200 to 2450 nm, may be connected to water, organics and clay minerals (Qi *et al*., 2017). Regarding clay minerals, 2200 and 2204 nm are assigned to montmorillonite, 2216 to illite, 2230 nm to the fundamental absorption bands of Al-OH bend of smectites (Viscara-Rossel and Behrens, 2010), 2308 and 2312 nm to kaolinite, 2336 nm to illite, 2372 and 2376 nm to kaolinite (Grove *et al*., 1992). Regarding organics, 2275 and 2279 nm are assigned to the overtone absorption bands of CH_2 and CH_3, 2307-2460 nm to the overtone absorption bands of methyl CH stretch, 2331 nm to the overtone absorption bands of CH_2 and COO (Ben-Dor *et al.*, 1997; Viscara-Rossel and Behrens, 2010).

In the case of the VIP algorithm computed from raw spectral data, key wavelengths were selected from 600 to 1150 nm for the four properties. In addition to the 600-1150 nm region, peaks at ~2200 nm and ~2300-2450 nm were selected for clay content and pH. VIP selected a broad interval because the algorithm is more sensitive to noise (Bangelesa, 2017).

Thomasson *et al*. (2001), found that 19 wavebands are important for clay content prediction in the Vis-NIR range (375, 475, 625, 675, 725, 1025, 1125, 1225, 1275, 1475, 1525, 1675, 1875, 2075, 2175, 2275, 2375, 2425 and 2475 nm). On the other hand, it has been shown by Tümsavaş *et al.* (2018), that key wavelengths for clay content predictions in laboratory conditions are 519, 966, 1141, 1525, and 1639 nm. For SOC predictions, Wang *et al*. (2015) found 440, 560, 625, 740, and 1336 nm as the key spectral wavelengths. Nocita *et al*. (2015) suggested that to predict SOC, the spectral portion between 580, 570 and 680 nm was sufficient. Bangelesa (2017) on the other hand, suggested the whole 400-700 nm range for SOC predictions using PLSR and RF.

For TN predictions, Xu *et al.* (2018) found key wavelengths at ~480, 600, 660, 720, 1290, 1400, 1900, 2200, and 2300 nm. Other studies also reported that 486, 607, 650, 1700 and 2050 nm were important to predict TN (Dalal and Henry, 1986). Henderson *et al.* (1992), Chang *et al*. (2001), and Reeves and McCarty (2001) identified the broad range of 1100-2498 nm as successful predictors of both SOC and TN in NIR spectroscopy. In the case of pH predictions, Tümsavaş (2017) identified 459, 709, 930, 2086 and 2205 nm as important wavelengths, with 460, 2086 and 2205 nm as the most prominent wavelengths. Thomasson *et al.* (2001) also found good results by using 425, 475, 525, 575, 625, 775, 825, 1025, 1075, 1175, 1225, 1325, 1625, 1975, 2325, 2425, and 2475 from the Vis-NIR range to predict pH among other soil properties.

It follows that many authors found similar but also relatively contradictory results regarding the determination of key wavelengths. In this study, the key wavelengths of the four soil properties were located in the same spectral regions. Thus, the properties were all influenced by the same features that contributed to the performance of the models. Our results are in agreement with those of Pinherho *et al.* (2010) who found that the 400-830 nm and 2150-2230 nm regions are the most important in particular for clay, SOC and pH predictions (excluding the regions related to structural water). For TN, our results are in line with the findings of Reeves and McCarty (2001) and Xu *et al.* (2018) who identified key wavelengths in both visible and NIR ranges.

In this study, the 400-700 nm region was important for predictions most likely due to hematite which influences the soil colour that in turn influences soil reflectance. This would make sense particularly for clay content and SOC since hematite would exist on soil particles as a coating agent (Pinherho *et al.*, 2010). On the other hand, the importance of the 2200-2450 nm region

is likely due to clay mineral features (-OH from kaolinite and illite) which are acting as a proxy predictor for soil properties (Clark *et al.*, 1990).

5.3 Performance of PLSR and SVMR in predicting clay content, SOC, TN, and pH

In the scientific literature, many studies compare the performance of multivariate techniques and their application in Vis-NIR spectroscopy (e.g., Viscarra Rossel *et al*., 2006; Stevens *et al*., 2010; Viscarra Rossel and Behrens, 2010; Vohland *et al*., 2011; Were *et al*., 2015). However, a model that performs well for one application or area may not work for another because of the specificity of the study area (e.g. geology and soil type) (Bayer *et al*., 2012), laboratory procedures (Nduwamungu *et al*., 2009) and multicollinearity and noise in spectral data (Vohland *et al*., 2011). Thus, none of the proposed multivariate methods has achieved universal acceptance (Chang *et al.*, 2001).

Few studies directly compare the performance of PLSR and SVMR for the prediction of soil properties. Xu *et al.* (2018) compared PCR, PLSR, BPNN, and SVMR for the prediction of SOM, TN, total P, and total K. They found that SVMR models outperformed other models. Furthermore, they showed that SVMR provides better performance than PLSR. Our results are similar for clay content, SOC and pH since the best predictions were provided by SVMR when implemented with the Savitzky-Golay first derivative transformation. According to Peng *et al.* (2014) and Li *et al.* (2015), the first derivative pre-processing method is the best in improving the performance of predictive models in spectroscopy.

