

https://dspace.org/

Predicting soil organic carbon in a small farm system using in situ spectral measurements and the random forest regression

Bangelesa, Freddy Fefe

2017

http://hdl.handle.net/10539/23485

Bangelesa, Freddy Fefe (2017) Predicting soil organic carbon in a small farm system using in situ spectral measurements and the random forest regression, , University of the Witwatersrand, Johannesburg, <http://hdl.handle.net/10539/23485>

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



UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

Predicting soil organic carbon in a small farm system using *in situ* spectral measurements and the random forest regression

Student: Bangelesa Fefe Freddy (1482954)

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 (Geographical Information Sciences and Remote Sensing)

Supervisor: Dr. Elhadi Adam

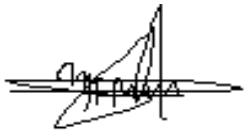
Co-supervisor: Prof. Jasper Knight

Johannesburg, 2017

DECLARATION

I, Freddy Fefe Bangelesa, 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 Sciences 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.



24th day of March 2017 in Johannesburg

Abstract

Soil organic carbon is considered as the most determining indicator of soil fertility. The purpose of this research was to predict the soil organic carbon in the Mokhotlong region, eastern of Lesotho using *in situ* spectral measurements and random forest regression. Soil reflectance spectra were acquired by a portable field spectrometer.

The performance of random forest regression was assessed by comparing it with one of the most popular models in spectroscopy, partial least square regression. Laboratory spectroscopy measurements of the soil samples were analysed for assessing the accuracy of *in situ* spectroscopy based-models. The effect of the Savitzky−Golay first derivative in improving partial least square regression and random forest regression in both spectral data was also assessed.

The results indicated that the random forest regression could accurately predict the soil organic carbon contents on an independent dataset using *in situ* spectroscopy data (RPD = 3.77, R_p^2= 0.88, RMSEP = 0.64%). The overall best predictive model was achieved with the derivative laboratory spectral data using random forest with the optimum number of key wavelengths (RPD = 3.77, R_p^2= 0.88, RMSEP = 0.64%). In contrast, partial least square regression was likely to overfit the calibration dataset. Important wavelengths to predict soil organic contents were localised around the visible range (400-700 nm). An implication of this research is that soil organic carbon can accurately be estimated using derivative *in situ* spectroscopy measurements and random forest regression with key wavelengths.

Keys word: soil organic carbon, spectroscopy, regression models.

Acknowledgements

Above all, I would like to sincerely thank my supervisor, Doctor Elhadi Adam, for his great supervision, brilliant ideas and orientation toward the present topic. My special acknowledgement goes to my co-supervisor, Professor Jasper Knight for his commitments, effort and patience throughout this research.

I also express my gratitude to all my lecturers for providing me enough background on remote sensing. My thanks also go to all the staff members involved in the Masters program of Geographic Information Sciences and remote sensing.

My thanks also to Professor Gerhard Bringmann and all his teams for granting me the international excellence scholarship BEBUC (Bourse d Excellence Bringmann aux Universite Congolaises) for a Master degree.

Table of Contents

Abstract ..... **i**

Acknowledgements ..... **ii**

Lists of Figures ..... **vi**

List of Tables ..... **viii**

List of Abbreviations ..... **ix**

CHAPTER I: INTRODUCTION **1**

I. 1. Background 1

I.2. Problem statement 3

I. 3. Research questions 5

I. 4. Aim and objectives 6

I.5. Structure of the thesis 6

I.6. Scope of study 6

CHAPTER II: LITERATURE REVIEW **7**

II.1. Soil organic carbon 7

II.1.1. Role of SOC in agriculture 7

II.1.2. SOC and climate change 8

II.2. Spectral reflectance and SOC properties 9

II.2.1. Characteristics of the soil spectral reflectance 9

II.2.2. SOM and soil reflectance 10

II.3. Remote sensing of SOC 12

II.4. Spectroscopy modelling methods of SOC 14

II.4.1. Linear models 15

II.4.2. Non-linear models 16

III.1. Study area 19
III.2. Soil sampling and spectral measurements 20
III.2.1. Spectral measurement at the field level 21
III.2.2. Spectral measurement at the laboratory level 21
III.3. SOC analysis 22
II.4. Spectra pre-processing and transformation 23
III.5. Statistical analysis 23
III.5.1. PLSR and variable selection 24
III.5.2. Random forest regression and variable selection 26
III.5.3. Model validation 27
CHAPTER IV: RESULTS 30
IV.1. SOC sample descriptive analysis 30
IV.2. Comparison of field and laboratory spectra 31
IV.3. Key wavelength selection 32
IV.3.1. VIP algorithms 32
IV.3.2. Recursive feature selection and percent increase in MSE 33
IV.3.3. Position of key wavelengths and interpretation 34
IV.4. Model development 35
IV.4.1. PLSR optimum number of component 36
IV.4.2. PLSR model performances within the laboratory dataset 37
IV.4.3. PLSR model performances within the field dataset 38
IV.4.4. RF regression model performances within the laboratory dataset 39
IV.4.5. RF regression models performance within the field dataset 40
IV.5. Comparison between RF regression and PLSR 41
CHAPTER V: DISCUSSION 43

V.1. Suitable spectral bands for SOC modelling 43

V.2. Performance of the RF regression compared to the PLSR regression in predicting SOC 44

V.3. Performance of field spectral measurements and laboratory spectral measurements in predicting SOC.....45

V.4. Influence of spectra first derivative on the performance of different models 45

V.5. Significant and limitations of the models.....46

CONCLUSIONS 47

REFERENCES.....48

APPENDICES 59

Lists of Figures

Figure 1. Spectral reflectance of three types of soil, giving examples of the impact of different biochemical properties on feature absorption (Rossel and Behrens, 2010) 10

Figure 2. Average spectral reflectance curves of three different depths with different OM of two soil types: (a) lithic distrochrept, and (b) typic fluvent (Mathew *et al.*, 1973). 11

Figure 3. Site location showing the position of field samples taken. 19

Figure 4. Field based spectroscopy for SOC content measurement: (a) Spectral measurement by ASD; (b) site localisation (Pictures captured by Dr. Elhadi Adams). 21

Figure 5. SOC analysis: (a) crushing; (b) sieving and (c) burning. (Pictures captured by Freddy Bangelesa).....22

Figure 6. Spectral data: (a) original spectra; (b) with noisy regions removed (1350–1640, 1790–1960, 2350–2500 nm); (c) and first derivative spectra. 23

Figure 7. Conceptualization of (a) MLR and (b) PLRS (Martens and Martens, 1986). 25

Figure 8. Hotelling's T2 for multivariate outlier detection. 30

Figure 9. Average laboratory and field VIS-NIR spectra; (black): field spectra data; (blue): laboratory spectra data. 31

Figure 10. Variable important in projection for key wavelength selection: (a) field raw spectra, (b) laboratory raw spectra; (c) field first derivative spectra; (d) laboratory first derivative spectra. 32

Figure 11. Recursive feature selection: (a) field raw spectra, (b) laboratory raw spectra; (c) field first derivative spectra; (d) laboratory first derivative spectra. 33

Figure 12. Percent increase in MSE: (a) field raw spectra (b) laboratory raw spectra; (c) field first derivative spectra; (d) laboratory first derivative spectra. 34

Figure 13. PLSR optimum number of components: (a) field PLSR raw model with all wavelengths; (b) laboratory PLSR raw model with all wavelengths; (c) the field first derivative PLSR model; (d) the laboratory first derivative PLSR model; (e) field PLSR model with key wavelengths; (f) the laboratory PLSR model with key wavelengths; (g) the field first derivative PLSR model with key wavelengths; (h) the laboratory first derivative PLSR model with key wavelengths. 37

Figure 14. Performance of PLSR in predicting SOC on an independent laboratory spectral dataset : (a) laboratory PLSR raw model with all wavelengths; (b) the laboratory first derivative PLSR model; (c) the laboratory PLSR model with key wavelengths; (d) the laboratory first derivative PLSR model with key wavelengths 38

Figure 15. Performance of PLSR in predicting SOC on an independent field spectral dataset: (a) field PLSR raw model with all wavelengths; (b) the field first derivative PLSR model; (c) field PLSR model with key wavelengths; (d) the field first derivative PLSR model with key wavelengths. 39

Figure 16. Performance of RF regression in predicting SOC on an independent laboratory dataset: (a) laboratory RF raw model with all wavelengths; (b) the laboratory first derivative RF model; (c) the laboratory RF model with key wavelengths; (d) the laboratory first derivative RF model with key wavelengths. 40

Figure 17. Performance of RF regression in predicting SOC on an independent field spectral dataset: (a) the field RF raw model with all wavelengths; (b) the field first derivative RF model; (c) field RF model with key wavelengths; (d) the field first derivative RF model with key wavelengths. 41

List of Tables

Table 1. The most frequent NIR bands of organic compounds (Stuart, 2004). 12

Table 2. Examples of different sensors/platforms performance in prediction SOC content, including calibration and predicted models. 14

Table 3. Waveband range of Analytical Spectral Devices (ASD, 2005). 20

Table 4. Interpretation of R^2 in calibrating SOC using NIR spectroscopy (Williams, 2001). 28

Table 5. Descriptive statistics of SOC contents (in %) within three different datasets. 31

Table 6. Comparison of VIP algorithm and recursive feature selection combined with percent decrease in MSE in selecting key wavelengths. 35

Table 7. Performance of all RF regression and PLSR models in the calibration and validation datasets. 42

List of Abbreviations

ANN:	Artificial Neural Networks
ASD:	Analytical Spectral Devices
BT:	Boosted Trees
CARS:	Competitive Adaptive Reweighted Sampling
DW:	Dry Weight
F	Field
FD	First order Derivative
Gt:	Gigatone
K	Key wavelength
L	Laboratory
LOI:	Loss Ignition Method
MARS:	Multivariate Adaptive Regression Spline
Mg:	Megagrams
MIR:	Mid Infrared
MLR:	Multilinear Regression
NIR:	Near Infrared
p.p.m:	parts er million
PCA:	Principal Component Analysis
PCR:	Principal Component Regression
Pg:	Petagrams
PLSR:	Partial Least Square Regression
R^2:	coefficient of determination
RF:	Random Forest
RMSE:	Root Mean Square Error
SMLR:	Stepwise Multiple Linear Regression.
SOC:	Soil Organic Carbon

SOM	Soil Organic Matter
SVM:	Support Vector Machine
SWIR:	Shortwave Infrared
UV:	Ultra Violet
VIS:	Visible
VIS-NIR:	Visible and Near Infrared

CHAPTER I: INTRODUCTION

I. **1. Background**

Soil organic Carbon (SOC) is a core constituent of the soil organic matter (SOM) which plays a primary role in soil physical and chemical properties. In an agricultural perspective for example, SOC content has a huge impact on bulk density, nutrient availability, water retention, structural stability, hydraulic conductivity, and soil biodiversity. SOC is the most important terrestrial global carbon pool with approximately 1600 Pg (Petagrams) for 1 m of the soil depth (Novara *et al.*, 2011). It also constitutes a very important parameter in climate regulation and support of primary production of ecosystems (Brevik *et al.*, 2015).

Under healthy and non-eroded landscapes, soil organic content remains stable over time and the mineralization of the soil carbon is compensated by the production of SOM (Novara *et al.*, 2011), and an important quantity of SOC is stored as partially decomposed SOM (Schulze *et al.*, 2000). The remaining SOC which is not mineralized is slowly oxidized and stabilized as humic substances. The dynamics of SOC is induced by both microbial activities (Guénon *et al.*, 2013), and abiotic process enhanced by external factors, among which erosion is the most fundamental one. Soil erosion is perceived as a major cause of SOC depletion on the arable layer (Lal, 2005), especially in uncovered mountainous regions where the slope is steep, transported by runoff. Both the kinetic energy of the impacting raindrops and water runoff force separate aggregates and expose SOM (Lal, 2005). Therefore, SOC which is concentrated in the top soil is preferentially removed because of its low density (Lal, 2005). In the United State for example, soil erosion induced by water is estimated at 15 Pg per year (La *et al.*, 1998). The lateral movement of soil can drastically modify the spatio-temporal variability of SOC within a landscape or field (Lal, 2009).

SOC depletion has a negative impact on food security. Lal (2009) underlined that the poor quality of the soil coupled with the diminution of the SOC, were major factors in food insecurity. The situation seems to be alarming by looking at the trend of soil degradation and food demand. A variety of researchers estimated that current production must be doubled by 2050 in order to respond to the world population expansion (Lal, 2009). Consequently, a need for fertile soil is relevant and SOC monitoring becomes more and more important.

Accurate measurement of SOC at different spatio-temporal scales is a big challenge (Kuhn *et al.*, 2009) because there is not yet any conventional method approved by scientists. However, many techniques of estimating SOC are based on grid sampling of SOC, repeated overtime. The drawback of this grid sampling technique is that it is very costful, time consuming and cannot produce a spatially-continuous map of SOC (Goidts and van Wesemael, 2007). Many researchers concluded that a sampling interval of less than 50 m is required to capture SOC spatial heterogeneity. More spatial advanced technologies need to be implemented (Stevens *et al.*, 2008).

In this respect, remote sensing is considered as a low cost, reproducible and rapid method of offering quantitative and continuous maps of SOC (Gehl and Rice, 2007). SOC modelling using remote sensing data is possible through the correlation between soil reflectance, soil colour and soil organic content. Satellite and airborne remote sensing have largely contributed to the assessment of SOC, but their major drawback is that they cannot discriminate SOC from partially covered vegetation (Rossel and Behrens, 2010). They generate a mixed pixel from soil and vegetation together. In addition, it is difficult to estimate the SOC using satellite and airborne remote sensing when concentrations are small because it results in very weak signal (Rossel and Behrens, 2010).

Spectroscopy approach in visible and near infrared bands (VIS-NIR, 400-2500 nm) is commonly used to relate the SOC contents to laboratory spectral measurements and *in situ* spectral measurements (field spectroscopy measurements). This approach is rapid and non-destructive (Guénon *et al.*, 2013). Under laboratory conditions, SOC contents are measured with high precision and accurate (Cohen *et al.*, 2007), while field spectroscopy measurements) are disturbed by atmospheric conditions, but are practical when laboratory facilities are not available (Reeves, 2010).

Many statistical regressions have been previously tested to relate the SOC concentration to spectral data (Brown *et al.*, 2006) using spectroscopy approach. Based on the root mean square error (RMSE) and coefficient of determination (R^2), most investigations reveal that, among statistical regressions, PLSR performs best, followed by principal component regression (PCR), multi linear regression (MLR) and PCR (Vasques *et al.*, 2008) when compared to other linear models. Machine learning algorithms exhibit good results when dealing with non-linear trends and more complex data (Rossel and Behrens, 2010). However,

results from these models vary according to the type of soil, thus cannot be applied everywhere. Rossel and Behrens (2010) explained it by the fact that soil absorption features are likely to overlap and shift in the location. That is why, it is very important to develop local models according to different regions. To the best of our knowledge, there are not investigations addressing SOC modelling using either laboratory or *in situ* measurements in southern Africa, especially in mountainous regions where SOC is vulnerable to erosion and where spectral reflectance is most variable.

I.2. Problem statement

SOM is of highest importance to local farmers because of its primary importance in the fertility of the soil (Van-Camp *et al.*, 2004). In the Lesotho highlands, southern of Africa, more than 86% of the Basotho population dwell in rural areas and depend on subsistence agriculture (Eash *et al.*, 2013). The main crops are maize, sorghum, peas, beans, wheat, oil seeds, nuts, soya, and potatoes.

However, agricultural productivity and its share in gross domestic product have been declining in Lesotho (Sicili, 2010). The country produces less than 30% of food consumed by its population as compared to 50% produced in the 1980's. This situation is largely explained by high soil erosion rates.

Lesotho is characterized as one of the African countries with the highest eroded landscape (Showers, 2005). Soil erosion is a significant problem due to a combination of geologic, climatic, ecological and human factors (Grab and Nüsser, 2001; Mbata, 2001; Meadows and Hoffman, 2002). Weathering of the underlying Jurassic basalts has produced a low-strength mixture of silica and expansive clay minerals (Bell and Haskins, 1997). Plagioclase within the basalts has been affected by zeolitization and chloritization, and olivine in particular has been replaced by iron oxides, serpentine and clays (mainly montmorillonite) (Garzanti *et al.*, 2014). These weathering products make the resulting soil susceptible to erosion by surface sheetflow, subsurface clay expansion, slaking and soil piping, and by landslide/debris flow activity caused by subsurface waterlogging and failure (e.g. Edwards *et al.*, 2016). An important quantity of arable soil, estimated at 40 Pg, is brought annually by soil erosion, resulting in SOC depletion resulting in loss of soil fertility.

This situation is amplified by the vulnerability of the country towards climate change. In 2007 for example, Lesotho was affected by a very severe drought generating global food insecurity (Showers, 2005).

Hence, precise and conservation agriculture which imply a good management of SOC as a proxy for soil fertility is very crucial in order to enhance agricultural activities by maintaining soil fertility. This can be done by estimating the concentration of SOC.

However, SOC content estimation using field sampling is difficult for Lesotho highlands due to its poor accessibility (mountains) and heterogeneity of the region, which would probably require many soil samples. Laboratory based measurements are limited because of the lack of facilities all over the region. Therefore, *in situ* spectral measurements are the most suitable approaches because they are practical and fast (Reeves, 2010).

We assume that RF regression will prove to be good in predicting SOC under the field conditions because of its high performance in other remote sensing applications, can handle non-linear data well (Rossel and Behrens, 2010) and is not complex to be optimized. In order to assess its performance, it will be compared with PLSR.

I. 3. Research questions

The research questions for this study are:

- What is the performance of the RF regression compared to the PLSR regression in predicting SOC in highlands of Lesotho?
- Which are the most suitable spectral bands for SOC modelling using different regression equations?
- What is the performance of field spectral measurements compared with laboratory spectral measurements in predicting SOC?
- What is the impact of spectra first derivative on the performance of different models?

I. 4. Aim and objectives

The aim of this research is to evaluate the performance of field spectral measurements in predicting SOC in an agricultural system by comparing two regression models.