The good performance of SVMR compared to PLSR in this study could also be attributed to the fact that SVMR generally outperforms PLSR in the presence of noise and outliers (Peng *et al*., 2014). Furthermore, SVMR performs better because of the non-linear behaviour documented for soil variables, which is overcome by the ability of SVMR to solve non-linear problems (Viscarra Rossel and Behrens, 2010). Morellos *et al*. (2016) showed that non-linear relationships between the Vis-NIR spectral data and soil variables inevitably emerge, and various external or internal factors (e.g., measurement conditions and characteristics of the analysed components) may enhance non-linear relationships.

Regarding TN predictions, our results showed that for all the models, PLSR outperformed SVMR. This is in agreement with the results of Shi *et al.* (2012) who found that PLSR was the most suitable method for estimating TN contents compared to SVMR. This could be because, aside from the complexity of the data, PLSR can also model the nonlinear relationship between

spectral data and soil properties by using enough principal components, although SVMR has an advantage in handling such a relationship (Vohland *et al.*, 2011).

Compared to other studies conducted in agricultural environments, the accuracy of most predictive models presented in this research is low. On the total of 32 models developed (8 models/property), only 14 achieved fair/good predictions. This is likely a result of the large size and the highly variant characteristics of the study area caused by non-agricultural environments, differences in geology and soil types (Bayer *et al*., 2012). The wide variety of soils obtained at the regional scale may introduce noise and nonlinearity, and thus reduce prediction accuracy (Zeng *et al.*, 2016).

The difficulty to achieve prediction models for large areas with changing conditions (referred to as global calibrations) was also addressed in previous studies (e.g., Stevens *et al*., 2010; Bayer *et al*., 2012). To improve the prediction accuracy, Zeng *et al.* (2016) proposed that calibrations should be performed with a set of samples taken in the same area (in a small farm for example) because the samples being geographically close, they should therefore have similar properties and spectral responses.

5.4 Limitations of the study

In this study, 18 out of 32 models were unsuccessful (very poor/poor) and 14 were fair/good/very good. Given the fundamentals of precision agriculture, the developed models may not be a suitable replacement for laboratory analyses when excellent accuracy is required. The limitation of the low model performance of this study could be explained by the inaccuracy of clay content, SOC, TN, and pH values extracted from the AfSIS digital soil maps since no conventional wet chemistry was done for the determination of reference soil properties. Also, the effect of different geology and soil types on model performance was not evaluated. However, this study validated the prospect of using spectroscopy for soil quality monitoring in the study area by predicting soil properties at a low cost and within a reasonable time frame.

CONCLUSIONS

In this study, the potential of Vis-NIR spectroscopy to predict clay content, SOC, TN, and pH in dried and stored soil samples from agricultural fields of the Eastern Cape Province was evaluated using two multivariate techniques, PLSR and SVMR. Qualitative characteristics of the collected spectra were analysed, key wavelengths were selected and the two regression models were compared. The effects of the Savitzky-Golay first derivative spectra transformation were also assessed. The following conclusions are drawn according to the results:

- Vis-NIR spectroscopy can be successfully used to predict soil clay content, SOC, TN and pH in the study area;
- SVMR models were the best to predict clay content, SOC, and pH when performed on first derivative data and first derivative data with key wavelengths. However, PLSR outperformed SVMR for the prediction of TN.
- The impact of the first derivative transformation was more evident for SVMR because the best models were obtained on processed data.
- Key wavelengths to predict clay content, SOC, TN, and pH were identified around 400-700 nm (in the Vis range) and 2200-2450 nm (in the NIR range), corresponding to iron oxides and clay minerals found in the study area.

Although variably good predictions were obtained for clay content, SOC, TN, and pH using SVMR and PLSR models, for future practical applications, the robustness of these models require better validation accuracy. For future studies, the overall accuracy can be improved by (1) reducing the study area to lessen the impact of differences in geology and soil types, (2) using the standard wet chemical methods to determine reference values of soil properties.

Several perspectives for future research arise from this study. More investigation needs to be oriented on the application of Vis-NIR spectroscopy and different linear and non-linear regression models to predict various soil properties at the farm level (in laboratory and *in situ* conditions), or the province scale by considering the impact of environmental factors (e.g. geology and soil types). It is also feasible to include soil spectroscopy into socioeconomic household surveys. A larger application of Vis-NIR spectroscopy in agricultural research in South Africa could unlock further understanding of the effects of farm management practices and changes in soil health over time.

REFERENCES

Abdi, D., Tremblay, G.F., Ziadi, N., Bélanger, G. and Parent, L.E. (2012). Predicting soil phosphorus-related properties using near-infrared reflectance spectroscopy. *Soil Science Society of America Journal,* 76: 2318-2326.

Adnan, N., Ahmad, M.H. and Adnan, R. (2006). A comparative study on some methods for handling multicollinearity problems. *Matematika,* 22: 109-119.

Al-Abbas, A.H., Swain, P.H. and Baumgardner, M.F. (1972). Relating organic-matter and clay content to multispectral radiance of soils. *Soil Science,* 114: 477-485.

Alibuhtto, M.C. and Peiris, T.S.G. (2015). Principal component regression for solving multicollinearity problem. *5th International Symposium 2015-SEUSL,* 231-238.

Andersson, N. and Galt, K. (1998). The wild coast spatial development initiative (SDI): community needs of development; community information empowerment and transparency (CIET) International: Bisho, South Africa.

Araujo, S.R., Wetterlind, J., Dematte, J.A.M. and Stenberg, B. (2014). Improving the prediction performance of a large tropical vis-NIR spectroscopic soil library from Brazil by clustering into smaller subsets or use of data mining calibration techniques. *European Journal of Soil Science,* 65: 718-729.