The specific objectives of this study are:

- To identify key wavelengths for developing a reliable SOC estimation model using RF and PLSR methods;
- To test the performance of the RF and PLSR regression models in predicting SOC;
- To test the effects of spectra first derivative on the performance of different regression models.

I.5. Structure of the thesis

Apart from the introduction and conclusions, this research report is divided into four chapters. The second chapter reviews different topics related to this research, SOC, spectral reflectance, remote sensing of SOC and spectroscopy modelling. Historical and contemporary literatures on the topic under investigation will be discussed in order to identify the knowledge gaps. The third chapter will explain the methodological approaches by describing the site location, sampling design, spectra measurements, chemical and statistical analysis. The fourth chapter will present different findings of the research. The discussion of the results and its wider implication will be addressed in chapter five.

I.6. Scope of study

This study is limited at developing spectral model using random forest and PLS regressions to quantify SOC concentration. Mapping consideration was not covered in this research. Due to the brevity of the time, the investigation was conducted in a specific farm system located at the east of Lesotho.

CHAPTER II: LITERATURE REVIEW

II.1. Soil organic carbon

Soils contain carbon both in organic and inorganic forms. Inorganic soil carbon is a product of both carbonic acid and of weathering of rocks in the soil, precipitating as carbonate minerals (Lal, 2009), while organic soil carbon comes from soil organisms, manure, branches, plant roots and leaf litter (Walcott *et al.*, 2009). SOC is the carbon occurring in the SOM and represents almost 58 percent of the SOM on average (Corsi *et al.*, 2012). SOC is strongly influenced by human activities and environmental conditions like shape, geology, climate and time (Walcott *et al.*, 2009)

II.1.1. Role of SOC in agriculture

SOC storage is presently one of the most topical research fields because of greenhouse gases increase, food demand expansion and severity of human-induced soil degradation (Brevik *et al.*, 2015). SOC increase has a positive impact on agricultural productivity because many organisms (insects, spiders, snails, mites, nematodes and some mammals) and microorganisms (bacteria, fungi, algae and protozoa) use the SOM as food (Walcott *et al.*, 2009). Researchers have qualified SOC like a 'universal keystone variable' in the management of soil quality (Loveland and Webb, 2003) making it the most important indicators for managing soil fertility in sub-Saharan Africa. In this region, chemical fertilizers are sometimes not accessible because of the farmers' low income (Hossain, 2001).

Compared to chemical fertilizers, SOC is well integrated in smallholder farmers' communities, not only because it is affordable, but also farmers are aware of its implications for fertility (Hossain, 2001). Farmers already know that dark or black soil is associated with a high agricultural productivity (Hossain, 2001).

SOM participates in nutrient storage and exchange, constitutes an important parameter of cation exchange capacity, and improves permeability, aeration, infiltration, aggregate stability and structure. It is the source of phosphorus (Walcott *et al.*, 2009), and reservoir of water and nutrients. In relation with the erosion control, SOC contributes to the stabilization of other parts of the soil and formation of aggregates which make the soil more resistant to erosion. The SOC participates also in the absorption of many pesticides and buffers the soil against pH changes (Walcott *et al.*, 2009).

Concerning its interaction with soil water, SOC increases the infiltration rate, as well as the holding capacity of the soil water (Walcott *et al.*, 2009).

II.1.2. SOC and climate change

SOC is not only important for agriculture but, also plays a crucial role in climate change. According to the Intergovernmental Panel on Climate Change (2014), the concentration of greenhouse gases has increased from 280 ppm (parts per million) to 349 ppm for CO_2 in the atmosphere between the pre-industrial era and 2005. As a consequence, the global average temperature has risen (from 13.6 °C to 14.4 °C) all over the 20^{th} century, as well as the sea level (15.2 cm to 22.9 cm). The arctic average cover sea ice has decreased at the rate of 2.7% per decade.

Between 1850 and 2000, the fossil fuel combustion was the major source of CO_2 in the atmosphere, but early scientists proved that from the 1940s to now, a big quantity of CO_2 was released by terrestrial sources rather than from fossil fuels (Lal, 2009). The soil contains more than 1500 Gt of carbon and is known as the greatest terrestrial carbon pool (Smith, 2008). A small release of CO_2 and CH_4 from the decomposition of SOC in the atmosphere will have adverse impacts in the carbon cycle.

In Europe for example, SOC storage on farms can approximate 20% of the global reduction needed during the first commitment period of the Kyoto Protocol (8% of reduction between 2008 and 2012 from a 1990 base) (EU Soil Thematic Strategy, 2004). This role makes it a good proxy for land degradation assessment.

Currently, the post-Kyoto agreements are attempting to consider SOC in the carbon trade (United Nations, 2015), but much still has to be done, among which, the ability to monitor, report and verify the levels of SOC are the major concerns (Walcott *et al.*, 2009). The fact that SOC may be considered in carbon trading is a great opportunity for developing countries to be involved in the carbon trade by selling carbon credits from sustainable soil management through precision agriculture.

II.2. Spectral reflectance and SOC properties

II.2.1. Characteristics of the soil spectral reflectance

Soil spectral reflectance is mainly affected by biochemical determinants such as SOM, soil moisture and soil mineralogy, and physical structure such as particle size and surface roughness (Lobell and Asner, 2002; Shepherd and Walsh, 2002). Infrared spectroscopy is governed by the principle of the radiation absorbance at molecular vibration frequencies (Soriano-Disla *et al.*, 2014). Soil spectral signatures are explained by the reflectance of the electromagnetic spectrum as a function of wavelength (Ben-Dor *et al.*, 1997). The vibrational stretching and bending structure of atoms and their electronic transitions define the spectral absorption features.

Vibrations of atoms mostly occur in the thermal and mid-infrared bands (2500-25000 nm), with weaker signals located at the VIS-NIR (Soriano-Disla *et al.*, 2014). Carboxyl, hydroxyl and amine functional groups are biochemical groups which are related to absorption features (Al-Abbas *et al.*, 1972). NIR works according to the absorption of solar radiation in the NIR (780-2500 nm) and interacts with C-H, N-H and O-H (Workman and Shenk, 2004). Atoms such as C-O, C-C and O-H do not exhibit absorption of the soil spectra in the NIR region, except when stronger soil absorbers like C-H, N-H and O-H do not dominate weak absorbance (Kramer *et al.*, 2004).

Molecular electronic transition is another phenomenon caused by the excitation of electrons from lower to higher energy levels. According to Rossel and Behrens (2010), the phenomenon is moreover correlated with iron oxides and enhances absorption features in the visible region. It is important to underline that absorption features can slightly differ according to the type of soil. Figure 1 describes different types of soil with the location of soil spectral absorption features. In the visible portion, broad absorption feature are correlated with SOM and iron oxide (FeO_2). For upland peat, broad absorption features are between 677 and 1108 nm as shown by Rossel and Behrens (2010). Specific absorption features around the NIR and SWIR are correlated with peak constituents such as lignin and cellulose at 1120 nm and 2100 nm, respectively, as noted by McMorrow *et al.* (2004). Fine absorption features of soil in SWIR are related to clay and carbonates, at around 2200 nm and 2300 nm,

respectively. Strong absorption caused by water can be perceived around 1400 nm and 1900 nm, but also slightly around 950 nm and 1200 nm (Rossel and Behrens, 2010).

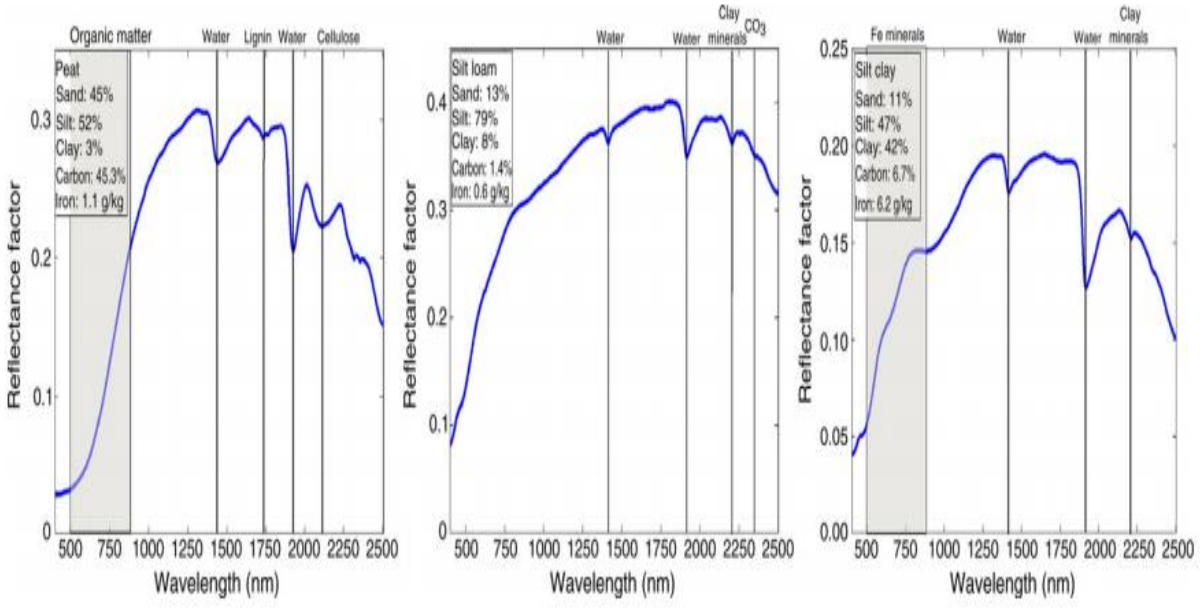


Figure 1. Spectral reflectance of three types of soil, giving examples of the impact of different biochemical properties on feature absorption (Rossel and Behrens, 2010)

Many researchers have used diverse portions of the electromagnetic spectrum to predict SOC, but the most common are the VIS (visible), NIR and MIR (mid infrared) regions. The MIR provides better results compared to VIS, NIR, VIS-NIR because it contains much more soil spectral information (Soriano-Disla *et al.*, 2014). However, MIR spectrometers are not commonly used because they are easily affected by water content and sample heterogeneity (Soriano-Disla *et al.*, 2014). However, the MIR is not always superior to NIR to predict SOC. Some studies reported that both are similarly powerful to predict the SOC (Stevens *et al.*, 2008; Vohland *et al.*, 2014).

II.2.2. SOM and soil reflectance

It is unanimously recognized that there is soil reflectance which is correlated with the SOM (Aber *et al.*, 1990). According to Hoffer and Johannsen, (1969), SOM in the portion of 400-2500 nm, is inversely proportional to the total reflectance. Dematte *et al.* (2003) found higher reflectance around 450-2500 nm after removing the SOM component using 30% H_2O_2. Similarly, Mathew *et al.* (1973) confirmed the same results after removing SOM component using 10% H_2O_2. Figure 2 illustrates how different layers of the soil associated with different

proportions of SOM can modify the spectral reflectance of the soil. The conclusion is that the mode the depth is big corresponct to the

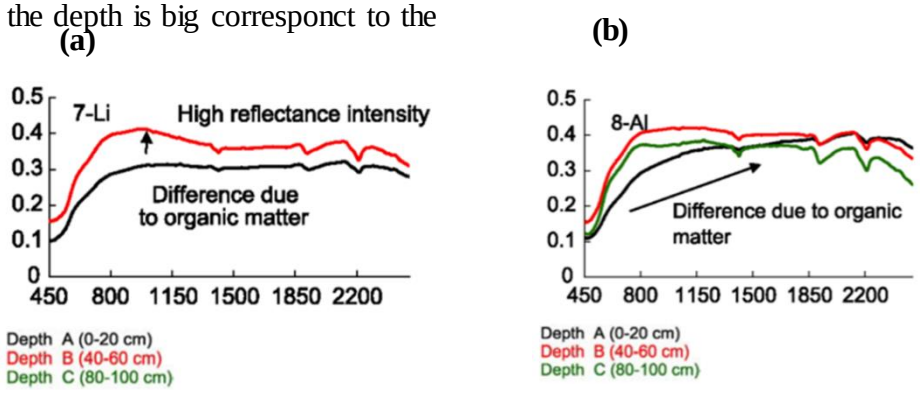


Figure 2. Average spectral reflectance curves of three different depths with different SOM of two soil types: (a) lithic distrochrept, and (b) typic fluvent (Mathew *et al.*, 1973).

In addition to the quantity of SOM in soil, different decomposition stages of SOM can change the spectral reflectance of the soil. Stoner and Baumgardner (1981) explained it by comparing three organic soils with different ages, sapric (completely decomposed), hemic (intermediately decomposed), and fibric (slightly decomposed). They found that fibric material which has a high level of (84.8%) shows the highest reflectance, followed by hemic (54.4%) and sapric (76.4%). Similar results were also found by Ben-Dor *et al.* (1997). This can be explained by the presence of many absorption features occuring in the functional groups. The most important difference observed in the slope value between 400-1100 nm and this slope decreases with increasing decomposition age.

Some scientists have oriented their studies more on the percentage below or above which soil may affect spectral reflectance. Different researchers came up with different results. Baumgardner *et al.* (1970) found that below 2.0%, SOM had a non-significant effect on soil reflectance. Montgomery (1976) found a content of 9% did not inhibit the effect of some soil characteristics on spectral reflectance. Nonetheless, He *et al.* (2005) could accurately predict soil content with a range between 1.06 and 1.65% using PLSR in NIR region. This result can be justified because the spectral resolution was high (Ben-Dor, 2002). The most important portions of the NIR for predicting SOC content are presented in Table 1. Visible portions are more related to the third overtone N-H stretching.

Table 1. The most frequent NIR bands of organic compounds (Stuart, 2004).

Wavelength (nm)	Assignment
2200-2450	Combination C-H stretching
2000-2200	Combination N-H stretching, combination O-stretching
1650-1800	First overtone C-H stretching
1400-1500	First overtone N-H stretching, first overtone O-H stretching
1300-1420	Combination C-H stretching
1100-1225	Second overtone C-H stretching
950-1100	Second overtone N-H Stretching, second overtone O-H stretching
850-950	Third overtone C-H stretching
775-850	Third overtone N-H stretching

II.3. Remote sensing of SOC

Many remote sensing approaches have been used to estimate SOC in the last couple of decades. They are mostly based on remote spectroscopy (hyperspectral), on satellite or airborne mounted, and field and laboratory spectroscopy.

Laboratory measurements have been widely used to predict SOC contents (McCarty *et al.*, 2002). This is because they are less expensive and faster than traditional estimation of SOC. However, steps like collection, grinding, sieving and drying of soil which are crucial during this process, makes it a little bit slow when compared to field measurements (Stevens *et al.*, 2008). The laboratory method is not only the most widely used, but the most accurate due to its high analytical precision. Laboratory measurements is recognized as an alternative for traditional SOC content estimation.

In case where laboratory facilities are not available, field spectroscopy measurements are preferable (Reeves, 2010). Field spectroscopy measurements are mostly used to quantify SOC content within a field (small scale) and offers many advantages for application like precise agriculture (Barnes *et al.*, 2003). Accommodated with high sampling interval, field spectroscopy is also useful for temporal change of SOC over short time periods. More information about the application of field spectroscopy are provided by Milton *et al.* (2009). In general, analytical spectral devices (ASD) such as AgriSpec and Fieldspec are mostly used as measuring instruments (Rossel *et al.*, 2009). In precise agriculture for exemple, ASD is mostly mounted on tractors (Bricklemyer and Brown, 2010) to measure the soil properties.

Field spectroscopy measurements are generally less accurate compared to laboratory measurements because of the surface roughness and moisture contents (Christy, 2008; Morgan *et al.*, 2009). However, Stevens *et al.* (2010) demonstrated that field spectroscopy measurements can be as accurate as laboratory measurements. All of these results are specific to different characteristics of the study area. Stevens *et al.* (2010) compared the efficiency of the laboratory, field and airborne spectroscopy to predict the SOC using PLSR. They have concluded that the RMSE of the field spectroscopy was similar to that of the Walkley and Black method and airborne spectroscopy was inaccurate.

Satellite sensors as well as airborne ones can be seen as a great opportunity to monitor SOC because of the temporal repetitiveness of the satellite and the large field of view, but few studies have addressed the contribution of satellite images in the assessment of SOC. In most cases, empirical models which integrate phenomena (land management, clay, topography and moisture) that affect the spatiotemporal dynamic of SOC are used as covariates (Croft *et al.*, 2012). Those models are applied on large scales. Physical based models which use the spectral information are limited by some factors such as the signal noise/ratio, soil type, soil moisture, soil roughness, bidirectional reflectance distribution function and variable spatial resolution and atmospheric disturbances (Vasques *et al.*, 2008). Satellite images are not effective to model SOC. Gomez *et al.* (2008) evaluated the performance of the Hyperion sensor to predict the SOC; they found an R^2 of 0.51 and suggested to investigate on the EnMAP hyperspectral satellite, a German sensor which can hold a good Signal to Noise Ratio. In contrast to satellite sensors, airborne platforms have shown a good performance with R^2 between 0.62 and 0.97. Selige *et al.* (2006) developed a multivariate statistical regression to model SOC concentration using EnMAP sensor, and found a result of R^2 = 0.89. An example of the SOC model developed with different platforms/sensors associated with accuracy assessment are provided in Table 2.

Table 2. Examples of different sensors/platforms performance in prediction SOC content, including calibration and predicted models.