ASD, Analytical Spectral Devices, Inc. (2005). Handheld Spectroradiometer: User's Guide, Version 4.05. Boulder, USA.

Babaeian, E., Homaee, M., Vereeckem, H., Montzka, C., Nouzouri, A.A. and van Genuchten, M. (2015). A comparative study of multiple approaches for predicting the soil-water retention curve: hyperspectral information vs. basic soil properties. *Soil Science Society of America Journal,* 79: 1043-1058.

Balasubramanian, A. (2017). Chemical properties of soils. Center for Advanced Studies in Earth Science, University of Mysore.

Bangelesa, F.F. (2017). *Predicting soil organic carbon in a small farm system using in situ spectral measurements and the random forest regression*. MSc thesis, Faculty of Science, University of the Witwatersrand, Johannesburg.

Barnes, E.M., Sudduth, K.A., Hummel, J.W., Lesch, S.M., Corwin, D.L., Yang, C., Daughtry, C.S.T. and Bausch, W.C. (2003). Remote and ground-based sensor techniques to map soil properties. *Photogrammetric Engineering & Remote Sensing,* 69: 619-630.

Baumgardner, M. F., Silva L. F., Biehl, L. L., and Stoner, E. R. (1985). Reflectance properties of soils. *Advances in Agronomy,* 38: 1-44.

Baxter, N. and Williamson, J. (2001). *Know your soils*. Part I. Introduction to soils. Department of Natural Resources and Environment, Center for Land Protection Research, Vic.

Bayer, A., Bachmann, M., Muller, A. and Kaufmann, H. (2012). A comparison of feature-based MLR and PLS regression techniques for the prediction of three soil constituents in a degraded South African ecosystem. *Applied and Environmental Soil Science,* DOI: 10.1155/2012/971252.

Bembridge, T.J. (1984). *A systems approach study of agricultural development problems in Transkei*. PhD Dissertation, University of Stellenbosch, Stellenbosch, South Africa.

Ben-Dor, E., Chabrillat, S., Demattê, J.A.M., Taylor, G.R., Hill, J., Whiting, M.L. and Sommer, S. (2009). Using Imaging Spectroscopy to study soil properties. *Remote Sensing of Environment,* 113: S39-S55.

Ben-Dor, E., Granot, A. and Notesco, G. (2017). A simple apparatus to measure soil spectral information in the field under stable conditions. *Geoderma,* 306: 73-80.

Ben-Dor, E., Inbar, Y. and Chen, Y. (1997). The reflectance spectra of organic matter in the visible near-infrared and short-wave infrared region (400-2500 nm) during a controlled decomposition process. *Remote Sensing of Environment,* 61: 1-15.

Ben-Dor, E., Patkin, K., Banin, A. and Karnieli, A. (2002). Mapping of several soil properties using dais-7915 hyperspectral scanner data - a case study over clayey soils in Israel. *International Journal of Remote Sensing,* 23: 1043-1062.

Bogrekci, I., and W.S. Lee. (2005). Spectral soil signatures and sensing phosphorus. *Biosystems Engineering*, 92(4): 527-533, DOI: 10.1016/j.biosystemseng.2005.09.001.

Boulesteix, A.N. and Strimmer, K. (2007). Partial least squares: A versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioinformatics,* 8: 32-44.

Bowers, S.A. and Hanks, R.J. (1965). Reflection of radiant energy from soils. *Soil Science,* 100: 130-138.

Bowers, S.A. and Smith, S.J. (1972). Spectrophotometric determination of soil water content. *Soil Science Society of America Proceedings,* 36: 978-980.

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

Budak, M. and Gunal, H. (2016). Visible and Near-Infrared spectroscopy techniques for determination of some physical and chemical properties in Kazova Watershed. *Advances in Environmental Biology,* 10: 61-72.

Bushnell, T.M. (1932). A new technique in soil mapping. *American Soil Survey Association Bulletin,* 13: 74-81.

Butkuté, B. and Slepetiene, A. (2004). Near-Infrared reflectance spectroscopy as a fast method for simultaneous prediction of several soil quality components. *Chemija,* 15: 15-20.

Chan, C. W., Schueller, J. K., Miller, W. M., Whitney, J. D. and Cornell, J. A. (2004). Error sources affecting variable rate application of nitrogen fertilizer. *Precision Agriculture,* 5: 601-616.

Chang, C.W., Laird, D.A., Mausbach, M.J. and Hurburgh, C.R. (2001). Near-Infrared reflectance spectroscopy–Principal components regression analyses of soil properties. *Soil Science Society of America Journal*, 65(2): 480-490, DOI: 10.2136/sssaj2001.652480x.

Chang, C.W., Laird, D.A., Mausbach, M.J. and Hurburg, C.R.J. (2001). Near-infrared reflectance spectroscopy - principal component regression analysis of soil properties. *Soil Science Society of America Journal,* 65: 480-490.

Chong, I.G. and Jun, C.H. (2005). Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intelligent Laboratory Systems,* 78: 103-112.

Cierniewski, J. and Kusnierek, K. (2010). Influence of several soil properties on soil surface reflectance. *Quaestiones Geographicae,* 29, DOI: 10.2478/v10117-010-0002-9.

Clark, R.N., King, T.V., Klejwa, M., Swayze, G. and Vergo, N. (1990). High spectral resolution reflectance spectroscopy of minerals. *Journal of Geophysical Research,* 95: 12653-12680.