Lab/ground/ air/sat	Geographic origin	Sensor	Predictive model	Wavelengths/ Ranges (nm)	Var R^2	RMSE	References
Laboratory	NSW, Australia	AgriSpec VIS-NIR Spectrometer	PLSR	350-2500	0.82	0.96	Rossel and Behrens, (2010)
			MARS		0.80	1.02	
			SVM		0.84	0.92	
			BT		0.62	1.49	
			RF		0.71	1.23	
Laboratory	Santa Fe river, Florida, USA	Quality Spect pro SpectroRadiometer	SMLR	350-2500	0.82	0.199	Vasques *et al.* (2008)
			PCR		0.76	0.224	
			PLSR		0.82	0.193	
			RT		0.67	0.191	
Laboratory	Turkey and UK	LabSpec2500 Near Infrared Analyzer	ANN	350-2500	0.92	0.82	Mouazen (2014)
Laboratory	Qinghai-Tibet, China	ASD FieldSpect Portable SpectroRadiometer	PLS	400-2500	0.85	7.28	Li *et al.* (2015)
Laboratory	Hubei and Jiangsu	ASD FieldSpect Pro FR	SPA	350-2500	0.73	2.78	Peng *et al.*(2014)
Ground	Ortho and Attert, Belgium	ASD FieldSpect Pro FR	PLS	400-2500	0.82	1.00	Stevens *et al.* (2008)
Ground	Qinghai-Tibet, China	ASD FieldSpect Pro FR	LS-SVM	400-2500	0.81	8.40	Li *et al.* (2015)
Aerial	Ortho and Attert, Belgium	CASI	PLSR	405-950	0.85	-	Stevens *et al.* (2010)
Aerial	Luxemburg	AHS-160 sensor	PLSR	430-2540	0.89	-	Stevens *et al.* (2010)
Satellite	NS, Australia	Hyperion	PLSR	400-2500	0.51	-	Gomez *et al.* (2008)

II.4. Spectroscopy modelling methods of SOC

There are various calibration methods to estimate the SOC but the most useful in the literature review are addressed here: partial least square regression, support vector machine regression, random forest regression, artificial neural network regression, principal component regression, least square regression and stepwise linear regression.

II.4.1. Linear models

Initially, (multi linear regression) MLR was applied to predict SOC using spectral data, but the issue related to multicollinearity was a major problem. Multicollinearity is a source of many uncertainties in model interpretations and decreases the model performance (Martens and Martens, 1986). This problem can be overcome by using least square regression like variable subset selection, ridge regression, PCR, and PLSR. These methods assumed that there is lack of correlation among predictors (Adnan *et al.*, 2006). In addition to the multicollinearity problem, in much software, multiple linear models cannot be computed when the number of variables is superior to the number of observations.

Stepwise multiple linear regressions also can deal with highly correlated predictors. A subset of predictors that correctly explain the response variable can be identified. However, the interpretability is generally decreased (Martens and Martens, 1986). Hruschka (1987) noticed that when the sample size is great enough, the model leads to the problem of over fitting.

Principal component analysis (PCA) is one of the statistical analyses applied to reduce the dimensionality of variables. PCA reduces it by creating uncorrelated new latent variables or components (Adnan *et al.*, 2006). Latent variables form linear combinations of the original variables so that the first component explains the most variation of the new data, followed by the second component which is orthogonal to the first one. Latent variables are uncorrelated to each other (Adnan *et al.*, 2006). This technique was for the first time used for the analysis of soil spectral reflectance by Condit (1970). In contrast, Adnan *et al.* (2006) stated that it performs less well when data are uncorrelated and when one is dealing with a great amount of data, as in spectroscopy. The manual selection of variable becomes very difficult.

Currently, PLSR is commonly used to predict SOC (Li *et al.*, 2015). Its popularity is explained by the fact that it handles well with accessible and easy-manipulated software and is not complex to understand and interpret (Soriano-Disla *et al.*, 2014). PLSR has an ability to reduce multi-dimensional data, mainly when there are a greater number of predictor variables than observations (Boulesteix and Stimmer, 2007). Furthermore, PLSR is also the best among linear regression models compared to the stepwise multiple linear regression (SMLR), principal components regression tree and committee trees (Vasques *et al.*, 2008). The superiority of PLSR over other linear regressions is attributed to the fact that PLSR: (1) selects predictors automatically, (2) handles well with a diverse of works, and (3) is

recommended for large datasets because the computation is very fast (Boulesteix and Stimmer, 2007).

PLSR requires linearity between spectral data and chemical components, but this is not always the case (Peng *et al.*, 2014). This prediction method can be improved by using for example the local PLSR when samples are collected from many land uses. For instance, Nocita *et al.* (2014) performed local PLSR by using the laboratory based VIS-NIR spectroscopy to model the SOC all over the European Union. The local regression was improved by adding some covariates (geomorphological and texture information). The prediction was applied in cropland, grassland and woodland. The prediction was accurate under cropland and grassland and less accurate under woodland and organic soils. The prediction under woodland and organic soils was largely improved by adding the sand content as a covariate. The final result showed the ability of the local PLSR to deal with large datasets. Many researchers have shown that the use of key spectrum can make the calibration model more robust. Vohland *et al.* (2014) tested the performance of the competitive adaptive reweighted sampling (CARS) to select suitable bands that would be calibrated in PLSR in the laboratory; they concluded that the CARS-PLSR calibration model was better than the one obtained by PLSR alone.

II.4.2. Non-linear models

Linear datasets are rarely found in the nature, especially in the domain of spectroscopy. In analysing non-linear data, many researchers have applied machine learning models such as support vector machine, artificial neural network, multivariate adaptive regression spline, and boosted regression tree and compared them with PLSR. Mouazen (2014) for example, compared ANN to PLSR in terms of model efficiency. The results revealed that the ANN performed better than PLSR. The algorithm also copes well with a non-linear trend and complex data (Croft *et al.*, 2012).

MARS was first applied by Friedman (1991). The model generates a recursive partitioning regression approach like classification and regression (Breiman *et al.*, 1984), which produces piece-wise linear models instead of piece-wise constant models. Boosted regression trees are less similar to multivariate adaptive regression spline. The algorithm generates an additive

regression model (Friedman, 2001). It is based on multiple modelling integrating both resampling and weighing approaches.

SVM uses an implicit mapping of the input into high dimensional feature space defined by a kernel function (Karatzoglou *et al.*, 2008). ANN is based on the back propagation algorithm (Rumelhart *et al.*, 1986). Back propagation is a technique for computing the gradient of the case-wise error function with respect to the network to minimize the overall network error (Sarle, 1995).

RF is a recent and robust algorithm used in data mining (Breiman, 2001). What makes it different from other models is its robustness to deal with many covariates, outliers, few samples and noise. It does not over fit, does not require any data pre-selection before, handles continuous predictors as well as categorical data, and the output is independent to monotone transformations of the predictors (Díaz-Uriarte and Alvarez de Andrés, 2006).

Rossel and Behrens (2010) have compared all of these models, linear and non-linear. They found that SVM were more powerful than multiple linear regressions, partial least square regression, multiple adaptive regression spline, random forest, boosted tree and artificial neural network. Li *et al.* (2015) found that the least-squares SVM performed better compared to PLSR. However, these authors did not show the effect of derivatives on the robustness of SVM and all other algorithms. The soil spectra were resampled at 10 nm spatial resolution. No models have been implemented with the whole spectral data, which seems to be complex to handle with. The question remains if all these models, especially SVM, will still be valid if all spectral information is considered. Other studies have also mentioned the robustness of SVM. SVM is a very complex model and it can be simplified once combined with a selection method like principal component analysis (PCA) or the successive projections algorithm (SPA). Peng *et al.* (2014) assessed the performance of SPA-SVM compared to PLSR in laboratory conditions. The result was that the SVM model can perform better in the presence of noise and outliers than PLSR (Peng *et al.*, 2014). In the same view, Stevens *et al.* (2010) compared the SVM to PLSR to predict SOC using the Airborne Hyperspectral Sensor 160; they concluded that the SVM is more suitable for large data sets. RF algorithm has not been much used in SOC prediction. Rossel and Behrens (2010) evaluated its performance for SOC prediction in laboratory conditions, the model does not perform well compared to other data mining algorithms. Although none of these calibration methods have achieved universal

acceptance, SVM seems to be the best performing. Also it is worth noting that these models were tested under laboratory conditions where all disturbances are controlled. Few studies investigated on field spectroscopy measurements or have compared both field and laboratory measurements.

CHAPTER III: MATERIALS AND METHODS

III.1. Study area

The research was conducted in the east of Lesotho, in the region of Mokhotlong. Lesotho covers an area of 30 588 km^2 and has approximately 1 876 633 inhabitants (Coburn *et al.*, 2013) (Fig. 3). Four morphological zones characterizes Lesotho, the lowlands (17%), foothills (15%), mountains (59 %) and Senqu River valley (9%) (Bureau of Statistic and Planning, 2007). In the eastern part of the country, Thabana-Ntenyana, the mountains rise to 3 482 m above sea level and is dominated by some river valleys with stunted peach trees near homesteads and denuded grassland (Bureau of Statistic and Planning, 2007). The mean annual rainfall is 775 mm and the average temperature is 11°C (Saha, 2011). The peak of precipitation occurs during the summer season. Almost 85% of annual rainfall occurs between the months of October and March, but during winter, the Monthly mean minimum temperatures range from -6.3 °C to 5.1 °C and the monthly mean maximum temperature occurs from November to February 16.5 °C at high altitudes and 29 °C in the lowlands) (Mason, 1996).

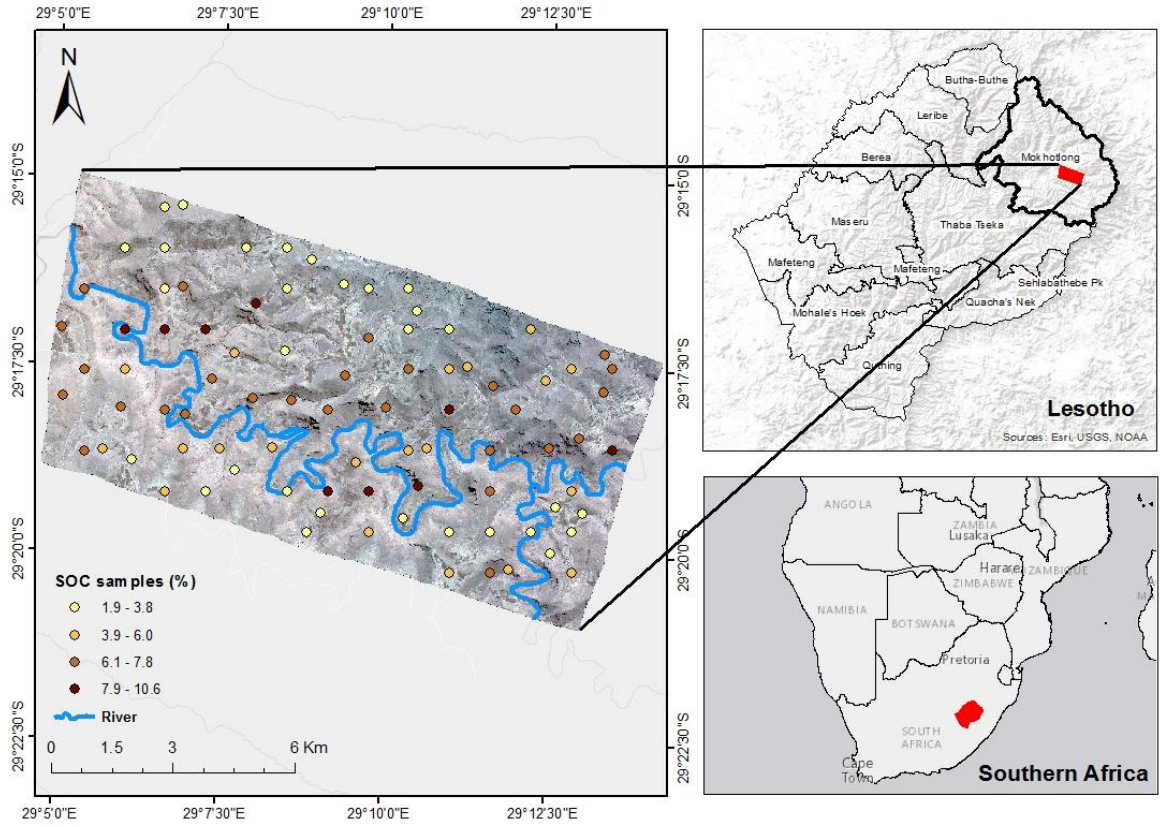


Figure 3. Site location showing the position of field samples taken.

III.2. Soil sampling and spectral measurements

Fieldwork was conducted along the Mokhotlong River in eastern Lesotho (Fig. 3) in October 2015 (spring/summer) between 0800 and 1200 hours. This river flows east to west and has incised through Jurassic basalts, giving rise to a highly meandering river pattern with bedrock spurs and steep valley sides. The river floodplain is very narrow with small strip agricultural fields located adjacent to the river channel and, where slopes sediments are available, terraced fields are present along the lower valley slope. Higher elevation areas at the tops of valley sides and on plateau summits (~2600–2900 m asl) are not enclosed and are characterised by tussocky, xerophytic, low-nutrient grasslands (Fig. 3).

The reflectance of the soil spectra was measured by the ASD FieldSpec Pro FR spectrometer (Analytical Spectral Devices Inc., Boulder CO, USA) at the field and laboratory levels. The wavelength of the instrument stretches from 350 to 2500 nm and with 3–10 nm of spectral resolution. Spectra were recorded for the region 350-1000 nm with a sampling interval of 1.4 nm and 2 nm for the region 1000-2500 nm (Table 3).

Table 3. Waveband range of Analytical Spectral Devices (ASD, 2005).

Region Names in Optical Electromagnetic Radiation		Wavelength (nm)
Ultra Violet (UV)		350-400
Visible (VIS)	Blue light	400-525
	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

III.2.1. Spectral measurement at the field level

Purposive sampling was conducted due to the landscape of the sites. Purposive sampling is based entirely on the operator's judgment in purposely or deliberately, selecting "representative" sample sites (McCoy, 2005). Sampling points at all locations (n=109) were generated randomly using Hawth's Analysis Tool within ArcMap. All points were converted to latitude longitude and a handheld GPS was used to navigate to the site (Fig. 4). Once the sampling point was located, a plot of 10 m by 10 m was drawn and the coordinates of the sampling point which were considered as the centroid of the plot. Within each plot, 3 sub-plots of 2 m by 2 m dimensions were randomly selected in order to take into consideration all the variability within the plot. Five spectral measurements from the nadir at about 1 m and with 5° field of view above soil were scanned in each sub-plots and averaged, in total 15 spectral measurements for each plot. For every measurement, the white reference panel was used to calibrate atmosphere conditions and irradiance of the sun. After the field was completed, spectral measurements were done at the laboratory level.

(a) **(b)**



Figure 4. Field based spectroscopy for SOC content measurement: (a) Spectral measurement by ASD; (b) site localisation (Pictures captured by Dr. Elhadi Adams).

III.2.2. Spectral measurement at the laboratory level

A surface (top 5 cm) soil sample of ~400 g was taken from each of these three sub-plots for laboratory analysis. Measurements were done under a black plate and plastic a day after the field sampling. The soils were scanned using a contact probe of the ASD FieldSpect Pro FR spectrometer. As for the spectral measurements at the field level, we referred to the spectralon for white referencing.

III.3. SOC analysis

In the laboratory, soil samples were first dried and crushed before being sieved. The Madison test sieve was used to sieve the samples. Only grain sizes of less than 2000 microns were used to estimate concentration of SOC. The loss ignition method (LOI) was chosen in this research to quantify SOC concentration. LOI is one of the most useful methods for SOC concentration assessment, fast, reliable and constitutes an inexpensive alternative method in contrast to automated carbon–nitrogen–sulfur (Konen *et al.*, 2002). This method involves only physical destruction of SOM. LOI is more precise when compared to other commonly used method, like weight oxidation which also involved chemical destruction, resulting in partial oxidation of SOC. However, it is important to underline that there is no standard procedure in LOI. In general, three steps are used: (1) incomplete combustion of soil at low temperatures (Ball, 1964); (2) the structural water removal from clay minerals (Sun *et al.*, 2009); and (3) the soil carbonate decomposition at large temperatures (Kasozi *et al.*, 2009). In this study, approximately 200 mg (milligrams) of sieved soil were added to a crucible. The empty porcelain crucibles were first weighed and then the crucible plus the soil. The crucible plus the soil were placed in a muffle furnace for 8 hours at 430 °C (Fig. 5). After ignition, the crucible plus the soil were reweighed. LOI was estimated as the difference between the oven-dry soil mass and the soil mass after combustion, divided by the oven-dry soil mass weighted (Schulte and Hopkins 1996), according to the following equation:

$$LOI_{430} = ((DW–DW_{430})/DW)*100 \quad (1)$$

Where LOI_{430} represents LOI at 430 °C, DW represents the weight of sample of; DW_{430} represents the dry weight of sample after combustion at 430 °C.

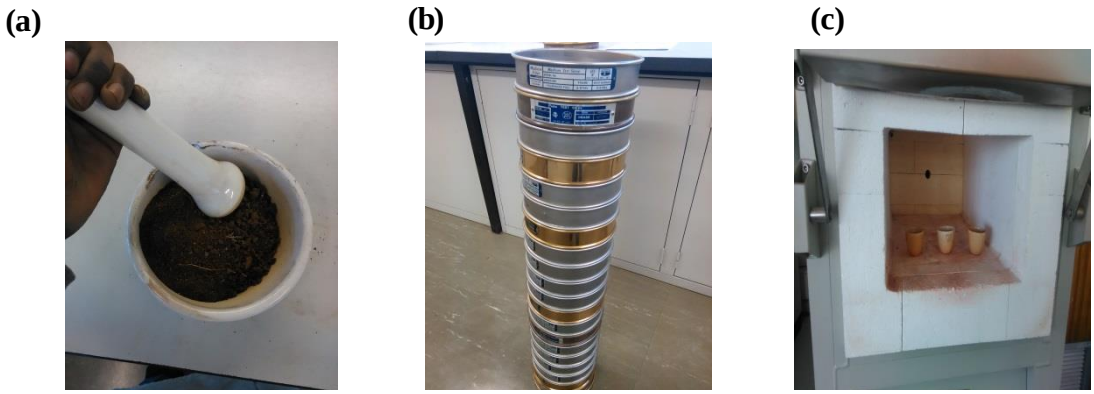


Figure 5. SOC analysis: (a) crushing; (b) sieving and (c) burning. (Pictures captured by Freddy Bangelesa)

II.4. Spectra pre-processing and transformation

Most of the time, the spectrum is affected by a low intensity of incoming light, the baseline variation and the overlapping peaks (Kooistra *et al.*, 2003). That is why, before modelling, the noisy ends were removed in order to correct the low intensity radiation appearing at the spectra edge. The spectrum below 400 nm was removed. The water vapour absorption features ranging between 1350-1460, 1790-1960 and 2350-2500 nm which can affect the model were also removed (Smith *et al.*, 2003; Kooistra *et al.*, 2003) (Fig. 6b). The laboratory and field spectra first derivative transformation was applied in order to enhance the spectral signal (Fig. 6c). The Savitzky−Golay (SG) derivative was used because it is the most useful. Both spectra pre-processing and transformation were implemented in R statistical package version 3.1.3 (R Development Core Team, 2012).

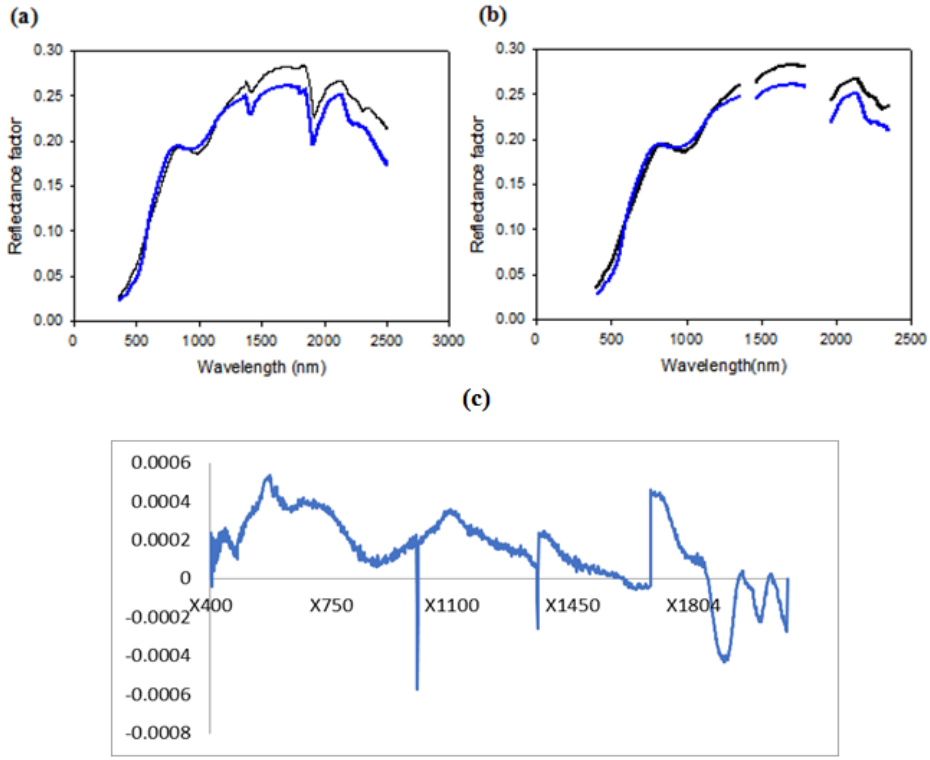


Figure 6. Spectral data: (a) original spectra; (b) with noisy regions removed (1350–1640, 1790–1960, 2350–2500 nm); (c) and first derivative spectra.

III.5. Statistical analysis

Before modelling, 70% of the data (training dataset) was randomly selected to perform the calibration model and the rest (30%) as the testing dataset (Efron and Tibshirani, 1993). The choosing of a random selection is justified by the fact that we wanted to prevent any

systematic bias like spatial correlation between nearest observations which will eventually affect the robustness of the models. To be evident that there is no any significant variability within the training dataset and testing dataset, the Kruskal-Wallis test was performed. The Kruskal-Wallis test evaluates if two or more samples which do not come from a normal distribution are statistically significant. The null hypothesis states that samples are the same. Before the implementation of the Kruskal-Wallis test, the normality of the both datasets (training and testing) was checked using the Kolmogorov-Smirnov goodness of fit test. The Kolmogorov-Smirnov test assumes that if the two samples are identical, the histograms of the two samples should be very similar (Shorack and Wellner, 2009). The test is based on the cumulative distribution function of the histograms, which shows the percentage of observations lying below certain points, and is obtained by adding the numbers in successive categories of the histogram (Shorack and Wellner, 2009). Hotelling's T2 for multivariate analysis for outliers detection was performed (Jackson, 1991). This test is based on the use of the squared Mahalanobis distance.

III.5.1. PLSR and variable selection

The PLSR method was implemented in order to construct predictive models when predictors are many, noisy and highly collinear such as hyperspectral reflectance data (Wold *et al.*, 1995). The logic behind the PLSR is almost the same with the one of the PCR. Both regression methods reduce the data dimensionality. The PCR converts the predictor variables to principal components and ranks them according to their respectively variances (Li *et al.*, 2015). The selection of predictors in PCR is done manually. With PLSR by contrast, this arbitrary selection is avoided. Instead, the number of predictor variables is reduced by selecting successive orthogonal factors from the variance-covariance matrixes in a manner to maximize the covariance between the predictors and the dependent variables (Li *et al.*, 2015). The concepts explaining MLR and PLSR are described by Martens and Martens (1986) (Fig 7.) With MLR, Y is directly modelled by all X variables, while in PLSR1, the latent variable (T) from X is used to model Y.

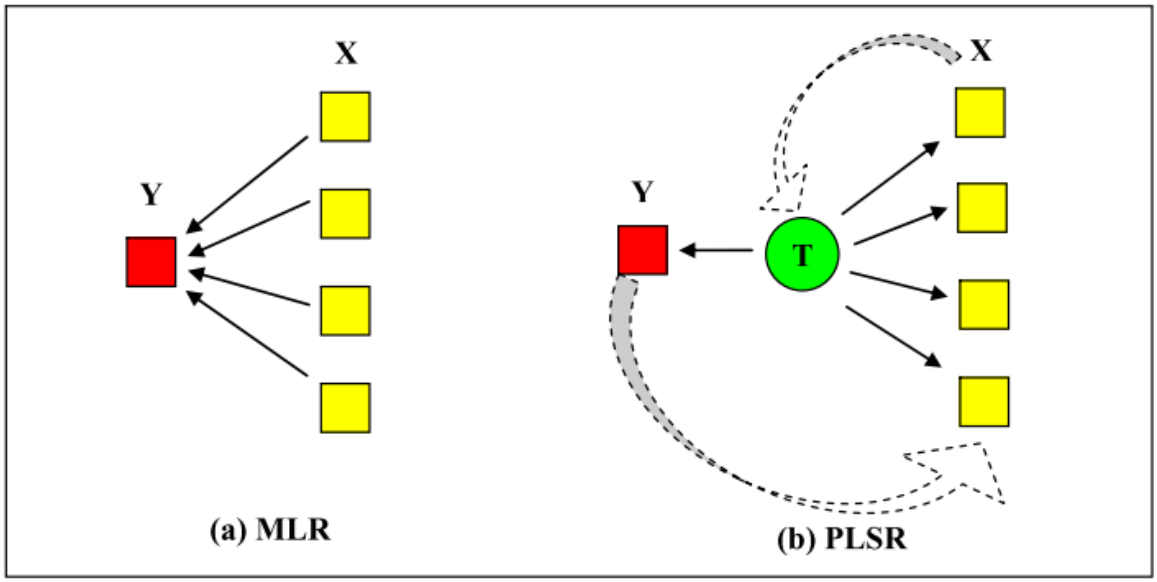


Figure 7. Conceptualization of (a) MLR and (b) PLRS (Martens and Martens, 1986).

The regression model (PLSR) identifies factors from the independent variables that are simultaneously important for the dependent variables and selects the most important ones (Wold, 1995). In case of the over-fitting or under-fitting problem, the leave-one-out cross-validation (LOOCV) method was applied to find the optimal number of components that reduces the prediction error variance (Li *et al.*, 2015). The logic behind it is that a single sample is removed and the operation is repeated many times (Gomez-Carracedo *et al.*, 2007). It is important to mention that LOOCV sometimes selects unnecessarily large numbers of components. To get round this situation, an external validation set is of highest importance (Gomez-Carracedo *et al.*, 2007).

They are many algorithms developed in PLSR to select the most important predictors (wrapper, filter and embedded). The filter method which is subdivided into the loading weight coefficient, the regression coefficient and variable important in projection (VIP) method, was used because it is computationally fast and easy to manipulate. The VIP method was finally chosen to reduce the spectral dimensionally in PLSR. VIP was first introduced by Wold *et al.* (1995). The logic is to assemble the relevance of every variable being reflected by the weight from every component. The variance importace mathematic model is described by Wold *et al.* (1995).

$$FS_{VIPk}\ (a) = K \sum_a w * w_{ak} * (\frac{SSY_a}{SSY_t}) \quad \textbf{(2)}$$

Where VIPk (*a*) refers to the importance of the K predictor variable based on a model with *a* factors. K is the total number of predictor variables. W_{ak} is the corresponding loading weight of the K variable in the *a*th PLSR factor, SSY_a is the explained sum of squares of the response variable, by a PSLR model with *a* factors, SSY_t is the total sum of squaress of the response variable, and K is the total number of predictor variables (Rossel and Behrens, 2010). The variable can be eliminated if the value is under the defined threshold. A more relevant threshold is between 0.83 and 1.21 (Mehmood *et al.*, 2012). A threshold of 1 was used in this study because it is commonly used in different studies (Chong and Jun, 2005).

III.5.2. Random forest regression and variable selection

The RF regression is a machine learning algorithm and a bagging method based on the CART regression tree (Breiman, 2001). The model was implemented in the RF package (Liaw and Weiner, 2002). The model uses recursive partitioning to split the data (spectra) into different homogeneous groups named regression trees (*ntree*). Each tree is individually grown to its optimum size based on a bootstrap sample from the training data set (approximately 70%) without any pruning (a continuous selection of input variables at every node). In terms of RF regression, a random subset of variables (*mtry*) are selected to determine the split at each node (Breiman, 2001). The model uses a deterministic algorithm to select the number of random samples and variable from the training dataset. In each tree, the data that are not in the tree (the out-of-bag: OOB data, approximately 30%) are predicted and the OOB error is produced in term of mean square errors through the difference between OOB data, therefore, it can be used to grow the regression trees (Breiman, 2001; Maindonald and Braun, 2006). The OOB error provides an estimation of the important variables by calculating how much OOB error of estimate increases when a variable is permuted, whilst all others remained unchanged (Archer and Kimes, 2008). Therefore, can be used like feature selection method.

In this study, the RF algorithm was implemented for both spectra and laboratory data sets. Because of a big number of variables, the model optimization which is computationally intense was not implemented. The default setting for *mtry* (1/3 of the total number of wavelength) and *ntree* (500) was used. The wavelengths were ranked according to their importance in predicting SOC concentration.

A recursive feature selection (Kohavi and John, 1997; Guyon and Elisseeff, 2003) was performed to determine the least number of wavelengths that predict SOC concentration with great accuracy. The recursive feature elimination algorithm is a wrapper feature selection method that uses all features (variables) as a starting point (Kittler, 1978). Models with low accuracy are removed from current subset. The procedure ends when the given numbers of variables are dropped (Kittler, 1978). One of the drawbacks of the method is that it is very slow compared to other methods, but provides accurate results (Kittler, 1978).

III.5.3. Model validation

Many parameters can be used to evaluate the efficiency of models in spectroscopy. Spectral model are generally evaluated using the coefficient of determination (R^2) and root mean square error (RMSE). Other parameters like the Akaike Information Criterion (AIC), the ratio of prediction to deviation (RPD), the root mean square error of calibration (RMSEC) and validation (RMSEP) are also useful for assessing model performances. RMSEC is the standard deviation of the difference between the measured and the estimated values for samples in the training set, while the RMSEP is computed within the validation set. RMSEC is expressed as:

$$RMSEC = \sqrt{\frac{\sum(y_m - y_p)^2}{N}} \text{ and } RMSEP = \sqrt{\frac{\sum(y_m - y_v)^2}{N}} \quad (\mathbf{3})$$

Where y_m are measured the direct values of SOC obtained from the laboratory measurement, y_p are predicted values derived from spectral data using either PLSR or RF, y_v are predicted values estimated using the validation set, and N refers to the number of samples. The model with a lowest coefficient root mean square error is the best.

The overall variation accounted for the regression model (PLSR or RF) was measured by the coefficient of determination (R_c^2). The predicted coefficient determination (R_p^2) measures the proportion of total variation accounted for the model using validation analysis. The model with a high coefficient of determination is the best (Table 4).

Table 4. Interpretation of R^2 in calibrating SOC using NIR spectroscopy (Williams, 2001).

Value of coefficient of determination	Interpretation
< 0.25	Not usable in NIR calibration
0.26-0.49	Poor
0.50-0.64	Good for rough screening
0.66-0.81	Good for screening and some other calibration
0.83-0.90	Usable with caution for most application, including research.
0.92-0.96	Usable for most applications, including quality assurance
> 0.98	Usable in any application

Another parameter to compare calibration models in these studies is the AIC. AIC compromises between prediction accuracy and parsimony (Akaike, 1973). By model parsimony, we mean least complexity, more information gain and least number of variables. AIC was estimated by:

$$AIC = n\ lnRMSE + 2p \quad (\mathbf{4})$$

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

At least, the RPD were also used to compare the performance of model of different datasets, as well as their practicability. In this study, RPD was taken as the most important indicator to compare different models because it computes the accuracy by integrating the training and testing datasets. RPD is the ratio of standard error of prediction to the standard deviation of reference data in the testing set:

$$RPD = \frac{STDEV(y)}{RMSEP} \quad (\mathbf{5})$$

Where STDEV(y) refers to the standard deviation of the training data and RMSEP refers the root mean square error of prediction.

The six categories of interpretation as recommended by Rossel *et al.* (2006) were given as follows:

- RPD greater than 2.5 implies excellent models/predictions;
- RPD between 2.0 and 2.5 implies very good, quantitative models/ prediction;
- RPD between 1.8 and 2.0 implies good models/predictions, where quantitative predictions are possible;
- RPD between 1.4 and 1.8 implies fair models/ predictions, which can be used for assessment and correlation;
- RPD between 1.0 and 1.4 implies poor models/predictions, where only high and low values are distinguishable;
- And RPD less than 1.0 implies very poor models/predictions, and their uses are not distinguishable.

CHAPTER IV: RESULTS

IV.1. SOC sample descriptive analysis

Of 109 soil samples collected in the study area, only 94 were analysed. Outlier's and incomplete samples were removed. Hotelling's test detected 4 outliers with a horizontal cut off limit of 6.32845 and vertical cut off limit of 3.98695 (Fig. 8).

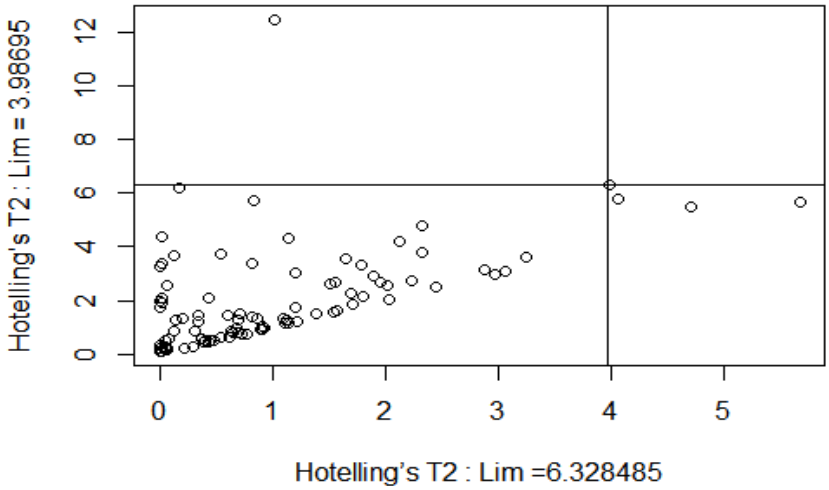


Figure 8. Hotelling's T2 for multivariate outlier detection.

From the 94 samples analysed, 70% (65) were randomly assigned to calibration dataset and 30% (29) to validation dataset. The descriptive statistic of SOC for the calibration dataset, validation dataset and the whole dataset are described in Table 5. The variation of SOC ranged from 1.93% to 10.66% with a mean value of 5.14% and the coefficient of variation of 41.45%. The coefficient of variation within the calibration and validation are almost the same, 40% and 43%, respectively. The Kolmogorov-Smirnov test revealed that all datasets were not normally distributed at 5% significant level with p-values of 0.01266, 0.002 and 0.21 for the whole dataset, calibration dataset and validation dataset, respectively. All subsets have a skewed distribution. The Kruskal-Wallis test for a skewed distribution indicated that there is no significant difference among the three datasets at 5% significant level (p-value = 0.64). Thus, both calibration and validation datasets statistically represent the total dataset.

Table 5. Descriptive statistics of SOC contents (in %) within three different datasets.

Data set	N	Min	Max	Mean	Median	Std
Whole dataset	94	1.93	10.66	5.14	4.97	2.13
Calibration dataset	65	2.172	9.570	4.97	4.87	1.99
Validation dataset	29	1.936	10.665	5.55	5.55	2.42

IV.2. Comparison of field and laboratory spectra

Figure 9 shows mean reflectance of field and laboratory measurements for all 94 samples. Spectral absorption feature positions are also shown. Because bands were averaged, the interpretation becomes complex. Nevertheless, only water absorption feature located around 1400 and 1900 nm can be perceived. In general, the reflectance of the laboratory spectral measurements is visually higher than that of the field. This is verified by one tailed student's t test, which shows that laboratory spectral measurements are significantly greater than field ones at 5% level (p value = 0.024). The Pearson correlation test reveals that both spectral measurements are strongly correlated at 5% significant level (R= 99, 97%, p-value < 0.00000000000000002).