Clark, R.N., King, T.V.V., Klejwa, M., Swayze, G.A. and Vergo, N. (1990). High spectral resolution reflectance spectroscopy of minerals. *Journal of Geophysical Research,* 95: 12653-12680.

Cohen, Y., Alchanatis, V., Meron, M., Saranga, Y. and Tsipris, J. (2005). Estimation of leaf water potential by thermal imagery and spatial analysis. *Journal of Experimental Botany,* 56: 1843-1852.

Condit, H. R. (1970). The spectral reflectance of American soils. *Photogrammetric Engineering,* 36: 955-966.

Coulson, K.L. and Reynolds, D.W. (1971). The spectral reflectance of natural surfaces. *Journal of Applied Meteorology,* 10: 1285-1295.

Council for Geoscience of South Africa (CGS). (2008). Simplified geological map of the Republic of South Africa and the Kingdoms of Lesotho and Swaziland. Available online: https://www.geoscience.org.za/images/DownloadableMaterial/RSA_Geology.pdf

Croft, H., Kuhn, N.J. and Anderson, K. (2012). On the use of remote sensing techniques for monitoring spatio-temporal soil organic carbon dynamics in agricultural systems. *Catena,* 94: 64-74.

Dalal, R.C. and Henry, R.J. (1986). Simultaneous determination of moisture, organic carbon and total nitrogen by near-infrared reflectance spectroscopy. *Soil Science Society of America Journal,* 50: 120-123.

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

Darst, B.F., Malecki, K.C. and Engelman, C.D. (2017). Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. *BMC Genetics,* 19: 65, DOI: 10.1186/s12863-018-0633-8.

Delgado, A. and Gomez, J.A. (2016). The Soil physical, chemical and biological properties. In: Villabos F., Feres E. (eds) *Principles of Agronomy for sustainable agriculture*. Springer, Cham, DOI: 10.1007/978-3-319-46116-8_2.

Demattê, J.A.M., Dotto, A.C., Bedin, L.G., Sayão, V.M. and Souza, A.B. (2019). Soil analytical quality control by traditional and spectroscopy techniques: Constructing the future of a hybrid laboratory for low environmental impact. *Geoderma,* 337: 111-121.

Demattê, J.A.M., Pereira, H.S., Nanni, M.R., Cooper, M. and Fiorio, P.R. (2003). Soil chemical alterations promoted by fertilizer application assessed by spectral reflectance. *Soil Science,* 168: 730-747.

Dobos, E., Montanarella, L., Nègre, T. and Micheli, E. (2001). A regional-scale soil mapping approach using integrated AVHRR and DEM data. *International Journal of Applied Earth Observation and Geoinformation,* 3: 30-42.

Dwivedi, R.S. (2017). *Remote sensing of soils*. Springer, Germany.

Erasmus, J. (1996). Eastern Cape: A Human development profile development. Paper 108, Centre for Policy and Information of the Development Bank of Southern Africa, Halfway House, South Africa.

FAO (Food and Agriculture Organization of the United Nations). (2005). The importance of soil organic matter. Key to drought-resistant soil and sustained food production. FAO Soils Bulletin 80, Rome, Italy.

FAO (Food and Agriculture Organization of the United Nations). (2006). Plant nutrition for food security. A guide for integrated nutrient management. FAO Fertilizer and Plant Nutrition Bulletin 19, Rome, Italy.

FAO (Food and Agriculture Organization of the United Nations). (2008). Guide to laboratory establishment for plant nutrient analysis. FAO Fertilizer and Plant Nutrition Bulletin 19, Rome, Italy.

FAO (Food and Agriculture Organization of the United Nations). (2015). *Healthy soils for a healthy life*. Available online: http://www.fao.org/soils 2015/news/news-detail/en/c/277682/.

FAO and ITPS (Food and Agriculture Organization of the United Nations and Intergovernmental Technical Panel on Soils). (2015). Status of the World's Soil Resources (SWSR) - Main Report. Rome, Italy.

Forkuor, G., Hounkpatin, O.K.L., Welp, G. and Thiel, M. (2017). High-resolution mapping of soil properties using remote sensing variables in South-Western Burkina Faso: A comparison of Machine Learning and Multiple Linear Regression Models. PLoS ONE 12(1): e0170478. DOI: 10.1371/journal.pone.0170478.

Foster, S., Schultz, B., McCuin, G., Neibling, H. and Shewmaker, G. (2013). *Soil Properties*. Fact Sheet-13-02, University of Nevada.

Gates, J.R. (2018). *A comparison of VNIR and MIR spectroscopy for predicting various soil properties*. MSc dissertation, University of Nebraska.

Gee, G.W. and Bauder, J.W. (1986). Particle-size analysis. In *Methods of Soil Analysis*, Part 1; Klute, A., Ed.; ASA and SSSA: Madison, WI, USA.

Gholizadeh, A., Carmon, N., Klement, A., Ben-Dor, E. and Boruvka, L. (2017). Agricultural soil spectral response and properties assessment: effects of measurement protocol and data mining technique. *Remote Sensing*, 9: 1078-1091.

Gourlay, S., Aynekulu, E., Carletto, C. and Shepherd, K. (2017). Spectral soil analysis & household surveys - A guidebook for integration. Washington, DC, World Bank.

Gregorutti, B., Michel, B. and Saint-Pierre, P. (2017). Correlation and variable importance in random forests. *Statistics and Computing*, 27: 659-678.

Grove, C., Hook, S. and Paylor, E. (1992). Laboratory Reflectance Spectra of 160 Minerals, 0.4 to 2.5 Micrometers, Jet Propulsion Laboratory, National Aeronautics and Space Administration, JPL publication 92-2, Pasadena, California, USA.

Hamann, M. and Tuinder, V. (2012). Introducing the Eastern Cape: A quick guide to its history, diversity and future challenges. Stockholm Resilience Centre (SRC), South Africa.

He, Y., Huang, M., García, A., Hernández, A. and Song, H.Y. (2007). Prediction of soil macronutrients content using near-infrared spectroscopy. *Computers and Electronics in Agriculture,* 58: 144-153.

Henderson, T.L., Baumgardner, M.F., Franzmeier, D.P., Stott, D.E. and Coster, D.C. (1992). High dimensional reflectance analysis of soil organic matter. *Soil Science Society of America Journal,* 56: 865-872.

Hengl, T., Heuvelink, G.B.M., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Shepherd, K.D., *et al.* (2015). Mapping Soil Properties of Africa at 250 m Resolution: Random Forests Significantly Improve Current Predictions. PLoS ONE 10(6): e0125814. DOI: 10.1371/journal.pone.0125814

Hensley, M., Le Roux, P., Du Preez, C., Van Huyssteen, C., Kotze, E. and Van Rensburg, L. (2006). Soils: The Free State's Agricultural Base. *South Africa Geographical Journal,* 88: 11-21.

Hoffer, R.M. and Johannsen, C.J. (1969). Ecological potentials in spectral signature analysis. In: Johnson, P.L. *Remote sensing in ecology*. University of Georgia Press: Athens, pp. 1-16.

Hosu, S.Y., Cishe, E.N. and Luswazi, P.N. (2016). Vulnerability to climate change in the Eastern Cape Province of South Africa: what does the future hold for smallholder crop farmers? *Agricultural Economics Research, Policy and Practice in Southern Africa,* 55: 133-167, DOI: 10.1080/03031853.2016.1157025.

Hunt, G. and Salisbury, J. (1970). Visible and near-infrared spectra of minerals and rocks: Silicate minerals. *Modern Geology,* 1: 283-300.

Jamil, N., Sajjad, N. and Ashraf, H. (2016). Physical and chemical properties of soil quality indicating forests productivity: A review. *American-Eurasian Journal of Toxicological Sciences,* 8: 60-68.

Jensen, J.R. (2007). *Remote Sensing of the Environment*: An Earth Resource Perspective, 2nd ed., Prentice-Hall: Upper Saddle River, NJ, USA.

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

Karltun, E., Lemenih, M. and Tolera, M. (2011). Comparing farmers' perception of soil fertility change with soil properties and crop performance in Beseku, Ethiopia. *Land Degradation and Development,* DOI: 10.1002/ldr.1118.

Kohavi, R. and John, G.H. (1997). Wrappers for Feature Subset Selection. *Artificial Intelligence,* 97: 273-324.

Kopackova, V. and Ben-Dor, E. (2016). Normalizing reflectance from different spectrometers and protocols with an internal soil standard. *International Journal of Remote Sensing,* 37: 1276-1290.

Kremer, J., Pedersen, K.S. and Igel, C. (2014). *Active learning with support vector machines*. University of Copenhagen, Denmark.

Kursa, M.B. and Rudnicki, W.R. (2010). Feature selection with the Boruta package. *Journal of Statistical Software,* 36, DOI: 10.18637/jss.v036.i11.

Lal, R. (2009). Challenges and opportunities in soil research. European. *Journal of Soil Science,* 60: 158-169.

Letey, J. (1958). Relationship between soil physical properties and crop production. *Advances in Soil Science,* 1: 77-293.

Li, S., Shi, Z., Chen, S.C., Ji, W.J., Zhou, L.Q., Yu, W. and Webster, R. (2015). In situ measurements of organic carbon in soil profiles using vis-NIR spectroscopy on the Qinghai-Tibet Plateau. *Environmental Science & Technology,* 49: 4980-4987.

Lohry, R. (2007). *Micronutrients: functions, sources and applications methods*. Indiana CCA Conference Proceedings. Nutra Flo Company, Iowa.

Ludwig, B., Murugan, R., Parama, V.R.P. and Vohland, M. (2019). Accuracy of estimating soil properties with Mid-Infrared spectroscopy: Implications of different chemometric approaches and software packages related to calibration sample size. *Soil Science Society of America Journal,* 83: 1542-1552.

Madari, B.E., Machado, P., Reeves, J. and Guimaraes, C.M. (2006). Mid- and Near-Infrared Spectroscopy assessment of soil composition parameters and structural indices in two ferralsols. *Geoderma*, 136: 245-259.

Martens, H. and Næs, T. (1989). *Multivariate Calibration*. John Wiley & Sons Ltd., New York.

Martens, M. and Martens, H. (1986). Partial least square regression. In: Piggot, J.R. (ed.). *Statistical procedures in food research*. Elsevier Applied Science, London, pp. 69-98.

McBratney, A., Field, D.J. and Koch, A. (2014). The dimensions of soil security. *Geoderma*, 213: 203-213, DOI: 10.1016/j.geoderma.2013.08.013.

McCauley, A. (2005). *Basic soil properties*. Soil & Water management module 1. Montana State University.

Mevik, B.H. (2016). VIP.R: Implementation of VIP (Variable Importance in Projection) for the 'pls' Package. Available online: http://mevik.net/work/software/VIP.R/.