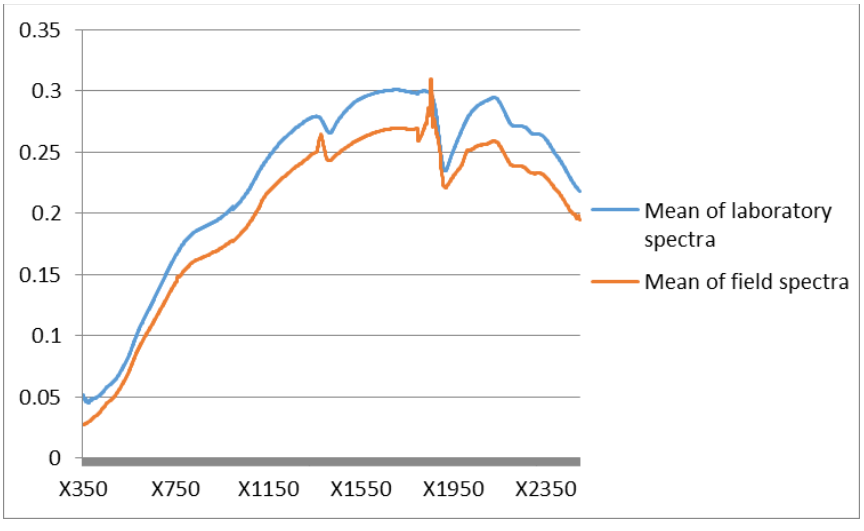


Figure 9. Average laboratory and field VIS-NIR spectra; (orange): field spectra data; (blue): laboratory spectra data.

IV.3. Key wavelength selection

IV.3.1. VIP algorithms

VIP algorithms were computed with PLRS for both field and laboratory spectra. Values where peak maximum above 1 were taken as key wavelengths to predict SOC contents (Fig 10). The algorithm has selected 80, 720, 57 and 881 key wavelengths for the laboratory first derivative laboratory spectra, the laboratory raw spectra, the field first derivative spectra and raw field spectra, respectively. Key wavelengths were distinctively selected with derivative spectra (field and laboratory) with two peaks, around 750 nm and 950 nm for the field first derivative spectra (Fig. 10c) and one around 680 nm for the laboratory first derivative spectra (Fig. 10d). For the raw spectra data (field and laboratory), key wavelengths were broadly selected from 1200 nm to 2200 nm and some others around 800 nm (Fig. 10a; Fig 10b).

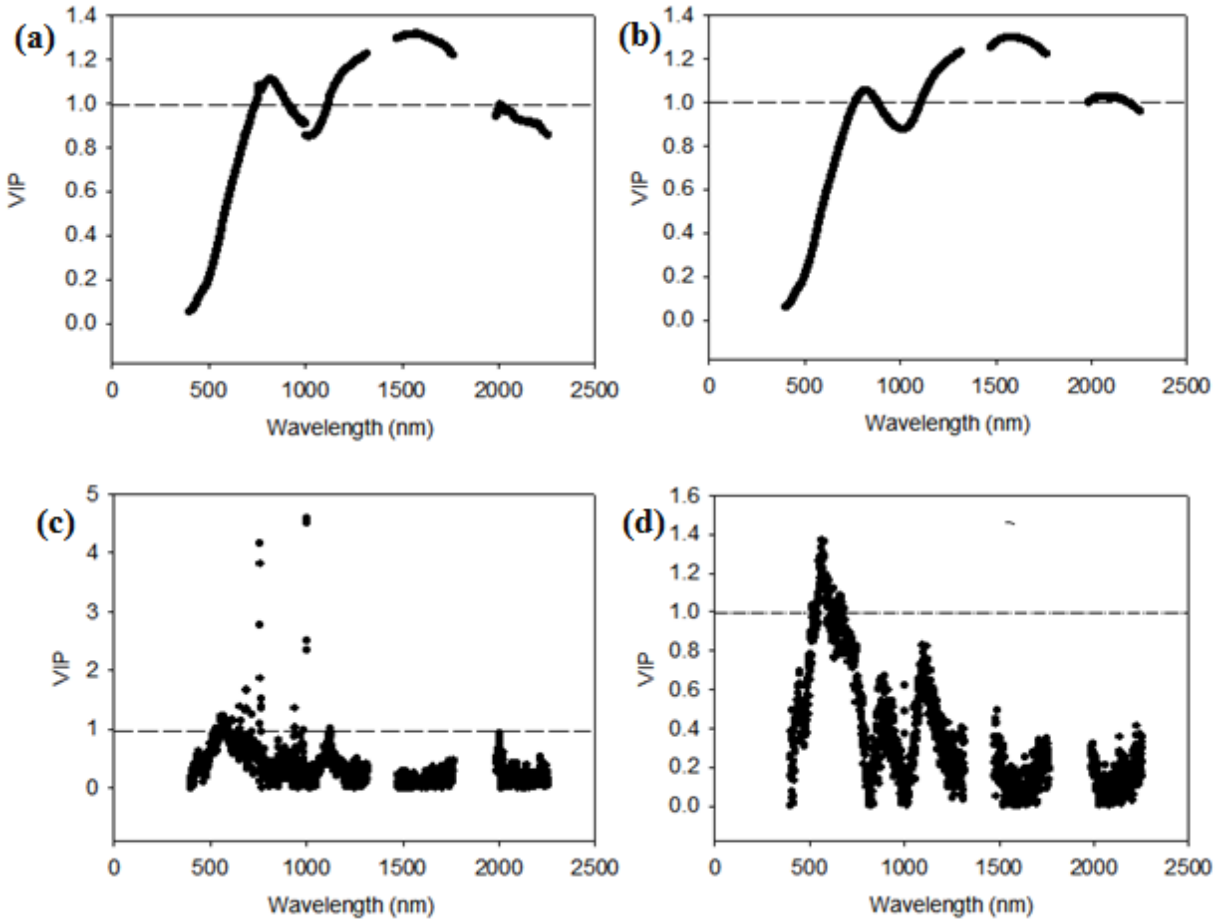


Figure 10. Variable important in projection for key wavelength selection: (a) field raw spectra, (b) laboratory raw spectra; (c) field first derivative spectra; (d) laboratory first derivative spectra.

IV.3.2. Recursive feature selection and percent increase in MSE

Figure 11 shows the outcome of the recursive feature selection in both laboratory and field spectral data. The algorithm was performed with 5 fold cross validation for selecting the most important wavelengths. The smallest average numbers of wavelength for each model that would offer the best predictive of random forest were identified. The lowest number of key wavelengths (49) was obtained with the first derivative field spectra with a RMSE of 0.61%. The use of 105 wavelengths produced the lowest RMSE (0.62%) for the laboratory first derivative spectral data. For the laboratory raw spectral data, the use of 105 variables produced the lowest RMSE (0.6109%). For the field raw spectral data, the use of 789 variables produced the lowest RMSE (1.059%).

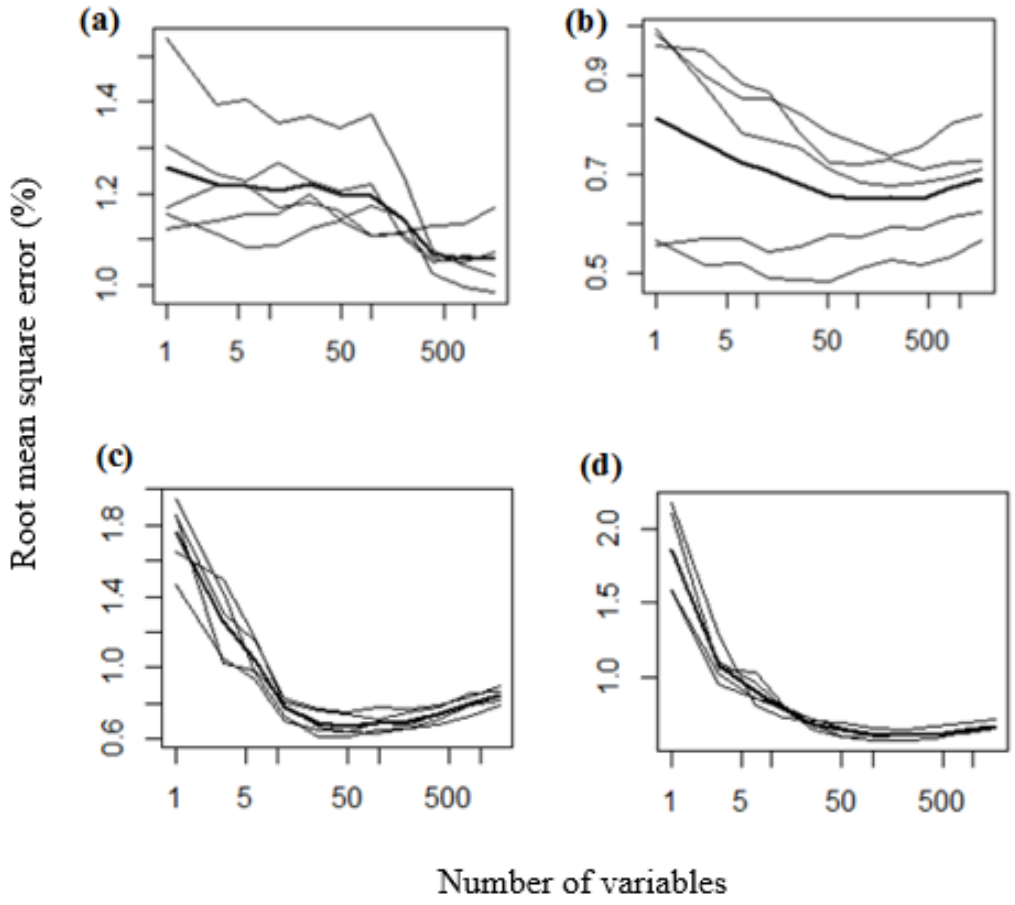


Figure 11. Recursive feature selection: (a) field raw spectra, (b) laboratory raw spectra; (c) field first derivative spectra; (d) laboratory first derivative spectra.

Figure 12 below shows the position of wavelengths according to the percent increase in mean MSE implemented with the RF regression algorithm. The recursive feature selection

combined with the percentage increase in MSE selects around half of key wavelengths in all datasets between the range of 600 and 900 nm. Important wavelengths are selected distinctively within the raw spectra (laboratory and field) and are mostly located around the visible range. One peak around 780 nm is identified for the field spectral data (Fig. 12a) and two for the laboratory spectral data, around 780 nm and 900 nm (Fig. 12b). In the field first derivative spectra, key wavelengths are located around 780 nm (Fig. 20c), the same with the raw field spectral data. The same location of important wavelengths with the laboratory field spectral data is observed within the first derivative laboratory spectral data (12.d)

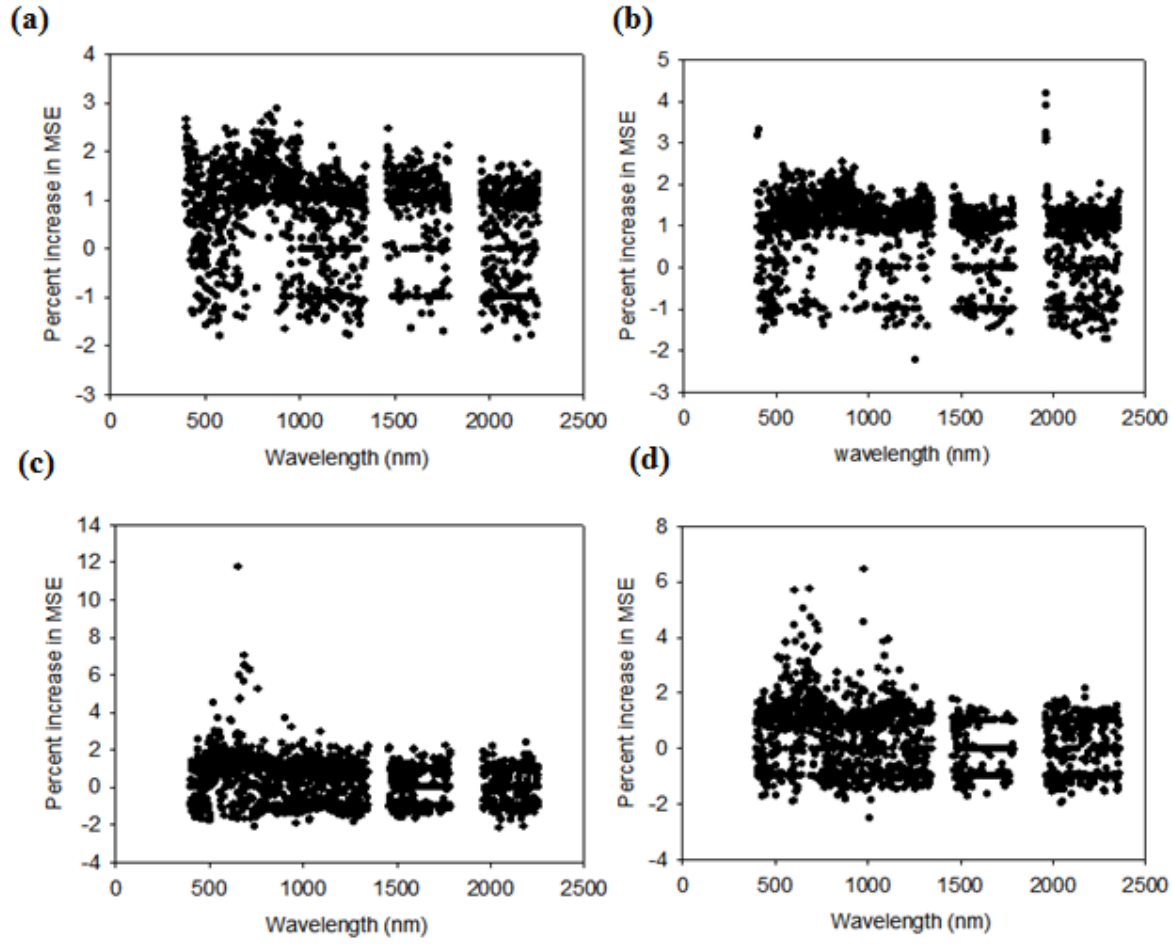


Figure 12. Percent increase in MSE: (a) field raw spectra (b) laboratory raw spectra; (c) field first derivative spectra; (d) laboratory first derivative spectra.

IV.3.3. Position of key wavelengths and interpretation

Table 6 compares in more detail the performance of the VIP algorithm and recursive feature selection combined with the percent decrease in MSE in term of selecting key wavelengths.

For interpretation purposes, the functional group and vibration mode of wavelength as suggested by Stuart (2004) are also presented. The recursive feature selection combined with the percent decrease in MSE has selected around half of key wavelengths in all datasets between the range of 400 and 700 nm. The VIP algorithm implemented under the first derivative spectral data (laboratory and field) has also selected most key wavelengths in the same range. Only the VIP algorithm implemented in the derivative datasets (laboratory and field) did not select most of wavelengths within the visible range.

Table 6. Comparison of VIP algorithm and recursive feature selection combined with percent decrease in MSE in selecting key wavelengths.

		VIP (PLSR)				**RF**			
Wavelength (nm)	Possible Assignment	Lab		Field		Lab		Field	
		Raw	FD	Raw	FD	Raw	FD	Raw	FD
2200-2450	Comb C-H stret	-	-	-	-	+	-	-	-
2000-2200	Comb N-H stret, comb O-stret	++	-	-	-	+	+	-	+
1790-1960	Water	+	+	+	+	+	-	+	-
1650-1780	1st overt C-H stret	+	-	++	-	-	-	+	-
1400-1500	1st overt N-H stret and O-H stret	++	-	++	-	-	-	++	-
1300-1420	Comb C-H stret	+	-	-	-	+	-	+	-
1350-1460	Water	+	-	++	+	-	-	-	+
1100-1225	2nd overt C-H stret	+	-	+	-	+	++	+	+
950-1100	2nd overt N-H stret and O-H stret	-	-	-	+	+	+	+	+
850-950	3rd overt C-H stret	+	-	+	+	++	+	++	+
775-850	3rd overt N-H stret	+	-	+	+	++	+	+	+
400-700	Mineral (Fe oxides)	-	+++	+	+++	+++	+++	+++	+++

∗ str = stretching vibration mode; comb= combination vibration mode; Overt = overtone; FD = first order derivative. The relative importance is indicated by '+', '++' and '+++', where '+++' indicates wavelength > 40 percent, '++' wavelength between 40 and 20 percent, '+' wavelength < 10 percent and '-' indicates and absence of wavelengths.

IV.4. Model development

In total, 16 models were developed using different datasets (field and laboratory) and spectral pre-processing data (raw, Savitzky-Golay derivative, key wavelengths and the combination of Savatzky-Golay and key wavelengths), 8 with PLRS and 8 for Radom forest regression. Models developed with PLRS are:

- F-PLSR: the field PLSR raw model with all wavelengths;
- F-FD-PLSR: the field first derivative PLSR model;
- F-K-PLSR: the field PLSR model with key wavelengths;
- F-FD-K-PLSR: the field first derivative PLSR model with key wavelengths;
- L-PLSR: the laboratory PLSR raw model with all wavelengths;
- L-FD-PLSR: the laboratory first derivative PLSR model;
- L-K-PLSR: the laboratory PLSR model with key wavelengths;
- L-FD-K-PLSR: the laboratory first derivative PLSR model with key wavelengths.

Model developed with RF regression are:

- F-RF: the field RF raw model with all wavelengths;
- F-FD-RF: the field first derivative RF model;
- F-K-RF: the field RF model with key wavelengths;
- F-FD-K-RF: the field first derivative RF model with key wavelengths;
- L-RF: the laboratory RF raw model with all wavelengths;
- L-FD-RF: the laboratory first derivative RF model;
- L-K-RF: the laboratory RF model with key wavelengths;
- L-FD-K-RF: the laboratory first derivative RF model with key wavelengths.

The best prediction model was selected according to its performance with respect to fitting to the validation dataset. The RPD was considered as the most important criteria.

IV.4.1. PLSR optimum number of component

PLSR requires the optimization of the number of components in order to minimize the overfitting problem. Figure 13 indicates the optimum component to be chosen based on the value of the root mean square error of cross validation (RMSECV) implemented with 65 leave-one-out segments. The numbers of components with a lowest RMSECV were used to fit the model. The L-FD-K-PLSR selects the fewer numbers of components, 3 with RMSECV of 0.79%. L-FD-PLSR used, 4 (RMSECV=0.785%) compared to others. The L-PLSR used 5 components (RMSECV = 0.719%). The F-FD-K-PLSR used 6 components (RMSECV = 0.89%). The F-FD-PLSR used 8 components (RMSECV = 0.695%), the F-K-PLRS used 8

components (0.766%). The L-K-PLSR used 12 number of components (RMSECV = 0.762%) and the F-PLSR used 15 number of components (0.497%).

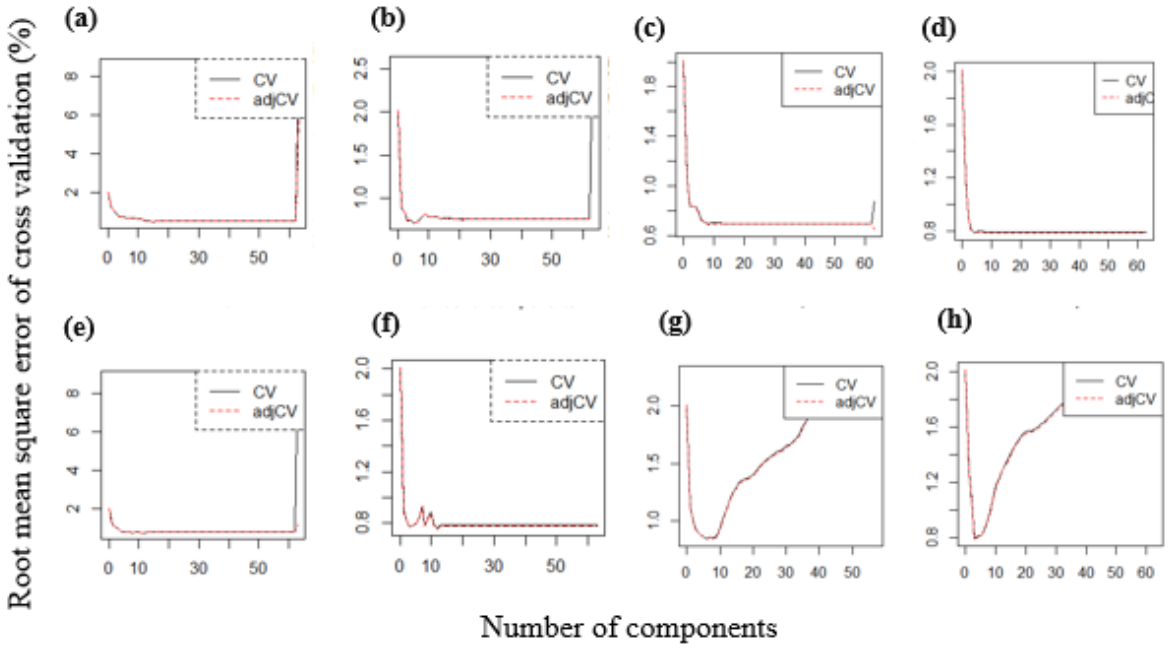


Figure 13. PLSR optimum number of components: (a) field PLSR raw model with all wavelengths; (b) laboratory PLSR raw model with all wavelengths; (c) the field first derivative PLSR model; (d) the laboratory first derivative PLSR model; (e) field PLSR model with key wavelengths; (f) the laboratory PLSR model with key wavelengths; (g) the field first derivative PLSR model with key wavelengths; (h) the laboratory first derivative PLSR model with key wavelengths.

IV.4.2. PLSR model performances within the laboratory dataset

All PLSR models developed under laboratory conditions (L-PLSR, L-FD-PLSR, L-K-PLSR, L-FD-K-PLSR) show a very good model prediction (2.0 <RPD<2.5) (Fig. 14). The best predictive model was achieved with L-PLSR (RPD = 2.27, Rp^2= 0.86, RMSEP = 0.87%). Followed by L-K-PLSR (RPD = 2.10, Rp^2 = 0.84, RMSEP = 0.95%). Derivative models, L-FD-PLSR (RPD = 2.00, Rp^2 = 0.82, RMSEP = 0.99%) and L-FD-K-PLSR (RPD = 2.00, Rp2= 0.82, RMSEP = 0.99%) have poorly performed.

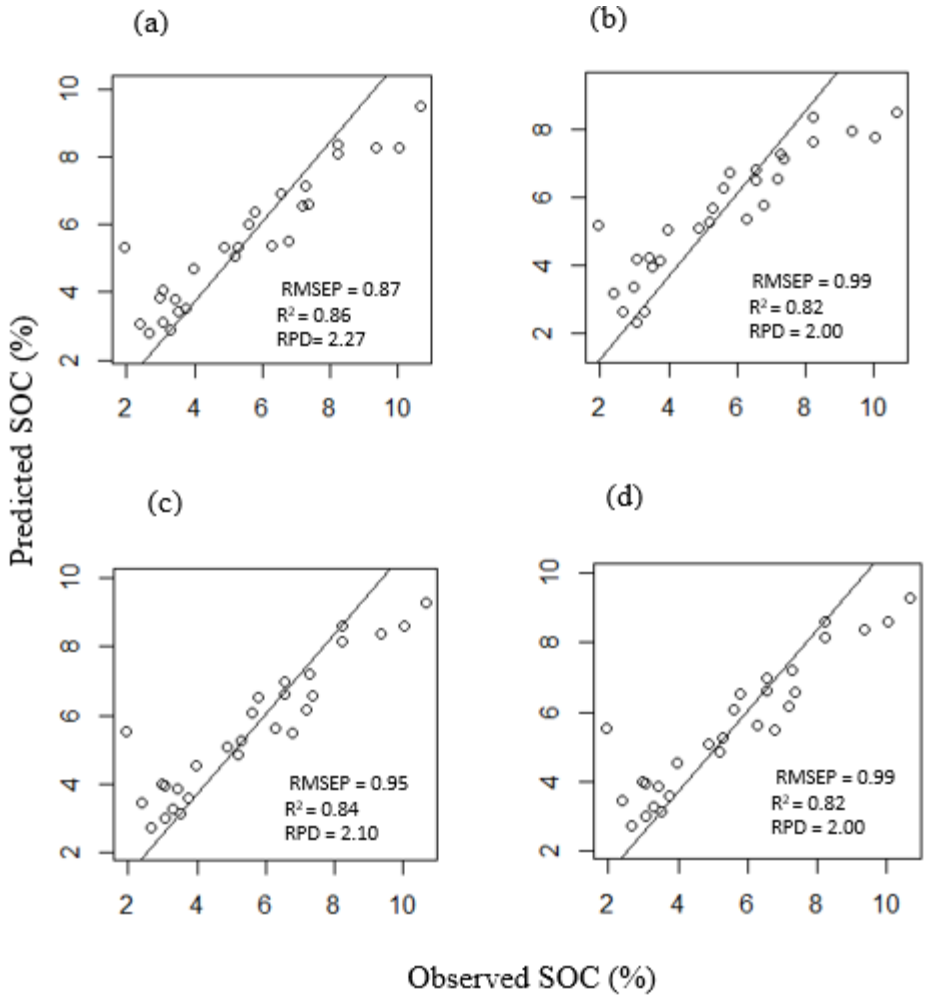


Figure 14. Performance of PLSR in predicting SOC on an independent laboratory spectral dataset : (a) laboratory PLSR raw model with all wavelengths; (b) the laboratory first derivative PLSR model; (c) the laboratory PLSR model with key wavelengths; (d) the laboratory first derivative PLSR model with key wavelengths

IV.4.3. PLSR model performances within the field dataset

Figure 15 shows PLSR models developed under the field dataset. Only L-PLSR (RPD = 1.88, Rp^2= 0.80, RMSEP = 1.05%) and L-FD-K-PLSR (RPD = 1.90, Rp^2 = 0.80, RMSEP = 1.04%) show good model prediction (1.8 <RPD<2.0). The two others, F-K-PLSR and F-FD-PLSR showed a very good model prediction (2.0 <RPD<2.5). In overall, the best predictive PLSR model within the field dataset was achieved with F-K-PLSR (RPD = 2.97, Rp^2= 0.86, RMSEP = 0.75%), and followed by F-FD-PLSR (RPD = 2.26, Rp^2= 0.86, RMSEP = 0.88%).

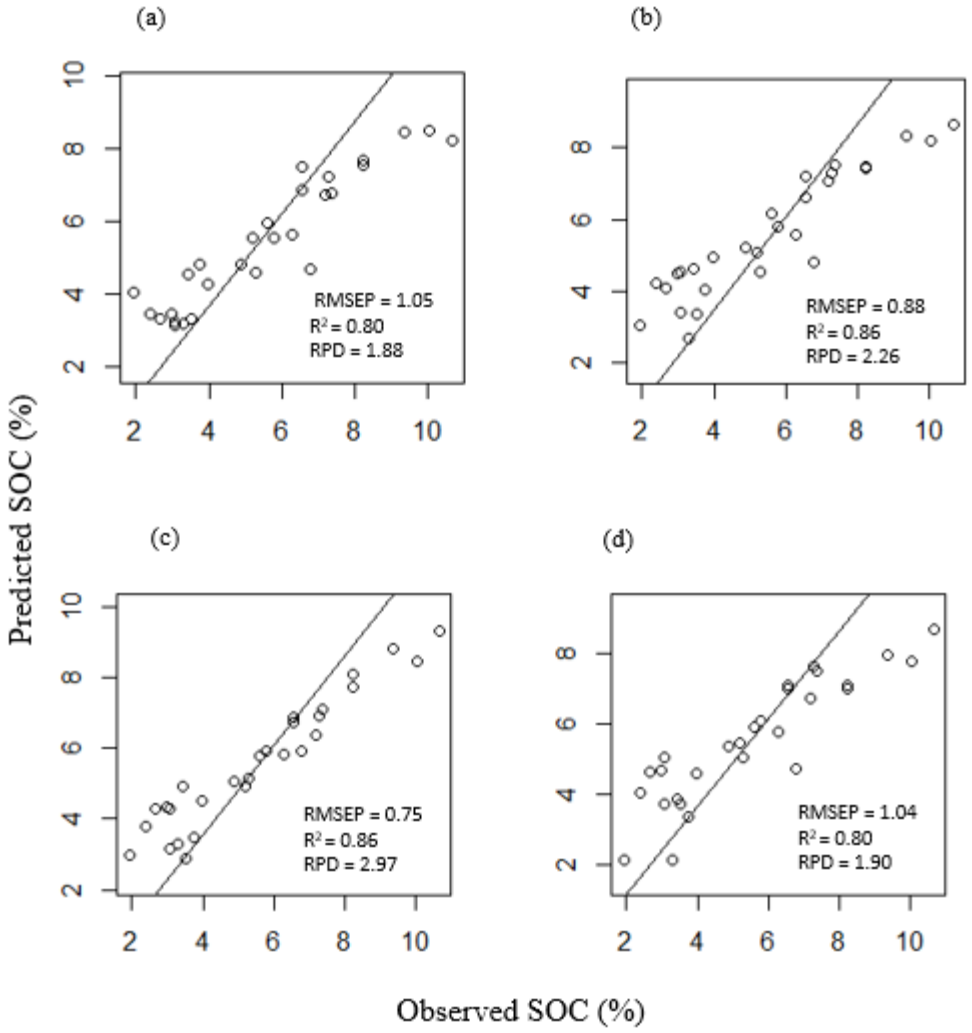


Figure 15. Performance of PLSR in predicting SOC on an independent field spectral dataset: (a) field PLSR raw model with all wavelengths; (b) the field first derivative PLSR model; (c) field PLSR model with key wavelengths; (d) the field first derivative PLSR model with key wavelengths.

IV.4.4. RF regression model performances within the laboratory dataset

Figure 16 shows the performance of RF regression models to predict SOC contents from an independent laboratory dataset. All RF laboratory models exhibit an excellent model prediction (RPD>2.5). The combination of first derivative and key wavelength improved significantly the raw spectral model because the best predictive model was obtained with L-FD-K-RF (RPD = 3.77, R_p^2= 0.88, RMSEP = 0.64%). The worse performance prediction model was obtained with F-RF (RPD = 3.03, R_p^2 = 0.76, RMSEP = 0.79%).

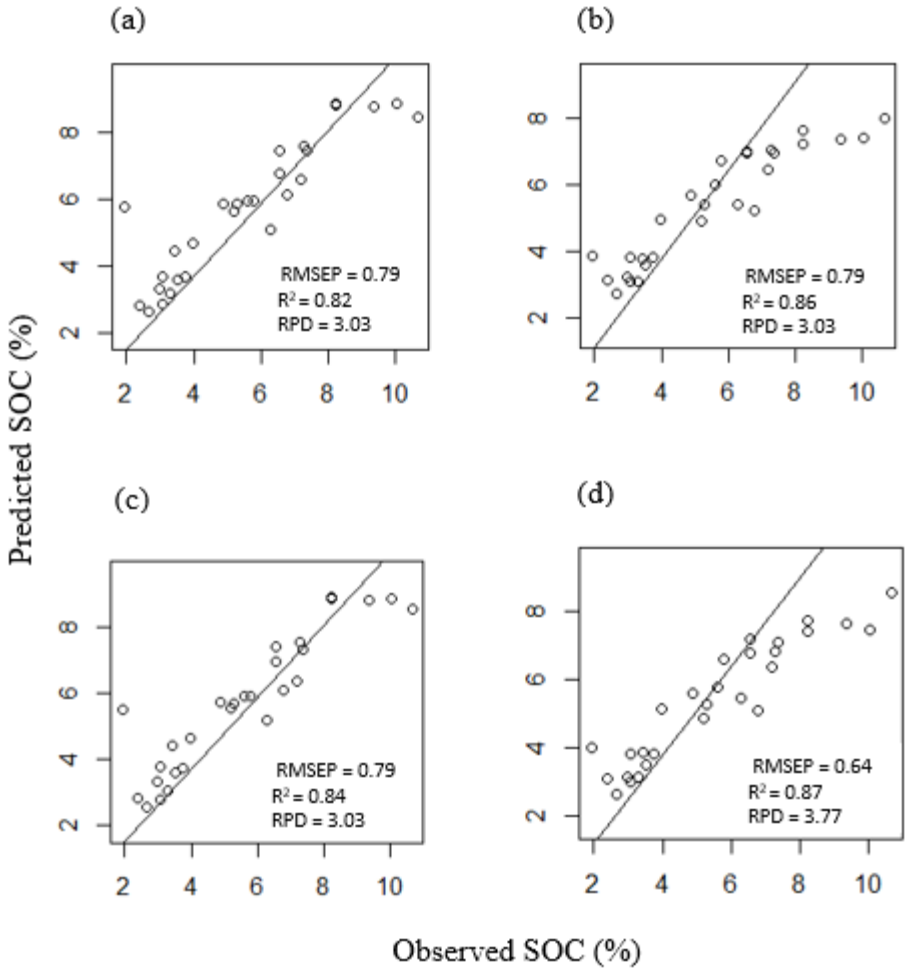


Figure 16. Performance of RF regression in predicting SOC on an independent laboratory dataset: (a) laboratory RF raw model with all wavelengths; (b) the laboratory first derivative RF model; (c) the laboratory RF model with key wavelengths; (d) the laboratory first derivative RF model with key wavelengths.

IV.4.5. RF regression models performance within the field dataset

The performance of RF regressions to predict SOC contents from an independent dataset is presented in Figure 17. All the predictive models were excellent (RPD>2.5). The best predictive model was obtained by combining the first order derivative and key wavelengths, F-FD-K-RF (RPD = 3.77, R_p^2= 0.87, RMSEP = 0.64 %). And after come the F-FD-RF (RPD = 3.03, R_p^2= 0.89, RMSEP = 0.79 %).

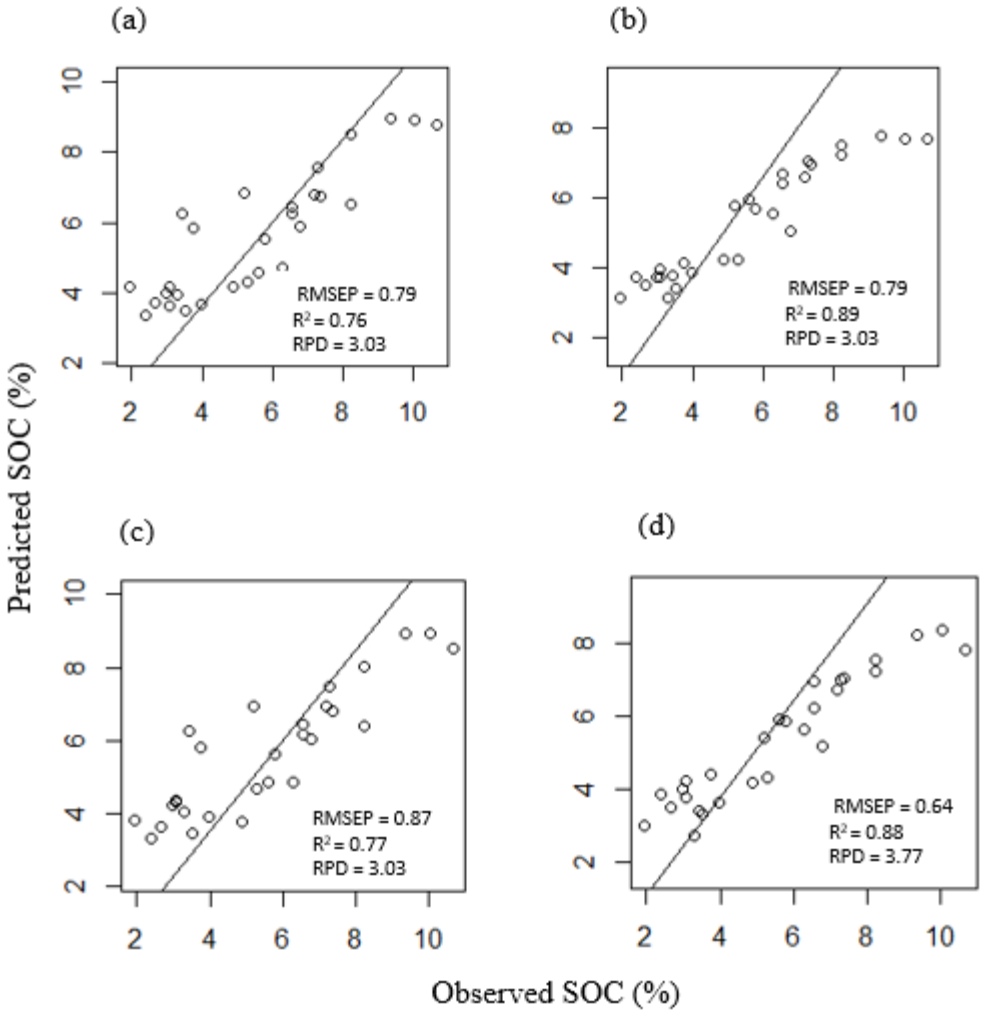


Figure 17. Performance of RF regression in predicting SOC on an independent field spectral dataset: (a) the field RF raw model with all wavelengths; (b) the field first derivative RF model; (c) field RF model with key wavelengths; (d) the field first derivative RF model with key wavelengths.