Miller, J.N. and Miller, J.C. (2010). *Statistics and Chemometrics for Analytical Chemistry*. Sixth edition, pp. 296.

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

Mnkeni, P.N.S., Mandiringana, O.T., Mkile, Z., van Averbeke, W., Van Ranst, E. and Verplancke, H. (2005). Mineralogy and fertility status of selected soils of the Eastern Cape Province, South Africa. *Communications in Soil Science and Plant Analysis*, 36: 2431-2446.

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

Mulder, V.L., de Bruin, S., Weyermann, J., Kokaly, R.F. and Schaepman, M.E. (2011). Characterizing regional soil mineral composition using spectroscopy and geostatistics. *Remote Sensing of Environment*, 139: 415-429

Mulla, D. (2013). Twenty-five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps. *Biosystems Engineering,* 114: 358-371.

Nalepa, J. and Kawulok, M. (2018). Selecting training sets for support vector machines: a review. *Artificial Intelligence Review,* 52: 857-900.

Nduwamungu, C., Ziadi, N., Parent, L. E., Tremblay, G.F. and Thuriès, L. (2009). Opportunities for, and limitations of, near-infrared reflectance spectroscopy applications in soil analysis: A review. *Canadian Journal of Soil Science,* 89: 531-541.

Newman, A.C.D. (1984). The significance of clays in agriculture and soils. *Philosophical Transactions of the Royal Society of London,* 311: 375-389.

Nocita, M., Stevens, A., van Wesemael, B., Aitkenhead, M., Bachmannj, M., Barth, B., Ben Dor, B., Brown, D.J., Clairotte, M., Csorba, A., Dardenne, P. *et al.* (2015). Soil Spectroscopy: An alternative to wet chemistry for Soil Monitoring. *Advances in Agronomy,* 132, DOI: 10.1016/bs.agron.2015.02.002.

Novara, A., Gristina, L., Bodì, MB. and Cerdà, A. (2011). The impact of fire on redistribution of soil on a Mediterranean hillslope under maquia vegetation type. *Land Degradation & Development,* 22: 530-536.

O'Rouke, S.M., Stockman, U., Holden, N.M., McBratney, A.B. and Minasny, B. (2016). An assessment of model averaging to improve predictive power of portable vis-NIR and XRF for the determination of agronomic soil properties. *Geoderma,* 279: 31-44.

Paterson, G., Turner, D., Wiese, L., Van Zijl, G., Clarke, C. and Van Tol, J. (2015). Spatial soil information in South Africa: Situational analysis, limitations and challenges. *Spatial soil information in South Africa,* 111, DOI: 10.17159/sajs.2015/20140178.

Pei, X., Sudduth, K.A., Veum, K.S. and Li, M. (2018). Improving *In-situ* estimation of soil profile properties using multi-sensor probe. *Sensors,* 19: 1011, DOI: 10.3390/s19051011.

Peng, X., Shi, T., Song, A., Chen, Y. and Gao, W. (2014). Estimating soil organic carbon using VIS/NIR Spectroscopy with SVMR and SPA methods. *Remote Sensing,* 6: 2699-2717.

Penny, D. (2004). The micronutrient and trace element status of forty-three soil quality benchmark sites in Alberta. Alberta Agriculture, Food and Rural Development, Conservation and Development Branch.

Pinheiro, E.F.M., Ceddia, M.B., Clingensmith, C.M., Grunwald, S. and Vasques, G.M. (2017). Prediction of soil physical and chemical properties by Visible and Near-Infrared diffuse reflectance spectroscopy in the Central Amazon. *Remote Sensing,* 9: 293, DOI: 10.3390/rs9040293.

Qi, H., Paz-Kagan, T., Karnieli, A. and Li, S. (2017). Linear multi-task learning for predicting soil properties using field spectroscopy. *Remote Sensing* 9: 1099, DOI: 10.3390/rs9111099.

Ramirez-Lopez, L., Wadoux, A.M., Franceschini, M.H.D., Terra, F.S., Marques, K.P.P, Sayo, V.M. and Demattê, J.A.A. (2019). Robust soil mapping at the farm scale with vis-NIR spectroscopy. *European Journal of Soil Science,* 70: 378-393.

Ray, D.K., West, P.C. and Mueller, N.D. (2013). Yields trends are insufficient to double global production by 2050. PLoS ONE 8(6): e66428, DOI: 10.1371/journal.pone.0066428.

Reeves, J.B. (2010). Near- versus mid-infrared diffuse reflectance spectroscopy for soil analysis emphasizing carbon and laboratory versus on-site analysis: Where are we and what needs to be done? *Geoderma,* 158: 3-14, DOI: 10.1016/j.geoderma.2009.04.005.

Reeves, J.B. and McCarty, G.W. (2001). Quantitative analysis of agricultural soils using near-infrared reflectance spectroscopy and a fibre-optic probe. *Journal of Near Infrared Spectroscopy*, 9: 25-34.

Schirrmann, M., Gebbers, R. and Kramer, E. (2013). Performance of automated near-infrared reflectance spectrometry for continuous in situ mapping of soil fertility at field scale. *Vadose Zone Journal*, 12, DOI: 10.2136/vzj2012.0199.

Shepherd, K.D. and Walsh, M.G. (2002). Development of reflectance spectral libraries for characterization of soil properties. *Soil Science Society of America Journal,* 66: 988-998.