IV.5. Comparison between RF regression and PLSR

Table 7 summarizes the performance of all RF and PLRS models in the prediction of SOC contents. Model performances were assessed in both calibration and validation datasets. The results indicate that the PLSR is more likely to overfit the calibration set than RF with an R_c^2 going up to 0.99. In terms of prediction of the independent dataset, RF regression outperforms PLSR. According to the model classification based on the RPD, all RF regression models have provided an excellent models prediction (RPD>2.5) whereas PLSR models provide either very good model prediction (2.0 <RPD<2.5) or good model prediction (1.8 <RPD<2.0). The overall best performance model was obtained with RF with the

laboratory dataset, L-FD-K-PLSR (RPD = 3.77, R_p^2= 0.88, RMSEP = 0.64%), followed by L-FD-K-PLSR (RPD = 3.77, R_p^2= 0.87, RMSEP = 0.64%) with the field spectral data.

Table 7. Performance of all RF regression and PLSR models in the calibration and validation datasets.

Models			Calibration set			Validation set		
	Model	pretreatment	RMSEC	R_c^2	AIC	RMSEP	R_p^2	RPD
Field spectra	PLSR	none	0.89	0.99	-123	1.05	0.80	1.88
		FD	0.550	0.99	-186	0.88	0.86	2.26
		K	0.67	0.86	63	0.75	0.86	2.97
		FD-K	0.64	0.89	132	1.04	0.80	1.90
	RF	none	0.9	0.79	-	0.79	0.76	3.03
		FD	0.74	0.85	-	0.79	0.89	3.03
		K	0.87	0.80	-	0.87	0.77	3.03
		FD-K	0.60	0.90	-	0.64	0.88	3.77
Lab spectra	PLSR	none	0.60	0.90	126	0.87	0.86	2.27
		FD	0.46	0.94	90.98	0.99	0.82	2.00
		K	0.12	0.99	82	0.95	0.84	2.10
		FD-K	0.63	0.89	131.8	0.99	0.82	2.00
	RF	none	0.65	0.89	-	0.79	0.82	3.03
		K	0.80	0.84	-	0.79	0.86	3.03
		FD	0.64	0.89	-	0.79	0.84	3.03
		FD-K	0.59	0.90	-	0.64	0.87	3.77

*RF = random forest; PLSR = partial least square regression; FD = first derivative; K = Key wavelength selection; none = spectra without transformation; RMSEC = root mean square error of calibration in %; RMSEV = root mean square error of validation in %; R_p^2 = coefficient of determination in the validation dataset; R_c^2 = coefficient of determination in the calibration dataset; AIC = Akaike Information Criterion; RPD = Ration of prediction to deviation.

CHAPTER V: DISCUSSION

V.1. Suitable spectral bands for SOC modelling

The recursive feature selection combined with the percent decrease in MSE selected around half of the key wavelengths in all datasets between the range of 400 and 700 nm (Table 6). The VIP algorithm implemented under the first derivative spectral data (laboratory and field) also selects most of key wavelengths in the same range. According to Rossel *et al.* (2015), the visible portion of the soil spectrum is mostly linked to Fe oxides. According to Garzanti *et al.*, (2014), the Lesotho highlands are dominated by iron oxides resulted from the weathering of olivine. Different soil chromophores may explain the high correlation between spectral data and SOC in the visible region (Mouazen *et al.*, 2007). Referring to Stuart (2004) in Table 1, this range of spectra can be attributed to the third overtone N-H stretching. This result is in accordance with what different researchers have found (Rossel and Behrens, 2010; Vohland and Emmerling, 2011). Rossel *et al.* (2006) discovered important wavelengths around 410, 570, and 660 nm in the visible region of the spectra. Under laboratory conditions, Wang *et al.* (2010) reported 440, 560, 625, 740, and 1336 nm as principal spectral bands to predict the SOC. Nocita *et al.* (2014) suggested that the spectral portion between 580, 570 and 680 nm was sufficient to predict SOC.

However, this was not the case with the VIP algorithm computed from raw spectral data (field and laboratory) where most key wavelengths were broadly selected around 1200-2200 nm with a peak at 1500 nm (Fig.10a, 10b). This may be explained by the fact that the VIP algorithm could be sensitive to noise. Some scientists reported the possibility of finding key wavelengths in those range of spectra (Rossel and Behrent, 2010; Rossel *et al.*, 2015). According to Stuart (2004) in Table 1, 2000-2200 nm corresponds to the combination N-H stretching and combination O-stretching. Rossel *et al.* (2015) attributed it to the range of carbonyl C=O/CH stretch vibration. But it also can be attributed to the type of clay mineral according to Nayak and Singh (2007). Rossel and Behrent (2010) founds also important wavelength in the range 2000-2200 nm using VIP and explains it by the presence of Kaolin. Spectra around 1455 nm are mostly attributed to water but with the phenomena of spectral overlapping as mention by Rossel and Behrens (2010), the range between 1400 and 1500 nm can also be affected. Nevertheless, Stuart (2004) in Table 1 attributed the spectra portion between 1400 and 1500 nm to the First overtone N-H stretching, and first overtone O-H

stretching. Other portion of the soil spectra are also preferably selected by both feature selection algorithms (850-950 nm, 1110-1225 nm). The range 850-950 nm is mostly attributed to the second overtone N-H stretching, and second overtone O-H stretching according to Stuart (2004) (Table 1), but also to Fe oxides according to Nayak and Singh (2007). The range 1110-1225 nm corresponds to the second overtone C-H stretching (Stuart, 2004). Most of wavelengths around the NIR are correlated to organic functional groups (Rossel *et al.*, 2015). Brown et al. (2006) identified also SOC key wavelength around NIR (960 nm, 1100 nm, 1400 nm, 1900 nm, 2309 nm, 2180 nm, 1744 nm, and 1870nm).

V.2. Performance of the RF regression compared to the PLSR regression in predicting SOC

There is not much literature comparing the performance of RF and PLSR in predicting SOC contents. However, Rossel and Behrens (2010) have assessed the performance of many data mining algorithms to predict SOC, including PLSR and RF regression. They have found that PLSR was slightly better than RF. Although our results proved the contrary, RF regression outperforms PLRS in predicting SOC contents in an independent dataset. This difference can be attributed to the complexity of our data, collected in a mountainous area where different altitudes, slopes, soil thicknesses may have different influences on soil composition and spectral characteristics. In such conditions, data are not linear and machine learning algorithms such as RF are likely to better predict because of their robustness to non-linear trends (Rossel and Behrens, 2010). In addition, our research takes all the spectral data as collected by the ASD. Rossel and Behrens (2010) resampled the data at 10 nm spatial resolution, making it more complex than ours. Diaz-Uriate and Andres (2006) reported the ability of RF regression to better fit the independent dataset. The weakness of PLSR to accurately predict SOC content in a new dataset as demonstrated in this study was also reported by other scientists (Chang *et al.*, 2001; Stevens *et al.*, 2006).

Comparing our overall models to previous studies, the RF models are more accurate than what Rossel and Behrens (2010) have found using 72 wavelengths under laboratory conditions. PLSR model performances are similar to what Rossel and Behrens (2010); Mouazen (2014) and Li *et al.* (2015) found under laboratory conditions (Table 2), and Stevens *et al.* (2008) and Rossel and Behrens (2010) under field conditions using the same regression.

V.3. Performance of field spectral measurements and laboratory spectral measurements in predicting SOC

Before comparing the performance of the field and laboratory spectra, the one tailed student's t test was implemented to show the dissimilarity between the field and the laboratory spectra. The results shows that laboratory spectral measurements are significantly different to field spectral measurements at 5% level (p value = 0.024). This observation can be justified by the presence of fresh moist soil in the field spectral measurements (Kuang and Mouazen, 2011). The moist soil increases the effect of forward scattering of light and enhances the abortion at all portions of the wavelength (Lobell and Asner, 2002). However, Rossel *et al.* (2009) found no significant difference between field and laboratory measurements. This is because they have removed all effects of water absorption.

The comparison of the performance of both laboratory and field RF models seems to be difficult in this research because the best models for both dataset have almost the same performance. It is to be noted that the laboratory spectral measurements in our study were not dried and sieved as supported by many authors. Consequently, the influence of moisture is not totally removed and may reduce its accuracy at the field spectra level. The best model in the laboratory dataset was achieved with L-FD-K-RF (RPD =3.77, Rp^2 = 0.87, RMSEP = 0.64%), while for the field was achieved with F-FD-K-RF (RPD = 3.77, Rp^2 = 0.88, RMSEP = 0.64%).

However, PLSR models developed with the laboratory dataset are slightly better than the ones developed in the field. Li *et al.* (2015) also found that PLRS model achieved better accuracy with the laboratory spectral data than the field data. The reason why our results are different from the ones obtained by Li *et al.* (2015) may be because PLSR is not robust to predict both laboratory and field spectral (noisy) data with the same accuracy.

V.4. Influence of spectra first derivative on the performance of different models

The impact of spectra derivatives and key wavelength selection in the raw spectral data (laboratory and field) is effective in improving RF regression models, as well as PLSR models. This is because the best predictive models for RF regression (RPD = 3.77, R_p^2= 0.88, RMSEP = 0.64%) and PLSR-K (RPD = 2.97, Rp^2= 0.86, RMSEP = 0.75%) with transformed spectral data, were L-FD-K-RF and F-K-PLSR, respectively. The effect of spectra derivatives and key wavelength selection on improving the raw model has been indicated by many other

scientists. Peng *et al.* (2014) assessed the impact of 8 pre-processing methods in improving the raw model and the first derivative was the best. Li *et al.* (2015) demonstrated the superiority of the first derivative with SG smoothing to improve PLSR. Vasques *et al.* (2008) also used the first SG smoothing to improve SOC models. Rossel and Behrens (2010) also improved the RF method by using a discrete wavelet transform algorithm like a feature selection method.

V.5. Significant and limitations of the models

The implication of this research is that field spectroscopy measurements offer a better perspective to predict SOC contents in mountainous landscapes with low price and reasonable time. RF can be used to monitor the quality of soil using SOC as a proxy. Furthermore, decision makers can use these models as tools to easily monitor SOC contents and integrate it in the soil carbon market which is not yet implemented because of the lack of cost effective methods. In addition to what has been obtained so far by different researchers, these findings have brought more insight in RF regression in term of prediction SOC in southern Africa mountainous landscapes. No previous research has been oriented in this way.

Field spectroscopy measurements as demonstrated in this work show more advantages to predict SOC than satellite images because of the complex morphology of the region. On mountainous landscapes, satellite images are limited to estimate SOC because of topographic shading, thickness, slope processes, microclimate effects, water retention and orographic precipitation which have a huge impact on SOC estimation.

Results from different models need to be applied carefully regarding the heterogeneity of the site location. The limitation of this investigation is that it provides overall models of SOC carbon in mountainous areas. The effects of different microclimate and altitude were not tested on the model accuracy. Also, it is not evident that the models will hold during another period of time, knowing that samples were collected in October, the rainy season. More investigation needs to be done in order to generalize our models.

CONCLUSIONS

The performance of RF regression in predicting SOC contents using field measurements was assessed. The model was compared to the most commonly used regression, PLSR. Field measurements from both regressions were compared to the predictions with results from the same spectra measured in the laboratory. Key wavelength selection and spectra first derivative effects were assessed on both regression models. The results indicated that:

- The best models in predicting SOC on an independent dataset were found with RF regressions, and PLSR models are more likely to overfit the calibration dataset.
- The accuracy of PLSR models developed with laboratory spectral data were slightly superior to those developed with field spectra. However, with RF models, both dataset exhibit the same accuracy.
- The derivative spectral models improved the raw spectral models because the best model with both laboratory and field spectral data were obtained with transformed spectral data.
- Key wavelengths to predict SOC contents were mostly localised around the visible range (400-700 nm).

However, this study offers many perspectives to conduct future research. More investigation need be oriented on the heterogeneity of the site location (mountains dominated) which is expressed here by a high coefficient of variation of SOC concentration. Therefore, we suggested further research on:

- The effect of altitude variation to the overall model performance in order to increase the accuracy.
- Testing the effect of the season change and different soil moisture variation on the model performance.

For a wider implication to local community, we would like to orient future researches on interpolating SOC concentration obtained from spectroscopy by using different geostatistical methods in order to map SOC contents.

REFERENCES

Aber, J., Wessman, C.A., Peterson, D.L., Mellilo, J.M. and Fownes, J.H.1990. Remote sensing of litter and soil decomposition in forest ecosystems. In: Hobbs, R.J. and Mooney, H.A. (Eds). *Remote sensing of biosphere functioning*, Springer, New York, pp. 87-101.

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

Akaike, H. 1973. Information theory and an extension of maximum likelihood principe. In: Petrov, B.N., Csaki, F. (Eds.), *Second International Symposium on Information Theory*. Akademia Kiado, Budapest, Hungary, pp. 267-281.

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 (6), 477–485.

Archer, K. J. and R. V. Kimes. 2008. Empirical Characterisation of Random Forest Variable Importance Measures. *Computional Statistics and Data Analysis*, 52, 2249-60.

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

Ball, D.F. 1964. Loss-on-Ignition as an estimate of and organiccarbon in non-calcareous soils. *Journal of Soil Science*, 15, 84–92.

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 and Remote Sensing* 69 (6), 619–630.

Bell, F.G. and Haskins, D.R. 1997. A geotechnical overview of Katse Dam and Transfer Tunnel, Lesotho, with a note on basalt durability. *Engineering Geology*, 46, 175–198.

Ben-Dor, E. 2002. Quantitative remote sensing of soil properties. *Advances in Agronomy*, 75, 173-243.

Ben-Dor, E., Inbar, Y. and Chen, Y. 1997. The reflectance spectra of 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.

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

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

Breiman, L., Friedman, J., Olshen, R. and Stone, C. 1984. *Classification and Regression Trees*. Wadsworth, Belmont (CA).

Brevik, E.C., Cerdà, A., Mataix-Solera, J., Pereg, L., Quinton, J.N., Six, J. and Van Oost, K. 2015. *The interdisciplinary nature of soil*, 1, 117–129.

Bricklemyer, R.S. and Brown, D.J. 2010. On-the-go VisNIR: potential and limitations for mapping soil clay and organic carbon. *Computers and Electronics in Agriculture*, 70 (1), 209–216.

Brown, D.J., Shepherd, K.D., Walsh, M.G., Dewayne, Mays, M. and Reinsch, T.G. 2006. Global soil characterization with VNIR diffuse reflectance spectroscopy. *Geoderma*, 32 (3-4), 273-290.

Bureau of Statistics and Planning. 2007. Lesotho agricultural situation report (1982-2004). Maseru, Lesotho.

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.

Chong, I. G. and Jun, C. H. 2005. Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intelligent Laboratory Systems*, 78(1), 103-112.

Christy, C.D. 2008. Real-time measurement of soil attributes using on-the-go near infrared reflectance spectroscopy. *Computers and Electronics in Agriculture*, 61 (1), 10–19.

Coburn, B.J., Okano, J.T. and Blower, S. 2013. Current drivers and geographic patterns of HIVin Lesotho: implications for treatment and prevention in Sub-Saharan Africa. *BMC Medicine*, 11 (224), 1-11.

Cohen, M., Mylavarapu, R.S., Bogrekci, I., Lee, W.S. and Clark, M.W. 2007. Reflectance spectroscopy for routine agronomic soil analyses. *Soil Science*, 172 (6), 469-485.

Condit, H. R. 1970. The spectral reflectance of American soils. *Photogrammetric Engineering*.

Corsi, S., Friedrich, T., Kassam, A., Pisante, M. and Sà, J. D. M. 2012. *Soil organic carbon accumulation and greenhouse gas emission reductions from conservation agriculture: a literature review*. Food and Agriculture Organization of the United Nations (FAO). Rome, 89 pp.

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.

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.

Díaz-Uriarte, R. and Alvarez de Andrés, S. 2006. Gene selection and classification of microarray data using random forest. *BioMedical Center Bioinformatics*, 7, 13-23.

Eash, N., Lambert, D.M., Marake, M., Thierfelder, C., Walker, F.R. and Wilcox, M.D. 2013. *Small-holder Adoption of Conservation Agriculture in Lesotho and Mozambique*. Huazhong Agricultural University Wuhan, China.

Edwards, R.J., Ellery, W.N. and Dunlevey, J. 2016. The role of the *in situ* weathering of dolerite on the formation of a peatland: The origin and evolution of Dartmoor Vlei in the KwaZulu-Natal Midlands, South Africa. *Catena*, 143, 232–243.

Efron, B. and Tibshirani, R. 1993. *An Introduction to Bootstrapping*. Chapman andHall/CRC, Boca Raton, FL/New York.

EU Soil Thematic Strategy. 2004. *TWG work package 2: status and distribution of soil in Europe*. Final report version 1.0, 31 March 2004.

Friedman, J.H. 1991. Multivariate adaptive regression splines (with discussion). *Annals of Statistics*, 19, 1-67.

Garzanti, E., Padoan, M., Setti, M., López-Galindo, A. and Villa, I.M. 2014. Provenance versus weathering control on the composition of tropical river mud (southern Africa). *Chemical Geology*, 366, 61–74.

Gehl, R. and Rice, C. 2007. Emerging technologies for *in situ* measurement of soil carbon. *Climatic Change*, 80 (1), 43–54.

Goidts, E. and van Wesemael, B. 2007. Regional assessment of soil organic carbon changes under agriculture in Southern Belgium (1955–2005). *Geoderma*, 141 (3–4), 341–354.

Gomez, C., Viscarra Rossel, R.A. and McBratney, A.B. 2008. Soil organic carbon prediction by hyperspectral remote sensing and field VIS-NIR spectroscopy: an Australian case study. *Geoderma*, 146 (3–4), 403–411.

Gomez-Carracedo, M.P., Andrade, J.M., Rutledge D.N. and Faber, N.M. 2007. Selecting the optimum number of partial least squares components for the calibration of attenuated total reflectance-mid-infrared spectra of unsersigned kerosene samples. *Analytica Chimica Acta*, 585, 253-265.

Grab, S. and Nüsser, M. 2001. Towards an integrated research approach for the Drakensberg and Lesotho mountain environments: A case study from the Sani plateau region. *South African Geographical Journal*, 83, 64–68.

Guénon, R., Vennetier, M., Dupuy, N., Roussos, S., Pailler, A. and Gros, R. 2013. Trends in recovery of Mediterranean soil chemical properties and microbial activities after infrequent and frequent wildfires. *Land Degradation & Development*, 24, 115–128.

Guyon, I. and A. Elisseeff. 2003. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, 3, 1157-1182.

He, Y., Song, H., García, P.A. and Hernández, G.A. 2005. Measurement and analysis of soil nitrogen and content using near-infrared spectroscopy techniques. *Journal of Zhejiang University of Science*, 6 (11), 1081-1086.

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

Hossain, M. Z. 2001. Farmer's view on soil depletion and its management in Bangladesh. *Nutrient Cycling in Agroecosystems*, 61, 197-204.

Hruschka, W.R. 1987. Data analysis: Wavelength selection methods. In: Williams, P., Norris, K (eds), *Near-Intrared Technology in Agriculture and Food Industries*, pp. 35-55.

IPCC. 2014: *Climate Change 2014: Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Geneva, Switzerland.

Jackson, J.E. 1991. *A user guide to principal components*. John Wiley and Sons Inc, 41-50

Karatzoglou, A., Smola, A. and Hornik, K., 2008. Kernlab: Kernel-based Machine Learning Lab. (At: http://cran.r-project.org/web/packages/kernlab/index.html. Accessed: 24/09/2016).

Kasozi, G.N., Nkedi-Kizza, P. and Harris, W.G. 2009. Varied carbon content of in Histosols, spodosols, and carbonatic soils. *Soil Science Society of America Journal*, 73, 1313–1318.

Kittler, J. 1978. Feature set search algorithms. *Pattern recognition and signal processing*, 41, 60.

Kohavi, R. and John G.H. 1997. Wrappers for features subset Selection. *Artificial Intelligence*, 97, 273-324.

Konen, M. E., Jacobs, P. M., Burras, C. L., Talaga, B. J. and Mason, J. A. 2002. Equations for predicting soil organic carbon using loss-on-ignition for north central US soils. *Soil Science Society of America Journal*, 66(6), 1878-1881.

Kooistra, L., Wanders, J., Epema, G.F., Leuven, S., Wehrens, R. and Buydens, L.M.C. 2003. The potential of field spectroscopy for the assessment of sediment properties in river floodplains. *Analytica Chimica Acta*, 484, 189-200.

Kramer. R., Workman. J.J. and Reeves, J.B. 2004. Qualitative analysis. In: Al-Amoodi, L., Roberts,C.A., Workman, J.J., Reeves, J.B., Barbarick, K.A. (Eds), *Dick Near-infrared*

spectroscopy in agriculture, Agronomy Monograph no. 44'. American Society of Agronomy, Inc., Crop Science Society of America, Inc., and Soil Science Society of America, Inc.: Madion, Wisconsin, USA, pp 85-102.

Kuang, B. and Mouazen, A. M. 2011. Calibration of visible and near infrared spectroscopy for soil analysis at the field scale on three European farms. *European Journal of Soil Science*, 62(4), 629-636.

Kuhn, N.J., Hoffmann, T., Schwanghart, W. and Dotterweich, M. 2009. Agricultural soil erosion and global carbon cycle. *Earth Surface Processes and Landforms*, 34 (7), 1033–1038.

Lal, R. 2005. Soil erosion and carbon dynamics. *Soil & Tillage Research*, 81, 137–142.

Lal, R. 2009. Challenges and opportunities in soil research. European. *Journal of Soil Science*, 60 (2), 158–169.

Li, S., Shi, Z., Chen, S., Ji, W., Zhou, L., Yu, W. and Webster, R. 2015. *In Situ* Measurements of Organic Carbon in Soil Profiles Using VIS-NIR Spectroscopy on the Qinghai. *Environmental Science and Technology*, 49, 4980−4987.

Liaw, A., Weiner, M., 2002. Classification and regression by random Forest. *R News*, 2, 18–22.

Lobell, D.B. and Asner, G.P. 2002. Moisture effects on soil reflectance. *Soil Science Society of America Journal*, 66 (3), 722–727.

Loveland, P. and Webb, J. 2003. Is there a critical level of in the agricultural soils of temperate regions? *A review*. *Soil & Tillage Research*, 70, 1-18.

Maindonald, J. and Braun, J. 2006. *Data analysis and graphics using R: an example-based approach* (Vol. 10). Cambridge University Press. London.

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

Mason, S.J. and Jury, M.R. 1996. Climatic variability and change over southern Africa: a reflection on underlying processes. *Progress in Physical Geography*, (21), 23-50.

Mathews, H.L., Cunningham, R.L., Cipra, J.E. and West, T.R. 1973. Application of multispectral remote sensing to soil survey research in Southeastern Pennsylvania. *Soil Science Society of America Proceedings*, 37, 88-93.

Mbata, J.N. 2001. Land use practices in Lesotho: Implications for sustainability in agricultural production. *Journal of Sustainable Agriculture*, 18, 5–24.

McCarty, G.W., Reeves, J.B., Follet, R.F. and Kimble, J.M. 2002. Mid-infrared and near-infrared diffuse reflectance spectroscopy for soil carbon measurement. *Soil Science Society of America Journal*, 66, 640-646.

McCoy, M.R. 2005. Field Methods in Remote Sensing. The Guilford press, New York London.

McMorrow, J.M., Cutler, M.E.J., Evans, M.G. and Al-Roichdi, A. 2004. Hyperspectral indices for characterizing upland peat composition. *International Journal of Remote Sensing*, 25 (2), 313–325.

Meadows, M.E. and Hoffman, M.T. 2002. The nature, extent and causes of land degradation in South Africa: legacy of the past, lessons for the future? *Area*, 34, 428–437.

Mehmood, T., Liland, K.H., Snipen, L. and Solve, S. 2012. A review of variable selection methods in Partial least Squares Regression. *Chemometrics and Intelligent Laboratory Systems*, 113, 62-69.

Milton, E.J., Schaepman, M.E., Anderson, K., Kneubühler, M. and Fox, N. 2009. Progress in field spectroscopy. *Remote Sensing of Environment*, 113, 92–109.

Montgomery, O.L. 1976. An investigation of the relationship between spectral reflectance and the chemical, physical and genetic characteristics of soils. Ph.D. Thesis, Purdue University.

Morgan, C.L.S., Waiser, T.H., Brown, D.J. and Hallmark, C.T. 2009. Simulated *in situ* characterization of soil organic and inorganic carbon with visible near-infrared diffuse reflectance spectroscopy. *Geoderma*, 151 (3–4), 249–256.

Mouazen, A.M. 2014. Comparing the artificial neural network with parcial least squares for prediction of soil organic carbon and pH at different moisture content levels using visible and near-infrared. *Biosystems Engineering*, 3, 1794–1804.

Nocita, M., Stevens, A., Noon, C. and Wesemael, B.V. 2013. Prediction of soil organic carbon for different levels of soil moisture using Vis-NIR spectroscopy. *Geoderma*, 199,37–42.

Nocita, M., Stevens, A., Toth, G., Panagos, P., van Wesemael, B. and Montanarella, L. 2014. Prediction of soil organic carbon content by diffuse reflectance spectroscopy using a local partial least square regression approach. *Soil Biology and Biochemistry*, 68, 337-347.

Novara, A., Gristina, L., Bodì, MB., 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.

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.

R Development Core Team, 2012. R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing. Vienna: Austria. http://www.r-project.org.

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 (1–2), 3–14.

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

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

Rumelhart, D.E., Hinton, G. E. and Williams, R. J. 1986. Learning internal representations by error propagation. In Rumelhart, D.E., McClelland, J.L. and the PDP Research Group (Eds.),

Parallel Distributed Processing: Explorations in the Microstructure of Cognition. MIT Press, London, pp. 318-362.

Saha, T.K. 2011. Impact of climate change on agricultural production in Lesotho: A case study. *African Crop Science Conference Proceedings*, 10, 273-277.

Sarle, W. 1995. Stopped training and other remedies for overfitting. 27th Symposium on the Interface Computing Science and Statistics, Pittsburgh, PA.

Schulte, E. E. and Hopkins, B. G. 1996. Estimation of soil by weight loss-on-ignition. *Soil : analysis and interpretation*, 21-31.

Schulze, E.D., Wirth. C., Heimann, M. 2000. Climate change-managing forests after Kyoto. *Science*, 289, 2058–2059.

Selige, T., Böhner, J. and Schmidhalter, U. 2006. High resolution topsoil mapping using hyperspectral image and field data in multivariate regression modeling procedures. *Geoderma*, 136(1), 235-244.

Shepherd, K.D. and Walsh, M.G. 2002. Development of reflectance spectral libraries for characterization of soil properties. *Soil Science Society of America Journal*, 66 (3), 988–998.

Shorack, G. R. and Wellner, J. A. 2009. *Empirical processes with applications to statistics*. Society for Industrial and Applied Mathematics.

Showers, K. 2005. *Imperial Gullies: Soil Erosion and Conservation in Lesotho*. Ohio University Press. Ohio.

Sicili, L. 2010. *Conservation agriculture and sustainable crop intensification in Lesotho*. Food and Agricultural Organization of the United Nations. Rome.

Smith, M. L., Martin, M.E., Plourdes, L. and Ollinger, S.V. 2003. Analysis of Hyperspectral Data for Estimation of Temperate Forest Canopy Nitrogen Concentration: Comparison between an Airborne (AVIRIS) and Spaceborne (Hyperion) Sensor. *IEEE Transactions on Geoscience and Remote Sensing*, 41,1332-1347.

Smith, P. 2008. Land use change and soil organic carbon dynamics. *Nutrient Cycling in Agroecosystems*, 81(2), 169–178.

Soriano-Disla, J.M., Janik, J., Rossel, R.A., Macdonald, L.M. and Mclaughilin, 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.

Ssali, H. 2000. *Soil resources of Uganda and their relationship to major farming systems*. Resource paper. Soils and Soil Fertility Management Programme, Kawanda, NARO, Uganda.

Stevens, A., Udelhoven, T., Denis, A., Tychon, B., Lioy, R., Hoffmann, L., van Wesemael, B. 2010. Measuring soil organic carbon in croplands at regional scale using airborne imaging spectroscopy. *Geoderma*, 158 (1–2), 32–45.

Stevens, A., Van Wesemael, B., Vanderschrick, G., Toure, H. and Tychon, B. 2006. Detection of carbon stock change in agricultural soils using spectroscopic techniques. *Soil Science Society. America Journal*, 70, 844–850.

Stoner, E.R. and Baumgardner, M.F. 1981. Characteristic variation of reflectance of surface soils. *Soil Science Society of America Journal*, 45, 1161-1165.

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

Sun, H., Nelson, M., Chen, F. and Husch, J. 2009. Soil mineral structuralwater loss during loss on ignition analyses. *Canadian Journal of Soil Science*, 89, 603–610.

United Nations/Framework Convention on Climate Change. 2015. Adoption of the Paris Agreement, 21st Conference of the Parties, Paris: United Nations.

Van-Camp, L., Bujarrabal, B., Gentile, A. R., Jones, R. J. A., Montanarella, L., Olazabal, C. and Selvaradjou, S. K. 2004. Volume III– and biodiversity. *Reports of the technical working groups established under the thematic strategy for soil protection*. Office for Official Publications of the European Communities, Luxembourg. 872p.

Vasques, G.M., Grunwald, S. and Sickman, J.O. 2008. Comparison of multivariate methods for inferential modeling of soil carbon using visible/near-infrared spectra. *Geoderma*, 146 (1–2), 14–25.

Viscarra Rossel, R. A. and Hicks, W. S. 2015. Soil organic carbon and its fractions estimated by visible–near infrared transfer functions. *European Journal of Soil Science*, 66(3), 438-450.

Vohland, M. and Emmerling, C. 2011. Determination of total soil organic C and hot water-extractable C from VIS-NIR soil reflectance with partial least squares regression and spectral feature selection techniques. *European Journal of Soil Science*, 62(4), 598-606.

Vohland, M., Ludwig, M., Thiele-Bruhn, S. and Ludwig, B. 2014. Determination of soil properties with visible to near and mid infrared spectroscopy: Effects of spectral variable selection. *Geoderma*, 1, 223-225.

Walcoot, 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, Departement of Agriculture, Fisheries and Forestry, Canberra.

Wang, Y., Fu, B., Lü, Y., Song, C. and Luan, Y. 2010. Local-scale spatial variability of soil organic carbon and its stock in the hilly area of the Loess Plateau, China. *Quaternary Research*, 73(1), 70-76.

Willimas, P.C. 2001. Implementation of near-infrared technology. In Williams, P. and Norris, K. (Eds), *Near-infrared technology in the agriculture and food industries*. America Association of Cereal Chemist, Minnesota, pp. 145-169.

Wold, S. 1995. PLS for multivariate linear modeling. In: van de Waterbeemd, H. (Ed.), *Chemometric Methods in Molecular Design*. VCH, Weinheim, Gemany, 195–218.

Workman, J.J. and Shenk, J. 2004. Understanding and using the near-infrared spectrum as an analytical method. In: Al-Amoodi, L., Roberts,C.A., Workman, J.J., Reeves, J.B., Barbarick, K.A. (Eds). *Dick Near-infrared spectroscopy in agriculture, Agronomy Monograph no. 44'*. American Society of Agronomy, Inc., Crop Science Society of America, Inc., and Soil Science Society of America. Madion, Wisconsin, USA, pp. 78-99.

APPENDICES

PLSR code

```
# Partial least square regression code compiled by Freddy
Bangelesa, November 2016.

##############################################################
#############

# remove previous datasets and clean up the workspace

rm(list=ls())

##############################################################
#############

# set working directory

setwd("C:/Users/user/Desktop/statproject/stat project")

# laoding packages

library(car)

library(pls)

library(caret)

library(plsVarSel)

# laoding the data

labreduced <- read.csv("C:/Users/user/Desktop/statproject/stat
project/labreduced.csv")

#random splitting of the data (70/30)

smp_size <-floor(0.70 * nrow(labreduced))

set.seed(123)

train_ind<-sample(seq_len(nrow(labreduced)), size = smp_size)

train2<-labreduced[train_ind, ]

test2<-labreduced[-train_ind, ]

testy<-subset(test2, select = carbone)

trainy<-subset(train2, select = carbone)
```

```
#fitting PLSR model

m.pls <- plsr(carbone ~.,data=train2, validation="LOO", method
= "oscorespls")

summary(m.pls)

# optimizating the number of components

comp <- which.min(m.pls$validation$PRESS)

# feature selection using variable important in selection
algorithm

vip <- VIP(m.pls, 1)

matplot(vip)

matplot(scale(cbind(vip)), type = 'l')

write.table(vip,        "C:/Users/user/Desktop/statproject/stat
project/labderivVIP.csv",sep = ","

            ,row.names = T,col.names = T)

#plotting the variance explained

par(mar = c(4, 4, 2.5, 1) + 0.1)

plot(RMSEP(m.pls), legendpos = "topright", ylab = "Variance
explained (RMSE)")

#plotting the fiting val models

par(mar = c(4, 4, 2.5, 1) + 0.1)

plot(m.pls, ncomp = comp, asp = 1, line = TRUE)

# subrtracting PLSR coefficients

PLSRcoefficients <- coef(m.pls, ncomp = comp, intercept = T)

par(mar = c(4, 4, 1.5, 1) + 0.1)

coefplot (m.pls, ncomp = 1, legendpos = "topleft",intercept =
FALSE, type = "l",

          xlab = "variable", ylab = "regression coefficient")

#PLSR calibration accuracy

m.pred2<-predict(m.pls, train2, ncomp=comp)
```

```
pls.eval2<-data.frame(obs2=trainy, pred2=m.pred2[,1,1])
model2 = lm(Argile~pred2, data = pls.eval2)
summary(model2)$r.squared
summary(model2)$adj.r.squared
summary(model2)$ sigma
MSE2 <- mean(residuals(model2)^2)
MSE2
RMSE2 <-sqrt(MSE2)
RMSE2
AIC(model2)
# PLSR validation accuracy
m.pred<-predict(m.pls, test2, ncomp=comp)
obs<-testy
pls.eval<-data.frame(obs=testy, pred=m.pred[,1,1])
model = lm(carbone~pred, data = pls.eval)
summary(model)$r.squared
summary(model)$adj.r.squared
summary(model)$sigma
MSE <- mean(residuals(model)^2)
MSE
RMSE <-sqrt(MSE)
RMSE
AIC(model)
# plotting predicted Vs observed
obs<-pls.eval$carbone
pred<-pls.eval$pred
par(mar = c(4, 4, 2.5, 1) + 0.1)
```

```
op <- par(c(lty="solid", col="black"))
plot(obs, pred, asp=1, xlab="Observed", ylab="Predicted")
prline <- lm(carbone~ pred, data= pls.eval)
abline(prline)
par(c(lty="dashed", col="black"))
abline(0, 1)
# computing RPD
SD<-sd(trainy$carbone)
RPD<-SD/RMSE
RPD
```