Shi, T., Cui, L., Wang, J., Fei, T., Chen, Y. and Wu, G. (2012). Comparison of multivariate methods for estimating soil total nitrogen with visible/near-infrared spectroscopy. *Plant and Soil,* 366: 363-375, DOI: 10.1007/s11104-012-1436-8.

Shorack, G. R. and Wellner, J. A. (2009). Empirical processes with applications to statistics. *Applied Mathematics,* 59, ISBN: 9780898716849.

Silva, S.H.G., Silva, E.A., Poggere, G.C., Guilherme, L.R.G. and Curi, N. (2018). Tropical soils characterization at low cost and time using portable X-ray fluorescence spectrometer (pXRF): Effects of different sample preparation methods. *Ciência e Agrotecnologia*, 42: 80-92, DOI: 10.1590/1413-70542018421009117.

Slessarev, E. W., Lin, Y., Bingham, N. L., Johnson, J. E., Dai, Y.; Schimel, J. P. and Chadwick, O. A. (2016). Water balance creates a threshold in soil pH at the global scale. *Nature,* 540: 567-569, DOI: 10.1038/nature20139.

Smola, A.J. and Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing,* 14: 199-222.

Soriano-Disla, J.M., Janik, L.J., Viscarra Rossel, R.A., MacDonald, L.M. and McLaughlin, M.J. (2014). The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties. *Applied Spectroscopy Reviews,* 49: 139-186.

StatsSA (Statistics South Africa). (2003). *Census 2001: Census brief*. Report no. 03-02-03 (2001). Pretoria.

Stenberg, B., Viscarra Rossel, R., Mouazen, M.A. and Wetterlind, J. (2010). Visible and Near-Infrared spectroscopy in soil science. In Sparks, D.L. *Advances in Agronomy*, vol. 107, Burlington: Academic Press, pp. 163-215.

Stevens, A., Udelhoven, T., Denis, A., Tychon, B., Lioy, R., Hoffmann, L. and van Wesemael, B. (2010). Measuring soil organic carbon in croplands at regional scale using airborne imaging spectroscopy. *Geoderma,* 158: 32-45.

Stoner, E.R. and Baumgardner, M.F. (1981). Characteristic variations in reflectance of surface soils. *Soil Science Society of American Journal,* 45: 1161-1165.

Stoppiglia, H., Dreyfus, G., Dubois, R. and Oussar, Y. (2003). Ranking a Random Feature for Variable and Feature Selection. *Journal of Machine Learning Research*, 3: 1399-1414.

Stuart, B. (2004). *Infrared spectroscopy: fundamentals and application*. John Wiley & Sons: Chichester, England.

Thomasson, J.A., Cox, M.X. and Al-Rajehy, A. (2001). Soil Reflectance Sensing for determining Soil Properties in Precision Agriculture. *American Society of Agricultural Engineers,* 44: 1445-1453.

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

Todorova, M., Atanassova, S., Lange, H. and Pavlov, D. (2011). Estimation of total N, total P, pH and electrical conductivity in soil by near-infrared reflectance spectroscopy. *Journal of Agricultural Science and Technology*, 3(1): 50-54.

Tümsavaş, Z. (2017). Possibility of determining soil pH using visible and near-infrared (Vis-NIR) spectrophotometry. *Journal of Environmental Biology,* 38: 1095-1100.

Tumsavas, Z., Tekin, Y., Ulusoy, Y. and Mouazen, A.M. (2018). Prediction and mapping of soil clay and sand contents using visible and near-infrared spectroscopy. *Biosystems Engineering*, DOI: 10.1016/j.biosystemseng.2018.06.008.

Vâgen, T.G., Shepherd, K.D. and Walsh, M.G. (2006). Sensing landscape level change in soil fertility following deforestation and conversion in the highlands of Madagascar using Vis-NIR spectroscopy. *Geoderma,* 133: 281-294.

Vapnik, V.N. (2000). The nature of statistical learning theory. In: Jordan, M., Lauritzen, S.L., Lawless, J.F., Nair, V. (Eds.), *Statistics for Engineering and Information Science*. Springer Verlag, New York, pp. 1-314.

Veum, S.K., Parker, P.A., Sudduth, K.A. and Holan, S.H. (2018). Predicting profile soil properties with Reflectance spectra via Bayesian Covariate-Assisted External Parameter Orthogonalization. *Sensors,* 18: 3869, DOI: 10.3390/s18113869.

Viscarra Rossel, R.A. and Behrens, T. (2010). Using data mining to model and interpret soil diffuse reflectance spectra, *Geoderma,* 158: 46-54.

Viscarra Rossel, R.A., Cattle, S.R., Ortega, A. and Fouad, Y. (2009). In situ measurements of soil colour, mineral composition and clay content by vis-NIR spectroscopy. *Geoderma,* 150: 253-266.

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

Vohland, M., Besold, J., Hill, J. and Fründ, H. (2011). Comparing different multivariate calibration methods for the determination of soil organic carbon pools with visible to near-infrared spectroscopy. *Geoderma,* 166: 198-205.

Walcott, J., Bruce, S. and Sims, J. (2009). Soil carbon for carbon sequestration and trading: a review of issues for agriculture and forestry. Bureau of Rural Sciences, Department of Agriculture, Fisheries & Forestry, Canberra.

Wang, Y., Huang, T., Liu, J., Lin, Z., Li, S., Wang, R. and Ge, Y. (2015). Soil pH value, organic matter and macronutrients contents prediction using optical diffuse reflectance spectroscopy. *Computers and Electronics in Agriculture,* 111: 69-77.

Warkentin, B.P. (1995). The changing concept of soil quality. *Journal of Soil and Water Conservation,* 58: 226-228.

Were, K., Bui, D.T., Dick, Ø.B. and Singh, B.R. (2015). A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. *Ecological Indicators,* 52: 394-403.

Wetterlind, J., Stenberg, B. and Soderstrom, M. (2008). The use of near-infrared (NIR) spectroscopy to improve soil mapping at the farm scale. *Precision Agriculture,* 9: 57-69.

Wold, S., Martens, H., and Wold, H. (1983). *The multivariate calibration problem in chemistry solved by the PLS method*. Springer, pp. 286-293.

Wold, S., Ruhe, A., Wold, H. and Dunn, W.J. (1984). The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing,* 5: 735-743.

Wright, M.N., Ziegler, A. and Konig, I.R. (2017). Do little interactions get lost in dark random forests? *BMC Bioinformatics,* 17: 145.

WWF. (2015). *Agriculture: Facts & Trends – South Africa*. Available online: http://awsassets.wwf.org.za/downloads/facts_brochure_mockup_04_b.pdf.

Xu, S., Zhao, Y., Wang, M. and Shi, X. (2018). Comparison of multivariate methods for estimating selected soil properties from intact soil cores of paddy fields by Vis-NIR spectroscopy. *Geoderma,* 310: 29-43.

Yanli, L., Youlu, B., Liping, Y. and Hongjuan, W. (2010). Hyperspectral extraction of soil organic matter content based on principal component regression. *New Zealand Journal of Agricultural Research,* 50: 1169-1175, DOI: 10.1080/00288230709510399.

Zelikman, E. and Carmina, E. (2013). The spectral response characteristics of the soils and their possible estimation by using partial least square regression (PLSR) analysis. *International Journal of Geomatics and Geosciences,* 3: 438-453.

Zeng, R., Zhao, Y., Li, D., Wu, D., Wei, C. and Zhang, G. (2016). Selection of "local" models for prediction of soil organic matter using a regional soil Vis-NIR spectral library. *Soil Science,* 181: 13-19.

Zhang, Y., Biswas, A., Ji, W. and Adamchuk, V. (2017). Depth-specific prediction of soil properties in situ using vis-NIR spectroscopy. *Soil Science Society of America Journal,* 81: 993-1004.

Zhou, P., Zhang, Y., Yang, W., Li, M., Liu, Z. and Liu, X. (2019). Development and performance test of an in-situ soil total nitrogen-soil moisture detector based on near-infrared spectroscopy. *Computers and Electronics in Agriculture,* 160: 51-58.

Zornoza, R., Guerrero, C., Mataix-Solera, J., Scow, K.M., Arcenegui, V. and Mataix- Beneyto, J. (2008). Near-infrared spectroscopy for determination of various physical, chemical and biochemical properties in Mediterranean soils. *Soil Biology and Biochemistry*, 40(7): 1923-1930, DOI: 10.1016/j.soilbio.2008.04.003.

APPENDICES

Appendix A – PLSR-VIP code

```
#removing previous datasets
rm(list=ls())
#setting working directory
setwd("C:/")
#laoding packages
library(car)
library(pls)
library(caret)
library(plsVarSel)
#laoding the datasets
mydata <- read.csv("C:/Data.csv")
#creation of the calibration and validation datasets
smp_size <-floor(0.70 * nrow(mydata))
set.seed(123)
train_ind<-sample(seq_len(nrow(mydata)), size = smp_size)
training<-mydata[train_ind, ]
testing<-mydata[-train_ind, ]
testy<-subset(testing, select = TN)
trainy<-subset(training, select = TN)
#fitting PLSR model
m.pls <- plsr(TN ~.,data=training, validation="LOO", method = "oscorespls")
summary(m.pls)
#optimizing the number of components
comp <- which.min(m.pls$validation$PRESS)
#key wavelengths selection using VIP
vip <- VIP(m.pls, 1)
matplot(vip)
matplot(scale(cbind(vip)), type = 'l')
write.table(vip, "C:/labderivVIP.csv",sep = "," ,row.names = T,col.names = T)
```

Appendix B – RF-Boruta code

```
#remove previous datasets and clean up the workspace
rm(list=ls())
#loading libraries
library(Boruta)
library(mlbench)
library(caret)
library(randomForest)
#loading the data
data <- read.csv("C:/TN.csv")
str(data)
#feature selection
set.seed(111)
boruta <- Boruta(TN ~ ., data = data, doTrace = 2, maxRuns = 500)
print(boruta)
#plot Boruta
plot(boruta, las = 2, cex.axis = 0.7)
#plot importance history
plotImpHistory(boruta)
#decision about tentative attributes
bor <- TentativeRoughFix(boruta)
print(bor)
#list all the important and unimportant variables
attStats(boruta)
#list the important variables
getConfirmedFormula(boruta)
#test to see if extracted features help to improve accuracy
#load training and testing datasets
train<-read.csv("C:/training.csv")
test<-read.csv("C:/testing.csv")
```

```
#random forest model for classification
set.seed(333)
rf1794 <- randomForest(TN~., data = train)  #1794 to indicate that this model will be based on all the 1794 wavelengths.
rf1794
summary(rf1794)
#prediction test
p <- predict(rf1794, test)
p
RMSE(p, test$TN)
#random forest prediction on important variables only
rf3<- randomForest(TN ~  X446 + X450 + X451, data=train) #3 to indicate three important variables.
p2 <- predict(rf3, test)
RMSE(p2, test$TN)
```